A machine learning approach to taking EEG-based brain-computer interfaces out of the lab

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

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

vorgelegt
von

Vinay Jayaram
aus Farmington Hills, United States of America

Juli - 2018

Tag der mündlichen Prüfung: 15.11.2018

Dekan der Math.-Nat. Fakultät: Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät: Prof. Dr. I. B. Autenrieth

1. Berichterstatter: Prof. Dr. Moritz Grosse-Wentrup

2. Berichterstatter: Prof. Dr. Martin Giese

Prüfungskommission: Prof. Dr. M. Grosse-Wentrup

Prof. Dr. M. Giese

Prof. Dr. B. Schölkopf

Prof. Dr. W. Rosenstiel

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

„A machine learning approach to taking EEG-based brain-computer interfaces out of the lab“

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 “……”, 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

Contents

Abstract **5**

Acknowledgements **6**

1 Synopsis **7**
1.1 Introduction 7
1.2 Transfer learning: leveraging many, small datasets 9
1.2.1 Multi-session approaches 10
1.2.2 Multi-task learning 11
1.3 Frequency modulation: more information from less electrodes . 14
1.3.1 Alpha peak shift 15
1.3.2 Frequency shifts in BCIs 16
1.4 How to validate? Principled use of open-access data 17
1.5 Conclusion: Moving forwards 19
1.5.1 Transfer Learning 20
1.5.2 Frequency Modulation 21
1.5.3 Outlook 22

2 Collected works **31**
2.1 Contributions 31
2.2 Paper 1: Transfer Learning in Brain-Computer Interfaces . . . 33
2.3 Paper 2: Task-induced frequency modulation features for brain-computer interfacing 54
2.4 Paper 3: MOABB: Trustworthy algorithm benchmarking for BCIs 68
2.5 Paper 4: Multi-Task Logistic Regression for Brain-Computer Interfaces 78

Abstract

Despite being a subject of study for almost three decades, non-invasive brain-computer interfaces (BCIs) are still trapped in the laboratory. In order to move into more common use, it is necessary to have systems that can be reliably used over time with a minimum of retraining. My research focuses on machine learning methods to minimize necessary retraining, as well as a data science approach to validate processing pipelines more robustly. Via a probabilistic transfer learning method that scales well to large amounts of data in high dimensions it is possible to reduce the amount of calibration data needed for optimal performance. However, a good model still requires reliable features that are resistant to recording artifacts. To this end we have also investigated a novel feature of the electroencephalogram which is predictive of multiple types of brain-related activity. As cognitive neuroscience literature suggests, shifts in the peak frequency of a neural oscillation – hereafter referred to as frequency modulation – can be predictive of activity in standard BCI tasks, which we validate for the first time in multiple paradigms. Finally, in order to test the robustness of our techniques, we have built a codebase for reliable comparison of pipelines across over fifteen open access EEG datasets.

Acknowledgements

It is terrifying, as it no doubt is for all doctoral students, to be writing the acknowledgements section. I thought to do it first, since it seemed easier than writing about everything I've learned in the last four years, but then I realized that trying to comprehend how far you've come, and who all has helped you to get there, is much worse.

My foremost gratitude goes to Moritz, who saw the potential in a biology major to enter this highly mathematical field, and who has guided me in this journey. Without his advice on how to think about my projects, and his feedback on my ideas, I may well have gained the quantitative skills I have now, but I would have never gotten this view of science and how it can be done. Looking forwards I believe it is the latter which will have the greater impact on my future. I am also very grateful to Bernhard for supporting me over the course of my PhD and always being available for guidance if I asked.

Next I must thank my labmates – Matthias, Tatiana, Atalanti, and Sebastian – for their input, both professional and personal, over the last years. I can't imagine having been with another set of colleagues, and I don't want to.

Lastly, but certainly not least, I would like to thank my family, who stood by me even as I announced my decision to disappear for years into the depths of Germany to study a field none of them had ever heard of. I could not have done so much of what I have done, if not for you, and your support will continue to be behind anything else I might publish or create, even if I don't get another chance to say it publicly.

Chapter 1

Synopsis

1.1 Introduction

One of the most surprising things about the field of brain-computer interfacing (BCI) is that it has been around for almost thirty years now, even though few outside academia realize it is any more than science fiction. In some labs, it even manages to work impressively well, once it is set up and calibrated. The mystery is why it is still almost exclusive to academia. As researchers, we explain it away by pointing out how difficult a problem it is. An outsider to the field, however, might see it somewhat differently. If you're telling me that we had a system in 1994 to control a cursor on a computer screen *with only your mind*, they might ask, then why isn't it in my iPhone yet? And they have a point. Many fields we've known about for decades now – brain-based control, muscle-based control – have surfaced once or twice in the popular imagination but invariably slunk back into the lab, while at the same time other fields – facial recognition, voice generation – are being deployed in an array of applications. Deep learning was the innovation that brought image and audio machine learning into widespread use; the question my doctorate explores is, how can we move towards those advances in BCIs?

It is hard, with two working thumbs and an iPhone, to appreciate the urgency of this gap in application, but physically abled people are not the target population for a BCI. A keyboard and mouse may not be the ideal means of communicating with a computer, but for someone with two functioning hands they serve quite well– the problems begin for those without. For people with paralysis, as well as for people with various neurological conditions, communicating with the world is not something to be simply taken for granted. For these people, even a good deal of inconvenience would be acceptable if it granted them a way of reliably communicating with a min-

imum of movement. Here, at least, there has been progress: eye-tracking technology, muscle-based systems, and smart typing systems are used by hundreds of individuals with severe movement disabilities to communicate with loved ones and caregivers (and, in some cases, to write famous French novels or publish famous English theoretical physics papers). Even so, there remain patients for whom none of this is good enough. At the end stage of Amyotrophic Lateral Sclerosis (ALS), patients enter the so-called completely locked-in state, in which every form of muscle activity is extinguished [1]. The only way to ensure that everyone who is not incapable of thought is still capable of expressing themselves is to have a BCI available. Yet true, reliable brain-based communication remains out of reach, even for individuals willing to undergo inconvenience.

For a device to be usable in daily life, it must have the following properties: It must fulfill a need, it must be reliable, and it must be simple to setup and use. Even in patient populations, state-of-the-art brain-computer interfaces only fulfill the first of these three properties. For people without muscle control, it is by definition the only way to communicate. However, the reliability and the simplicity represent a difficult tradeoff. For a BCI to translate thoughts into actions, it must have a model which converts the raw electroencephalogram (EEG) signal into estimates of brain activity. Any model requires data to be properly fitted to the intended application, and an ideal model performs reliably under real-life conditions. Unfortunately, no such ideal model exists, and the best current models for BCI applications require significant amounts of data from every recording session to perform optimally. Further, transient changes in environmental factors can also strongly affect the predictive accuracy of current models. What this leads to is a situation in which models fitted with good data and applied in the lab can be very effective, but reducing the signal quality or training data size can often lead to reductions in performance.

In order to overcome these limitations we may consider how image recognition overcame its own problems: better models with bigger data. The field of BCI does not have models that scale well to large offline datasets, and so my work has focused on building and validating models which allow the field to effectively use big data. For this to happen, there are two important hurdles to be overcome: Models must be able to deal with data from many (possibly short) recording sessions, and they must be robust to variable data quality and few recording channels. To explore my contribution to the literature, the rest of this thesis is structured as follows: First there is a brief overview of the specific contributions in transfer learning, feature generation, and offline validation, followed by a section to explore the potential of these contributions in combination and to look to future experiments.

1.2 Transfer learning: leveraging many, small datasets

While effective in lab settings, machine learning in BCIs quickly runs into the problem of signal non-stationarity. Between two different recording sessions, the properties of the signal can vary significantly, causing the commonly used variance features to have different distributions [2], [3]. These factors are both random, such as the electrical noise in the room and the impedences of the electrodes, and based on the mental and physical state of the user. While it can be relatively simple to learn a model to predict certain brain responses within a single recording session, this model very rarely generalizes perfectly to new recordings. Because of this, the majority of machine learning within the realm of BCIs focused on the problem of within-recording-session machine learning and simply assumed that users would be willing to record a small calibration dataset at the beginning of each recording session. In the lab setting this is a reasonable assumption, but for more practical application the requirement quickly becomes onerous. Having to teach a computer over and over again what seems – to a human – to be the same thing is a very frustrating experience. Further, this approach is, from a machine learning point of view, incredibly wasteful. Ignoring large amounts of data that were gathered earlier and only using session-specific data forces any model to use only a small subset of the full possible samples, which necessitates simple models to avoid overfitting. Furthermore, it makes session usefulness entirely dependent on the user: If the training data is bad for any reason, even if the user is normally proficient in the system, then the rest of the session will not work.

Over the last 20 years of machine learning research there has been a great deal of progress on suitable models for single-session BCIs. In particular spatial filtering via common spatial patterns (CSP) showed that the number of discriminative brain regions during a BCI task is usually much smaller than the number of recording channels, and that recovering the signals from exactly these regions is far more robust than using features generated from the channels themselves [4], [5]. However, this approach suffers greatly from overfitting to the training data: spatial filters trained on one session are not likely to transfer to other recording sessions (for a review of the need for regularization see Lotte et al [5]).

Since BCIs are interactive systems, the difficulty of calibration-free or plug-and-play BCIs has also been tackled via adaptive fitting. This is an approach in which an initially poor model is trained through use of the system online, instead of first requiring a calibration dataset that is recorded without

feedback to the user. These approaches use sequential supervised samples to iteratively update the parameters of a linear classifier [6], [7]. Although this is an effective way of eliminating a calibration session, it still requires many labelled samples, and a small number of features, for the classifier to stabilize.

1.2.1 Multi-session approaches

Standard classification or regression models assume a single training and test set; transfer learning assumes there are also multiple datasets that were previously recorded. These could either correspond to different subjects or the same subject at previous points in time, or both. The goal of these methods is to leverage the previous recordings – with or without data from the current recording session – to fit a classification model for the current recording. This has mostly been considered in the multi-subject case, as over time the assumption was that a single subject would learn to generate data that corresponds to the initially trained model.

The early approach to the problem was to try pooling all the data and using a more complicated classification model. As this method was incapable of taking the statistics of a new recording into account, it was not particularly successful. These early models were attempted on channel-level features however, and so many people considered how to use CSP in a multi-subject context in order to get more stationary and robust features. Since a group of people doing motor imagery are activating the same area of cortex, albeit with slightly differing individual anatomies, it is reasonable to expect that the optimal spatial filters for all of them will be close to each other in some sense. Starting with Devlaminck et al. [8] and continuing on until the present day [9] many groups have invented ways of combining offline data when fitting CSP to a new subject. This has had some success, but also two major drawbacks: First, it is incapable of selecting features that are not based on variance. Second, it cannot be done in an adaptive manner, but rather requires a fixed amount of data from both classes be recorded in each new subject.

New approaches based on the geometry of the covariance matrix manifold have shown themselves to be more predictable and robust than spatial filtering [10], [11]. This is done by using all the variance information in the EEG, regardless of whether the underlying source is neural data or a correlated noise source, leading to better performance at the cost of model introspection. Particularly in the case of BCIs, however, the ability to determine what features of the recorded signal are predictive of the desired brain activity is very important. Invariant features solve the classification problem, but without knowing what part of the signal is predictive there is the danger that

the learned features do not actually correspond to neural data, but rather correlated noise sources such as muscles. Though this is mostly harmless in cases when the system is optimizing for accuracy in healthy subjects, it can be problematic in patient populations where certain types of features are lost over disease progression.

In order to get both a cross-subject transfer learning approach and to recover a representation of the variation over individuals, generative modelling has been used. Generative approaches are approaches that attempt to model the underlying recorded signals or features, instead of simply attempting to find a function that reliably predicts class membership. The most typical way this is done is via covariate shift adaptation [12], in which the most relevant data points are chosen from the offline pool for each new task based on their overlap with the data of the new recording. While effective, this approach has the downside that it requires a reasonably large pool of unlabeled data from each new task in order to do density fitting. It is also difficult to faithfully model data distributions in high dimensions, forcing this approach to limit the number of features the classifier can use. Further approaches via hierarchial Bayesian modelling of the recorded time-series have also been done [13]–[15]. These have the advantage of being able to encapsulate the information from the pool of offline subjects in hyper-priors that can be used to efficiently find parameters for new tasks. By modelling the projections of the sources directly, they allow for easy introspection to confirm that sources are neural in origin. However, the optimization strategy for this approach requires many independence assumptions in order to be tractable via more efficient methods such as variational inference. Further, incremental additions of data, such as those used in the adaptive calibration techniques, is not possible.

1.2.2 Multi-task learning

Instead of generative Bayesian modelling, our approach was to attempt a Bayesian treatment of the problem in the discriminative sense, in order to both efficiently recover a classification function for new subjects and a representation of the subject-invariant knowledge. Rather than attempting to model hidden sources that generated the observed data, we modeled the classification hyperplane in order to bias the recovered classifier towards the classifiers that are optimal for previous recordings [16]. The idea behind this method is reasonably simple: Instead of using a generic regularization method, such as ridge regression in the case of a linear classifier, with multiple datasets it is possible to determine data-driven regularization parameters. By modelling a regression vector as a random variable, as seen in Figure 1.1,

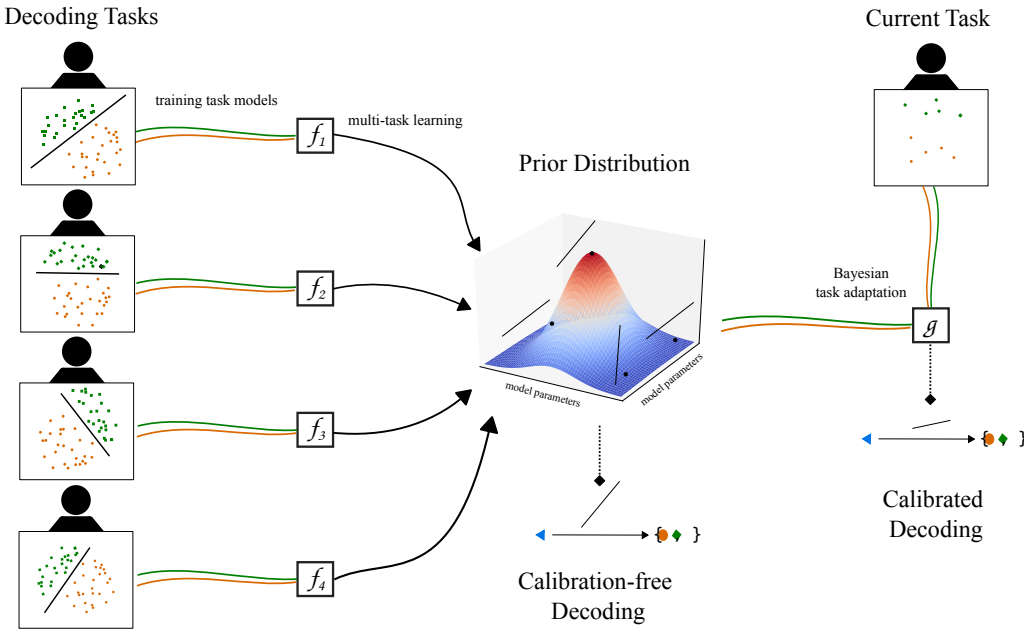

Figure 1.1: Schematic for modelling regression vectors in multi-task learning, reprinted from Jayaram et al. [17]

it is possible to estimate the parameters of the distribution over regression vectors and use them to bias the solution in future tasks.

The signals that are recovered by classification models in BCIs are usually projecting from one or few parts of the brain to the electrodes. Because of this, it makes little sense to take features computed on all the channels and simply stack them into one large feature vector. Instead, we determined that we could treat channels and features-per-channel differently via a bilinear model, in which two independent regression vectors are derived and combined to form the classifier: one which learns a weighting over channels and one which learns a weighting over features. In this way it is possible to consider many channels and many features per channel without a multiplicative increase in the number of parameters to be fit. This approach is a marked improvement over standard spatial filtering because it is capable of taking multiple frequency bands, as well as non-amplitude-based features, into account easily. It is also possible to visualize the prior distributions over the features and channels, which can lead to insight on which features are most important, as seen in Figure 1.2.

One major benefit of our proposed method is its scalability. Most approaches in BCI decoding are poorly suited to very large datasets, for example datasets with more than 100 subjects or recordings. Approaches that

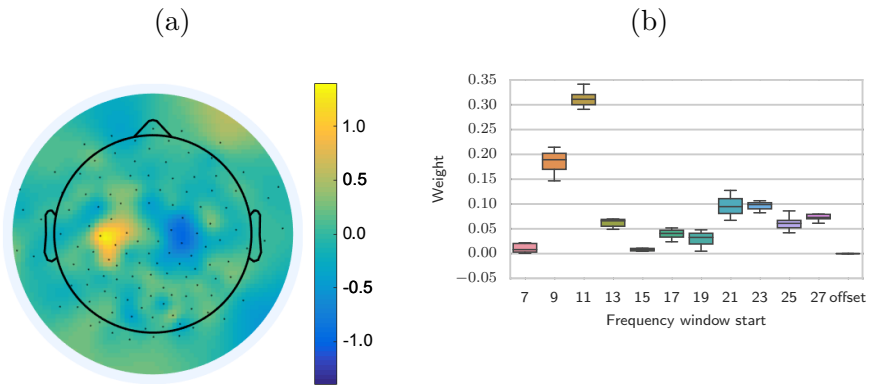

Figure 1.2: Figure, reprinted from [18], which shows the learned spatial and spectral prior means for a trained multi-task model on motor imagery data. The spatial weights are largest around the two motor cortices (a) and the spectral weights are higher around the frequencies corresponding to the anatomical α and β bands(b)

pool or that use covariate shift compensation require iterating through the entire offline pool in order to calibrate a classifier for a new subject, which causes the computation time to increase for the new recording proportionally to the number of recordings in the current pool. In contrast, multi-task learning requires an optimization to learn the parameters of the priors, but once those are fixed the update for a single subject is simply the solution to a set of linear equations. Further, while other optimizations require all the data to be loaded into memory at once, this approach is easily parallelized as each recording is dealt with independently in the maximization step of the algorithm.

A second consideration with big datasets is the possibility of negative transfer in the case that some subjects either have poor data or do the task very differently. This is a problem for most transfer learning techniques as they have no capacity to tell whether a particular task is good or bad. However in the context of a probabilistic model, negative transfer can be viewed as outliers in the samples that are used to build the prior distribution. Because of this they can be taken care of via standard outlier rejection techniques, or dealt with in the expectation step of the algorithm directly via the choice of a prior distribution with a larger tail. This results in the multi-task learning procedure's stability as the number of input datasets increases, under the mild assumption that there are comparatively few bad datasets.

One final benefit to this model, in comparison with generative approaches, is its compatibility with adaptive training. Since this approach corresponds

to a linear model with a fixed regularization term, it is possible to use sequential labeled data points to adapt both the linear and bilinear models in real time. This allows it to be used seamlessly in adaptive BCIs such as those described in Vidaurre et al [7] and Faller et al. [6].

1.3 Frequency modulation: more information from less electrodes

The vast majority of all data recorded for BCIs was recorded in the confines of a laboratory, using technicians to set up the equipment. As a result, most of it is of an appropriately high quality, with low electrode impedences and a minimum of noise from movement and other artifactual sources. This does not, however, describe a reasonable use case. Even for locked-in patients, it is infeasible to spend a half hour setting up electrodes; in the case of consumer electronics, this becomes even more ridiculous. Home use of BCI systems requires something that can be used in daily life, and that often means a minimum in setup time as well as number of electrodes.

Unfortunately, most established methods perform significantly worse with lower-quality signals and fewer electrodes. An illustrative example of this is the ongoing debate between dry and wet electrodes. Wet electrodes have historically been considered a gold standard for EEG signal quality [19], but must be manually optimized to achieve acceptable signals. In contrast, dry electrodes do not require gel application, as the name suggests, and therefore are both faster to apply and can withstand longer recordings, in exchange for a loss in signal quality and possibly comfort. The unreliability of standard methods in the face of a lower signal-to-noise ratio is such that making BCI paradigms work on non-wet systems is sufficiently novel to be published time and time again [20]–[22].

Another important hurdle in moving BCIs out of the laboratory is the physical environment. Labs can be set up to minimize possible sources of electrical noise, and participants are instructed to keep still and blink as little as possible. Moving into the real world, however, neither of these is easy to get, especially if one considers a medical environment like a hospital. As reviewed in Minguillon et al. [23], EEG recorded outside of the lab setting face many artifacts, and unless one wants to engineer a pipeline of complicated artifact-removal techniques, it is necessary to look for features that are less affected by them.

The first section of this thesis detailed a framework for transfer learning which allows for the effective leveraging of large amounts of offline data.

However, in order for that framework to be effective, it requires a large amount of data – which is difficult to attain when one must spend a large amount of time on ensuring that electrodes and recording conditions are perfect. In order to obtain enough data to fit robust classifiers, compromises in signal quality are unavoidable. Therefore, it is important to find features that can be computed which are more stable to signal quality differences. Via the bilinear approach detailed in Section 1.2, it is possible to easily add new channel-wise features to the classification model.

1.3.1 Alpha peak shift

In order to study what new features may be usable in BCI paradigms, one may begin by looking into the neuroscientific literature on brain rhythms. In non-invasive recordings, task-related differences in amplitude are best characterized, but invasive neuroscience has long known of many other properties of neural oscillations that are modulated by activity. One of these properties is the specific frequency at which a circuit oscillates. Bragin et al. [24] showed via invasive recordings of the rat hippocampus that while the amplitude of γ oscillations was tied to the phase of θ oscillations, both of these oscillations varied significantly in frequency throughout the recording period. This was further verified in humans in which it was shown that peak frequency changes could reliably predict working memory loads where amplitude-based measures were not predictive [25].

Most human studies of frequency shift are centered around the α peak. In the literature, this has been considered in both a static and dynamic manner. Studies of the average location of the α peak among individuals has showed a great deal of variance in frequency, due to both genetic [26], [27] and environmental factors [28]. These individual differences have been correlated with measures of intelligence [29], [30]. However, Aurlien et al [31] showed that the peak location can vary by up to 1Hz within a single recording session. Looking into the dynamics of this oscillation as recorded via EEG, it was determined that the α peak location is related to cognitive readiness on a shorter timescale, and that it can be modulated in a task-dependent manner in addition to the aforementioned static correlations [32]. More recent work has also confirmed that the frequency is modulated consciously in the case of cued and uncued stimulus perception[33]. In addition to these results on healthy, awake subjects, α peak frequency has been researched in sleep as well as epilepsy research, which has led to many algorithms to reliably extract it from the brain [34], [35].

The α peak is also well-known in the BCI literature, though it has only been viewed through phase or amplitude measures. The spectral power in

the alpha range is predictive of some mental activities, and this is sufficiently robust on a single-trial basis to be used within BCI paradigms [36]. More importantly, however, spectral power was shown to be predictive in ALS patients across a range of disease severities [37], which offers hope that it may be predictive into the completely locked-in state as well.

1.3.2 Frequency shifts in BCIs

While extensive, the work on characteristic features of the oscillations in invasive recordings is rarely used in the non-invasive sphere. Since muscle activity projects strongly to frequencies above 20Hz, the isolation of gamma rhythms from the non-invasive EEG is very difficult in practice, causing phase-amplitude coupling to be difficult to detect. As phase-amplitude coupling is much more well-researched than frequency shifts, it is not surprising that both are mostly ignored in the EEG literature. Instead, there is a large amount of work on phase-related measures, both in order to generate features based on pairwise coupling [38] and in order to refine estimates of spatial filters [39].

In the space of BCIs, the frequency drift of the EEG signal has been explored via adaptive filtering, which has been employed in order to improve the fidelity of amplitude-based measures. The location of the true α frequency in relation to the passband of the spectral filter is crucial to determining the signal-to-noise ratio of the power of the bandpassed signal. Therefore, methods that can resize and shift the passband based on the location of the central frequency can more reliably estimate the power over time.

While not yet in common use, there are many possible benefits to frequency shift features. The influence of artifacts on the power spectrum is well-known and, in setups with few channels, very difficult to get rid of. Transient, high-amplitude artifacts in particular are problematic since they have a disproportionate effect on the variance of a signal. It may be, however, that the frequency shift is less affected by this issue. We proposed to look at the change in peak frequency of the EEG as a feature for a classifier, both in motor imagery and in a cognitive paradigm [40]. This idea only showed up once before, and was validated on a small motor imagery dataset [41], but we were interested in seeing if this is a valid approach across multiple paradigms.

In order to isolate the peak frequency on a per-trial basis, we elected to use the analytic signal. The analytic signal is computed via the Hilbert transform and leads to a representation of the signal that includes both amplitude and phase information. By extracting the phase and taking the difference between every pair of neighboring timepoints, it is possible to

extract an instantaneous frequency for a given time series at every time point. Since this estimate of the instantaneous frequency is noisy due both to edge artifacts and discretization issues, we chose the median of the instantaneous frequencies within each trial as the estimate of the peak frequency [40]. A similar approach was used in Cohen et al [42] to show the perceptual relevance of instantaneous frequency in EEG and artifical network models.

Our results showed that the location of the frequency peak as computed via the median instantaneous frequency for each trial is a noisy but helpful predictor in both motor imagery and cognitive paradigms. In particular in the case of low-channel settings, we show that adding the median instantaneous frequency can reliably help classification in a variety of frequency bands [43]. In reference to the signal processing literature, we refer to this feature as the FM feature.

While the success in cognitive tasks can be explained by the literature on alpha peak location, the results on motor imagery are a new finding. Our results with CSP suggest that neural populations that show event-related desynchronization also have a shift downwards in their peak alpha frequency[40], which has not yet been described even in invasive studies. While work exists in monkeys that shows how waves of activity at different frequencies propagate across the primary motor cortex [44], there is as of yet no description of how the frequency of these waves is modulated. It is further quite interesting to ask why the frequency and power are correlated, as this may have implications for the circuit properties underlying the α rhythm.

1.4 How to validate? Principled use of open-access data

It is rarely clear, when developing methods for BCIs, how well these methods transfer to other labs or hardware. Since there is so much variance in everything from sampling rate to electrode number to dry versus wet electrodes, positive results on one platform don't necessarily imply positive results on another. In extreme cases, results that work on the vast majority of datasets can still fail for one particular lab, as seen in Figure 1.3. Unfortunately, this particular question is very difficult to phrase within the field of BCIs. Most labs have their own preferred setup, developed over generations of doctoral students, and their own pipeline for processing and saving that data. Trying to compare across many labs can be quite difficult given these methodological differences. This has resulted in the condition that there is little incentive for newly proposed algorithms to be taken up by the community, as even after

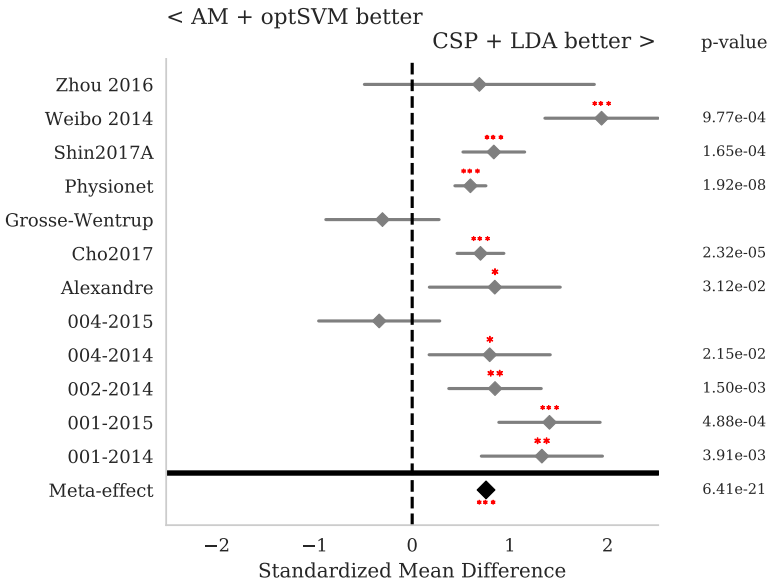

Figure 1.3: Figure, taken from Jayaram et al. [45]. This shows a comparison of log-variances per channel combined with an SVM versus CSP and LDA as a method for classifying motor imagery across 13 open-access datasets. Even though CSP performs significantly better in most datasets, there are still some–such as Grosse-Wentrup et al. 2009[46] –which go against the trend. This highlights the need for validation over many hardware types to convince labs that a new algorithm truly will work better.

the work of implementing them for a lab's specific setup there is a reasonable chance that they will not improve performance.

The closest previous attempt to standardize how BCI algorithms are compared is the BCI competitions datasets [47]–[50], which are as of now over 10 years old. Thousands of papers have been written using only these datasets, which makes it very likely that the results by now are heavily overfitted to that specific combination of electrodes, paradigm, and sensor number. Over this same period, however, thousands of hours of BCI data have been recorded all over the world, and hundreds of those hours are available freely online – the issue is simply that nobody has spent the time to gather them and run analyses over all of them. This is the goal of Jayaram et al. [45], in which we introduce the MOABB, an open-source codebase that allows for replicable result generation, combined with a statistical procedure in order to reliably rank compared algorithms. Given the results shown in Figure 1.3, we hope that offering this codebase to the community will allow for future

algorithic advances to be better received.

The first analysis with this method is centered around single-recording classification performance, as that is the field in which the majority of previous work has been done. By verifying this procedure with well-known pipelines from the field, done in such a way that anyone can reproduce the results, we hope to convince the community that the rankings that come out of our analysis are trustworthy. With trust comes two things: The ability to conduct further studies, and hopefully also an increase in publically available data. If our results in a more well-studied subfield of EEG classification pass peer review, we can continue on to test transfer learning and novel feature spaces safe in the knowledge that the results thereof are reliable. Ifthe community is not satisfied with our processing or our statistical analysis, then there is no point in running more experiments with the same analysis. Further, and more optimistically, we hope that providing a platform for labs around the world to easily use open-access data makes experimental labs more willing to share their data, since they will know that groups can work on it. With more data we could expand beyond questions of algorithm and ask more interesting questions about how hardware, paradigm, and even personal characteristics affect the EEG across hundreds of subjects.

1.5 Conclusion: Moving forwards

When looking at how machine learning could help EEG-based BCIs step out of the lab, two key areas stood out: developing methods for using pools of offline data, and searching for new features that are robust to less ideal recording conditions. In order to validate the results of these explorations, we took advantage of the large amount of open-access data and built a validation framework for comparing different pipelines. However, although the methods introduced in this thesis have many interesting properties, comparing them directly against modern variants of CSP, or modern amplitude estimation methods, is very likely to show that their performance is worse. This may be because the ideas they suggest are less effective, but it is equally likely that it is because so much more work has been expended on the other directions. Because of this, the remainder of this section will first detail, for the transfer learning and frequency modulation, technical possibilities for improving these methods, such that this latter possibility can be dealt with. Finally, a more unified future outlook will be presented.

1.5.1 Transfer Learning

Both the bilinear regression model and the choice of a Gaussian distribution for the probabilities are simplifications that allow for easy inference and interpretation. As such, these can both be extended into more powerful approaches, often losing little in terms of simplicity.

A bilinear model as described in Jayaram et al. [18] can equivalently be thought of as a linear model in the feature space where the regression matrix is constrained to be of rank 1. As fixed-rank manifolds are not convex, this means that optimization over them is also non-convex, which is the reason that a nested iterative approach is described in [18]. Relaxing this constraint has two major benefits. First, it allows for multiple spatial and feature weightings. It was determined that in CSP, the optimal number of spatial filters is six [51], and it is therefore likely that one is not the optimal number of spatial weightings in this model. By using a rank-based penalty instead of a rank constraint as used in Farquhar et al. [52], a flexible number of spatial and feature weightings can be estimated. In order to visualize the individual pairs as was done in Figure 1.2, a matrix decomposition can be used.

Next, we might consider the Gaussianity assumption on both the task-specific noise and the prior over the weight vectors. Two ways of extending this approach are the use of hyper priors and the use of other distributional forms. One simple example of a hyper prior would be one that assumes the mean vector comes from a zero-mean, isotropic Gaussian. A downside of the approach described in Jayaram et al. [18] is that the regularization penalizes deviation from a mean vector, but the norm of this mean vector is not penalized. Via a zero-mean prior over the computed mean, this could be remedied. In addition, the current approach penalizes the Mahalanobis distance between task weight vectors and the prior; instead of this, we may want to choose to penalize a difference distance between tasks and the mean, in order to promote properties such as sparsity in our solution. There is already work by Argyriou et al. [53] in which a convex multi-task approach with sparsity is derived, and this can be very easily substituted into the current algorithm.

Using different distributional assumptions can also be easily done. Assuming a logistic model, in which a squared loss function is traded for a cross-entropy loss, has been shown to lead to increases in performance when compared to a fully linear model [54]. It may also be possible to derive a solution when the loss function is assumed to be the hinge loss, effectively turning each task-specific problem into a maximum-margin optimization. Another approach would be to use different assumptions about the prior, such as a Laplace distribution, in order to more robustly compute the mean and co-

variance within each iteration.

A larger concern in relation to this method is that it requires pre-computed features per channel. As work with ICA and CSP shows, linear filters applied to the electrodes in the time domain can lead to very predictive features. Indeed, while we showed that using multi-task learning it was possible to find an optimal classifier in a fixed feature space with very little data, post-hoc comparisons with subject-specific CSP showed that a linear model over the channel-wise features is not ideal for within-session classification. If instead of pre-computing the features, we were to extend this model to automatically determine relevant features from the time domain, it would be significantly more powerful. However, it is not entirely clear how this could be done. One possibility would be to extend the model from Farquhar et al. [52] with the probabilistic multi-task approach.

When comparing so many different possible models, the possibility of overfitting to a single dataset becomes more and more dangerous. Via the MOABB codebase, comparisons of all these methods can be done with a minimal risk of overfitting [45].

1.5.2 Frequency Modulation

We validated the idea that neural oscillations change their oscillatory frequency in response to both cognitive tasks and motor ones, and more crucially that this difference is recognizable via the instantaneous frequency per channel. Ideally, this finding reflects a property of the underlying neural signals. However, the Hilbert transform of a mixed signal is not well-defined, in the sense that it is unclear which of the underlying components is being reflected in the estimated instantaneous phase. If the recovered frequency and amplitude were independent then this would not be a problem, but our results showed that they correlate. Therefore, more robust model-based methods of estimating the instantaneous frequency need to be tested as well, in order to determine if this is simply an artifact of neural amplitude change or if the underlying signal is also frequency modulated. Cognitive psychology has good evidence of this shift in cognitive tasks, but frequency shift has yet to be described in vivo for motor imagery.

One important goal of the work with frequency modulation is to see if it is more robust to differences in signal quality than amplitude modulation. Unfortunately, the current study was done on data recorded under laboratory conditions with wet electrodes, and so another comparison with a noisier dataset is required. Re-running the analysis from Jayaram et al. [40] with the framework provided by the MOABB codebase is a first step in this direction. Further, while there were no conclusive findings in the analysis of ALS patient

data, a larger sample size may help account for the confounding effect of disease progression.

The fact that frequency and amplitude information are correlated in both of the tested paradigms leads to another interesting question, of whether this can be used as a way of characterizing neural data. There are few reliable ways to characterize data that is neural in origin from just a time-series or spectrum – a problem that has been heavily investigated in the field of ICA for EEG [55]. If there are reliable amplitude-frequency couplings in neural signals, and if these are different from muscle or noise signals, this could be exploited to realize new signal decomposition approaches.

Lastly, it has been shown that feedback on the power in the upper α band can decrease cognitive impairments in some cases [56]. Giving feedback on the power in only the upper part of the α band may be equivalent to giving feedback that pushes the α frequency higher, and it may be that using the frequency location instead of the half-width bandpower is a more reliable signal.

1.5.3 Outlook

One of the major problems causing BCIs to remain in the laboratory is their inconsistency and the frustration of trying to use them habitually. Transfer learning represents a path towards optimal use of past data, such that machine learning models perform more stably. In addition, new features may also be less susceptible to the major sources of noise within the signal. And, finally, it is very expensive to test new models in a closed-loop BCI in terms of the manpower involved. Better validation via precollected data allows one to conduct online experiments with provably robust machine learning, allowing users and developers to concentrate on the other areas of the system.

Looking out from the work of my doctorate, there are both methodological and experimental directions to be followed. The model for transfer learning can be extended in many ways, and with a robust procedure for validating these ways it is possible to determine the optimal way of using this scheme to aid in BCI decoding. On the other side, the implications of frequency modulation in task-based experiments have not yet been explored, and there is therefore a great deal of room to see what the benefits of using these features over traditional amplitude features are. In addition to both of these improvements done independently, however, there is also an obvious question of what happens when both the transfer learning and frequency modulation are combined. While a simple idea in theory, in practice this requires a large pool of subjects doing recordings, ideally longitudinally, and that is still a difficult proposition in brain-computer interfacing. No low-cost hardware

has been optimized to the point where such a study is feasible outside a doctoral project, but the future is bright in this regard. There has already been some work showing how commercial, low-cost EEG may be suitable for BCI applications [57], [58]. In combination with some of the innovations presented here, it may be soon that BCIs finally leave the lab after all.

Bibliography

[1] A. R. Murguialday, J. Hill, M. Bensch, *et al.*, "Transition from the locked in to the completely locked-in state: A physiological analysis," *Clinical Neurophysiology*, vol. 122, no. 5, pp. 925–933, 2011, ISSN: 1388-2457. DOI: http://dx.doi.org/10.1016/j.clinph.2010.08.019.

[2] P. Shenoy, M. Krauledat, B. Blankertz, *et al.*, "Towards adaptive classification for BCI," *Journal of Neural Engineering*, vol. 3, no. 1, R13–R23, Mar. 2006, ISSN: 1741-2560. DOI: 10.1088/1741-2560/3/1/R02.

[3] J. Millan, "On the need for on-line learning in brain-computer interfaces," in *2004 IEEE International Joint Conference on Neural Networks*, vol. 4, IEEE, pp. 2877–2882, ISBN: 0-7803-8359-1. DOI: 10.1109/IJCNN.2004.1381116.

[4] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topography*, vol. 2, no. 4, pp. 275–284, 1990.

[5] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[6] J. Faller, C. Vidaurre, T. Solis-Escalante, *et al.*, "Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 3, pp. 313–319, May 2012, ISSN: 1534-4320. DOI: 10.1109/TNSRE.2012.2189584.

[7] C. Vidaurre, C. Sannelli, K.-R. Müller, *et al.*, "Machine-learning-based coadaptive calibration for brain-computer interfaces," *Neural Computation*, vol. 23, no. 3, pp. 791–816, Mar. 2011, ISSN: 0899-7667. DOI: 10.1162/NECO_a_00089.

[8] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, *et al.*, "Multisubject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, 2011.

[9] M. Cheng, Z. Lu, and H. Wang, "Regularized common spatial patterns with subject-to-subject transfer of EEG signals.," *Cognitive neurodynamics*, vol. 11, no. 2, pp. 173–181, Apr. 2017, ISSN: 1871-4080. DOI: 10.1007/s11571-016-9417-x.

[10] A. Barachant, S. Bonnet, M. Congedo, *et al.*, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013, ISSN: 0925-2312. DOI: 10.1016/J.NEUCOM.2012.12.039.

[11] ——, "Common spatial pattern revisited by Riemannian geometry," in *2010 IEEE International Workshop on Multimedia Signal Processing*, IEEE, Oct. 2010, pp. 472–476, ISBN: 978-1-4244-8110-1. DOI: 10.1109/MMSP.2010.5662067.

[12] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *The Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[13] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, 2014.

[14] ——, "Bayesian common spatial patterns with Dirichlet process priors for multi-subject EEG classification," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2012, pp. 1–6.

[15] ——, "Bayesian multi-task learning for Common Spatial Patterns," in *2011 International Workshop on Pattern Recognition in NeuroImaging*, IEEE, May 2011, pp. 61–64, ISBN: 978-1-4577-0111-5. DOI: 10.1109/PRNI.2011.8.

[16] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 17–24.

[17] V. Jayaram, K.-H. Fiebig, J. Peters, *et al.*, "Transfer learning for BCIs," *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*, p. 425, 2018.

[18] V. Jayaram, M. Alamgir, Y. Altun, *et al.*, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[19] M. A. Lopez-Gordo, D. Sanchez-Morillo, and F. P. Valle, "Dry EEG electrodes," *Sensors*, vol. 14, no. 7, pp. 12 847–12 870, Jul. 2014, ISSN: 1424-8220. DOI: 10.3390/s140712847.

[20] M. Spüler, "A high-speed brain-computer interface (BCI) using dry EEG electrodes," *PLOS ONE*, vol. 12, no. 2, D. Zhang, Ed., e0172400, Feb. 2017, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0172400.

[21] A. Pinegger, S. C. Wriessnegger, J. Faller, *et al.*, "Evaluation of different EEG acquisition systems concerning their suitability for building a braincomputer interface: Case studies," *Frontiers in Neuroscience*, vol. 10, p. 441, Sep. 2016, ISSN: 1662-453X. DOI: 10.3389/fnins.2016.00441.

[22] F. Popescu, S. Fazli, Y. Badower, *et al.*, "Single trial classification of motor imagination using 6 dry EEG electrodes," *PLOS ONE*, vol. 2, no. 7, C. Miall, Ed., e637, Jul. 2007, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000637.

[23] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, Jan. 2017, ISSN: 1746-8094. DOI: 10.1016/J.BSPC.2016.09.005.

[24] A. Bragin, G. Jandó, Z. Nádasdy, *et al.*, "Gamma (40-100 Hz) oscillation in the hippocampus of the behaving rat.," *The Journal of Neuroscience*, vol. 15, no. 1 Pt 1, pp. 47–60, Jan. 1995, ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.15-01-00047.1995.

[25] N. Axmacher, M. M. Henseler, O. Jensen, *et al.*, "Cross-frequency coupling supports multi-item working memory in the human hippocampus.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 3228–33, Feb. 2010, ISSN: 1091-6490. DOI: 10.1073/pnas.0911531107.

[26] D. J. A. Smit, D Posthuma, D. I. Boomsma, *et al.*, "Heritability of background EEG across the power spectrum," *Psychophysiology*, vol. 42, no. 6, pp. 691–697, 2005.

[27] S. Bodenmann, T. Rusterholz, R. Dürr, *et al.*, "The functional Val158Met polymorphism of COMT predicts interindividual differences in brain α oscillations in young men," *The Journal of Neuroscience*, vol. 29, no. 35, pp. 10 855–10 862, 2009.

[28] S. Haegens, H. Cousijn, G. Wallis, *et al.*, "Inter-and intra-individual variability in alpha peak frequency," *Neuroimage*, vol. 92, pp. 46–55, 2014.

[29] E. Angelakis, J. F. Lubar, and S. Stathopoulou, "Electroencephalographic peak alpha frequency correlates of cognitive traits," *Neuroscience Letters*, vol. 371, no. 1, pp. 60–63, Nov. 2004, ISSN: 0304-3940. DOI: 10.1016/J.NEULET.2004.08.041.

[30] T. H. Grandy, M. Werkle-Bergner, C. Chicherio, *et al.*, "Individual alpha peak frequency is related to latent factors of general cognitive abilities," *Neuroimage*, vol. 79, pp. 10–18, Oct. 2013, ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2013.04.059.

[31] H Aurlien, I. O. Gjerde, J. H. Aarseth, *et al.*, "EEG background activity described by a large computerized database," *Clinical Neurophysiology*, vol. 115, no. 3, pp. 665–673, 2004.

[32] E. Angelakis, J. F. Lubar, S. Stathopoulou, *et al.*, "Peak alpha frequency: An electroencephalographic measure of cognitive preparedness.," *Clinical neurophysiology : Official journal of the International Federation of Clinical Neurophysiology*, vol. 115, no. 4, pp. 887–97, Apr. 2004, ISSN: 1388-2457. DOI: 10.1016/j.clinph.2003.11.034.

[33] R. Solís-Vivanco, O. Jensen, and M. Bonnefond, "Top-down control of alpha phase adjustment in anticipation of temporally predictable visual stimuli," *Journal of Cognitive Neuroscience*, vol. 30, no. 8, pp. 1157–1169, Aug. 2018, ISSN: 0898-929X. DOI: 10.1162/jocn_a_01280.

[34] D. P. Nguyen, M. A. Wilson, E. N. Brown, *et al.*, "Measuring instantaneous frequency of local field potential oscillations using the Kalman smoother," *Journal of Neuroscience Methods*, vol. 184, no. 2, pp. 365–374, 2009.

[35] M. J. Prerau, P. L. Purdon, and U. T. Eden, "Tracking non-stationary spectral peak structure in EEG data," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Jul. 2013, pp. 417–420, ISBN: 978-1-4577-0216-7. DOI: 10.1109/EMBC.2013.6609525.

[36] R. Scherer, J. Faller, E. V. C. Friedrich, *et al.*, "Individually adapted imagery improves brain-computer interface performance in end-users with disability," *PLOS ONE*, vol. 10, no. 5, L. Bianchi, Ed., e0123727, May 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123727.

[37] M. Hohmann, T. Fomina, V. Jayaram, *et al.*, "A cognitive braincomputer interface for patients with amyotrophic lateral sclerosis," *Progress in Brain Research*, vol. 228, pp. 221–239, Jan. 2016, ISSN: 0079-6123. DOI: 10.1016/BS.PBR.2016.04.022.

[38] B. Hamner, R. Leeb, M. Tavella, *et al.*, "Phase-based features for motor imagery brain-computer interfaces," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, pp. 2578–2581.

[39] O. Falzon, K. P. Camilleri, and J. Muscat, "The analytic common spatial patterns method for EEG-based BCI data," *Journal of neural engineering*, vol. 9, no. 4, p. 45 009, 2012.

[40] V. Jayaram, M. Hohmann, J. Just, *et al.*, "Task-induced frequency modulation features for brain-computer interfacing," *Journal of Neural Engineering*, vol. 14, no. 5, p. 056 015, Oct. 2017, ISSN: 1741-2560. DOI: 10.1088/1741-2552/aa7778.

[41] A. Médl, D. Flotzinger, and G. Pfurtscheller, "Hilbert-transform based predictions of hand movements from EEG measurements," in *Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE*, IEEE, vol. 6, 1992, pp. 2539–2540.

[42] M. X. Cohen, "Fluctuations in oscillation frequency control spike timing and coordinate neural networks.," *The Journal of Neuroscience*, vol. 34, no. 27, pp. 8988–98, Jul. 2014, ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.0261-14.2014.

[43] V. Jayaram, B. Schölkopf, and M. Grosse-Wentrup, "Frequency peak features for low-channel classification in motor imagery paradigms," in *Proceedings of the 8th IEEE EMBS Conference on Neural Engineering*, 2017.

[44] D. Rubino, K. A. Robbins, and N. G. Hatsopoulos, "Propagating waves mediate information transfer in the motor cortex," *Nature Neuroscience*, vol. 9, no. 12, pp. 1549–1557, Dec. 2006, ISSN: 1097-6256. DOI: 10.1038/nn1802.

[45] V. Jayaram and A. Barachant, "MOABB: trustworthy algorithm benchmarking for BCIs," May 2018. arXiv: 1805.06427.

[46] M. Grosse-Wentrup, C. Liefhold, K. Gramann, *et al.*, "Beamforming in non-invasive brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1209–1219, Apr. 2008, ISSN: 0018-9294. DOI: 10.1109/TBME.2008.2009768.

[47] M. Tangermann, K.-R. Müller, A. Aertsen, *et al.*, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, p. 55, Jul. 2012, ISSN: 1662-4548. DOI: 10.3389/fnins.2012.00055.

[48] B Blankertz, K. R. Müller, D Krusienski, *et al.*, "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 153–159, 2006.

[49] A Schloegl, "Results of the BCI competition 2005 for data set IIIa and IIIb," Institute for Human-Computer Interfaces - BCI Lab, University of Technology Graz, Austria, Tech. Rep., 2005.

[50] G Blanchard and B Blankertz, "BCI competition 2003 - data set IIa: Spatial patterns of self-controlled brain rythm modulations," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1062–1066, 2004.

[51] B. Blankertz, R. Tomioka, S. Lemm, *et al.*, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008, ISSN: 1053-5888. DOI: 10.1109/MSP.2008.4408441.

[52] J. Farquhar, "A linear feature space for simultaneous learning of spatio-spectral filters in BCI," *Neural Networks*, vol. 22, no. 9, pp. 1278–1285, Nov. 2009, ISSN: 0893-6080. DOI: 10.1016/J.NEUNET.2009.06.035.

[53] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, Dec. 2008, ISSN: 0885-6125. DOI: 10.1007/s10994-007-5040-8.

[54] K.-H. Fiebig, V. Jayaram, J. Peters, *et al.*, "Multi-task logistic regression in brain-computer interfaces," in *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics*, 2016.

[55] B. W. McMenamin, A. J. Shackman, J. S. Maxwell, *et al.*, "Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG," *Neuroimage*, vol. 49, no. 3, pp. 2416–2432, 2010, ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2009.10.010.

[56] C. Escolano, M. Aguilar, and J. Minguez, "EEG-based upper alpha neurofeedback training improves working memory performance," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2011, pp. 2327–2330, ISBN: 978-1-4577-1589-1. DOI: 10.1109/IEMBS.2011.6090651.

[57] S. Debener, F. Minow, R. Emkes, *et al.*, "How about taking a low-cost, small, and wireless EEG for a walk?" *Psychophysiology*, vol. 49, no. 11, pp. 1617–1621, Nov. 2012, ISSN: 00485772. DOI: 10.1111/j.1469-8986.2012.01471.x.

[58] M. Duvinage, T. Castermans, M. Petieau, *et al.*, "Performance of the Emotiv Epoc headset for P300-based applications," *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 56, Jun. 2013. DOI: 10.1186/1475-925X-12-56.

Chapter 2

Collected works

2.1 Contributions

1. **Vinay Jayaram** , Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. IEEE Computational Intelligence Magazine 11, no. 1 (2016): 20-31.

   The idea of a hierarchial Gaussian model was introduced by Morteza Alamgir, Yasemin Altun, and Prof. Grosse-Wentrup and published at a conference in 2010 [16]. I added a task-specific bilinear model and derived the optimization strategy for this approach. I participated in one of the recordings for the ALS patient data and did not participate in the recording of the motor imagery data. All coding and analysis was performed by me, in addition to the writing of the manuscript in the lab of Prof. Schölkopf

2. **Vinay Jayaram**, Matthias Hohmann, Jennifer Just, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Task-induced frequency modulation features for brain-computer interfacing. Journal of Neural Engineering 14, no. 5 (2017): 056015.

   For this manuscript the original idea for investigating the frequency shift came from qualitative analysis of ALS patient data in conjunction with Prof. Grosse-Wentrup. All code and analysis was done by me, in addition to the decision to test this feature on healthy subject data in

both cognitive and motor paradigms. Matthias recorded the data used in the cognitive task analysis and Jennifer, Matthias, and I assisted with the recordings with ALS patients. All work was done with the support of Prof. Schölkopf

3. **Vinay Jayaram**, Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. arXiv preprint (2018). **Submitted to the Journal of Neural Engineering in May 2018**.

   The initial idea came from Dr. Alexandre Barachant, but the codebase as it is available was a fully joint project between the two of us. The idea for this manuscript, and the specific analysis we did within it, was initiated by me and done in collaboration with Dr. Barachant. For the statistical analysis we consulted with Dr. Marco Congedo.

4. Fiebig, Karl-Heinz, **Vinay Jayaram**, Jan Peters, and Moritz Grosse-Wentrup. Multi-task logistic regression in brain-computer interfaces. In Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, pp. 002307-002312. IEEE, 2016. ©2016 IEEE, reprinted with permission.

   The initial idea was proposed in a joint session between myself, Prof. Grosse-Wentrup, and Karl-Heinz. Karl-Heinz did the derivations and wrote the analysis and manuscript, while I supervised and reviewed all three. Karl-Heinz was also supervised by Prof. Peters.

2.2 Paper 1: Transfer Learning in Brain-Computer Interfaces

Transfer Learning in Brain-Computer Interfaces

Vinay Jayaram

IMPRS for Cognitive and Systems Neuroscience, University of Tübingen, Tübingen, Germany

Department of Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Morteza Alamgir

Department of Computer Science, University of Hamburg, Hamburg, Germany

Yasemin Altun

Bernhard Schölkopf

Moritz Grosse-Wentrup

Department of Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

The performance of brain-computer interfaces (BCIs) improves with the amount of available training data; the statistical distribution of this data, however, varies across subjects as well as across sessions within individual subjects, limiting the transferability of training data or trained models between them. In this article, we review current transfer learning techniques in BCIs that exploit shared structure between training data of multiple subjects and/or sessions to increase performance. We then present a framework for transfer learning in the context of BCIs that can be applied to any arbitrary feature space, as well as a novel regression estimation method that is specifically designed for the structure of a system based on the electroencephalogram (EEG). We demonstrate the utility of our framework and method on subject-to-subject transfer in a motor-imagery paradigm as well as on session-to-session transfer in one patient diagnosed with amyotrophic lateral sclerosis (ALS), showing that it is able to outperform other comparable methods on an identical dataset.

1 Introduction

It is often a problem in various fields that one runs into a series of tasks that appear - to a human - to be highly related to each other, yet applying the optimal machine learning solution of one problem to another results in poor performance. Specifically in the field of brain-computer interfaces (BCIs), it has long been known that a subject with good classification of some brain signal today could come into the experimental setup tomorrow and perform terribly using the exact same classification function. One initial approach to get over this problem was to fix the classification rule beforehand and train the patient to force brain activity to conform to this rule. For example, Wolpaw et al. in the early 90's chose weights for the α and μ rhythms and trained participants to modulate the bandpower in these frequency bands in order to control

Corresponding author: Vinay Jayaram (email: vjayaram@tuebingen.mpg.de)

a cursor [1], [2]. Similarly, Birbaumer et al. trained a patient to create large depolarizations of the electroencephalogram (EEG) over the course of several seconds, using a simple threshold on the bandpassed raw signal [3]. In their time, both approaches were successful, but took training time on the order of months to master. To overcome this limitation, several groups introduced machine learning techniques for adapting BCIs to their users [4]–[10]. They successfully managed to learn decoding rules with high acccuracy using only a fraction of the training trials required by the earlier approaches, allowing subjects to communicate consistently with a computer in a single session. Unfortunately, a training period had to be repeated at the beginning of each usage session as the learned discrimination rules were not immediately stable. A naive solution to this limitation was to pool training data from multiple recordings; however, the statistical distributions of these data varies across subjects as well as across sessions within individual subjects, giving this approach varying effectiveness. In recent years, several groups have started explicitly modelling such variations to exploit structure that is shared between data recorded from multiple subjects and/or sessions. In this article, we provide an overview of previous work on the topic and present a unifying approach to transfer learning in the field of BCIs. We demonstrate the utility of our framework on subject-to-subject transfer in a motor-imagery paradigm as well as on session-to-session transfer in one patient diagnosed with amyotrophic lateral sclerosis (ALS).

1.1 Previous work

Transfer learning describes the procedure of using data recorded in one task to boost performance in another, related task (for a more exhausive review of the machine learning literature, see [11]). That is to say, we assume *a priori* that there is some structure shared by these tasks; the goal, then, is to learn some representation of this structure so further tasks can be solved more easily. In the context of BCIs, transfer learning is of critical importance - it has long been known that the EEG signal is not stationary, and so in its strictest sense one can consider every trial a slightly new task. As such, long sessions of BCI usage present unique problems in terms of consistent classification [12]. The question is how to transfer some sort of knowledge between them: a question that can be answered in one of two general ways. Either we can attempt to find some structure in the data that is invariant across datasets or we can find some structure in how the decision rules differ between different subjects or sessions. We denote these as *domain adaptation* and *rule adaptation* respectively (Figure 1).

Looking at the literature, BCI has been almost exclusively dominated by domain adaptation approaches. One popular feature space in the field is the trial covariance matrices used both in Common Spatial Patterns (CSP) [4], [13] and other more modern methods [14]. Many transfer learning techniques have been attempted with CSP, mostly relying on an assumption that there exists a set of linear filters that is invariant across either sessions or subjects. An early example of session-to-session transfer of spatial filters is the work by Krauledat et al. [15], in which a clustering procedure is employed to select prototypical spatial filters and classifiers, which are in turn applied to newly recorded data. Using this approach, the authors demonstrate that calibration time can be greatly reduced with only a slight loss in classification accuracy. The problem of subject-to-subject transfer of spatial filters is addressed by Fazli et al. [16]: also building upon CSP for spatial filtering, the authors utilize a large database of pairs of spatial filters and classifiers from 45 subjects to learn a sparse subset of these pairs that are predictive across subjects. Using a leave-one-subject-out cross-validation procedure, the authors then demonstrate that this sparse subset of spatial filters and classifiers can be applied to new subjects with only a moderate performance loss in comparison to subject-specific calibration. Note that in both above approaches transfer

learning amounts to determining invariant spaces on which to project the data and learning classifiers in these spaces. This line of work has been further extended by Kang et al. [17], [18], Lotte and Guan [19], and Devlaminck et al. [20]. In these contributions, the authors demonstrate successful subject-to-subject transfer by regularizing spatial filters derived by CSP with data from other subjects, which amounts to attempting to find an invariant subspace on which to project the data of new subjects. Recently, a method of distance measures between trial covariance matrices has also been used to great effect in both motor imagery [21] and event-related potential paradigms [22] as a domain adaptation tool. Related to the spirit of the regularized CSP methods described above, they work by trying to find the best projection plane for the trial covariance matrices, invariant to subjects and sessions, and then run a classification algorithm. Other domain adaptation approaches include that by Morioka et al. [23], in which an invariant sparse representation of the data is learned using many subjects and then the transformation into that space is applied to new subjects, and the technique of stationary subspace analysis [24], [25], which attempts to find a stationary subspace of the data from multiple subjects and/or sessions.

A very related technique to domain adaptation is *covariate shift*, which has also found use in BCIs. Sugiyama et al. have used covariate shift adaptation to combine labeled training data with unlabeled test data [26]. Here, it is assumed that the marginal distribution of the data changes between the subjects and/or sessions, but the decision rule with respect to this marginal distribution remains constant. This assumption leads to a re-weighting of training data from other subjects and/or previous sessions based on unlabeled data from the current test set that corrects for covariate shifts–in essence, correcting for the difference in marginal distributions in the different subjects and/or sessions. In addition to their results, several other authors have also reported improvements in BCI decoding performance by using similar techniques for covariate shift adaptation [27]–[29]. Other techniques such as boosting [30] have also used re-labelling of offline data to increase performance [31].

The covariate shift and other methods presented in the previous paragraph represent a very different assumption about the tasks than the methods that attempt to find an invariant space to project the data. Instead of assuming that there exists some space where the data already lives that is invariant for all individuals or across all time, it attempts to model the variation between individuals and efficiently discover a transformation for new individuals to the known space (in comparison, an invariant subspace could be seen as applying an identical transform to all individuals). This approach of attempting to learn a representation of the variability is most naturally attempted in the space of possible rules, since it often offers a ready-made parametrization of the approximating function. One possibility for such modelling is to treat the parameters of a decoding model as random variables that are, for each subject and/or session, drawn from the same distribution. The prior distribution of the model parameters can then be used to link training data across multiple subjects and/or sessions, and be learned by a simultaneous optimization over previous subjects and/or sessions. Rule adaptation of this sort has been attempted in Kinderman et al. [32], which attempts to learn a classification prior in the P300 task, but restricts the covariance to multiples of the identity while it allows the mean to be determined by the distribution of subject weight vectors. A framework of *multitask learning* which attempts to learn a full distribution has been introduced to the field of BCIs by Alamgir et al. [33]. Specifically, the authors treat classification as a linear regression problem and model the regression weights as a random variable that is drawn from a multivariate Gaussian distribution with unknown mean and covariance matrix. By jointly estimating the parameters of this distribution and regression weights for multiple subjects, they demonstrate a substantial improvement in decoding performance in a motor-imagery paradigm. However,

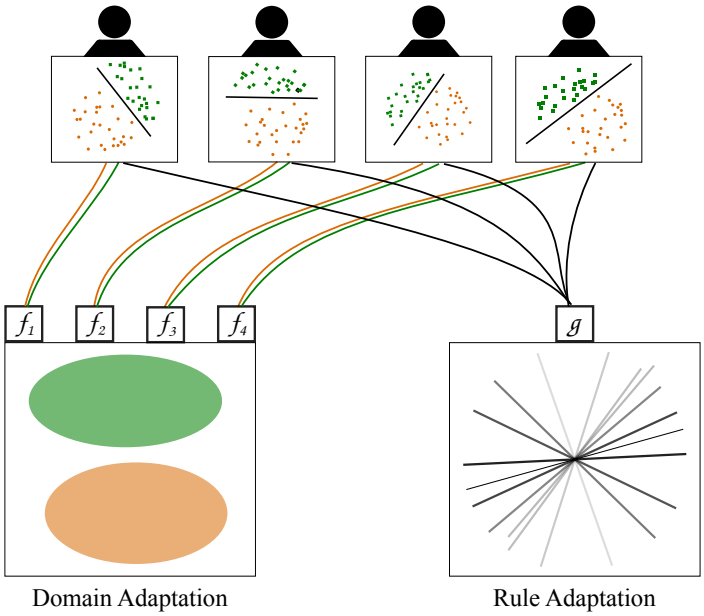

Fig. 1: Given a set of training datasets (top) there are two ways to model the similarities shared by them. *Domain adaptation* (left) refers to the strategy of attempting to find a transformation to a data space in which a single decision rule will classify all samples. Instead of learning a new rule for the new data, data is simply transformed to the invariant space. *Rule adaptation* (right) is the strategy of attempting to learn the structure of the classification rules. New datasets are faced with a much smaller search space of possible rules which allows for much faster learning of novel decision boundaries.

this work suffered from various limitations. In modelling each channel bandpower as a separate feature it became necessary to employ channel selection in a pre-processing step, and also to attempt to isolate and remove noisy subjects from the training pool. In this work we extend the previous results of multitask learning with a new technique that is robust both to subjects who perform poorly and to an extremely high-dimensional feature space.

2 A General Framework for Transfer Learning in BCIs

In this article, we build upon our prior work on multitask learning [33] to derive a general framework for transfer learning in BCI, applicable to any spatiotemporal feature space and able to be used on multi-session and multi-subject data equally, and further introduce a BCI-specific method for reducing the feature space dimension.

2.1 Preliminaries

In this section, we introduce the decoding model used throughout this work. We index multiple subjects or recording sessions by $s = \{1, \ldots, S\}$ and assume that for each subject/session we are given data from n_s trials, $D_s = \{(\mathbf{x}_s^i, y_s^i)\}_{i=1}^{n_s}$. Here, $\mathbf{x}_s^i \in \mathbb{R}^d$ refers to the features derived from the recorded brain signals of subject/session s during trial i, with d denoting the number of features. For the datasets presented in this article, $\mathbf{x}_s^i$ consists of EEG log-bandpower estimates at different scalp locations; however, it is equally applicable to timepoints after event onset if the signal of interest is an event-related potential. More specifically, if the number of electrodes is E and the number of EEG log-bandpower estimates is F, the number of features is $d = E \times F$. Variable y_s^i denotes the subject's stimulus, e.g., motor imagery of either the left

or right hand in trial i of session s. As we furthermore only deal with two-class paradigms, we let $y_s^i \in \{-1, 1\}$ for all i and s, though this framework is applicable to regression problems as well.

Assuming our model is linear with a noise term η, we can model our data by a linear function

$$y_s^i = \mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^i + \eta$$

associated to each subject/session s, where the parameters $\mathbf{w}_s$ constitute the weights assigned to the individual features that are used to predict the stimulus for trials in subject/session s. Given a new brain signal $\mathbf{x}$ for subject/session s, the stimulus is predicted by

$$\hat{y}_s^{i+1} = \text{sign}\{\mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^{i+1}\}. \tag{1}$$

We first investigate training $\mathbf{w}_s$ for each subject independently in Section 2.2 and extend this formalism to train $\mathbf{w}_s$ jointly on multiple subjects/sessions in Section 2.3.

2.2 Training Models for Subjects/Sessions Independently

When faced with some set of data and labels, the goal is to determine the parameters $\mathbf{w}_s$ that allow for the best prediction of the labels from the data. Mathematically speaking, for each subject/session s, the parameters $\mathbf{w}_s$ are determined such that the number of errors in the dataset of subject/session s, D_s, is small. The choice of how to define 'errors' for a given set of predictions can drastically influence both the values of the final parameters and the ease with which they can be found; in machine learning, this is called a *loss function* and by finding the minimum of this function we can recover the parameters that result in the lowest defined error. The most commonly used loss functions to calculate errors are convex proxies such as log-loss, hinge loss or least squares loss [34]; in this paper, we use least squares loss, which we arrive at naturally with the assumption that the error term η is distributed as $\mathcal{N}(0, \sigma^2)$.

To begin, let us consider a probabilistic interpretation of the problem. Using Bayes Rule, the probability of our parameters given our data decomposes as follows (note that we ignore the possible dependence of the prior $p(\mathbf{w}_s)$ on $\mathbf{x}_s^i$ or σ^2):

$$p(\mathbf{w}_s | y_s^i, \mathbf{x}_s^i, \sigma^2) \propto p(y_s^i | \mathbf{w}_s, \mathbf{x}_s^i, \sigma^2) p(\mathbf{w}_s). \tag{2}$$

With the model from the previous section and the assumption of Gaussian noise, $p(y_s^i | \mathbf{w}_s, \mathbf{x}_s^i, \sigma^2) \sim \mathcal{N}(\mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^i, \sigma^2)$, and assuming our samples $\mathbf{x}_s^i$ are independent, we may derive the negative log likelihood as follows:

$$p(y_s^1, ..., y_s^{n_s} | x_s^1, ..., x_s^{n_s}, \mathbf{w}_s, \sigma^2) = \prod_{i=1}^{n_s} \mathcal{N}(y_s^i; \mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^i, \sigma^2) \tag{3}$$

$$LL(\mathbf{w}_s; D_s, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n_s} \left(y_s^i - \mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^i \right)^2, \tag{4}$$

The negative log likelihood defines a convenient loss function as its value increases with the square of the difference between our prediction $\mathbf{w}_s^{\mathrm{T}} \mathbf{x}_s^i$ and the true label y_s^i for each data point. For notational convenience, we write the loss in matrix form by defining the input matrix $\mathbf{X} = [\mathbf{x}^{1\mathrm{T}}, \ldots, \mathbf{x}^{n\mathrm{T}}]^{\mathrm{T}}$ and the output vector $\mathbf{y} = [y^1, \ldots, y^n]^{\mathrm{T}}$. Then, the loss for subject/session s is given by $\|\mathbf{X}_s \mathbf{w}_s - \mathbf{y}_s\|^2$, where $\|\mathbf{v}\|$ is the ℓ^2 or Euclidean norm. If we ignore the prior and solve for $\mathbf{w}_s$ analytically from here, we end up with the equations for regular linear regression.

It is well known that complex models that are trained without a validation dataset can *over-fit*, leading to poor generalization to new data points. A classical technique to control over-fitting is adding a penalty term to the loss function that reduces the complexity of the model. A common choice for this regularizer is given by the sum of the squares of the weight parameters,

$$\Omega(\mathbf{w}_s) = \frac{\|\mathbf{w}_s\|^2}{2}. \tag{5}$$

Addition of $\Omega(\mathbf{w}_s)$ to the optimization problem is equivalent to assuming a Gaussian prior on $\mathbf{w}_s$ with $\mathbf{0}$ mean and unit covariance $\mathbf{I}$ and incorporating this prior in the log-scale.[1] If the variance of the prior is not assumed to be exactly the identity matrix but rather some matrix $\alpha\mathbf{I}$ then this formulation describes ridge regression.

However, the above assumption is rarely a reasonable one. If there exists some better prior information on the distribution of the weights that can be represented by a mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, this information can be used instead in the regularizer by assuming a Gaussian distribution with the corresponding mean and covariance term, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, as the prior and defining the regularizer as the negative log prior probability

$$\Omega(\mathbf{w}_s; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\left[(\mathbf{w}_s - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{w}_s - \boldsymbol{\mu})\right] + \frac{1}{2}\log\det(\boldsymbol{\Sigma}). \tag{6}$$

Note that the last term is constant with respect to $\mathbf{w}_s$ for fixed $\boldsymbol{\Sigma}$, and further that $\Omega(\mathbf{w}_s; \mathbf{0}, \mathbf{I})$ is equivalent to (5).

The new loss function can then be derived by taking the negative logarithm of the posterior of y_s:

$$p(y_s^i|\mathbf{w}_s, \mathbf{x}_s^i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda) \propto \mathcal{N}(y_s^i; \mathbf{w}_s^{\mathrm{T}}\mathbf{x}^i, \lambda)\mathcal{N}(\mathbf{w}_s; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7}$$

$$LP(\mathbf{w}_s; D_s, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda) = \frac{1}{\lambda}\|\mathbf{X}_s\mathbf{w}_s - \mathbf{y}_s\|^2 + \Omega(\mathbf{w}_s; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \mathrm{C}. \tag{8}$$

We replace σ^2 with λ to emphasize that in the loss function, the variance of the original noise model is equivalent to a term that controls the ratio of the importance assigned to the prior probability of the learned weight vector versus how well the learned vector can predict the labels in the training data. Put another way, the higher the variance of the noise in the model, the less we can trust our training data to lead us to a good solution; moving forwards, it is more convenient to think of the variable in terms of this trade-off than as purely a noise variance. From this point the actual optimization problem can be formulated as

$$\min_{\mathbf{w}_s} LP(\mathbf{w}_s; D_s, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda). \tag{9}$$

2.3 Training Models for Subjects/Sessions Jointly

In a standard machine learning setting, there is a single prediction problem or *task* to model and there is usually no prior information on the distribution of the model parameters $\mathbf{w}$. However, if there are multiple prediction tasks that are related to each other, it is possible to use information from all the tasks in order to improve the inferred model of each task. In particular, if the tasks share a common structure along with some task-specific variations, the shared structure can be used as the prior information $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in (6) in order to ensure that the solutions to all the tasks are close to each other in some space.

[1]Note that $LL(\mathbf{w}_s; D_s, \sigma^2) + \Omega(\mathbf{w}_s)$ gives the negative log posterior for $\mathbf{w}_s$ given D_s and the assumed prior.

In the BCI training problem, we treat each subject/session as one task and the shared structure $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the subject/session-invariant characteristics of stimulus prediction. More precisely, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the mean vector and covariance matrix of features. As such, $\boldsymbol{\mu}$ defines an out-of-the-box BCI that can be used to classify data recorded from a novel subject/session without any subject/session-specific calibration process. The divergence of a subject/session model from the shared structure, $\|\mathbf{w}_s - \boldsymbol{\mu}\|$, represents the subject/session-specific characteristics of the stimulus prediction.

Clearly, the shared structure is unknown in this setting. Our goal is to infer the shared structure, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, from all the tasks along with the model parameters $\mathbf{w}_s$ jointly. This can be achieved by combining the optimization problem of all tasks

$$\min_{\mathbf{W},\boldsymbol{\mu},\boldsymbol{\Sigma}} LP(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; D, \lambda) = \min_{\mathbf{W},\boldsymbol{\mu},\boldsymbol{\Sigma}} \frac{1}{\lambda} \sum_s \|\mathbf{X}_s \mathbf{w}_s - \mathbf{y}_s\|^2 + \sum_s \Omega(\mathbf{w}_s; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{10}$$

where $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_S]^{\mathrm{T}}$, $D = \{D_s\}_{s=1}^S$, and d is the dimension of each weight vector. Let us investigate each term of this optimization problem separately. The first term is the sum of the losses from each session, and by minimizing it we ensure all the sessions are well fitted. The second term controls the divergence of each subject/session model from the underlying mean vector $\boldsymbol{\mu}$ and penalizes the elements of the residual $\hat{\mathbf{w}}_s = \mathbf{w}_s - \boldsymbol{\mu}$ scaling with $\boldsymbol{\Sigma}^{-1}$. Expanding one of these terms,

$$\hat{\mathbf{w}}_s^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \hat{\mathbf{w}}_s = \sum_i \sum_j \boldsymbol{\Sigma}_{i,j}^{-1} \hat{\mathbf{w}}_{s,i} \hat{\mathbf{w}}_{s,j},$$

we observe that $\boldsymbol{\Sigma}_{i,j}^{-1}$ is proportional to the partial correlation between the i-th and j-th components of the weight vector, which is defined as the correlation between these after all other components have been regressed out. Thus, for a given matrix $\boldsymbol{\Sigma}^{-1}$, this term will be minimized when for each set of components with high partial correlation, the subject/session-specific weight vectors $\mathbf{w}_s$ allow only one of these to deviate greatly from the mean of that component. Hence, $\boldsymbol{\Sigma}^{-1}$ acts as an implicit feature selector. The final term, which is a constant in the independent setting of (8), controls the complexity of the covariance matrix.

We solve the minimization in (10) with respect to $\mathbf{W}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iteratively by alternating holding $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{W}$ constant. For fixed $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, optimization over $\mathbf{w}_s$ decouples across subjects/sessions and hence can be optimized independently. In each iteration we get the new $\mathbf{w}_s$ by taking the derivative with respect to $\mathbf{w}_s$ for all s and equating to 0. This yields the following closed form update for each $\mathbf{w}_s$:

$$\mathbf{w}_s = \left(\frac{1}{\lambda} \mathbf{X}_s^{\mathrm{T}} \mathbf{X}_s + \boldsymbol{\Sigma}^{-1} \right)^{-1} \left(\frac{1}{\lambda} \mathbf{X}_s^{\mathrm{T}} \mathbf{y}_s + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right). \tag{11}$$

Hence, the model parameters are a combination of the shared model contribution $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and the contribution of the individual subject/session data $\mathbf{X}_s^{\mathrm{T}}\mathbf{y}_s$. This combination is scaled with the inverse of covariance term which again comes from both the data $\mathbf{X}_s^{\mathrm{T}}\mathbf{X}_s$ and the shared model $\boldsymbol{\Sigma}^{-1}$. In order to avoid inverting $\boldsymbol{\Sigma}$, which is a $O(d^3)$ operation, we perform the equivalent update

$$\mathbf{w}_s = \left(\frac{1}{\lambda} \boldsymbol{\Sigma} \mathbf{X}_s^{\mathrm{T}} \mathbf{X}_s + \mathbf{I} \right)^{-1} \left(\frac{1}{\lambda} \boldsymbol{\Sigma} \mathbf{X}_s^{\mathrm{T}} \mathbf{y}_s + \boldsymbol{\mu} \right). \tag{12}$$

For fixed $\mathbf{W}$, the updates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given in Algorithm 1 and derived in the Supplementary Materials.

Algorithm 1 Multitask BCI training

Input: D, λ
Set $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} = (\mathbf{0}, \mathbf{I})$
repeat
  Update $\mathbf{w}_s$ using (12)
  Update $\boldsymbol{\mu}$ using $\boldsymbol{\mu}^* = \frac{1}{S}\sum_s \mathbf{w}_s$
  Update $\boldsymbol{\Sigma}$ using $\boldsymbol{\Sigma}^* = \frac{\sum_s (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^{\mathrm{T}}}{\mathrm{Tr}(\sum_s (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^{\mathrm{T}})} + \epsilon \mathbf{I}$
until convergence
Output: $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

2.4 Decomposition of Spatial and Spectral Features

The learning method described above can be applied to any feature representation where the features extracted from each electrode are appended together. Let E be the number of electrodes and F be the number of features extracted from each electrode. The final feature vector then is size EF, rendering the covariance matrix large and iterative updates expensive. It also causes the number of features to grow linearly with the number of channels and channel-specific features, an increase that can be avoided by taking advantage of the structure of the EEG. Specifically, we assume that the contribution of the features is invariant across electrodes but the importance of each electrode varies. Hence, the weights corresponding to the feature vector mentioned above can be decomposed into two components: the weight of each electrode $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_E)$ and the weights of features that are shared across all electrodes $\mathbf{w} = (w_1, \ldots, w_F)$. We note that though in this paper spectral features are used, there is no reason that temporal features such as ERP timepoints could not be used instead. With this formulation, the linear regression function is given by

$$f_s(X; \mathbf{w}_s, \boldsymbol{\alpha}_s) = \boldsymbol{\alpha}_s^{\mathrm{T}} X \mathbf{w}_s,$$

where $X \in \mathbb{R}^{E \times F}$ denotes the matrix of features for each channel for a given trial. This causes the number of parameters in the decoding model to be reduced from EF to $E + F$.

The new optimization problem is now over $\mathbf{W}, \mathbf{A}, \boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\mathbf{w}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$, where $\mathbf{A} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_S]^{\mathrm{T}}$. However, it can easily be seen that $\boldsymbol{\alpha}^{\mathrm{T}} X \mathbf{w} = \boldsymbol{\alpha}^{\mathrm{T}} \tilde{X}$, where $\tilde{X} = X\mathbf{w}$, and thus that y, instead of being a function of the features, can now be considered a function of the aggregated features for each electrode. As this formulation assumes that $\boldsymbol{\alpha}$ and $\mathbf{w}$ are indepedent, the prior over model parameters can be incorporated as the product of indepedent priors for both $\mathbf{w}$ and $\boldsymbol{\alpha}$. As such, the same arguments used to define a prior of $\mathbf{w}$ can be applied to $\boldsymbol{\alpha}$ to define a new distribution for y_s^i and a new optimization problem (for readability we define the parameters of the Gaussian priors over $\mathbf{w}$ and $\boldsymbol{\alpha}$ as $\theta_{\mathbf{w}}$ and $\theta_{\boldsymbol{\alpha}}$ respectively):

$$p(y_s^i | X_s^i, \mathbf{w}_s, \boldsymbol{\alpha}_s, \theta_{\mathbf{w}}, \theta_{\boldsymbol{\alpha}}, \lambda) \propto \mathcal{N}(y_s^i; \boldsymbol{\alpha}_s^{\mathrm{T}} X_s^i \mathbf{w}_s, \lambda) \mathcal{N}(\mathbf{w}_s; \theta_{\mathbf{w}}) \mathcal{N}(\boldsymbol{\alpha}_s; \theta_{\boldsymbol{\alpha}}) \tag{13}$$

$$LP(\mathbf{W}, \mathbf{A}, \theta_{\mathbf{w}}, \theta_{\boldsymbol{\alpha}} | D, \lambda) = \frac{1}{\lambda} \sum_s \sum_i \| \boldsymbol{\alpha}_s^{\mathrm{T}} X_s^i \mathbf{w}_s - y_s^i \|^2 + \sum_s \Omega(\mathbf{w}_s; \boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}) + \sum_s \Omega(\boldsymbol{\alpha}_s; \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) + \mathrm{C} \tag{14}$$

where again, $\Omega(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the negative log prior probability of the vector $\mathbf{x}$ given the Gaussian distribution parametrized by $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It is easy to see that $\mathbf{w}$ and $\boldsymbol{\alpha}$ function identically except for a transpose. The updates for the weights over the features and the channels are linked, so we first iterate until convergence within each subject/session before continuing on to update the prior parameters, which leads to the following algorithm (Algorithm 2):

Algorithm 2 Multitask BCI training with uninformative $\boldsymbol{\alpha}$

Input: D, λ
Set $\{(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}), (\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})\} = (\mathbf{0}, \mathbf{I})$
Set $\boldsymbol{\alpha}_s = \mathbf{1}$
repeat
  repeat
    Compute $\tilde{\mathbf{X}}_s = [\boldsymbol{\alpha}_s^{\mathrm{T}} X_s^1; \ldots; \boldsymbol{\alpha}_s^{\mathrm{T}} X_s^n]$
    Compute $\widehat{\mathbf{X}}_s = [X_s^1 \mathbf{w}_s, \ldots, X_s^n \mathbf{w}_s]$
    Update $\mathbf{w}_s$ using $\mathbf{w}_s^* = \left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\mathbf{w}}\tilde{\mathbf{X}}_s^{\mathrm{T}}\tilde{\mathbf{X}}_s + \mathbf{I}\right)^{-1}\left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\mathbf{w}}\tilde{\mathbf{X}}_s^{\mathrm{T}}\mathbf{y}_s + \boldsymbol{\mu}_{\mathbf{w}}\right)$
    Update $\boldsymbol{\alpha}_s$ using $\boldsymbol{\alpha}_s^* = \left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\widehat{\mathbf{X}}_s\widehat{\mathbf{X}}_s^{\mathrm{T}} + \mathbf{I}\right)^{-1}\left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\widehat{\mathbf{X}}_s\mathbf{y}_s + \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right)$
  until $\mathbf{W}$ and $\mathbf{A}$ converge for fixed $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  Update $\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}$ using $\boldsymbol{\mu}_w^* = \frac{1}{S}\sum_s \mathbf{w}_s, \boldsymbol{\mu}_{\boldsymbol{\alpha}}^* = \frac{1}{S}\sum_s \boldsymbol{\alpha}_s$
  Update $\boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ using $\boldsymbol{\Sigma}_{\mathbf{w}}^* = \frac{\sum_s (\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})(\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})^{\mathrm{T}}}{\mathrm{Tr}(\sum_s (\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})(\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})^{\mathrm{T}})} + \epsilon\mathbf{I}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^* = \frac{\sum_s (\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^{\mathrm{T}}}{\mathrm{Tr}(\sum_s (\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^{\mathrm{T}})} + \epsilon\mathbf{I}$
until convergence
Output: $(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$

This reduces the size of the feature space from EF to $E + F$, which simplifies learning the regression parameters and also reduces calculation speed. The more degrees of freedom, the more data a model requires to find a good fit, so by reducing the number of parameters we also reduce the number of necessary training trials. Also for the case of a model with EF parameters, the matrix inversion necessary to compute a decision rule is $O(E^3F^3)$, which is changed for a model with $E+F$ parameters to $O((E+F)^3)$. We also note that the initialization of Algorithm 2 shown above is non-informative. Our experiments have suggested that the alternative method shown below (Algorithm 3) works more effectively in some cases.

Algorithm 3 Multitask BCI training with $\boldsymbol{\alpha}$ initialization

Input: D, λ
Set $\{(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}), (\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})\} = (\mathbf{0}, \mathbf{I})$
Concatenate subject data in D into single pooled subject $\hat{D}$
Run ridge regression on $\hat{D}$ using the feature decomposition regression function
Set $\boldsymbol{\alpha}_s$ to the ridge regression spatial weights
repeat
  repeat
    Compute $\tilde{\mathbf{X}}_s = [\boldsymbol{\alpha}_s^{\mathrm{T}} X_s^1; \ldots; \boldsymbol{\alpha}_s^{\mathrm{T}} X_s^n]$
    Compute $\widehat{\mathbf{X}}_s = [X_s^1 \mathbf{w}_s, \ldots, X_s^n \mathbf{w}_s]$
    Update $\mathbf{w}_s$ using $\mathbf{w}_s = \left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\mathbf{w}}\tilde{\mathbf{X}}_s^{\mathrm{T}}\tilde{\mathbf{X}}_s + \mathbf{I}\right)^{-1}\left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\mathbf{w}}\tilde{\mathbf{X}}_s^{\mathrm{T}}\mathbf{y}_s + \boldsymbol{\mu}_{\mathbf{w}}\right)$
    Update $\boldsymbol{\alpha}_s$ using $\boldsymbol{\alpha}_s = \left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\widehat{\mathbf{X}}_s\widehat{\mathbf{X}}_s^{\mathrm{T}} + \mathbf{I}\right)^{-1}\left(\frac{1}{\lambda}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\widehat{\mathbf{X}}_s\mathbf{y}_s + \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right)$
  until $\mathbf{W}$ and $\mathbf{A}$ converge for fixed $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  Update $\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}$ using $\boldsymbol{\mu}_w^* = \frac{1}{S}\sum_s \mathbf{w}_s, \boldsymbol{\mu}_{\boldsymbol{\alpha}}^* = \frac{1}{S}\sum_s \boldsymbol{\alpha}_s$
  Update $\boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ using $\boldsymbol{\Sigma}_{\mathbf{w}}^* = \frac{\sum_s (\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})(\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})^{\mathrm{T}}}{\mathrm{Tr}(\sum_s (\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})(\mathbf{w}_s - \boldsymbol{\mu}_{\mathbf{w}})^{\mathrm{T}})} + \epsilon\mathbf{I}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^* = \frac{\sum_s (\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^{\mathrm{T}}}{\mathrm{Tr}(\sum_s (\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_s - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^{\mathrm{T}})} + \epsilon\mathbf{I}$
until convergence
Output: $(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$

Online resource for multitask learning

Supplementary materials, appendix, and MATLAB and Python implementations of all three algorithms described here can be found at http://brain-computer-interfaces.net/.

2.5 Adaptation to Novel Subjects

In Section 2, we outlined a simple yet effective approach to infer the subject-invariant BCI model, given by learning the parameters of a Gaussian distribution over the weights. This model can be used successfully on novel subjects immediately via $f(\mathbf{x};\theta) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}$ in the case of regular linear regression or $f(X;\theta_{\mathbf{w}},\theta_{\boldsymbol{\alpha}}) = \boldsymbol{\mu}_{\boldsymbol{\alpha}}^{\mathrm{T}} X \boldsymbol{\mu}_{\mathbf{w}}$ in the case of feature decomposition, though depending on the covariance of the learned priors this can result in poor performance. It is possible to further improve the performance of this model by adapting to the subject as more subject-specific data becomes available by simply using the learned priors and considering the problem independently as discussed in Section 2.2. The standard regression case is discussed there; for the feature decomposition method, we consider n trials X^i, where each $X^i \in \mathbb{R}^{E\times F}$ is a matrix with columns denoting features and rows denoting electrodes. In this setting the update equations are identical to the inner loop of Algorithm 2. We emphasize that $\mathbf{w}$ and $\boldsymbol{\alpha}$ are linked, so the update steps must be iterated until convergence. The parameter λ is determined in practice through cross-validation over the training data.

3 Experiments

We conducted two experiments with real-world data sets. The first used both the initial multitask learning algorithm as well as the version with decomposition of spectral and spatial features while the second only used the version with feature decomposition (hereafter referred to as FD). The first is an example of subject-to-subject transfer with a motor imagery dataset recorded for ten healthy subjects, and the second is an example of session-to-session transfer for a neurofeedback paradigm recorded in a single subject with ALS.

3.1 Subject-to-Subject Transfer

Paradigm

As an initial test of this algorithm, we considered how it performs on the most common paradigm in spectral BCIs: motor imagery. Specifically, subjects were placed in front of a screen with a centrally displayed fixation cross. Each trial started with a pause of three seconds. A centrally displayed arrow then instructed subjects to initiate haptic motor imagery of either the left or right hand, as indicated by the arrow's direction. After a further seven seconds the arrow was removed from the screen, marking the end of the trial and informing subjects to cease motor imagery.

Dataset

Ten healthy subjects participated in the study (two females, 25.6 ± 2.5 years old). One subject had already participated twice in other motor imagery experiments while all others were naïve to motor imagery and BCIs. EEG data was recorded from 128 channels, placed according to the extended 10-20 system with electrode Cz as reference, and sampled at 500Hz. BrainAmp amplifiers (BrainProducts, Munich) with a temporal analog high-pass filter with a time constant of 10s were used for this purpose. A total of 150 trials per class (left/right hand motor imagery) per subject were recorded in pseudorandomized order, with no feedback provided to the subjects during the experiment.

Feature Extraction

For feature extraction, recorded EEG data was first spatially filtered using a surface Laplacian setup [35]. We did not employ more sophisticated methods for spatial filtering, such as CSP or beamforming, in order to keep the spatial filtering setup data-independent. For each subject, trial and electrode, frequency bands of 2 Hz width, ranging from 7-29 Hz, were then extracted using a discrete Fourier transform with a Hanning window, computed over the length of the trial. Log-bandpower within the last seven seconds of each trial for each frequency band then formed the (128×12)-dimensional feature vector.

Classification Performance

Here we show the efficiency of the proposed algorithms by examining the effect of multitask learning and FD on classification performance. For all algorithms, one subject was successively chosen as the test subject and all other subjects were then used for training. Test-specific training data of between 10 and 100 trials per condition were then given to each algorithm, and the remaining trials out of 300 were used for testing. Multitask learning was done using Algorithm 1 and Algorithm 3 with a cross-validated λ. Note that for all tested algorithms the feature space was the full 128 channels, each with 12 feature bands.

We looked at two control algorithms to compare with the multitask learning approaches. The first was to consider ridge regression, which regularizes the regression method only by penalizing the magnitude of the resultant weight vector (see (5)) and can be seen as using an uninformed prior for the distribution of weight vectors; the second was to consider a support vector machine (SVM) trained on the same feature space. We further tested both control algorithms two ways: Once with pooled data and once with only subject-specific data. For the pooled condition, all data from the training subjects was concatenated to the training trials from the test subject to form a combined training set, on which the control algorithms were run. For the subject-specific condition, only training data from the test subject was used to train the control algorithms. All controls were compared to the multitask approaches, where the learned prior mean(s) and covariance(s) were used to regularize the least-squares regression method.

The following list summarizes the algorithms:

- MT_FD: multitask learning with Algorithm 3
- MT: multitask learning with Algorithm 1
- RR: standard ridge regression
- RR_FD: ridge regression using the FD regression method
- SVM: SVM with a linear kernel given the full 128×12 feature space

Results

The results for the pooled sub-condition can be found in Figure 2 and the results for the single-subject sub-condition can be found in Figure 3. Note that in both graph, the curves for the MT and MT_FD algorithms are identical.

The MT_FD algorithm consistently outperformed the other algorithms at nearly all levels of test subject data. In the pooled condition, it equalled the zero-training and low-data accuracy of the pooled data while also managing to more effectively use subject-specific data, leading to a higher mean accuracy than any other algorithms with 200 training trials.

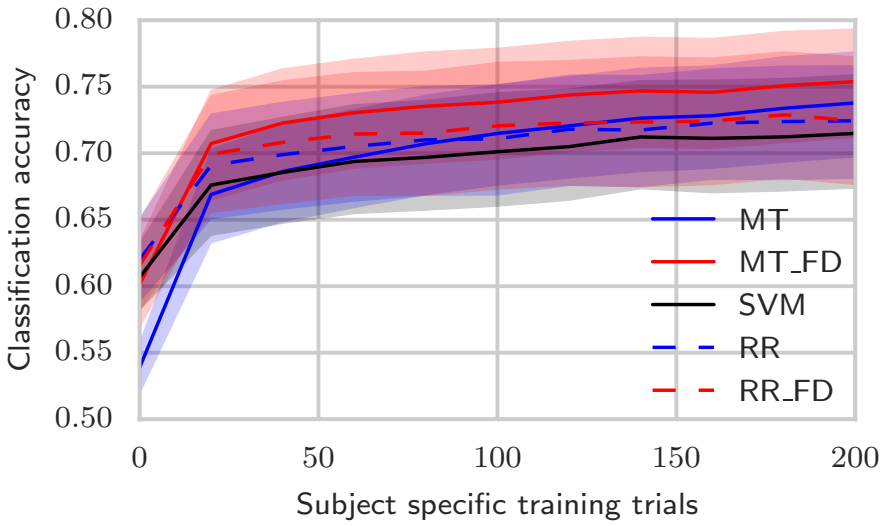

Fig. 2: Mean and STD (shaded) for classification accuracy of MT and pooled conditions across the ten subjects. The control algorithms were trained on data pooled across training subjects, and are compared against classification using Algorithm 1 (MT, solid blue) and Algorithm 3 (MT_FD, solid red). Displayed control algorithms are ridge regression using the standard regression method (RR, dashed blue), ridge regression using the FD regression method (RR_FD, dashed red) and SVM (SVM, solid black). The FD formulation of the multitask learning has comparable performance with few training trials to pooled regression and both multitask algorithms manage to improve more than the pooled controls given a larger number of training trials.

Interestingly, the MT algorithm without FD did more poorly than the pooled ridge regression without FD in the zero-training and low subject-specific trial cases, which we hypothesize is due to the fact that each individual subject had so few training trials compared to the size of the feature space (300 compared to 1400). Rule adaptation requires learning a rule for each subject, which is hampered by this low number. However, by concatenating the trials together in ridge regression, pooling manages to work better. Regardless of this, MT without FD is still able to more effectively use subject-specific data than any of the pooling algorithms as shown by the higher slope of the classification curve. The single-subject condition was used to determine whether this regularization could reduce the maximum accuracy: With large training data and no data from different subjects, the best subject-specific rule can be found, and so we consider the maximum single-subject accuracy as an approximation of the maximum achievable accuracy with a linear boundary. We find that the MT approaches at high numbers of trials achieve accuracies nearly identical to those achieved by only subject-specific training, showing that there is no reduction in maximum achievable accuracy for the MT approach. For subject-specific results please consult the Supplementary Materials.

To further confirm that our results are classifying on the signal we expect, we considered the mean of the spatial and spectral priors in the MT_FD condition (Figure 4). The learned topography is most strongly weighted around the electrodes directly over the motor cortices and the different cortices also have opposite signs, which is in agreement with spatial filters learned in CSP [4], [13] and beamforming [10]. Further, looking at the spectral weights, we see that the most important weight is on the μ band, which is consistent with previous results on the subject, suggesting that our classification accuracy is indeed due to training on a brain-derived signal and not any sort of artifact.

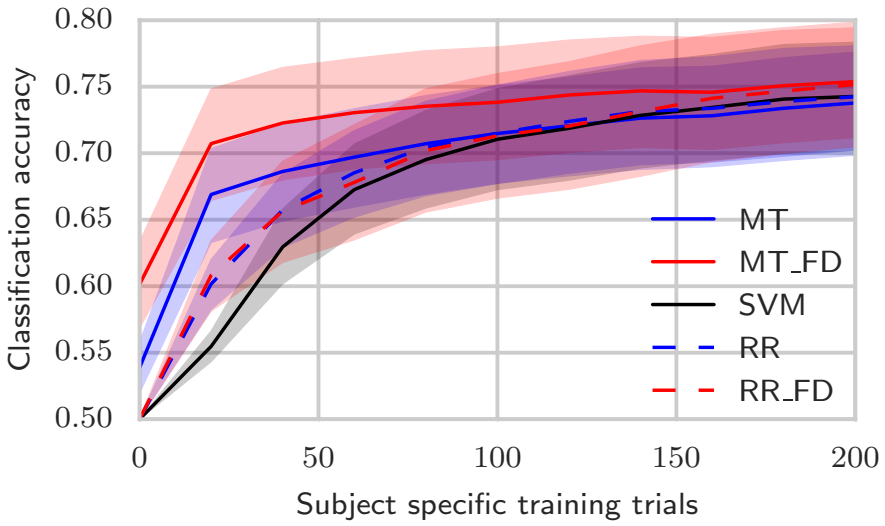

Fig. 3: Mean and STD (shaded) for classification accuracy of MT and single-subject conditions across the ten subjects. Classification values for the multitask algorithms are identical to those shown in Figure 2. The control algorithms were trained on data exclusively from the test subject, and are compared against classification using Algorithm 1 (MT, solid blue) and Algorithm 3 (MT_FD, solid red). Displayed control algorithms are ridge regression using the standard regression method (RR, dashed blue), ridge regression using the FD regression method (RR_FD, dashed red) and SVM (SVM, solid black). The multitask algorithm with FD regression estimation performs better on average regardless of the number of trials, though single-subject ridge regression with the FD regression method manages to equal its performance at 200 training trials.

3.2 Session-to-Session Transfer

A common issue in BCI paradigms, especially those used with patient populations, is the low number of trials per session. Given the success of our FD approach on the motor imagery data, we attempted it here on a 30-session dataset where each session had only ten training and between ten and twenty test trials for each condition.

Data Collection

We trained a patient diagnosed with ALS to modulate the δ-bandpower (1–5 Hz) in the precuneus in thirty sessions over the course of fifteen months. The patient's ALS-FRS-R [36] score decreased from 33 to 9 over the course of this time, an average of 1.6 points per month. The paradigm setup is identical to the setup in [37] except that the frequency band that received feedback was 1–5 Hz and the target area was changed from the superior parietal cortex to the precuneus. In brief, however: The subject learned through operant conditioning to modulate power in a beamformed signal pointed at the precuneus over the course of sixty seconds, deviating either up or down from a session-specific mean. Each minute-long interval was one trial, and each run was twenty trials (ten up and ten down). For each session, the subject completed between two and three runs. The first session was used entirely for training.

Throughout all sessions a 121-channel EEG was recorded at a sampling frequency of 500 Hz using actiCAP active electrodes and a QuickAmp amplifier (both provided by BrainProducts GmbH, Gilching, Germany). Electrodes were placed according to the extended 10-20 system with electrode Cz as the initial reference. All recordings were converted to common average reference.

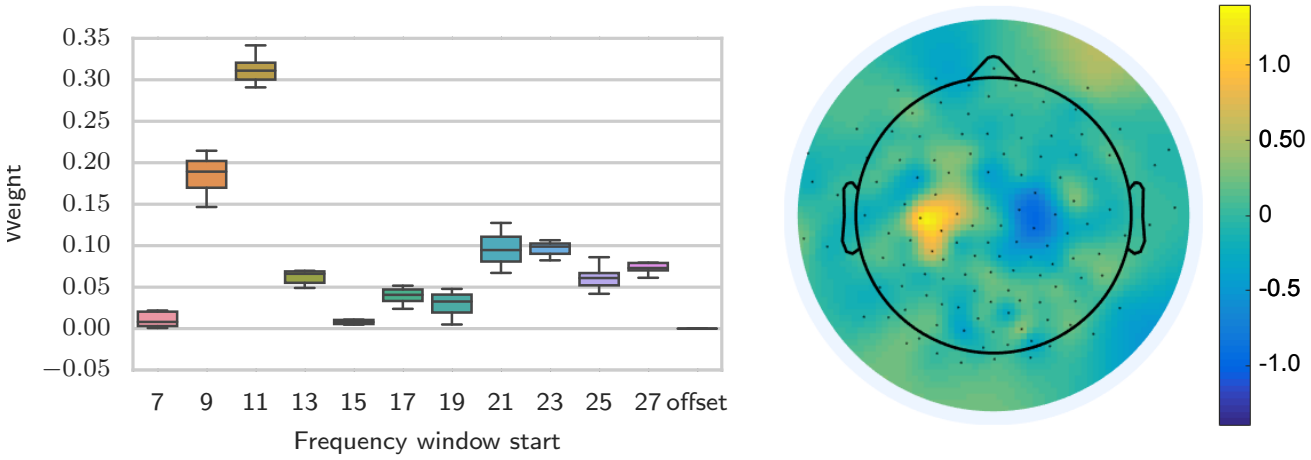

Fig. 4: (left) Box and whiskers plot of the absolute value of the learned prior means for all 10 leave-one-out executions of the MT algorithm, showing medians, first quartiles, and 1.5 times the IQR for the learned weights in each frequency range. The sign of the prior frequency mean is exchangeable with the sign of the spatial mean, and so the absolute value is used here to correct for that. X-axis shows the starting frequency of each 2 Hz window. Note that the highest weights are concentrated around the windows of the μ frequency range. (right) Sum of the learned spatial weights over training subject groups plotted topographically with respect to the head showing a concentration of high-magnitude weights over the motor cortices.

Feature extraction and training

To eliminate artifacts, independent component analysis (ICA) was performed on each session using the SOBI algorithm [38] and components corresponding to cortical features [39], [40] were manually chosen. The time-series of these components were then re-projected to the electrode space. For each trial and electrode, the log-powers in the frequency bands $\delta =$ 1-5 Hz, $\theta =$ 5-8 Hz, $\alpha =$ 8-12 Hz, $\beta_1 =$ 12-20 Hz, $\beta_2 =$ 20-30 Hz, $\gamma_1 =$ 30-70 Hz, $\gamma_2 =$ 70-90 Hz were computed using a discrete Fourier transform over the sixty seconds of the trial to create a 121×7 feature space. The first session was used for training, after which the first run of each session was used to update the classifier according to Section 2.5 and the updated weight vectors used to classify the data in the next one or two runs in a pseudo-online fashion. Between sessions Algorithm 2 was re-run with all data of the most recent session included, as we found experimentally that the non-initialized case performed better on these data. We compare results between the MT, RR, and SVM performance (Figure 5). The spatial and frequency weights learned by the MT algorithm are shown in Figure 6. Single and pooled were computed similarly to those presented in Section 3.1 except that instead of subjects we used sessions.

3.2.1 Results

We can see that the multitask and the pooled ridge regression have the highest median (85%) and show more density in higher classification accuracies. Both are significantly better than the single-session ridge-regression ($p < 0.0001$, Wilcox signed-rank test); as the SVM results are clearly bimodal a median comparison is not informative. Between the pooled and multitask FD conditions the differences are small, which may reflect the fact that inter-session differences are not as large as inter-subject differences. However, the multitask formulation has a higher minimum classification accuracy (65% vs 60%) than the pooled accuracy, suggesting that considering the sessions separately still adds a small benefit when attempting to test on sessions that are outliers for some reason. This may be related to why the SVM distributions are bimodal, as the SVM

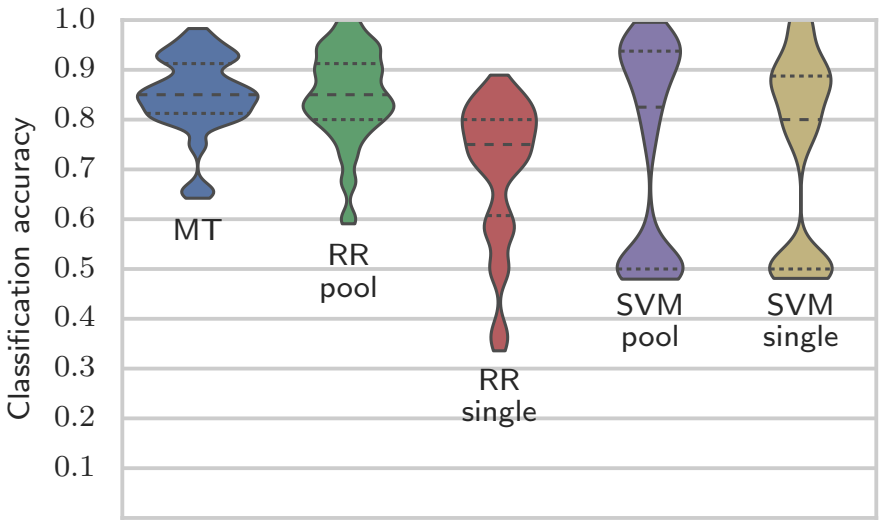

Fig. 5: Density plot of classification accuracy over sessions for each algorithm. MT corresponds to multitask learning using Algorithm 2 and RR corresponds to ridge regression using the FD regression method. Dashed line corresponds to median for the distribution and dotted lines show upper and lower quartiles. Classification accuracies using the pooled FD regression and multitask learning have a higher minimum classification accuracy than any other method.

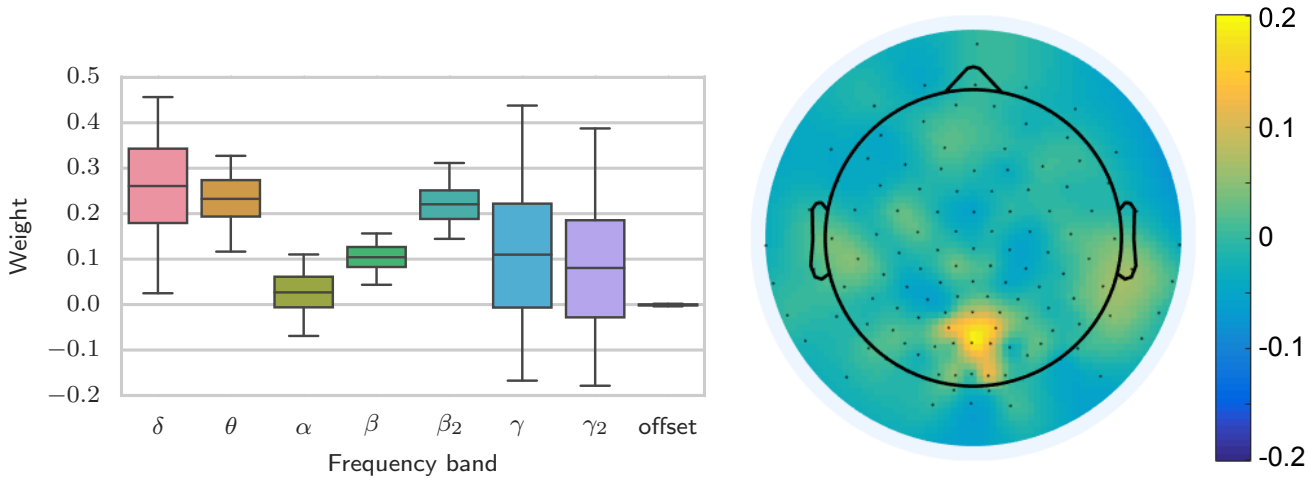

Fig. 6: (left) Box and whiskers plot with median, quartiles, and 1.5 times the IQR for frequency weights over 30 test sessions. Though this ignores the evolution of weights over time the δ range is highly weighted. (right) Sum of learned spatial mean weights after thirty sessions plotted topographically with respect to the head, showing that the area over the parietal cortex is emphasized.

either classifies excellently or at chance level in both the single-session and pooled cases. This also suggests that there are outlier sessions, in which the distribution of data in the feature domain is sufficiently different from past data to cause the cross-validation over the training data to poorly predict test data classification. Possibly the fact that there is no distinction made between sessions in the pooled case causes these methods to have lower minimum accuracies. Looking further to the spatial and spectral weights, we see that the weights are concentrated in the area directly above the precuneus. Instead of a smooth topography, however, we see that certain channels are strong and nearby channels are not, which is consistent with the feature selection aspect of the regularization as discussed in Section 2.

4 Discussion

We have introduced a framework for transfer learning, both across subjects and across sessions, that works across feature spaces and requires fewer training trials than other state-of-the-art methods for classification, representing an effective combination of pooled data and single-subject/session training. Previous work in transfer learning for BCI focuses on transforming the feature space of individual subjects/sessions such that one decision boundary generalizes well across subjects/sessions, here referred to as 'domain adaptation'. In contrast, our method treats the decision boundary as a random variable whose distribution is inferred from previous subjects/sessions. As a result, our method is complementary to domain adaptation methods. Further, we show that applying this formulation with an altered regression method that takes feature decomposition into account is effective at learning structure between both multiple subjects and multiple sessions in EEG-based BCI tasks. By assuming that the weights of the channels and the features are independent we are able to drastically reduce the size of the feature space. This method works better than an SVM trained on an equal feature space both in the zero-training transfer learning case and after a within-session training period. The prior parameters describing the distribution of the weight vectors can also be quickly used to see spatial and spectral topographies associated with a given task.

Though the proposed regression method appears to work well across datasets, it has some undesirable features. One such characteristic of the proposed regression method is the variable importance of initializing the spatial weights smartly. In the motor imagery paradigm, a lack of proper initialization resulted in very poor results; conversely, in the other experiment, using initialization was less effective. While there is no clear rule as to when it might be necessary, we can easily see a possible explanation for this problem when considering the regression method itself:

$$\hat{y} = \boldsymbol{\alpha}^{\mathrm{T}} X \mathbf{w} = (C\boldsymbol{\alpha})^{\mathrm{T}} X \left(\frac{1}{C} \mathbf{w} \right) \tag{15}$$

where C is an arbitrary real number. This symmetry means that the likelihood function is not actually convex, making the location at which it is initialized in its domain important to the predictivity of the results. When initialized poorly, it can fail to find predictive features. Further work may determine an appropriate criterion to make the regression method properly convex. For practical application, however, we found no obvious trend as to which paradigms work better with a non-informative initialization versus a ridge-regression initialization. Our suggestion with this method would be to test both empirically and choose the one that works best.

A second problematic result of using the FD regression is the addition of another loop in the algorithm, as now for each subject/session there must be iteration to determine an appropriate spatial and spectral combination. However, in practice we found this to run quite smoothly. The other option is to use the regular regression method, which results in a far larger matrix that has to be inverted for every session. We also found that the convergence in the case of the FD regression happened orders of magnitude faster than in the non-FD case, possibly due to the far more favorable ratio of training trials to features. Overall, though there is a second loop in the algorithm, the FD case is actually faster than the non-FD case, in practice, on high-dimensional datasets. Finally, we note that the restriction of a single spatial weight vector and frequency weight vector means that a single brain process can be classified at a time. Winning entries in the BCI competition IV mostly used multiple signals to achieve their high accuracies [41], a possibility that is not possible using this approach as

they would return conflicting regression weights.

Though ours is the first presentation of inference for the full distribution of weight vectors in BCI, this approach has been well-studied in the machine learning domain for a variety of different problems [42]. One possible future direction is to specify our priors as samples from a Dirichlet process and attempt to take advantage of any clustering as the number of subjects increases [43], as has been shown to be effective in CSP multi-subject learning [44]. It is also interesting to note that the multitask learning formulation is simply an additive convex term to the loss function, which suggests that it can be added to any algorithm as a cheap way of learning something about the shared structure of classification rules (though without some involvement of the shared parameters in the computation of the task-specific rules an iterative procedure would be impossible). Further work with this approach in SVMs or LDA should prove to be very interesting. Lastly, the current approach requires that the entire iterative scheme is re-run after the inclusion of any new subjects or sessions, which quickly becomes inefficient as the number of considered subjects or sessions increases. More research to help streamline the update rule of the priors would be invaluable in the age of big data.

It is likely that all the methods presented here would perform better if prior knowledge had been incorporated into choosing the feature space. For example, Alamgir et al. [33] use data only from the electrodes directly over the motor cortex. Indeed, given a small feature space and a separable problem, it is well known that optimizing the objective function of an SVM leads to better test classification than simple least-squares loss. The problem is simply that we do not always have so much prior information; further, in the case of newer paradigms such as the one the ALS patient was trained on, such information is currently unavailable, a problem that will only continue as more possible paradigms are experimented with. We hope that the multitask framework presented here will function as a way of quickly judging the efficacy and activation topography of new BCI paradigms. By training with feature decomposition we are able to get a picture of what channels and features are important to the task at hand, and can then possibly re-run with the non-FD algorithm in order to better capture the multitask structure in the smaller feature space. However, there are also instances in which the data has a very large number of dimensions that do actually contribute to the classification of the task at hand, and we have shown that multitask learning is robust to these sorts of datasets.

5 Conclusion

Previous approaches to transfer learning in BCI have ignored the possibilities of knowledge transfer within the feature space, constraining themselves mostly to spatial filtering and domain adaptation. Here, we present a method for learning that transfers knowledge from previous subjects to new ones in any desired spatiotemporal feature space, able to work both on its own and on top of other paradigms. Testing on both motor imagery and a novel cognitive paradigm, we find that our proposed methods better deal with both session-to-session and subject-to-subject variability as compared to simple pooling, achieving accuracies comparable to or better than single-session training with far fewer training trials. Further, this work presents a framework on top of which other objective functions can be used to determine priors for decision boundaries that minimize other sorts of error. Any parties interested in trying these algorithms for themselves will find implementations of all three algorithms in MATLAB at the following website: http://brain-computer-interfaces.net/.

ACKNOWLEDGMENT

The authors would like to thank Tatiana Fomina, Christian Förster, Natalie Widmann, Marius Klug, and Nadine Simon for their help in gathering the data used in the second experiment.

REFERENCES

[1] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An EEG-based brain-computer interface for cursor control," *Electroencephalography and Clinical Neurophysiology*, vol. 78, no. 3, pp. 252–259, 1991.

[2] J. R. Wolpaw and D. J. McFarland, "Multichannel EEG-based brain-computer communication," *Electroencephalography and Clinical Neurophysiology*, vol. 90, no. 6, pp. 444–449, 1994.

[3] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, Mar 1999.

[4] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[5] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying single trial EEG: Towards brain computer interfacing," in *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 157–164.

[6] T. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support vector channel selection in BCI," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1003–1010, 2004.

[7] M. Grosse-Wentrup, K. Gramann, and M. Buss, "Adaptive spatial filters with predefined region of interest for EEG based brain-computer-interfaces," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 537–544.

[8] N. Hill, T. Lal, M. Schröder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Schölkopf, A. Kübler, and N. Birbaumer, "Classifying EEG and ECoG signals without subject training for fast BCI implementation: Comparison of nonparalyzed and completely paralyzed subjects," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 183–186, June 2006.

[9] B. Blankertz, G. Dornhege, M. Krauledat, K. Müller, and G. Curio, "The non-invasive berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, 2007.

[10] M. Grosse-Wentrup and M. Buss, "Multi-class common spatial pattern and information theoretic feature extraction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, 2008.

[11] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[12] H. A. Abbass, J. Tang, R. Amin, M. Ellejmi, and S. Kirby, "Augmented cognition using real-time EEG-based adaptive strategies for air traffic controljikes," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1. SAGE Publications, 2014, pp. 230–234.

[13] Z. Koles, "The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 6, pp. 440–447, 1991.

[14] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *Latent Variable Analysis and Signal Separation*. Springer Berlin Heidelberg, 2010, pp. 629–636.

[15] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PLoS One*, vol. 3, no. 8, pp. 1–12, 2008.

[16] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[17] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, Aug. 2009.

[18] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, 2014.

[19] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[20] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multisubject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, 2011.

[21] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "BCI signal classification using a riemannian-based kernel," in *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2012)*. Michel Verleysen, 2012, pp. 97–102.

[22] M. Congedo and A. Barachant, "A special form of SPD covariance matrix for interpretation and visualization of data manipulated with riemannian geometry," in *MaxEnt 2014*, vol. 1641. AIP publishing LLC, 2015, p. 495.

[23] H. Morioka, A. Kanemura, J.-i. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, "Learning a common dictionary for subject-transfer decoding with resting calibration," *NeuroImage*, vol. 111, pp. 167–178, 2015.

[24] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, p. 214101, 2009.

[25] W. Samek, C. Vidaurre, K.-R. Mller, and M. Kawanabe, "Stationary common spatial patterns for braincomputer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[26] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *The Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[27] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1318–1324, 2010.

[28] R. Mohammadi, A. Mahloojifar, and D. Coyle, "Unsupervised short-term covariate shift minimization for self-paced BCI," in *Proc. of 2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, 2013, pp. 101–106.

[29] M. Arvaneh, I. Robertson, and T. E. Ward, "Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface," in *Proc. of 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2014, pp. 6501–6504.

[30] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th International Conference on Machine learning*. ACM, 2007, pp. 193–200.

[31] H. Yang, C. Guan, K. S. G. Chua, C. C. Wang, P. K. Soon, C. K. Y. Tang, and K. K. Ang, "Detection of motor imagery of swallow EEG signals based on the dual-tree complex wavelet transform and adaptive model selection," *Journal of Neural Engineering*, vol. 11, no. 3, 2014.

[32] P.-J. Kindermans, H. Verschore, D. Verstraeten, and B. Schrauwen, "A P300 BCI for the masses: Prior information enables instant unsupervised spelling," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 710–718.

[33] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 17–24.

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[35] D. McFarland, L. McCane, S. David, and J. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 386–394, 1997.

[36] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, no. 12, pp. 13–21, 1999.

[37] M. Grosse-Wentrup and B. Schölkopf, "A brain-computer interface based on self-regulation of gamma-oscillations in the superior parietal cortex," *Journal of Neural Engineering*, vol. 11, no. 5, p. 056015, 2014.

[38] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[39] J. Onton, M. Westerfield, J. Townsend, and S. Makeig, "Imaging human EEG dynamics using independent component analysis," *Neuroscience & Biobehavioral Reviews*, vol. 30, no. 6, pp. 808–822, 2006.

[40] M. Grosse-Wentrup, S. Harmeling, T. Zander, J. Hill, and B. Schölkopf, "How to test the quality of reconstructed sources in independent component analysis (ICA) of EEG/MEG data," in *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2013, pp. 102–105.

[41] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz *et al.*, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, pp. 1–31, 2012.

[42] K. Yu, V. Tresp, and A. Schwaighofer, "Learning gaussian processes from multiple tasks," in *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*. New York, NY, USA: ACM, 2005, pp. 1012–1019.

[43] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *J. Mach. Learn. Res.*, vol. 8, pp. 35–63, May 2007.

[44] H. Kang and S. Choi, "Bayesian common spatial patterns with Dirichlet process priors for multi-subject EEG classification," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–6.

2.3 Paper 2: Task-induced frequency modulation features for brain-computer interfacing

Task-induced frequency modulation features for brain-computer interfacing

OPEN ACCESS
IOP Publishing
Journal of Neural Engineering
J. Neural Eng. **14** (2017) 056015 (12pp)
https://doi.org/10.1088/1741-2552/aa7778

Task-induced frequency modulation features for brain-computer interfacing

Vinay Jayaram[1,2], Matthias Hohmann[1,2], Jennifer Just[3], Bernhard Schölkopf[1] and Moritz Grosse-Wentrup[1]

[1] Department of Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] IMPRS for Cognitive and Systems Neuroscience, University of Tübingen, Tübingen, Germany
[3] Hertie Institute for Clinical Brain Research, Hoppe-Seyler-Str. 3, Tübingen, Germany

E-mail: vjayaram@tue.mpg.de, mhohmann@tue.mpg.de, bs@tue.mpg.de, moritzgw@tue.mpg.de and jennifer.just@uni-tuebingen.de

Received 25 April 2017
Accepted for publication 6 June 2017
Published 19 September 2017

Abstract

Objective. Task-induced amplitude modulation of neural oscillations is routinely used in brain-computer interfaces (BCIs) for decoding subjects' intents, and underlies some of the most robust and common methods in the field, such as common spatial patterns and Riemannian geometry. While there has been some interest in phase-related features for classification, both techniques usually presuppose that the frequencies of neural oscillations remain stable across various tasks. We investigate here whether features based on task-induced modulation of the frequency of neural oscillations enable decoding of subjects' intents with an accuracy comparable to task-induced amplitude modulation. *Approach*. We compare cross-validated classification accuracies using the amplitude and frequency modulated features, as well as a joint feature space, across subjects in various paradigms and pre-processing conditions. We show results with a motor imagery task, a cognitive task, and also preliminary results in patients with amyotrophic lateral sclerosis (ALS), as well as using common spatial patterns and Laplacian filtering. *Main results*. The frequency features alone do not significantly out-perform traditional amplitude modulation features, and in some cases perform significantly worse. However, across both tasks and pre-processing in healthy subjects the joint space significantly out-performs either the frequency or amplitude features alone. This result only does not hold for ALS patients, for whom the dataset is of insufficient size to draw any statistically significant conclusions. *Significance*. Task-induced frequency modulation is robust and straight forward to compute, and increases performance when added to standard amplitude modulation features across paradigms. This allows more information to be extracted from the EEG signal cheaply and can be used throughout the field of BCIs.

Keywords: BCI, brain-computer interface, EEG, signal processing

(Some figures may appear in colour only in the online journal)

1. Introduction

Brain-computer interfaces (BCIs) are moving closer and closer to stable use, as evidenced by their increasing precision [1–3] and their inclusion in events like the Cybathlon BCI race [4]. Surprisingly, however, this enormous increase in usability has been achieved using only a very small part of the entire EEG signal. Neuroscience tells us that neural circuits tend to oscillate [5], and these oscillations project to the EEG signal where they can be isolated by bandpassing or other methods. Further, these oscillations can have different spectral locations depending on the makeup of the underlying circuits. The neural frequency bands used by the brain are well known, and so by taking advantage of the properties of

these neural oscillations, modern stimulus-independent BCIs have achieved their current performance [6–9]. However, we show in this paper that previously unconsidered properties of the EEG can easily and reliably be used to improve decoding performance.

The two major properties of an oscillation are amplitude and phase, both of which have long histories in the realm of BCIs. The easiest and most-used property of neural oscillations in the context of BCIs is the signal amplitude or power [10]. The average response of cortical ensembles to tasks, be they cognitive or motor, is an increase or decrease in synchrony of neighboring circuits [11], giving rise to a change in the measured power at those circuits' frequency of oscillation—which is a mathematically convenient property to calculate given an EEG signal. To borrow a term from signal processing, we designate this amplitude modulation (AM) as the time-domain amplitude is closely related to the power. After being computed, these powers can be used as features for standard machine learning algorithms. Over time, many fundamental methods in BCIs have been developed using task-related differences in spectral power such as CSP [12] and the more recent Riemannian approaches [13], both of which have led to substantial gains in performance.

Phase-related features, though more recently exploited, have also enjoyed great success. Mostly, this has happened through measures of phase synchrony, in which two channels are compared to see how often the phases are identical in both [14]. This value for a given time period is used as a feature for a classifier, and has been shown to be useful for BCIs both offline [15–17] and online [18]. Outside of identical phases, consistent phase differences [19] and the average instantaneous phase difference [20] have also been used and shown to add new information, allowing for more accurate classification of intentions.

As has been known for many years, however, BCIs based on these feature spaces simply do not work for some subset of even healthy individuals (for a comprehensive review, see [21]). Much work has been done on attempting to determine *a priori* who will not be able to use a rhythm-based BCI [22], but the existence of such people is simply a given. This is not even to consider various patient groups for whom a BCI is sometimes the only hope for communication with the world, such as amyotrophic lateral sclerosis (ALS) patients. Within just this group, traditional features have proved ineffective by the end stage of the disease despite over two decades of research [23].

One thing that ties phase and amplitude-based features together is that they are both restricted by one thing: the frequency. For all the methods above, a frequency range needs to be assumed to compute the AM or phase features. There is, unfortunately, no good range that works for all people. Differences in the spectral location of neural frequency bands have been attributed to many factors. In the particular case of the μ band, i.e. oscillations in sensorimotor areas that range from 8 to 13 Hz, it has been shown to vary with age [24], genetics [25, 26], and psychological factors [27]. For the parietal α, Haegens *et al* [28] determined that even within an individual and a single recording session the peak frequency of the α band can vary by up to one Hertz [28], and is related to such factors as cognitive load [28, 29]. These more transient changes in the location of neural peaks are nearly always dealt with as reasons to shift the spectral window for further processing, indeed the general recommendation for motor imagery is to always use a subject-specific μ frequency window [30, 31].

Neurophysiological research has given us ample evidence to believe that neural frequency peaks shift both location and size in response to tasks as well as innate factors. As BCI researchers, our natural question is then to ask whether this shift can be quantified and used as a feature for classification. Outside the field of BCI, this average frequency feature has already been attempted in LFP recordings [32, 33] as well as with EEG for non-BCI usage [34, 35]. Within the field, this has been dealt with by two works. Wu *et al* [36] showed that the frequency location of the trialwise empirical modes can be a robust feature for classification, but they did not look at task-induced frequency shifts. Rather, they used an SSVEP paradigm to induce the frequency shifts externally. A paper by Medl *et al* [37] did look at intrinsic, task-related frequency shifts over 25 years ago, in a visionary work that considered the instantaneous frequency and envelope as possible features for a BCI. However this work was never expanded past an exploratory analysis, an expansion which we attempt here. As we believe that the location of the neural oscillations in the frequency domain may be a feature for classification, we rechristen this approach frequency modulation (FM), as it measures the change in location of spectral peak.

In this paper we review a basic method for extracting the average peak frequency for a given BCI trial and show that this feature can be used across BCI paradigms as well as frequency bands to increase accuracy when added to traditional bandpower-based feature spaces.

1.1. Related work

The average frequency of a trial is equivalent to the average slope of the unwrapped phases at every point. Phase features are well-described in the BCI literature; however, they are exclusively related to phase synchronicity, and have never been considered in this way. The closest in BCIs is [20] that looks at the instantaneous phase difference, but between two different channels. The approach of analytic CSP [38] takes advantage of the phase values at every point in time, but uses them in a similar way to the synchrony measures used above. While the Hilbert transform does allow them to look at instantaneous phase, they compute the imaginary covariance matrix, which explicitly ignores the time dependencies in the phase which make up the average frequency.

Outside of BCIs but still in the realm of neurophysiology, the instantaneous frequency has in the last decade begun to come into use. In invasive recordings the instantaneous frequency has been estimated by AR modelling and used in local field potentials [32, 33]. In EEG, AR-based frequency estimation has been done in auditory ERP for clinical purposes [35] and also to detect sleep spindles [34]. Within the field, this has been dealt with by three works. Wu *et al* [36] showed that

Table 1. ALS patient data.

Subject	Age	Sex	ALSFRS-R[a]
1	59	F	0
2	75	M	42
3	54	M	48
4	N/A	M	33
5	51	F	12

[a]Revised amyotrophic lateral sclerosis functional rating scale [41]. The rating scale was filled out after the recording session by the experimenter.

the frequency location of the trialwise empirical modes can be a robust feature for classification, but they did not look at task-induced frequency shifts. Rather, they used an SSVEP paradigm to induce the frequency shifts externally. Park *et al* [39] also use a variant of empirical mode decomposition and the Hilbert transform to show that a version of instantaneous frequency can be used in motor imagery as well. However, they only attempt the Hilbert transform on the intrinsic modes themselves; we here show that it is unnecessary to use the decomposition first. Last, a paper by Medl *et al* [37] did look at intrinsic, task-related frequency shifts over 25 years ago, in a visionary work that considered the instantaneous frequency and envelope as possible features for a BCI. However this work was never expanded past an exploratory analysis, an expansion which we attempt here.

2. Methods

2.1. Datasets

To test whether FM features are useful across experimental paradigms, we use two datasets from different paradigms. The first is a motor imagery paradigm: subjects were placed in front of a screen with a centrally displayed fixation cross. Each trial started with a pause of three seconds. A centrally displayed arrow then instructed subjects to initiate haptic motor imagery of either the left or right hand, as indicated by the arrow's direction. After a further seven seconds the arrow was removed from the screen, marking the end of the trial and informing subjects to cease motor imagery. Ten healthy subjects participated in the study (eight males, two females, 25.6 ± 2.5 years old). One subject had already participated twice in other motor imagery experiments while all others were naïve to motor imagery and BCIs. EEG data was recorded from 128 channels, placed according to the extended 10–20 system with electrode Cz as reference, and sampled at 500 Hz. BrainAmp amplifiers (BrainProducts, Gilching, Germany) with a temporal analog high-pass filter at 0.1 Hz were used for recording. A total of 150 trials per class (left/right hand motor imagery) per subject were recorded in pseudorandomized order, with no feedback provided to the subjects during the experiment.

The second paradigm is a cognitive paradigm introduced by Hohmann *et al* [40]: Eleven healthy subjects (6 female, mean age 28 ± 7.5 years) as well as five ALS patients (demographics and ALS-FRS-R scores given in table 1) were instructed to either activate self-referential memories by thinking of a positive memory, or to focus on a mental subtraction task. Trials were 35 s long with a 5.5 ± 0.5 pause in between. Each healthy subject completed 60 trials without feedback, while the number of trials varied for ALS patients based on their ability to participate. Patients managed between 30 and 40 trials. EEG was recorded from 124 channels placed according to the extended 10–20 system with identical sampling and amplification to the previous task.

2.2. Pre-processing

To investigate whether the usefulness of FM features is robust with respect to various pre-processing steps, we tested three different types of spatial filtering: No spatial filtering, Laplacian spatial filtering [42], and common spatial patterns (CSP) [43]. In order to limit the variance when estimating classification accuracy in section 2.4, we designed each pre-processing step to give us a small, two-dimensional feature space. For no spatial filtering, we chose two channels above the left and right motor cortices (C3 and C4) and over frontal and parietal areas (Fz and Pz) for the motor imagery and the cognitive paradigm, respectively. For spatial filtering, we chose to limit our analysis to the motor imagery dataset, as the literature on BCI performance and neurophysiological interpretations of spatial filtering are more extensive for this paradigm. We chose one data-dependent and one data-independent filtering method: Laplace filtering and CSP. For the Laplace filtering we took the four closest channels in the extended 10-20 setup to compute the Laplace filter for channels C3 and C4 over left and right motor cortices, respectively. For CSP we used 10-fold cross-validation to ensure the filters were never applied to the same data used to generate them. No regularization aside from a small epsilon to ensure positive-definiteness were used for the computation. Two spatial filters were used for each subject corresponding to the two most extreme eigenvalues.

2.3. Feature extraction

The neural bands used in this analysis are the standard accepted ranges: 4–8 Hz for θ, 8–13 Hz for α (and μ), and 13–30 Hz for β. In our experience and within the data, the α range of the ALS patients varied (especially for late-stage patients) drastically and so, exclusively for these subjects, all frequency information was calculated using subject-specific bands. In order to ensure the features were compared as fairly as possible, we employed identical preprocessing to generate both AM and FM features within each subject. In all cases the signal was initially bandpassed offline with a 3rd order Butterworth filter, between the low and high values of the various bands of interest. To preserve phase, this was done once forwards and once backwards on the full data from each continuous recording.

2.3.1. Amplitude modulation. It is well known in the field of BCIs that the bandpower within certain bands of interest is linked to activity within the brain. From Parseval's theorem, we can compute this power by restricting frequencies in the

signal to those of interest and taking the variance of the result. As a result we can compute the bandpower, given a signal $x(t)$, by using a bandpass filter and then performing the following calculation:

$$\mathrm{AM}(x(t)) = \mathrm{E}\left[(x(t) - \bar{x})^2\right]. \quad (1)$$

2.3.2. Frequency modulation. To calculate the average frequency location for each trial, we used the analytic signal. The analytic signal is a complex-valued signal which has zero negative frequency components, constructed from a real-valued signal $x(t)$ and its Hilbert transform:

$$x_a(t) = x(t) + i\mathrm{H}[x(t)]. \quad (2)$$

This can equivalently be written in its polar form, $x_a(t) = A(t)\mathrm{e}^{j\phi(t)}$, in which $A(t)$ denotes the amplitude of the signal at time t and $\phi(t)$ denotes the phase. In previous BCI applications, both of these have been used as features for a classifier [20, 44, 45]. However, as with [37], we are not interested in the phase, but rather the instantaneous frequency. This can be computed in general as the derivative of the phase, though for discrete signals we can use the difference operator as follows:

$$\mathrm{IF}[n] = \frac{\phi[n] - \phi[n-1]}{2}. \quad (3)$$

This leads, for a time series with n samples, to $n-1$ values. As these values are individually quite noisy, we opt to take the median of them as our FM feature for each trial to best correct for the influence of outliers.

2.3.3. Joint feature space. For both paradigms and all types of pre-processing, we wanted to compare purely AM and FM features as well as the combination of them. To do this we simply concatenated the two-dimensional feature spaces of each feature type to create a joint four-dimensional feature space. One could also think of this as simply computing two features per channel per trial. As the number of trials is so much higher than the dimensionality of this joint feature space it is very unlikely that the concatenation described here affected results negatively.

2.4. Classification

In order to test the usefulness of the FM feature in comparison to the AM feature, we used a simple binary classification tool widely employed in BCIs: linear discriminant analysis (LDA). While more complicated methods could possibly get far better performance, LDA allows us to test how separable the data is in multiple dimensions under the most basic assumptions of homogeneous Gaussianity. In the single-feature cases each trial was represented by a two-dimensional vector and for the joint feature space each trial was represented by a four-dimensional vector. Accuracies were calculated by 10-fold cross-validation within each subject over paradigms and spatial filtering choices. However, for ALS patients it was slightly more complicated as the number of trials varied among subjects. To ensure all subjects could be included, we downsampled to the minimum number of 30. A stratified random split was implemented: 30 trials were chosen randomly with equal classes and a cross-validated accuracy was computed for this subset. This random split was iterated 1000 times and the average of these accuracies was used for each subject.

2.5. Statistical analyses

To test statistical significance of differences between conditions (AM, FM, or joint feature spaces) within a paradigm, a permutation test was used. We wanted to test two hypotheses: (1) that subject accuracies using only the FM features are significantly different on average from AM features, and (2) that the joint feature accuracies are significantly higher, on average, than the AM only accuracies. This corresponds to a two-tailed test for hypothesis (1) and a one-tailed test for hypothesis (2), as it should not be possible for the accuracy to decrease with a larger feature space. For both of these we took a pair of values from each subject (AM accuracy/FM accuracy or AM accuracy/joint accuracy) and computed the pairwise t-statistic over the subjects within a paradigm. We then shuffled the pairings 1000 times and computed the t-statistics for these to generate the empirical t-statistic null distribution. The computed p-value is the proportion of t-statistics from the null distribution higher than the t-statistic generated with the true feature set-accuracy pairings. In the case of the two-tailed test, instead of looking for the percentage of the null distribution that was larger we looked for the percentage that was more extreme.

3. Experiments

In all the different pre-processing and paradigm cases explained above, we conducted the same experiment: within each subject, we computed the accuracy using only AM features, only FM features, and with the joint feature space. We then compared these different methods across subjects within each experiment. In motor imagery the μ and β bands are both well-known for carrying discriminative information, and so we conducted one experiment in each band for the data without pre-processing.

Next, we investigated the effects of spatial filtering on these feature spaces. We limited our analysis for these experiments to only the μ band, and processed the data after spatial filtering identically to the analysis without spatial filtering.

For the cognitive paradigm data in the healthy subjects, literature shows the α and θ bands as being predictive for these tasks [27, 40, 46], and so we did the experiment within the α and θ bands. For the ALS patients, it was impossible to find a θ rhythm for some subjects and the α rhythms varied dramatically, and so we show only results in the subject-specific α. This was identified by recording two 5-minute resting state recordings, one with eyes open and one with eyes closed, and then using the established observation that α power is increased in the eyes closed state [27] to determine the extent of the α band.

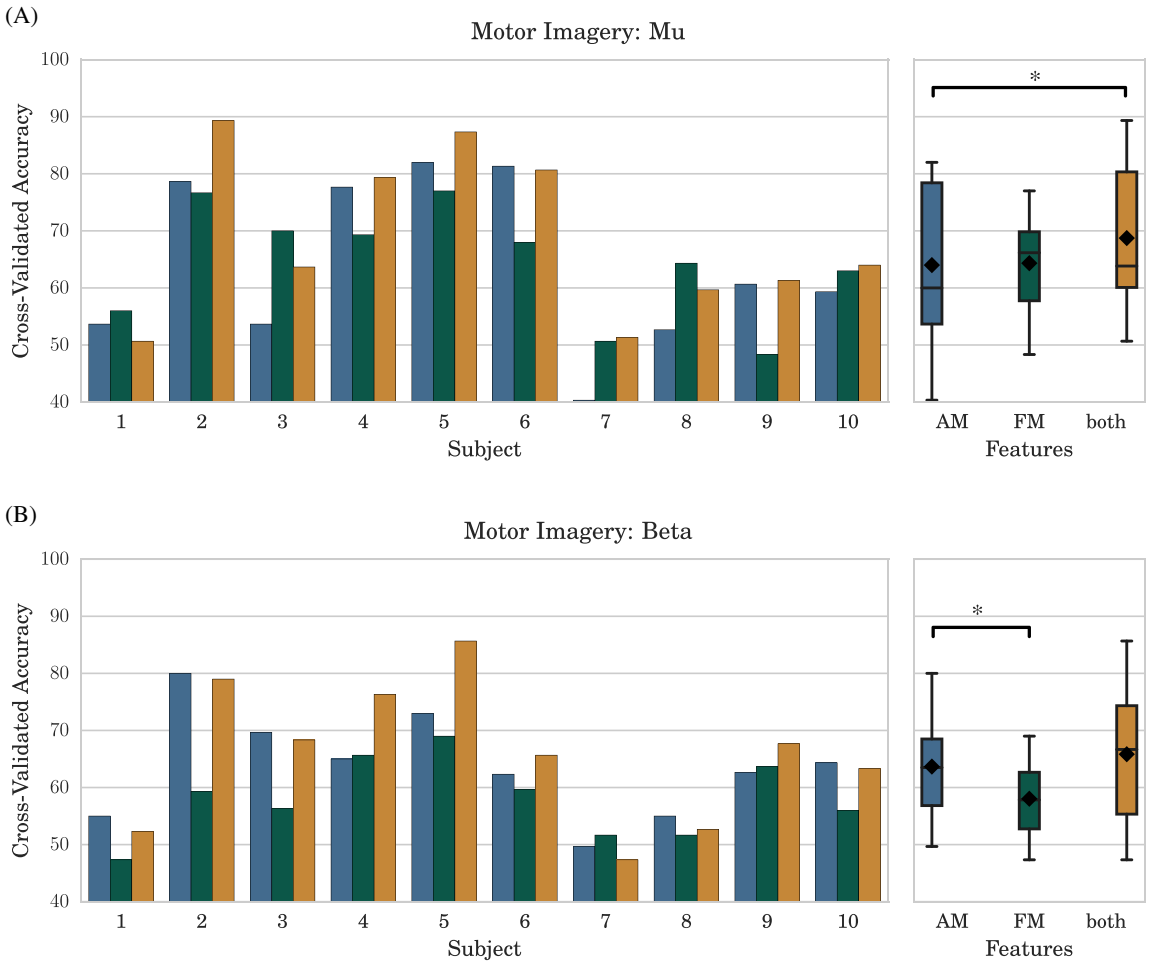

Figure 1. Comparison of cross-validated accuracies for different feature sets in the motor imagery data. Channels selected were C3 and C4. Blue shows accuracies using only AM features, green shows accuracies using only the FM features, and gold shows accuracies using the combined feature space. The left plot shows the per-subject cross-validated accuracies and the right plot shows a box-and-whiskers plot with the mean (diamond), median (line) and upper and lower quartiles over subjects; outliers are shown as small black diamonds; $*$: $p < 0.05$. (A) Accuracies computed in the μ band from 8–13 Hz. FM only and AM only do not significantly differ while the joint space is significantly better than AM (average 68.7% versus 64.0%, $p = 0.01$). (B) Accuracies computed in the β band from 13–30 Hz. FM only is significantly worse than AM only (58.0% versus 63.7%, $p = 0.02$) while the joint space is not significantly better (65.8% versus 63.7%, $p = 0.15$).

4. Results

Figure 1 shows the results of the experiments on the motor imagery data without pre-processing in both frequency bands. While the FM only features do worse than the AM only features in the β band (average 58.0% versus 63.7%, $p = 0.02$), there was no significant difference in the μ band (64.3% versus 64.0%, $p = 0.91$). The joint space was significantly better than the AM only features in the μ band (68.7% versus 64.0%, $p = 0.01$) but not the β band (65.8% versus 63.7%, $p = 0.15$). Figure 2 shows the results in the μ band for the two spatially pre-processed experiments. The FM only features were less predictive on average than the AM only features, though not significantly so (Laplacian: 65.4% versus 71.3%, $p = 0.18$, CSP: 68.5% versus 70.2%, $p = 0.42$). The joint space continued to improve on the AM only results with both a Laplacian filter (74% versus 71.3%, $p = 0.04$) and CSP filters (73% versus 70%, $p < 0.001$). Curiously, in half of the subjects Laplacian filtering actually decreased cross-validated accuracies using only FM features (subjects $2, 5, 6, 8, 9$) at the same time as it increased accuracies using AM alone (80% of subjects). Conversely, using CSP the FM and AM accuracies both increased as compared to without spatial filtering.

Figures 3 and 4 both concern the results on the cognitive task. Figure 3 shows results on experiments in two frequency bands for the cognitive paradigm in healthy subjects. In the α band the FM only features are significantly worse than the AM only features (72.7% versus 78.9%, $p = 0.01$) and the joint space does not out-perform the AM only features (79.4% versus 78.9%, $p = 0.39$). In the θ band the FM only features and AM only features do not differ significantly (77.1% versus 78.8%, $p = 0.47$), but the joint space significantly out-performs the AM only features (83.3% versus 78.8%, $p = 0.01$), showing that this effect is not isolated to the α band. Unfortunately, this result does not carry over to the ALS patients in figure 4, for whom the mean accuracy of the joint space is actually worse (67.4% versus 71.6%).

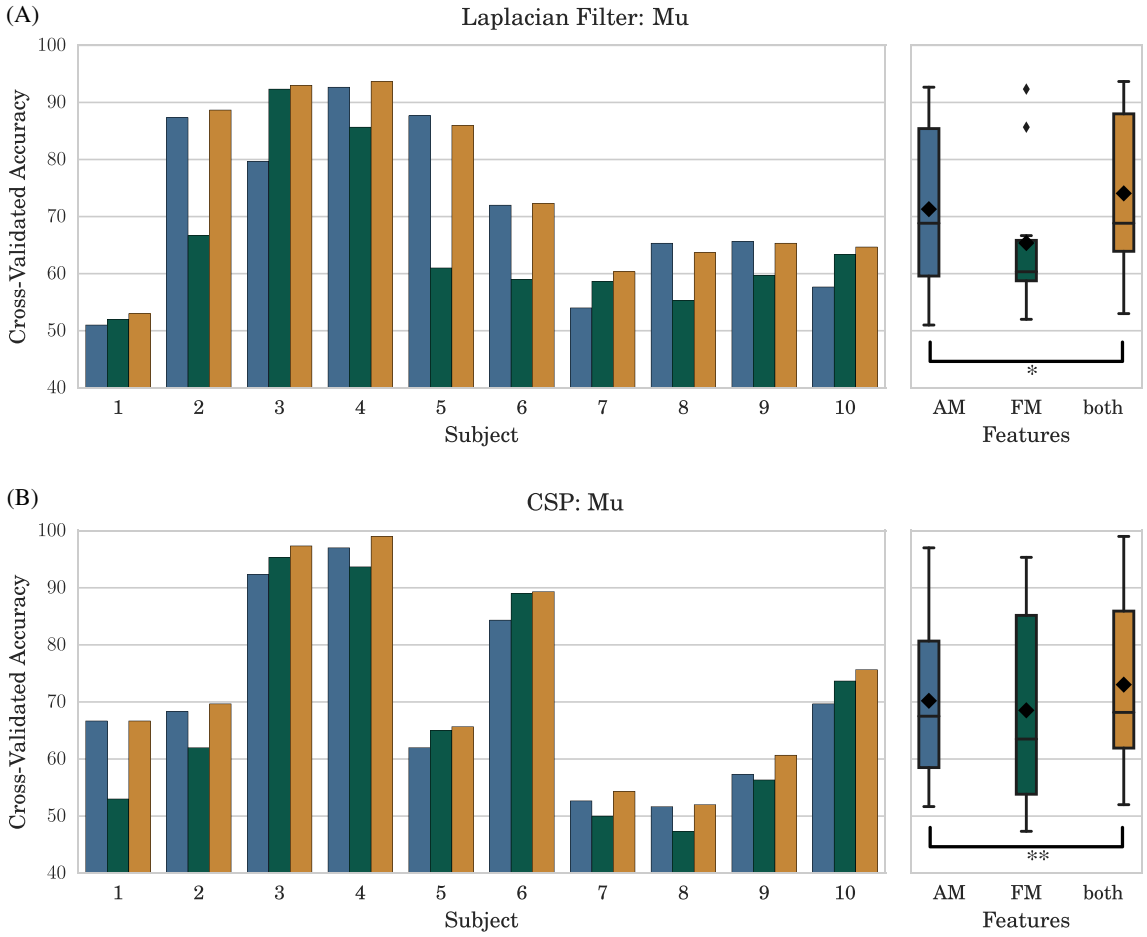

Figure 2. Comparison of cross-validated accuracies across different sorts of spatial filtering. The left plot shows the per-subject cross-validated accuracies and the right plot shows a box-and-whiskers plot with the mean (diamond), median (line) and upper and lower quartiles over subjects; outliers are shown as small black diamonds; $*$: $p < 0.05$, $**$: $p < 0.01$. (A) For Laplace filtering a discrete Laplacian filter for electrodes C3 and C4 was computed using the four closest electrodes to each. The joint space is significantly better than AM (74% versus 71.3%, $p = 0.04$). For CSP (B), two spatial filters were kept per subject. The joint space is significantly better than AM (73% versus 70%, $p < 0.01$). In both cases, the FM only accuracies do not differ significantly from the AM only accuracies.

While the joint feature space does not significantly out-perform AM in some cases, notably the motor β and the cognitive α, it still has a higher average accuracy than the AM feature space even in those cases, suggesting that while results are not significantly better there is at the very least no detriment. The only case in which it fails to out-perform is the case of the ALS patients-but given the low number of trials and the heterogeneity of the disease stages shown here, it is still too early to be sure that this feature is not of use.

These results show that the FM feature is surprisingly robust. In order to better understand the magnitude and variance of the feature across subjects, we plotted the two-dimensional feature space to see the separation within subjects for both paradigms. FM features were computed for all trials and de-meaned within subjects to control for different peak locations, then overlayed in order to see how much conditions differed on average. The one-dimensional projections are plotted on the side and top. In motor imagery (figure 5) we find that the frequency features for C3 and C4 are not very correlated across subjects ($\rho = 0.28$) and that the peak shift is relatively slight compared to the variance within channels. To check the spatial correlation we plot the correlation of the feature in channels C3 and C4 with the other 127 channels, averaged over classes and subjects (figure 5). The result shows a clear drop-off of correlation with distance on the scalp, which is consistent with the projections of the motor cortex.

In the cognitive paradigm, the frequency scatter plot (figure 6) shows a strong correlation between Fz and Pz ($\rho = 0.59$), which is interesting given how far apart they are on the head. When the correlations are plotted (figure 6) it can be seen that these two channels are, in general, more spatially correlated than C3 and C4. However, this correlation is more anteroposterior than it is lateral, which is consistent with the projection of fronto-parietal networks [47, 48].

In the ALS patients, however, the correlation structure is quite different (figure 7). In comparison with healthy subjects the separation between condition-specific median frequencies is quite pronounced in both channels, and further the channels appear comparatively less correlated ($\rho = 0.45$). This suggests that despite the lack of significant improvement in decoding accuracy, this feature is still quite robust in ALS patients. When we plot the correlations of the FM feature

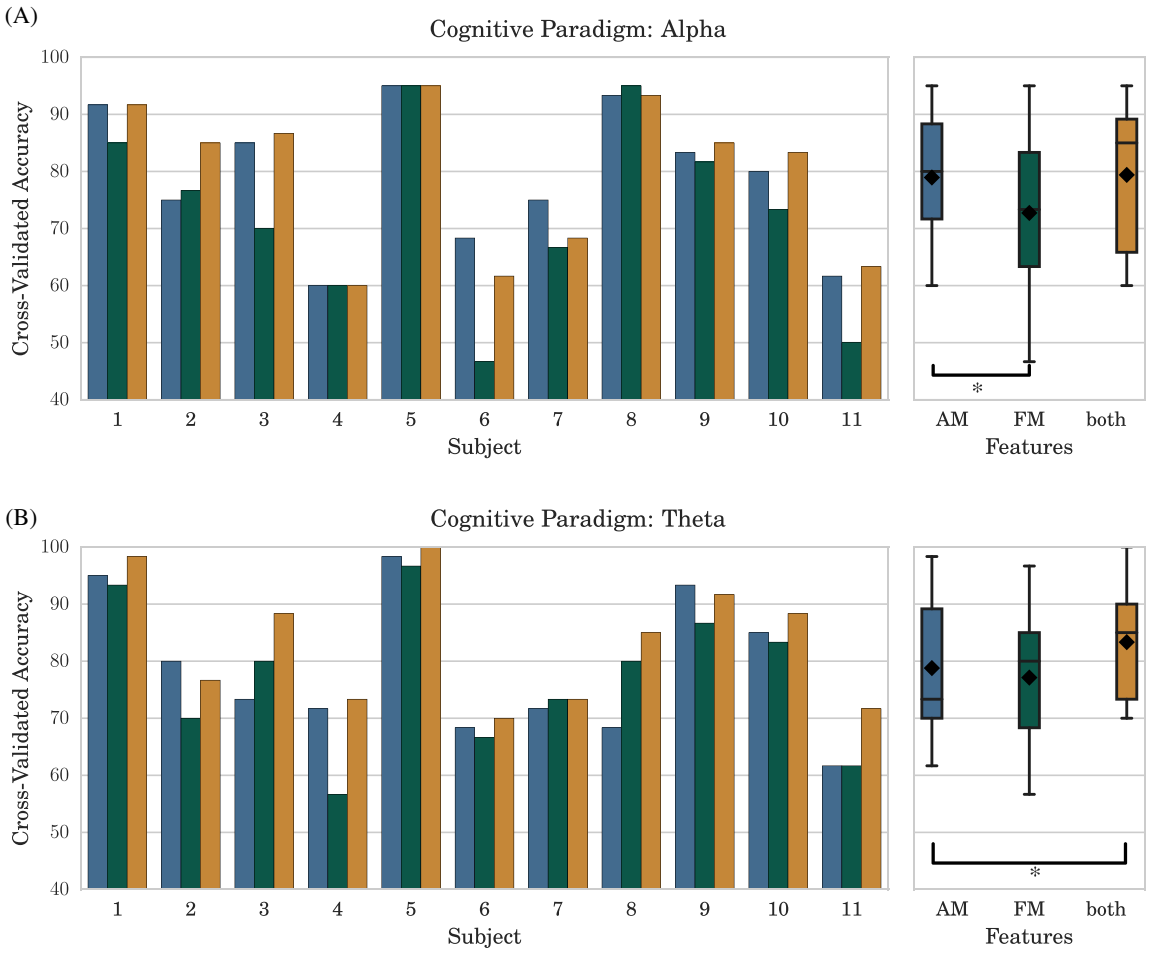

Figure 3. Comparison of cross-validated accuracies for different feature sets in the cognitive paradigm data. Channels selected were Fz and Pz. Data is plotted identically to figures 1 and 2; $*$: $p < 0.05$. (A) Accuracies computed in the α band from 8–13 Hz. FM only is significantly worse (72.7% versus 78.9% , $p = 0.02$) than AM only while the joint accuracy is not significantly better (79.4% versus 78.9% , $p = 0.39$). (B) Accuracies computed in the θ band from 4–8 Hz. There is no significant difference between the FM only and AM only results, and the joint space is significantly better than AM (83.3% versus 78.8%, $p = 0.01$).

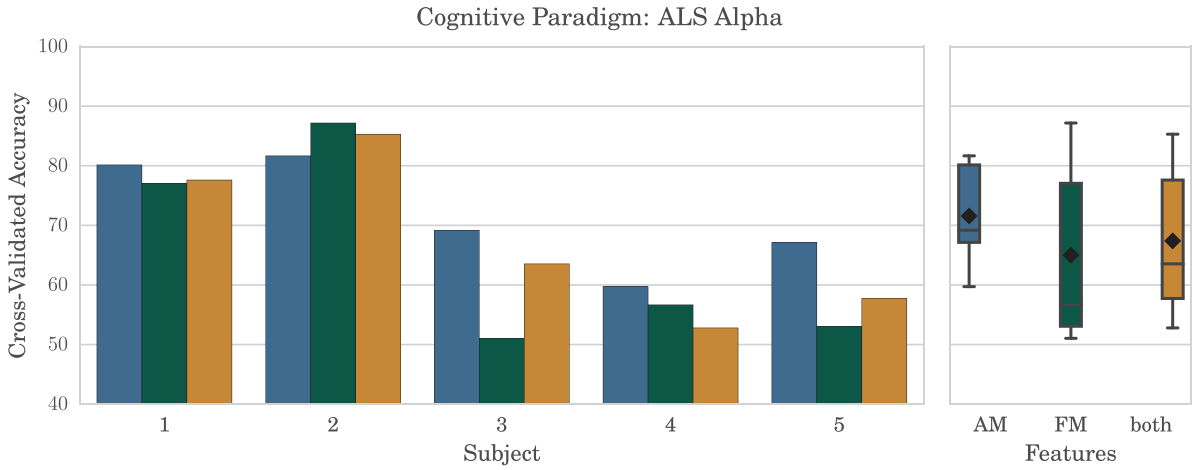

Figure 4. Comparison of cross-validated accuracies for different feature sets in the cognitive paradigm data for ALS patients. Channels selected were Fz and Pz. Blue shows accuracies using only AM features, green shows accuracies using only the FM features, and gold shows accuracies using the combined feature space. The left plot shows the per-subject cross-validated accuracies and the right plot shows the distribution over subjects. Neither FM (65.4% on average) nor the joint space (67.6%) was significantly different from the AM features alone (71.7%).

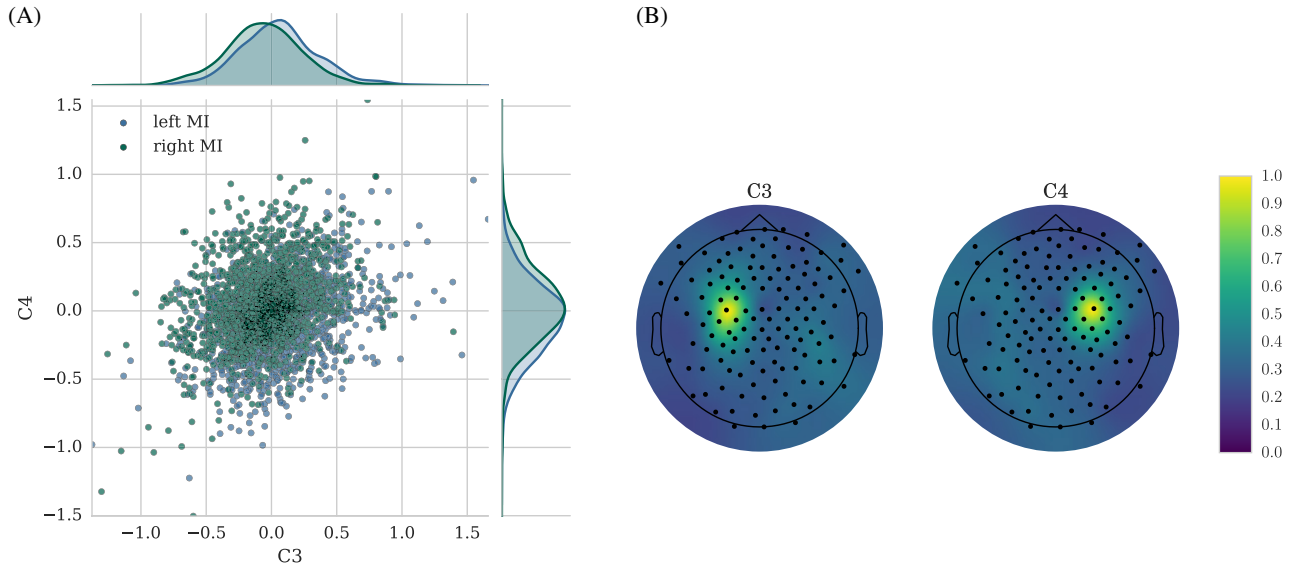

Figure 5. Characteristics of FM features in motor imagery. (A) Scatter plot of frequency figure in channels C3 and C4, plotted for all subjects after mean subtraction. (B) Plot of the correlation across trials of the frequency feature at channels C3 and C4 with the other 127 channels, averaged over subjects.

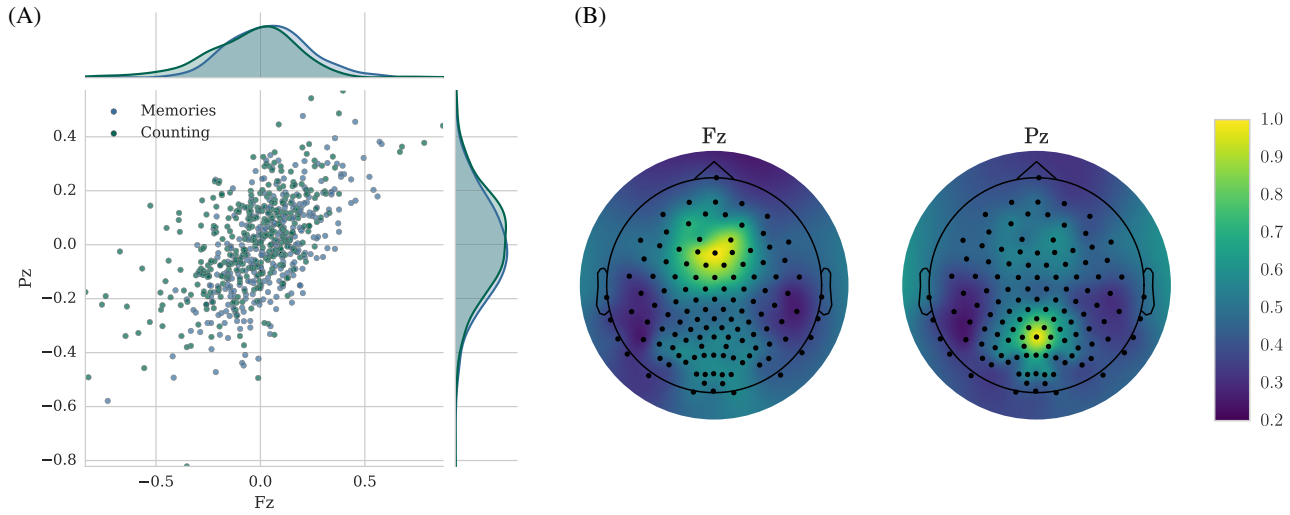

Figure 6. Characteristics of FM features in a cognitive paradigm for healthy subjects. (A) Scatter plot of frequency feature in channels Fz and Pz, plotted for all subjects after mean subtraction. (B) Plot of the correlation across trials of the frequency feature at channels Pz and Fz with the other 123 channels, averaged over subjects. Note that the colors go from 0.2 to 1 to emphasize the structure of the correlations.

with the other channels, the clear pattern we see in the healthy subjects on average is repeated, although the anteroposterior trend appears to be lessened.

5. Discussion

We extend the results of [37] with these experiments and show that the location of a neural frequency peak on the frequency axis can be used as a reliable feature for BCI decoding across paradigms and neural frequency bands, and further that it is stable to standard spatial filtering approaches. We found that across both paradigms, and independent of standard spatial filtering approaches, the joint feature space substantially increases the cross-validated accuracy over healthy subjects. Further, we find this effect is not limited to the obvious α peak but generalizes to less visually separable peaks such as the θ. Especially for studies with few electrodes, this has the potential to markedly improve performance by allowing another useful feature to be extracted without increasing the hardware; furthermore, this opens a new avenue for theoretical study, as processing techniques (whether in terms of temporal or spatial filtering) to find task-related frequency modulation in mixed signals do not, to our knowledge, exist.

Algorithmically, the method we use to compute the average frequency also leads to many questions. Most previous approaches to using instantaneous frequency, or even average frequency, have relied on autoregressive models and computed instantaneous frequency and amplitude from there [32, 33]. These require a separate optimization and must be given model orders in advance; we have shown that using a simple transform like the Hilbert is sufficient to get features for classification. However, that is not to say this is the only simple

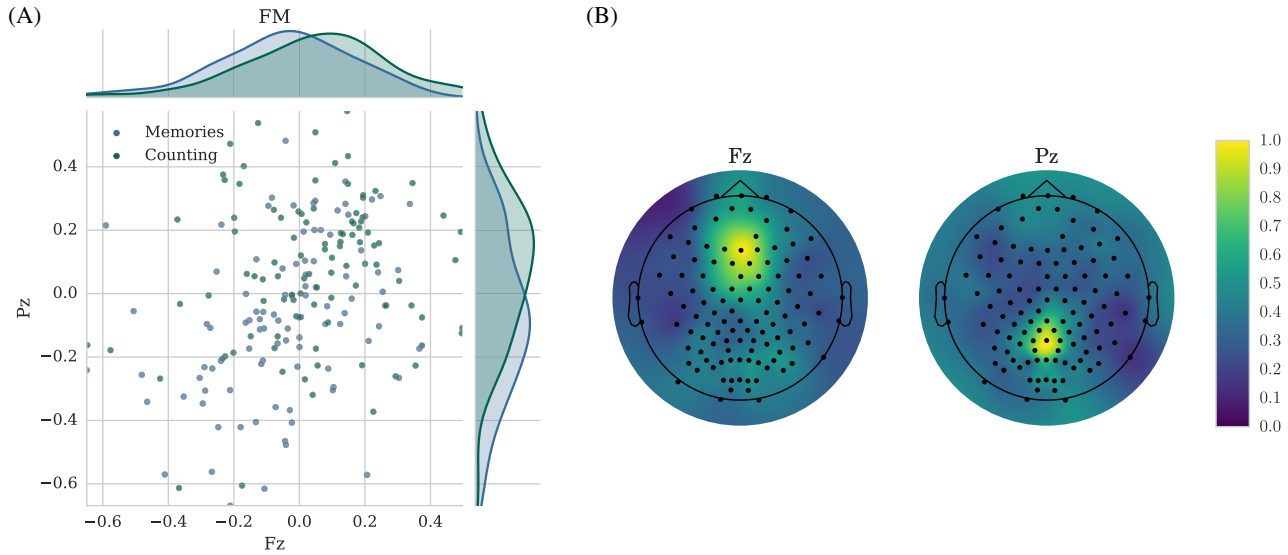

Figure 7. Characteristics of FM features in a cognitive paradigm in ALS patients. (A) Scatter plot of frequency feature in channels Fz and Pz, plotted for all subjects after mean subtraction. (B) Plot of the correlation across trials of the frequency feature at channels Pz and Fz with the other 123 channels, averaged over subjects.

option. The Hilbert transform on a bandpassed signal, and the median we use beyond it, were chosen for their stability to outliers of the instantaneous frequency within a trial. It is very likely that there are more stable approaches to determining the peak frequency yet to be discovered. It is further intriguing that the phase of a mixed signal is not a well-defined concept. What exactly does it mean that taking the Hilbert transform of a linearly mixed signal gives us a property of one of the component signals? How is this related to the frequency range or the number of signals being mixed?

Aside from such concerns, it is unclear how higher-level choices, like the location of the frequency band or the length of the mental task periods, affect the results. Standards for both of these were only set using AM features. As a preliminary *a posteriori* test of the influence of the range of the bandpass filter, we tested many different bandpasses for both features to determine which range performs best on average on the motor imagery data without spatial pre-processing. For both AM and FM features, we varied the frequency band between 6 and 17 Hz in 1 Hz steps for each subject. We then computed the cross-validated accuracy using the same procedure as explained in section 2 and averaged across subjects. The results can be seen in figure 8. As can be seen, the average best frequency band location is quite different for AM and FM features. AM features tend to prefer high bandpasses, while the FM features have their best average performance closer to the standard μ range. This could be due to many factors, but one to consider is that the β range is also predictive in motor imagery and begins, traditionally, at 13 Hz; perhaps, given the imperfections inherent in any discrete time bandpassing procedure, the AM features benefit from this bleeding over of the β range. Next, we tested how different trial lengths affected performance. Using the standard α band and varying the trial length between 1 and 7 s, we compared cross-validated performance on the same dataset, for which the results are shown in figure 9. Both feature spaces do better with longer trials, suggesting that the longer period of brain activity is helpful in finding the true median frequency as well as the true ERD.

It is also important to ask ourselves whether these results may be influenced by artifacts, and not represent features specifically coming from the brain. Looking at figures 5, 6 and 7 shows that the features from the chosen electrodes are almost uncorrelated with the FM features from frontal and peripheral channels that would be dominated by ocular or muscular artifacts, respectively. While such a group level analysis cannot guarantee that certain subjects were influenced by artifacts, it can make us very confident that at the group level, this analysis shows FM features to indeed be related to the brain, and slightly if at all to some sort of artifact.

Next, in a neuroscientific light, these findings are quite curious. The unique information that the FM feature appears to carry suggests that either it comes from the same neural source but has a different noise profile, or that it is generated from a different circuit than the one that generates task-related bandpower differences. Based on our preliminary CSP-based analysis, it appears to be the case that the same circuit generates both, at least some of the time, by optimizing spatial filters for power differences between conditions, we also see an increase in the FM feature predictability. Moreover, we find that optimizing for power differences actually causes the FM to become even more predictive than the AM in some cases. However, the fact that Laplacian filtering can sometimes change the ratio of the predictivity of FM versus AM features suggests that this is not the full story. Perhaps the robustness of FM features is due to the fact that they are generated by the same neural ensembles that generate AM features but also by other ensembles, as while the Laplace filter is well-known for narrowing the sensitivity of the electrode to the locations directly perpendicular to it, CSP filters have more freedom in how they optimize.

A major question for FM features is how robust they are to neurological disease. Patients represent the group most

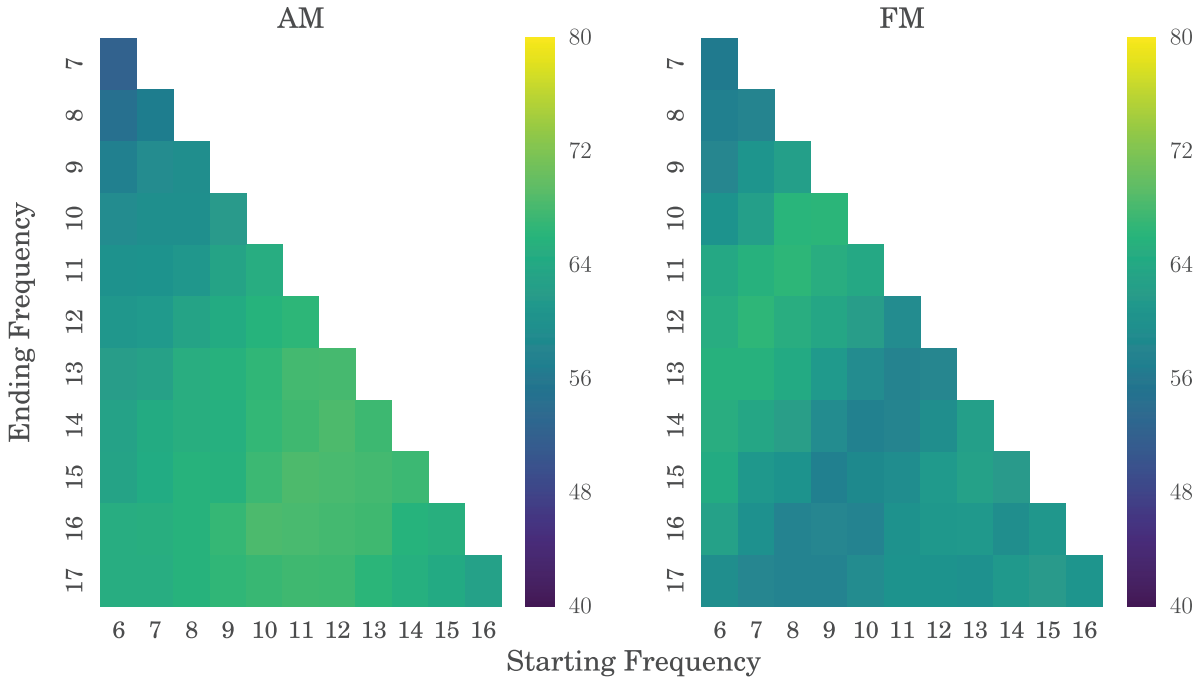

Figure 8. Cross-validated accuracy for the motor imagery data set using channels C3 and C4 for frequency bands of varying spectral location and size. The vertical axis shows ending frequency and the horizontal axis shows starting frequency for both AM (left) and FM (right) features. The colorbar shows the average cross-validated accuracy over all ten participants. Between bandpower and frequency features, the best band varies in both location and size.

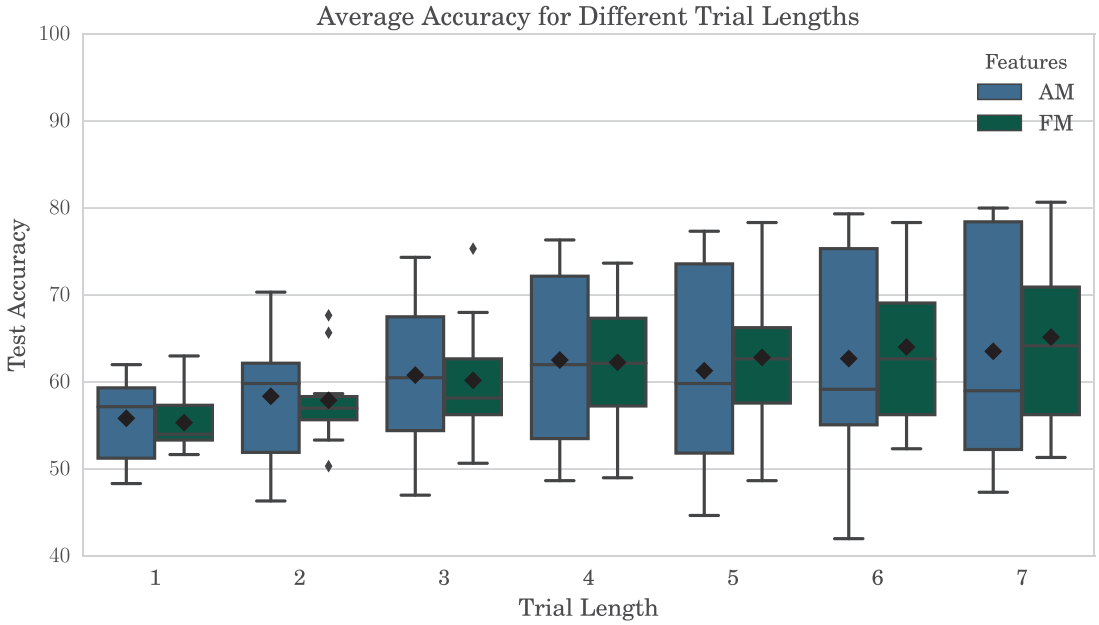

Figure 9. Cross-validated accuracy for the motor imagery data set using channels C3 and C4 for trial lengths between 1 and 7 s. The vertical axis shows cross-validated accuracy and the horizontal axis shows the length of the trial for AM (blue) and FM (green) features. Box plots are identically labelled to figures 1 and 2. For both feature spaces, average accuracy over subjects increases with trial length.

in need of the opportunities BCI offers, and so determining whether this feature is more or less robust for non-neurotypical individuals is a pressing question. The data we have presented does not show that FM is more robust than AM for ALS patients, but it also suggests that more research is required. All accuracies for ALS patients were computed using only 30 trials, half that of the healthy subjects, which makes the accuracy estimation particularly noisy. What we can see, however, is that in some patients the FM feature is quite predictive. Looking at the ALS-FRS-R scores of those patients, this includes one who is completely locked-in and one who is still in very early stages, suggesting that this feature at least needs to be looked into more. Given the amount of variance across healthy participants, attempting to average across both people and disease stages with five participants can give us a preliminary look at best.

Finally, and somewhat crucially, one might ask what the benefit of attempting to find a single feature type that is predictive across many tasks is, as most tasks in BCIs use very unique features to achieve optimum performance. However, this sort of task-specific feature engineering is limiting, in that it only allows for new paradigms to be discovered simultaneous with new feature development. The results shown here, as well as previous results in bandpower-related BCIs, are strong evidence to support the idea that there are common feature spaces across tasks. While it is probable that current optimal performance for a single subject on a single task requires specialized feature computation, the lack of scalability means the search for a generic method of computation is crucial to the development of BCIs in the future, leaving the task of optimizing for individual performance to more intelligent machine learning on these common feature spaces.

6. Conclusion

Our goal was to test whether FM features are usable for classification in a very simple way, and not to over-optimize. Given that we have shown this to be the case, the space for optimization is enormous. The projection of muscle and eye artifacts into the spectral power of measured signals is well known, but whether they have a similar task-related change in peak location is not yet studied. Perhaps FM features are more robust to artifacts than AM features, which would be a crucial step forward in the quest for more stable ways of reading the brain. Further, the cross-session and online reliability of these features must be shown.

ORCID iDs

Vinay Jayaram https://orcid.org/0000-0002-8581-4921

References

[1] Schwarz A, Scherer R, Steyrl D, Faller J and Müller-Putz G R 2015 A co-adaptive sensory motor rhythms brain-computer interface based on common spatial patterns and random forest *37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 1049–52

[2] Sannelli C, Vidaur C, Müller K-R and Blankertz B 2016 Ensembles of adaptive spatial filters increase BCI performance: an online evaluation *J. Neural Eng.* **13** 046003

[3] An H-S, Kim J-W and Lee S-W 2016 Design of an asynchronous brain-computer interface for control of a virtual avatar *4th Int. Winter Conf. on Brain-Computer Interface* (IEEE)

[4] CYBATHLON: Brain-computer interface race www.cybathlon.ethz.ch/cybathlon-2020/disciplines/bci-race.html

[5] Buzsáki G 2006 *Rhythms of the Brain* (Oxford: Oxford University Press)

[6] Aghaei A S, Mahanta M S and Plataniotis K N 2016 Separable common spatio-spectral patterns for motor imagery BCI systems *IEEE Trans. Biomed. Eng.* **63** 15–29

[7] Perdikis S, Leeb R and Millán J d R 2016 Context-aware adaptive spelling in motor imagery BCI *J. Neural Eng.* **13** 036018

[8] Steyrl D, Scherer R, Faller J and Müller-Putz G R 2016 Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier *Biomed. Eng.* **61** 77–86

[9] Fomina T, Lohmann G, Erb M, Ethofer T, Schölkopf B and Grosse-Wentrup M 2016 Self-regulation of brain rhythms in the precuneus: a novel BCI paradigm for patients with ALS *J. Neural Eng.*

[10] Pfurtscheller G and Neuper C 1997 Motor imagery activates primary sensorimotor area in humans *Neurosci. Lett.* **239** 65–8

[11] Pfurtscheller G and Lopes da Silva F H 1999 Event-related EEG/MEG synchronization and desynchronization: basic principles *Clin. Neurophysiol.* **110** 1842–57

[12] Koles Z J, Lazar M S and Zhou S Z 1990 Spatial patterns underlying population differences in the background EEG *Brain Topography* **2** 275–84

[13] Barachant A, Bonnet S and Congedo M 2010 Christian Jutten Riemannian geometry applied to BCI classification *Latent Variable Analysis and Signal Separation* (Berlin: Springer) pp 629–36

[14] Lachaux J-P *et al* 1999 Measuring phase synchrony in brain signals *Hum. Brain Mapp.* **8** 194–208

[15] Wang Y, Hong B and Gao X 2006 Shangkai gao phase synchrony measurement in motor cortex for classifying single-trial EEG during motor imagery *28th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 75–8

[16] Gysels E and Celka P 2004 Phase synchronization for the recognition of mental tasks in a brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **12** 406–15

[17] Loboda A, Margineanu A, Rotariu G and Lazar A M 2014 Discrimination of EEG-based motor imagery tasks by means of a simple phase information method *Int. J. Adv. Res. Artif. Intell.* **3**

[18] Brunner C, Scherer R, Graimann B, Supp G and Pfurtscheller G 2006 Online control of a brain-computer interface using phase synchronization *IEEE Trans. Biomed. Eng.* **53** 2501–6

[19] Onton J, Delorme A and Makeig S 2005 Frontal midline EEG dynamics during working memory *Neuroimage* **27** 341–56

[20] Hamner B, Leeb R, Tavella M and Millán J D R 2011 Phase-based features for motor imagery brain-computer interfaces *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 2578–81

[21] Ahn M and Jun S C 2015 Performance variation in motor imagery brain-computer interface: a brief review *J. Neurosci. Methods* **243** 103–10

[22] Blankertz B, Sannelli C, Halder S, Hammer E M, Kübler A, Müller K-R, Curio G and Dickhaus T 2010 Neurophysiological predictor of SMR-based BCI performance *Neuroimage* **51** 1303–9

[23] Marchetti M and Priftis K 2014 Brain-computer interfaces in amyotrophic lateral sclerosis: a metanalysis *Clin. Neurophysiol.* **126** 1255–63

[24] Aurlien H, Gjerde I O, Aarseth J H, Eldøen G, Karlsen B, Skeidsvoll H and Gilhus N E 2004 EEG background activity described by a large computerized database *Clin. Neurophysiol.* **115** 665–73

[25] Smit D J A, Posthuma D, Boomsma D I and de Geus E J C 2005 Heritability of background EEG across the power spectrum *Psychophysiology* **42** 691–7

[26] Bodenmann S, Rusterholz T, Dürr R, Stoll C, Bachmann V, Geissler E, Jaggi-Schwarz K and Landolt H-P 2009 The functional val158met polymorphism of comt predicts interindividual differences in brain α oscillations in young men *J. Neurosci.* **29** 10855–62

[27] Klimesch W 1999 EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis *Brain Res. Rev.* **29** 169–95
[28] Haegens S, Cousijn H, Wallis G, Harrison P J and Nobre A C 2014 Inter-and intra-individual variability in alpha peak frequency *Neuroimage* **92** 46–55
[29] Osaka M 1984 Peak alpha frequency of EEG during a mental task: task difficulty and hemispheric differences *Psychophysiology* **21** 101–5
[30] Doppelmayr M, Klimesch W, Pachinger T and Ripper B 1998 Individual differences in brain dynamics: important implications for the calculation of event-related band power *Biol. Cybern.* **79** 49–57
[31] Feshchenko V A 1994 The way to EEG-classification: transition from language of patterns to language of systems *Int. J. Neurosci.* **79** 235–49
[32] Nguyen D P, Wilson M A, Brown E N and Barbieri R 2009 Measuring instantaneous frequency of local field potential oscillations using the Kalman smoother *J. Neurosci. Methods* **184** 365–74
[33] Foffani G, Bianchi A M, Baselli G and Priori A 2005 Movement-related frequency modulation of beta oscillatory activity in the human subthalamic nucleus *J. Physiol.* **568** 699–711
[34] Gharieb R R and Cichocki A 2001 Segmentation and tracking of the electro-encephalogram signal using an adaptive recursive bandpass filter *Med. Biol. Eng. Comput.* **39** 237–48
[35] Piazza C, Cantiani C, Tacchino G, Molteni M, Reni G and Bianchi A M 2014 ERP and adaptive autoregressive identification with spectral power decomposition to study rapid auditory processing in infants *36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 4591–4
[36] Wu C-H, Chang H-C, Lee P-L, Li K-S, Sie J-J, Sun C-W, Yang C-Y, Li P-H, Deng H-T and Shyu K-K 2011 Frequency recognition in an ssvep-based brain computer interface using empirical mode decomposition and refined generalized zero-crossing *J. Neurosci. Methods* **196** 170–81
[37] Médl A and Flotzinger D 1992 Gert Pfurtscheller Hilbert-transform based predictions of hand movements from eeg measurements *14th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 6 (IEEE) pp 2539–40
[38] Falzon O, Camilleri K P and Muscat J 2012 The analytic common spatial patterns method for EEG-based BCI data *J. Neural Eng.* **9** 045009
[39] Park C, Looney D, Kidmose P, Ungstrup M and Mandic D P 2011 Time-frequency analysis of eeg asymmetry using bivariate empirical mode decomposition *IEEE Trans. Neural Syst. Rehabil. Eng.* **19** 366–73
[40] Hohmann M R, Fomina T, Jayaram V, Förster C, Just J, Synofzik M, Schölkopf B, Schöls L and Grosse-Wentrup M 2016 An improved cognitive brain-computer interface for patients with amyotrophic lateral sclerosis *Proc. of the 6th Int. BCI Meeting*
[41] Cedarbaum J M, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B and Nakanishi A 1999 The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function *J. Neurol. Sci.* **169** 13–21
[42] Hjorth B 1975 An on-line transformation of EEG scalp potentials into orthogonal source derivations *Electroencephal. Clin. Neurophysiol.* **39** 526–30
[43] Ramoser H, Müller-Gerking J and Pfurtscheller G 2000 Optimal spatial filtering of single trial EEG during imagined hand movement *IEEE Trans. Rehabil. Eng.* **8** 441–6
[44] Clochon P, Fontbonne J-M, Lebrun N and Etévenon P 1996 A new method for quantifying EEG event-related desynchronization: amplitude envelope analysis *Electroencephal. Clin. Neurophysiol.* **98** 126–9
[45] Knosche T R and Bastiaansen M C M 2002 On the time resolution of event-related desynchronization: a simulation study *Clin. Neurophysiol.* **113** 754–63
[46] Mu Y and Han S 2010 Neural oscillations involved in self-referential processing *Neuroimage* **53** 757–68
[47] Sauseng P, Klimesch W, Schabus M and Doppelmayr M 2005 Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory *Int. J. Psychophysiol.* **57** 97–103
[48] Fomina T, Hohmann M, Schölkopf B and Grosse-Wentrup M 2015 Identification of the default mode network with electroencephalography *37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE) pp 7566–9

2.4 Paper 3: MOABB: Trustworthy algorithm benchmarking for BCIs

MOABB: Trustworthy algorithm benchmarking for BCIs

Vinay Jayaram*†, Alexandre Barachant‡
* Max Planck Institute for Intelligent Systems, Department Empirical Inference, Tübingen, Germany
Email: vjayaram@tue.mpg.de
† IMPRS for Cognitive and Systems Neuroscience, University of Tübingen, Tübingen, Germany
‡ CTRL-Labs, New-York, USA
Email: alexandre.barachant@gmail.com

***Abstract*—BCI algorithm development has long been hampered by two major issues: Small sample sets and a lack of reproducibility. We offer a solution to both of these problems via a software suite that streamlines both the issues of finding and preprocessing data in a reliable manner, as well as that of using a consistent interface for machine learning methods. By building on recent advances in software for signal analysis implemented in the MNE toolkit, and the unified framework for machine learning offered by the scikit-learn project, we offer a system that can improve BCI algorithm development. This system is fully open-source under the BSD licence and available at https://github.com/NeuroTechX/moabb. To validate our efforts, we analyze a set of state-of-the-art decoding algorithms across 12 open access datasets, with over 250 subjects. Our analysis confirms that different datasets can result in very different results for identical processing pipelines, highlighting the need for trustworthy algorithm benchmarking in the field of BCIs, and further that many previously validated methods do not hold up when applied across different datasets, which has wide-reaching implications for practical BCIs.**

I. Introduction

Brain-computer interfaces (BCIs) have long presented the neuroscience methods community with a unique challenge. Unlike in vision research, where one has a database of images and labels, a BCI is defined by a signal recorded from the brain and fed into a computer, which can be influenced in any number of ways both by the subject and by the experimenter. As a result, validating approaches has always been a difficult task. Number of channels, requested task, physical setup, and many other features vary between the numerous publically available datasets online, not to mention issues of convenience such as file format and documentation. Because of this, the BCI methods community has long done one of two things to validate an new approach: Recorded a new dataset, or used one of few well-known, tried-and-true datasets.

Recording a new dataset, the ideal way to show that a proposed method works in practice, presents problems for post-hoc analysis. Without making data public, it is impossible to know whether offline classification results are convincing or due to some coding issue or recording artifact. Further, it is well-known that differences in hardware [21, 28], paradigm [2], and subject [2] can have large differences in the outcome of a BCI task, making it very difficult to generalize findings from any single dataset.

Over the years many datasets have been published online, and serve as an attractive option when time or hardware do not permit recording a new one. In the last year and a half, over a thousand journal and conference submissions have been written on the BCI Competition III [5, 27] and IV [32] datasets. Considering that these datasets have been available publically for over a decade, the true number of papers which validate results against them is likely much higher. While it is impossible to deny the impact these two datasets have had on the field, relying so heavily on a small number of datasets – with less than 50 subjects total – exposes the field to several important issues. In particular, overfitting to the setups offered there is likely.

Lastly, and possibly most problematically, the scarcity of available code for BCI algorithms old and new puts the onus on each individual lab to reproduce the code for all other competing methods in order to make a claim to be comparable with the 'state-of-the-art' (SOA). As a result, the vast majority of novel BCI algorithm papers compare either against other work from the same lab, or old, easily implementable standards such as CSP [19] or channel-level variances combined with a classifier of choice [12].

Computer vision has solved this problem with enormous datasets like Imagenet [10] bundled with machine learning packages (Tensorflow[1], PyTorch, and Theano[23]). However, generating BCI data is often a very taxing process both physically and mentally, and so it is not reasonable to create datasets of such size. Rather, the field requires many different people recording data in many contexts in order to create an appropriate benchmark. We propose our platform, the MOABB (Mother Of All BCI Benchmarks) Project, as a candidate for this application. The MOABB project consists of the aggregation of many publicly available EEG datasets, converted to a common format and bundled in the software package, as well as a collection of SOA algorithms. Using this system researchers can to automatically benchmark those algorithms and run an automated statistical analysis, making the process of validating new algorithms painless and reproducible. The source code is written in Python and publically available under the BSD licence at https://github.com/NeuroTechX/moabb.

As an initial validation of this project, we present results on the constrained task of binary classification in two-class imagined motor imagery, as that is the most widely used motor

imagery paradigm and allows us to demonstrate the process across the largest number of datasets. However, we note that this is only the first question we attempt to answer in this field. The format allows for many other questions, including different channel types (EEG, fNIRS, or other), multi-class paradigms, and also transfer learning scenarios as described in [17].

II. Methods

Any BCI analysis is defined by three elements: A dataset, a context, and a pipeline. Here we describe how all of these components are dealt with within our framework, and how specifically we set the options for the initial analyses presented here.

A. Datasets

Public BCI datasets exist for a wide range of user paradigms and recording conditions, from continuous usage to single-session to multiple-sessions-per-subject. Within the current MOABB project, we have unified the access to many datasets, described in Table I.

Adding new open-source datasets is also simple via the MNE toolkit [14, 15], which is used for all preprocessing and channel selection. Any dataset that can be made compatible with their framework can quickly be added to the set of data offered by this project. In addition, the project offers test functions to ensure candidate code conforms to the software interface.

B. Context

A *context* is the set of characteristics that defines the preprocessing and validation procedure. To go from a recorded EEG time-series to a pipeline performance value for a given subject or recording session, many parameters must be defined. First, trials need to be cut out of the continuous signal and pre-processed, which is possible in many different ways when taking into account parameters such as trial overlap, trial length, imagery type, and more. Once the continuous data is processed into trials, and these trials are fed into a pipeline, the next question of how to create training and test sets, and how to report performance, comes into play. We separate these two notions in our software and call them the *paradigm* and the *evaluation* respectively.

1) Paradigm: A paradigm defines how one goes from continuous data to trials for a standard machine learning pipeline to deal with. While not an issue in image processing, as each trial is just one image, it is crucial in EEG and biosignals processing because most datasets do not have exactly the same events defined in the continuous data. For example, many datasets with two-class motor imagery use left versus right hand, while some use hands versus feet; there are also many possible non-motor imageries. For any reasonable analysis the specific sort of imagery or ERP must be controlled for, as they all have different characteristics in the data and further are variably effective across subjects [2, 26]. After choosing which events or imageries are valid, the question comes to pre-processing of the continuous data, in the form of ICA cleaning, bandpass filtering, and so on. These must also be identical for valid comparisons across algorithm or datasets. Lastly, there are questions of how to cut the data into trials: What is the trial length and overlap; or, in the case of ERP paradigms, how long before and after the event marker do we use? The answers to all these questions are summed up in the paradigm object.

2) Evaluation: Once the data is split into trials and a pipeline is fixed, there are many ways to train and test this pipeline to minimize overfitting. For datasets with multiple subjects recorded on multiple days, we may want to determine which algorithm functions best in multi-day classification. Or, we may want to determine which algorithm is best for small amounts of training data. It is easy to see that there are many possibilities for splitting data into train and test sets depending on the question to be answered, and these must be fixed identically for a given analysis. Furthermore, there is the question of how to report results. Multiclass problems cannot use metrics like the ROC-AUC which provide unbiased estimates of classifier goodness in binary cases; depending on things like the class balance, various other metrics have their own benefits and pitfalls. Therefore this must also be fixed across all datasets, contingent on the class of predictions the pipelines attempt to make. We define this as our *evaluation*.

C. Pipeline

We define a *pipeline* as the processing that takes one from raw trial-wise data into labels, taking both spatial filtering and classification model fitting into account. A convenient API for dealing with this kind of processing is defined by scikit-learn [24], which allows for easily definable dimensionality reduction, feature generation, and model fitting. To maximize reproducibility we allow pipelines to be defined either by yaml files or through python files that generate the objects, but force all machine learning models to follow the scikit-learn interface.

In essence, the MOABB combines the preceding components into a procedure that takes a list of algorithms and datasets and trains each pipeline to each subject or recording session independently in order to generate goodness-of-fit scores such as accuracy or ROC-AUC. These scores can then be visualized and used for statistical testing.

III. Statistical Analysis

At the end of the MOABB procedure there are scores for every subject in every dataset with every pipeline. The goal of this project is to synthesize these numbers into an estimate of how likely it is that each pipeline out-performs the other pipelines. However, even if imagery type and channel number were held constant, differences in trial amount, sampling rate, and even location and hardware mean that we cannot expect subjects across datasets to be naively comparable. Therefore, we run independent statistical tests within each dataset and combine the p-values afterwards. A secondary problem is that the difference distribution for two algorithms within a given dataset is very unlikely to be Gaussian. It is well-known that

Name	Imagery	# Channels	# Trials	# Sessions	# Subjects	Epoch	Citations
Cho et al. 2017	Right, left hand	64	200	1	49	0-3s	[9]
Physionet	Right, left hand	64	40-60	1	109	1-3s	[13, 25]
Shin et al. 2017	Right, left hand	25	60	3	29	0-10s	[6, 29]
BNCI 2014-001	Right, left hand	22	144	2	9	2-6s	[32]
BNCI 2014-002	Right hand, feet	15	160	1	14	3-8s	[30]
BNCI 2014-004	Right, left hand	3	120-160	5	9	3-7.5s	[20]
BNCI 2015-001	Right hand, feet	13	200	2/3	13	3-8s	[11]
BNCI 2015-004	Right hand, feet	30	70-80	2	10	3-10s	[26]
Alexandre Motor Imagery	Right hand, feet	16	40	1	9	0-3s	[3]
Yi et al. 2014	Right, left hand	60	160	1	10	3-7s	[33]
Zhou et al. 2016	Right, left hand	14	100	3	4	1-6s	[35]
Grosse-Wentrup et al. 2009	Right, left hand	128	300	1	10	3-10s	[16]
Total:					**275**		

TABLE I: Dataset attributes

some subjects are BCI illiterate [2], which implies that no pipeline can reliably out-predict another one on that subset of subjects. Therefore, for large enough datasets, the distribution of differences in pipeline scores is very likely to be at least bimodal.

To deal with this issue while also keeping the framework running fast enough to execute on a normal desktop, we use a mixture of permutation and non-parametric tests. Within each dataset, either a one-tailed permutation-based paired t-test (for datasets with less than 20 subjects) or a Wilcoxon signed-rank test is run for each pair of pipelines, generating a p-value for the hypothesis that pipeline a is bigger than pipeline b for each pair of pipelines. These p-values are combined via Stouffer's method[31], with a weighting given by the square root of the number of subjects as suggested in [7], to return a final p-value for each hypothesis. Since each score is compared against $N_{pipelines} - 1$ other scores for the same subject, we also apply Bonferroni correction to protect against false positives. In order to determine effect size, we computed the standardized mean difference within datasets and combined them using the same weighting as was given to Stouffer's method.

IV. Experiment

To show off the possibilities of this framework, we ran various well-known BCI pipelines from across many papers in order to conduct the first big-data, side-by-side analysis of the state of the art in motor imagery BCIs.

A. Context

For the paradigm, we choose to look at datasets including motor imagery. Motor imagery is the most-studied sort of imagery for BCIs [34], and we further limit ourselves to the binary case as this has not yet been solved. For evaluations, we choose within-session cross-validation, as this represents the best-case scenario for any pipeline, with minimal non-stationarity.

1) Paradigm: As there are many methods that show that multiple frequency bands can lead to improved BCI performance[18], and further that discriminative data is concentrated in the anatomical frequency bands, we test two preprocessing pipelines: A single bandpass containing both the alpha and beta ranges, from $8 - 35$Hz, and another from $8 - 35$Hz in 4Hz increments. All data was also subsampled to 128Hz, as the memory requirements became prohibitive otherwise.

2) Evaluation: The evaluation was chosen to be within-session, as that minimizes the effect of non-stationarity. As this is a binary classification task, the ROC-AUC score was chosen as the metric to score 5-fold cross validation (the splits were kept identical for all pipelines in a given subject). In comparison with the more interpretable classification accuracy, the ROC-AUC is less sensitive to imbalanced classes, which is important in this case where the datasets vary heavily. In order to return a single score per subject, the scores from each session were averaged when multiple sessions were present.

B. Pipelines

We implement a selection of pipelines from the BCI literature, as well as the well-known standards of CSP + LDA and channel-level variances + SVM. Specific implemented pipelines are in Table II; all hyperparameters were set via cross-validation.

V. Results

Figure 1 shows all the results generated by this entire processing chain. Surprisingly, perhaps, the pipelines do not clearly cluster on the dataset level, making it unclear which ones perform best from simply this plot. What is very clear, however, is that different datasets have very different average scores independent of pipeline. This is particularly true when one considers the case of [35] versus [13]: Zhou et al [35] had pre-trained subjects, which compared to the naive sample in the Physionet database makes a drastic difference.

Figure 2 shows the difference between CSP and the channel log-variance and tangent space methods, as these are all well-known approaches and have been compared against each other often in the past. Based on this meta-analysis, CSP reliably out-performs channel log-variances across datasets – however, there are datasets such as [16] and [26] in which the opposite trend is shown. Similarly, while the tangent space projection method normally out-performs CSP, that is also not true for half of the sampled datasets. The confidence intervals also show why this is likely the case – for studies with very few subjects, such as [35], the confidence intervals make even very strong standardized effects quite untrustworthy.

Figure 3 compares CSP against commonly used variants. Here, the difference is heavily dependent on dataset and no clear trend is visible. It is interesting to note that in the

Name	Preprocessing	Classifier	Introduced in
CSP + LDA	Trial covariances estimated via maximum-likelihood with unregularized common spatial patterns (CSP). Features were log variance of the filters belonging to the 6 most diverging eigenvalues	Linear Discriminant Analysis (LDA)	[19]
DLCSPauto + shLDA	Trial covariances estimated by OAS [8] followed by unregularized CSP. Features were log variance on the 6 top filters.	LDA with Ledoit-Wolf shrinkage of the covariance term	[22]
TRCSP + LDA	CSP with Tikhonov regularization, features were log variance on the 3 best filters for each class	LDA	[22]
FBCSP + optSVM	Filter bank of 6 bands between 8 and 35 Hz followed by OAS covariance estimation and unregularized CSP. Log variance from each of the 4 top filters from each sub-band were pooled and the top 10 features chosen by mutual information were used.	A linear support vector machine was trained with its regularization hyperparameter set by a cross-validated grid-search from $[0.01 100]$.	[18]
TS + optSVM	Trial covariances estimated via OAS then projected into the Riemannian tangent space to obtain features	Linear SVM with identical grid-search	[4]
AM + optSVM	Log variance in each channel	Linear SVM with grid-search	N/A

TABLE II: Processing pipelines

case of filter-bank CSP, the BNCI 2014 datasets (which are included in the BCI Competition datasets used in [18]) show FBCSP to out-perform regular CSP while the opposite is true for others such as Physionet. We further confirm the result from [22] that regularizing the covariance estimates does not improve the results of CSP. However, somewhat surprisingly, the finding that Tikhonov weighting increases performance was not validated in this analysis.

The meta-effects shown in Figures 2 and 3 are summed up in Figure 4, which displays the meta-effect size in cases that the algorithm on the y-axis significantly out-performed the algorithm on the x-axis according to the statistical procedure outlined in Section III, as well as the significance denoted by the stars under the meta-effect size. Here we can see that all other algorithms out-performed log-variance features on average (though with significant variance over datasets as seen in the other figures) and that among CSP and its variants, tangent space projection is better.

VI. Discussion

We present a system for reliably comparing BCI pipelines that is both easily extended to incorporate new datasets and equipped with an automated statistical procedure for determining which pipelines perform best. Furthermore, this system defines a simple interface for submitting and validating new BCI pipelines, which could serve to unify the many methods that exist so far. To test that system, we present results using standard pipelines in contexts that have wide relevance to the BCI community. By looking across multiple, large datasets, it is possible to make statements about how BCIs perform on average, without any sort of expert tuning of the processing chain, and further to see where the major pitfalls still lie.

The results of this analysis suggest that many well-known methods do not reliably out-perform simpler ones, despite the small-scale studies done years ago to validate them. In particular, the world of CSP regularization literature does not appear to have the effect that was originally claimed. Rather, the major difference in BCI classification isn't actually the algorithm, as of now, but the recording and human paradigm characteristics. The two most clear findings to come out of this are that log variances on the channel level are almost never better than CSP or Riemannian methods, and that the tangent space classification pipeline is the best of the tested models for single-session classification.

In particular in the cases of FBCSP and the regularized approaches presented here, the results presented here are surprising finding as they go against the results reported in the original papers. In the case of FBCSP, we perform similarly to the results shown in [18]. BNCI 2014-001 and 2014-004 are originally from the BCI competitions and were used in the original paper, and our finding is that on these datasets FBCSP indeed out-performed regular CSP. In the case of the regularized variants DLCSPauto and TRCSP, our results on the BCI competition data do not actually follow the originally reported trend. Some possible for reasons are the following: our use of single-session recordings ignores the initial training and test distinctions given within the competition, and we also used the AUC-ROC instead of the accuracy that was reported in the initial analysis. The full code to replicate these results is available publically, and so we hope we can at least rule out improper coding as a source of error.

Looking at these findings, it is particularly interesting to look at the case of filter-bank CSP versus CSP, as in this analysis the significance goes in both directions depending on the dataset. Since datasets vary in many characteristics, such as channel number, imagery type, and trial time, it is hard to determine what exactly underlies this diverging performance – but it is likely that this is not purely by chance. With increasing numbers of available datasets, however, the answers to such differences become possible. If we have many different situations in which to test algorithms, we can determine what factors contribute to the differences in performance between them. It is also important to emphasize that the results shown here must be taken in context. All results were generated by cross-validation within single recording sessions, which limits the possible non-stationarity. Because of this, regularization is at its least useful – which means that it would be inappropriate to dismiss regularization in the case of CSP out of hand. Rather, this same analysis should be re-run in the case of cross-session classification, a task that is currently infeasible due to the number of multi-session datasets.

Fig. 1: Visualization of all generated scores, across all datasets. The dotted line corresponds to a chance level performance of 0.5

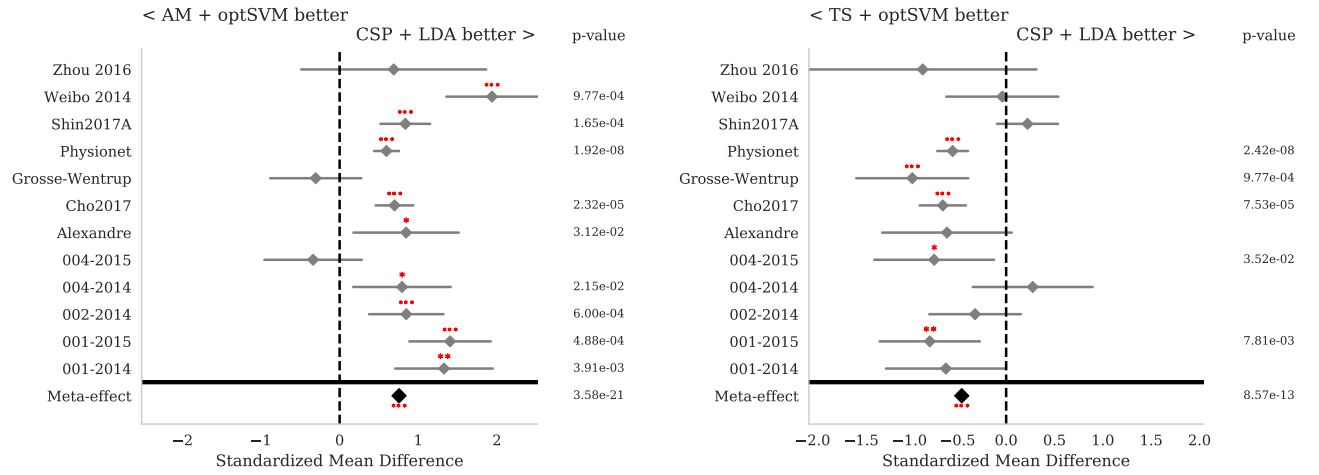

Fig. 2: Meta-analysis style plots showing the performance of log variance features (A) and tangent space features (B) both compared against CSP. The effect sizes shown are standardized mean differences, with p-values corresponding to the one-tailed Wilcoxon signed-rank test for the hypothesis given at the top of the plot and 95% interval denoted by the grey bar. Stars correspond to ***=p<0.001, **=p<0.01, *=p<0.05. The meta-effect is shown at the bottom of the plot. While there is a significant amount of variance between datasets–variance that could give contradictory results if these datasets were evaluated in isolation–the overall trend shows that CSP is on average better than channel log-variances and worse than tangent space projection.

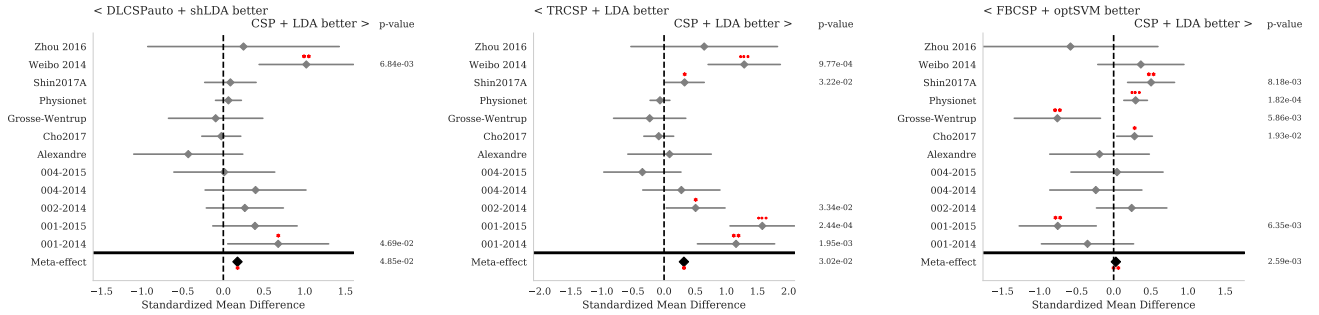

Fig. 3: Meta-analysis style plots showing the performance of CSP versus CSP variants: DLCSPauto(A), TRCSP (B), and filter-bank CSP (C). The effect sizes shown are standardized mean differences, with p-values corresponding to the one-tailed Wilcoxon signed-rank test for the hypothesis given at the top of the plot and 95% interval denoted by the grey bar. Stars correspond to ***=p<0.001, **=p<0.01, *=p<0.05. The meta-effect is shown at the bottom of the plot. While there is a significant amount of variance between datasets–variance that could give contradictory results if these datasets were evaluated in isolation–the overall trend shows that CSP out-performs the other algorithms in this setting.

VII. Conclusion

Meta-analysis is a well-described tool in other scientific fields to attempt to synthesize the effects of many different studies that all bear on the same, or very similar hypotheses. Though its use in BCIs has been hampered by the difficulties involved in gathering the data and algorithms in a single place, the MOABB project has the potential to offer a solution to this problem. The analysis here, though done with over 250 subjects, is still only a fraction of the number of subjects recorded for BCI publications over the years. With more papers that describe more varied setups, the power of this system can only grow, and what this analysis shows most clearly is that the sample size problem in BCIs is bigger than we might have expected. By gathering the data and offering a system for testing algorithms, we hope that this platform in the coming years can help to solve it.

Acknowledgements

We would like to extend our thanks to Dr. Marco Congedo for his valuable input regarding the appropriate statistical procedure for this analysis, and also to the NeuroTechX community for helping to get this project started.

corresponding author: Vinay Jayaram (email: vjayaram@tue.mpg.de)

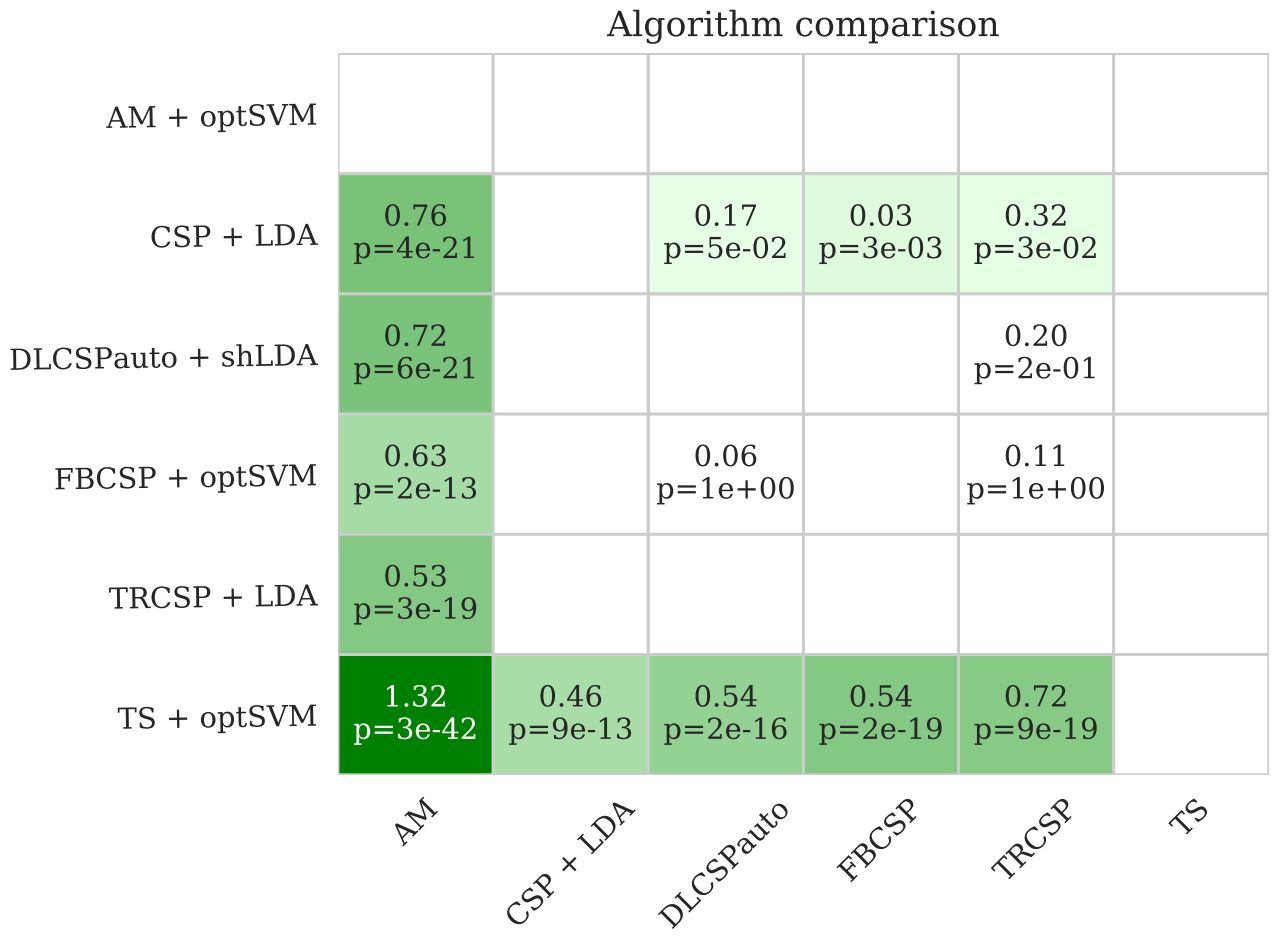

Fig. 4: Ranking of algorithms in performance across all datasets with statistics generated as defined in section III. As all p-values are single-sided, in the case that the effect goes in the opposite direction of the hypothesis the values are removed for clarity. The values correspond to the standardized mean difference of the algorithm in the y-axis minus that in the x-axis and the associated p-value

REFERENCES

[1] Martn Abadi et al. *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*. 2016. URL: https://www.tensorflow.org/.

[2] Brendan Allison et al. "BCI demographics: How many (and what kinds of) people can use an SSVEP BCI?" In: *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 18.2 (2010), pp. 107–116.

[3] Alexandre Barachant. "Commande robuste d'un effecteur par une interface cerveau machine EEG asynchrone". In: (2012). URL: https://tel.archives-ouvertes.fr/tel-01196752/.

[4] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. "Classification of covariance matrices using a Riemannian-based kernel for BCI applications". In: *Neurocomputing* 112 (2013), pp. 172–178. ISSN: 0925-2312. DOI: 10.1016/J.NEUCOM.2012.12.039. URL: https://www.sciencedirect.com/science/article/pii/S0925231213001574.

[5] B Blankertz et al. "The BCI Competition III: Validating Alternative Approaches to Actual BCI Problems". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14.2 (2006), pp. 153–159.

[6] B Blankertz, G Dornhege, M Krauledat, K R Müller, and G Curio. "The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects". In: *NeuroImage* 37.2 (2007), pp. 539–550.

[7] L Bovino et al. "On the combination of". In: *Analysis* 5.33 (2003), pp. 42–54.

[8] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. "Shrinkage Algorithms for MMSE Covariance Estimation". In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5016–5029. ISSN: 1053-587X. DOI: 10.1109/TSP.2010.2053029. URL: http://ieeexplore.ieee.org/document/5484583/.

[9] Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. "EEG datasets for motor

imagery brain-computer interface". In: *GigaScience* 6.7 (2017), pp. 1–8. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix034. URL: http://academic.oup.com/gigascience/article/6/7/1/3796323/EEG-datasets-for-motor-imagery-braincomputer.

[10] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009, pp. 248–255.

[11] J. Faller, C. Vidaurre, T. Solis-Escalante, C. Neuper, and R. Scherer. "Autocalibration and Recurrent Adaptation: Towards a Plug and Play Online ERD-BCI". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20.3 (2012), pp. 313–319. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2012.2189584. URL: http://ieeexplore.ieee.org/document/6177271/.

[12] D. Garrett, D.A. A Peterson, C.W. W Anderson, and M.H. H Thaut. "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11.2 (2003), pp. 141–144. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2003.814441. URL: http://ieeexplore.ieee.org/document/1214704/.

[13] Ary L Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet". In: *Circulation* 101.23 (2000), e215 LP –e220. URL: http://circ.ahajournals.org/content/101/23/e215.abstract.

[14] Alexandre Gramfort et al. "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroscience* 7 (2013), p. 267. ISSN: 1662453X. DOI: 10.3389/fnins.2013.00267. URL: http://journal.frontiersin.org/article/10.3389/fnins.2013.00267/abstract.

[15] Alexandre Gramfort et al. "MNE software for processing MEG and EEG data". In: *NeuroImage* 86 (2014), pp. 446–460. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2013.10.027. URL: http://www.ncbi.nlm.nih.gov/pubmed/24161808http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3930851http://linkinghub.elsevier.com/retrieve/pii/S1053811913010501.

[16] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss. "Beamforming in Noninvasive BrainComputer Interfaces". In: *IEEE Transactions on Biomedical Engineering* 56.4 (2009), pp. 1209–1219. ISSN: 0018-9294. DOI: 10.1109/TBME.2008.2009768. URL: http://ieeexplore.ieee.org/document/4694120/.

[17] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. "Transfer Learning in Brain-Computer Interfaces". In: *Computational Intelligence Magazine, IEEE* 11.1 (2016), pp. 20–31.

[18] Kai Keng Ang, Zhang Yang Chin, Haihong Zhang, and Cuntai Guan. "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 2390–2397. ISBN: 978-1-4244-1820-6. DOI: 10.1109/IJCNN.2008.4634130. URL: http://ieeexplore.ieee.org/document/4634130/.

[19] Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. "Spatial patterns underlying population differences in the background EEG". In: *Brain Topography* 2.4 (1990), pp. 275–284.

[20] R. Leeb et al. "Brain-Computer Communication: Motivation, Aim, and Impact of Exploring a Virtual Apartment". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15.4 (2007), pp. 473–482. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2007.906956. URL: http://ieeexplore.ieee.org/document/4359220/.

[21] Miguel Angel Lopez-Gordo, Daniel Sanchez-Morillo, and F Pelayo Valle. "Dry EEG electrodes". In: *Sensors* 14.7 (2014), pp. 12847–12870.

[22] F Lotte and C Guan. "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms". In: *IEEE Transactions on Biomedical Engineering* 58.2 (2011), pp. 355–362.

[23] Adam Paszke et al. *Automatic differentiation in PyTorch*. 2017. URL: https://openreview.net/forum?id=BJJsrmfCZ.

[24] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830. ISSN: ISSN 1533-7928. URL: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[25] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. "BCI2000: A General-Purpose Brain-Computer Interface (BCI) System". In: *IEEE Transactions on Biomedical Engineering* 51.6 (2004), pp. 1034–1043. ISSN: 0018-9294. DOI: 10.1109/TBME.2004.827072. URL: http://www.ncbi.nlm.nih.gov/pubmed/15188875http://ieeexplore.ieee.org/document/1300799/.

[26] Reinhold Scherer et al. "Individually Adapted Imagery Improves Brain-Computer Interface Performance in End-Users with Disability". In: *PLOS ONE* 10.5 (2015). Ed. by Luigi Bianchi, e0123727. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123727. URL: http://dx.plos.org/10.1371/journal.pone.0123727.

[27] A Schloegl. *Results of the BCI Competition 2005 for data set IIIa and IIIb*. Tech. rep. Institute for Human-Computer Interfaces - BCI Lab, University of Technology Graz, Austria, 2005.

[28] A Searle and L Kirkup. "A direct comparison of wet, dry and insulating bioelectric recording electrodes". In: *Physiological measurement* 21.2 (2000), p. 271.

[29] Jaeyoung Shin et al. "Open Access Dataset for EEG+NIRS Single-Trial Classification". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017), pp. 1735–1745. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2016.2628057. URL: http://ieeexplore.ieee.org/document/7742400/.

[30] David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R. Müller-Putz. "Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier". In: *Biomedical*

Engineering 61.1 (2016), pp. 77–86. ISSN: 1862-278X. DOI: 10.1515/bmt-2014-0117. URL: http://www.degruyter.com/view/j/bmte.2016.61.issue-1/bmt-2014-0117/bmt-2014-0117.xml.

[31] Samuel A. Stouffer, Edward A. Suchman, Leland C. Devinney, Shirley A. Star, and Robin M. Williams Jr. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II)*. Oxford, England: Princeton University Press, 1949. URL: http://psycnet.apa.org/record/1950-00790-000.

[32] Michael Tangermann et al. "Review of the BCI Competition IV". In: *Frontiers in Neuroscience* 6 (2012), p. 55. ISSN: 1662-4548. DOI: 10.3389/fnins.2012.00055. URL: http://journal.frontiersin.org/article/10.3389/fnins.2012.00055/abstract.

[33] Weibo Yi et al. "Evaluation of EEG Oscillatory Patterns and Cognitive Process during Simple and Compound Limb Motor Imagery". In: *PLoS ONE* 9.12 (2014). Ed. by Natasha M. Maurits, e114853. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0114853. URL: http://dx.plos.org/10.1371/journal.pone.0114853.

[34] Han Yuan and Bin He. "Brain-computer interfaces using sensorimotor rhythms: current state and future perspectives." In: *IEEE Transactions on Biomedical Engineering* 61.5 (2014), pp. 1425–35. ISSN: 1558-2531. DOI: 10.1109/TBME.2014.2312397. URL: http://www.ncbi.nlm.nih.gov/pubmed/24759276http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4082720.

[35] Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. "A Fully Automated Trial Selection Method for Optimization of Motor Imagery Based Brain-Computer Interface". In: *PLOS ONE* 11.9 (2016). Ed. by Bin He, e0162657. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0162657. URL: http://dx.plos.org/10.1371/journal.pone.0162657.

2.5 Paper 4: Multi-Task Logistic Regression for Brain-Computer Interfaces

Multi-Task Logistic Regression in Brain-Computer Interfaces

Karl-Heinz Fiebig*, Vinay Jayaram†, Jan Peters‡ and Moritz Grosse-Wentrup§
* Department of Computer Science
Darmstadt University of Technology, Darmstadt, Germany
Email: karl-heinz.fiebig@stud.tu-darmstadt.de
† Department of Empirical Inference
Max Planck Institute for Intelligent Systems, Tübingen, Germany
Email: vjayaram@tuebingen.mpg.de
‡ Intelligent Autonomous Systems
Darmstadt University of Technology, Darmstadt, Germany
Email: mail@jan-peters.net
§ Department of Empirical Inference
Max Planck Institute for Intelligent Systems, Tübingen, Germany
Email: moritzgw@tuebingen.mpg.de

***Abstract*—A Brain-Computer Interface (BCI) is used to enable communication between humans and machines by decoding elicited brain activity patterns. However, these patterns have been found to vary greatly across subjects or even for the same subject across sessions. Such problems render the performance of a BCI highly specific to subjects, requiring expensive and time-consuming individual calibration sessions to adapt BCI systems to new subjects.**

This work tackles the aforementioned problem in a Bayesian multi-task learning (MTL) framework to transfer common knowledge across subjects and sessions for the adaptation of a BCI to new subjects. In particular, a novel framework from previous work that is able to exploit structure of multi-channel Electroencephalography (EEG) will be extended by a Bayesian hierarchical logistic regression decoder for probabilistic binary classification. As such, the derived model will be able to explicitly learn spatial and spectral features in a paradigm that makes it further applicable for identification, analysis and evaluation of paradigm characteristics without relying on expert knowledge in this matter. The new decoder shows a significant improvement in performance on calibration-free decoding compared to previous MTL approaches for rule adaptation and uninformed models while also outperforming them as soon as subject-specific data becomes available. We will further demonstrate the ability of the model to identify relevant topographies along with signal band-power features that agree with neurophysiological properties of a common sensorimotor rhythm paradigm in BCIs.

I. Introduction

Second only to brain signal acquisition, the decoding of subject intention is fundamental to the practice of practical brain-computer interfaces (BCIs). Introduction of machine learning techniques for signal decoding shifted from extensively training subjects to control a BCI (e.g. [1]–[4]) into BCIs that are able to adapt to their users (e.g. [5]–[7]). Combining adaptive BCIs with the mobility of noninvasive electroencephalography (EEG) hardware to measure brain activity further enables research towards practical out-of-the-box applications in non-laboratory environments.

However, state-of-the-art BCIs suffer from high performance variations between subjects and even across sessions within the same subject [8], [9]. A calibration session with the subject prior to actual BCI usage is therefore still necessary, which poses a major hindrance to out-of-the-box applications. On top of that, machine learning based decoders have to deal with rather few data points gained from the calibration phase making them prone to overfitting and requiring very low dimensional feature spaces. Only a hand full of carefully selected features using expert knowledge of the paradigm are usually used to train decoders that generalize well to upcoming sessions [10]–[13].

Approaches to solve the problem of such performance variations in the past years have been dominated by domain adaptation techniques. These techniques use data acquired from different subjects and sessions to train decoders on invariant feature spaces, most commonly by preprocessing signals with common spatial patterns [14]–[16]. While domain adaptation is able to drastically reduced calibration time for new subjects and sessions with only slight performance losses, subject-specific variations are modeled exclusively by a fixed feature space. This makes it difficult to adapt BCIs to new subjects when calibration data becomes available. A more natural approach to this problem is given by rule adaptation techniques, which encode variations directly into the decision rule of decoding models.

Recent work in this area incorporate data fusion methods [17] and a Bayesian multi-task learning (MTL) framework first proposed by Alamgir et al. [18]. Being able to use more available data, Jayaram et al. generalized the framework to a Bayesian hierarchy with a novel feature space to decompose EEG structure into a spatial and frequency or time domain [19], [20]. This approach did not only enable the authors to transfer knowledge between subject and sessions, but further drop the need of expert knowledge by learning topographic

and broad-band features of the paradigm. In particular, they regarded the prediction problem of each subject and session as an individual task and assumed a common statistical distribution underlying the variations between tasks. In doing so, they developed a MTL algorithm to train prior distributions for the spatial and spectral parameter dimension of linear regression models that capture shared structure in different tasks and can be later used for task-specific adaptation. The resulting framework quickly produced reliable decoders (classification accuracies above 70%) based on a two-class sensorimotor rhythm (SMR) paradigm for new subjects using no or only few calibration trials while outperforming models trained solely from pooled or subject-specific data when calibration data was available.

However, linear regression models as used for decoding brain states in the above-mentioned framework assume Gaussian noise on the dependent variables. This may be a reasonable assumption for quantitative outcomes (i.e. regression problems), but BCI decoders usually state classification problems with categorical classes to represent different brain conditions. This work extends the proposed MTL framework for transfer learning by a more suitable assumption on binary dependent variables with a probabilistic model for two-class classification. Exploiting EEG structure for dimensionality reduction yields a bilinear logistic regression model that is able to learn relevant topographic and band-specific features itself instead of relying on a small set of manually selected features.

The rest of this paper is structured as follows. Section II will introduce the notation used throughout this work and derive a logistic regression model within a Bayesian MTL framework from previous work. The model is then extended for bilinear feature decomposition (FD) exploiting the structure of multi-channel EEG signals and the final learning algorithms are presented. After describing the SMR based experimental setup used to evaluate the models, section III will present the results and show that the derived model outperforms comparable models in calibration free decoding as well as in subject-adaptation. Section IV concludes this work with a summary, discussion on the results and future work.

II. Methods

A. Notation

Throughout this paper we will denote scalars with lower case, vectors with bold lowercase, matrices with uppercase letters and sets with calligraphic uppercase letters. We will regard the decoding problem for each subject or session as an individual task and denote the data set of m tasks with $\mathcal{T} = \left\{\mathcal{D}^{(t)}\right\}_{t=1}^{m}$. As we will work with binary dependent variables, each task data set $\mathcal{D}^{(t)} \in \mathcal{T}$ is formalized with

$$\mathcal{D}^{(t)} = \left\{\left(\boldsymbol{x}_i^{(t)}, y_i^{(t)}\right)\right\}_{i=1}^{n_t} \subset \mathbb{R}^d \times \{C_1, C_2\}$$

consisting of n_t data points with d-dimensional feature vectors extracted from EEG signals. The corresponding binary class label C_1 or C_2 represents one of two brain conditions of interest. In case of FD, each feature vector $\boldsymbol{x}_i^{(t)}$ is replaced by the corresponding feature matrix $X_i^{(t)} \in \mathbb{R}^{d\times k}$ that organizes d band-power features from k channels. Matrix calculus will follow denominator-layout notation.

B. Multi-task Logistic Regression

A very popular method for simple binary classification using probabilistic predictions is to pass the linear model through the logistic sigmoid activation. The hypothesis model is then given by

$$h(\boldsymbol{x}; \boldsymbol{w}) = \left(1 + \exp\left(\boldsymbol{w}^T \boldsymbol{x}\right)\right)^{-1} \in]0, 1[\tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is an input feature and $\boldsymbol{w} \in \mathbb{R}^d$ is the parameter vector of the model family. In our setting, the output of (1) can be interpreted as the probability $p\left(C_1 \mid \boldsymbol{x}\right) = h(\boldsymbol{x}; \boldsymbol{w})$ of feature $\boldsymbol{x}$ belonging to brain condition C_1. Likewise, the probability of $\boldsymbol{x}$ belonging to class C_2 is given by the complementary event $p\left(C_2 \mid \boldsymbol{x}\right) = 1 - h(\boldsymbol{x})$.

Assume that we have gathered a set $\mathcal{T}$ of m tasks. Following the MTL framework we will model each task with an individual hypothesis from (1). Hence, we get to train a set of m weight vectors $\mathcal{W} = \left\{\boldsymbol{w}^{(t)}\right\}_{t=1}^{m}$ from the corresponding task data sets in $\mathcal{T}$. By representing the classes with $\{C_1, C_2\} = \{0, 1\}$ and assume they follow a Bernoulli distribution parameterized with our hypothesis, we can define the *likelihood* of all our data (assuming iid feature samples) through

$$p\left(\mathcal{T} \mid \mathcal{W}\right) = \prod_{t=1}^{m} \prod_{i=1}^{n_t} \mathrm{Ber}\left(y_i^{(t)} \mid h\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{w}^{(t)}\right)\right) \tag{2}$$

where $\mathrm{Ber}\left(y \mid h\left(\boldsymbol{x}, \boldsymbol{w}\right)\right) = h\left(\boldsymbol{x}, \boldsymbol{w}\right)^y \left(1 - h\left(\boldsymbol{x}, \boldsymbol{w}\right)\right)^{1-y}$. Using Bayes rule we can further state the *posterior* distribution

$$p\left(\mathcal{W} \mid \mathcal{T}\right) = \frac{p\left(\mathcal{T} \mid \mathcal{W}\right) p(\mathcal{W})}{p\left(\mathcal{T}\right)} \tag{3}$$

over the weights given the task data sets based on the likelihood in (2), a *prior* distribution $p(\mathcal{W})$ and the *evidence* $p(\mathcal{T})$. This posterior is the entry point for the Bayesian MTL framework, were we assume that variations between tasks underly a common statistical distribution. In particular, we can capture common structure between the related tasks in a shared prior $p(\mathcal{W})$. Statistics of $p(\mathcal{W})$ can be used afterwards for out-of-the-box decoding on new subjects or improving decoders by combining shared knowledge with subject-specific data. The question remains how to obtain such a prior?

Following the MTL framework, we will model the shared prior with a general multivariate Gaussian density function $p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right)$ parameterized by a mean $\boldsymbol{\mu}_{\boldsymbol{w}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\boldsymbol{w}} \in \mathbb{R}^{d\times d}$. Hence, assuming that the task weights in $\mathcal{W}$ are iid, the prior for our posterior reads

$$p(\mathcal{W}) = \prod_{t=1}^{m} p\left(\boldsymbol{w}^{(t)}\right) = \prod_{t=1}^{m} \mathcal{N}\left(\boldsymbol{w}^{(t)} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) \tag{4}$$

Plugging (2) and (4) into (3) results in a parameterized model for the posterior distribution which yields

$$p(\mathcal{W} \mid \mathcal{T}) \propto \prod_{t=1}^{m} \prod_{i=1}^{n_t} \mathrm{Ber}\left(y_i^{(t)} \mid h\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{w}^{(t)}\right)\right) \prod_{t=1}^{m} \mathcal{N}\left(\boldsymbol{w}^{(t)} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right). \tag{5}$$

Our goal is to maximize $p(\mathcal{W} \mid \mathcal{T})$ w.r.t. the weights in $\mathcal{W}$ and the prior parameters $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$. Notice that instead of maximizing $p(\mathcal{W} \mid \mathcal{T})$ we can equivalently minimize $-\log p(\mathcal{W} \mid \mathcal{T})$ without loss of generalization. In fact, using (5), applying the negative logarithm and going through the math yields a loss minimization objective of the form

$$L(\mathcal{W}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}) = -\sum_{t=1}^{m} \sum_{i=1}^{n_t} E_{ce}\left(\boldsymbol{w}^{(t)}; \boldsymbol{x}_i^{(t)}, y_i^{(t)}\right) + \frac{1}{2} \sum_{t=1}^{m} \Omega\left(\boldsymbol{w}^{(t)}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) \tag{6}$$

where E_{ce} is the point-wise *cross-entropy error* function

$$E_{ce}(\boldsymbol{w}; \boldsymbol{x}, y) = y \log h(\boldsymbol{x}; \boldsymbol{w}) + (1 - y) \log\left(1 - h(\boldsymbol{x}; \boldsymbol{w})\right)$$

derived from the likelihood of the data and Ω is a *regularization* term

$$\Omega(\boldsymbol{w}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}) = (\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})^T \Sigma_{\boldsymbol{w}}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}}) + \log |\Sigma_{\boldsymbol{w}}|$$

arising from the prior distribution. Notice that Ω reduces to the same regularizer as for the Gaussian prior on the linear model presented in the previous work, and so we can interpret the minimization process of (6) in the same way: Ω penalizes weights that deviate too far from the prior mean while the covariance scaling acts as an implicit feature selector [19]. However, the squared error in the loss objective of the linear model switched with a more suitable error measurement for binary classification, namely the cross-entropy loss [21].

In order to train a prior model to capture common task knowledge, we want to minimize (6) w.r.t. the prior parameters $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$. It turns out that L is minimized by computing corresponding standard sample estimates of the Gaussian statistics from the optimal weights in $\mathcal{W}$, i.e. the mean is estimated with the average over all task weights

$$\mathrm{mean}(\mathcal{W}) = \frac{1}{m} \sum_{t=1}^{m} \boldsymbol{w}^{(t)} \tag{7}$$

and the covariates are estimated with the sample covariance matrix or some numerically more stable version like

$$\mathrm{cov}(\mathcal{W}; \boldsymbol{\mu}) = \frac{\sum_{t=1}^{m} \left(\boldsymbol{w}^{(t)} - \boldsymbol{\mu}\right)\left(\boldsymbol{w}^{(t)} - \boldsymbol{\mu}\right)^T}{\mathrm{Tr}\left(\sum_{t=1}^{m} \left(\boldsymbol{w}^{(t)} - \boldsymbol{\mu}\right)\left(\boldsymbol{w}^{(t)} - \boldsymbol{\mu}\right)^T\right)} + \varepsilon I, \tag{8}$$

where an appropriate $\varepsilon > 0$ ensures practicable condition numbers on the estimate. Unfortunately, the maximum a-posteriori (MAP) estimates for the optimal weights has to minimize the cross-entropy error term in L that has no closed form solution. However, L is differentiable w.r.t. each individual weight $\boldsymbol{w}^{(t)} \in \mathcal{W}$ yielding the gradient

$$\nabla L\left(\boldsymbol{w}^{(t)}; \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) = \sum_{i=1}^{n_t} \left(h\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{w}^{(t)}\right) - y_i^{(t)}\right) \boldsymbol{x}_i^{(t)} + \Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w}^{(t)} - \boldsymbol{\mu}_{\boldsymbol{w}}\right). \tag{9}$$

This vector can be used in gradient based optimization procedures (e.g. [22], [23]) to obtain optimal weight estimates given the prior parameters. The algorithm to learn the Gaussian prior is based on the observation that by fixing either the task weights in $\mathcal{W}$ or the prior parameters $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$ the cyclic dependency between them is broken. By repeatedly computing MAP estimates of the weights simultaneously given the prior and updating the prior parameters afterwards with Gaussian sample estimates from the task weights eventually converges to a solution for $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$. The learning procedure is outlined in Fig. 1

Algorithm 1: Multi-task Logistic Regression

Data: Training sets $\mathcal{T}$ from m related tasks
Result: $\boldsymbol{\mu}_{\boldsymbol{w}}$, $\Sigma_{\boldsymbol{w}}$

Initialize $\boldsymbol{\mu}_{\boldsymbol{w}} = \mathbf{0}$ and $\Sigma_{\boldsymbol{w}} = I$;
Arbitrary initialize $\mathcal{W} = \left\{\boldsymbol{w}^{(t)}\right\}_{t=1}^{m}$;
while $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$ not converged **do**
  for $\boldsymbol{w}^{(t)} \in \mathcal{W}$ **do**
    Compute the MAP estimate for $\boldsymbol{w}^{(t)}$ by minimizing (6) w.r.t. $\boldsymbol{w}^{(t)}$ holding $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$ fixed (e.g. b) ;
  Update $\boldsymbol{\mu}_{\boldsymbol{w}} = \mathrm{mean}(\mathcal{W})$ using (7);
  Update $\Sigma_{\boldsymbol{w}} = \mathrm{cov}(\mathcal{W}; \boldsymbol{\mu}_{\boldsymbol{w}})$ using (8);

Fig. 1. MTL Logistic Regression algorithm to train a Gaussian prior based on a set of different task data sets.

C. Spatio-spectral Feature Decomposition

EEG signals are recorded with k electrodes placed at specific locations on the scalp. A very popular method to train decoders in BCI is to use band power features in d frequency bands of the signal, making up a general feature space of kd dimensions. Because expert knowledge of neurophysiological and -psychological properties of the paradigm are used to determine the relevant electrodes and frequency bands, a subset of the full kd features is chosen and therefore applicable to smaller data sets obtained from calibration sessions in BCIs.

In order to not rely on expert knowledge of the paradigm, we have to use many more electrodes to cover the scalp and small frequency bins over a broad spectral range. In this case, the feature space of kd dimensions easily exceeds the number of available data samples and decoders are often no longer able to rely on feature statistics. Hence, BCI research has to keep relying heavily on expert knowledge.

Jayaram et al. [19] proposed FD as a spatio-spectral feature space for EEG signals that significantly reduces the feature

dimensionality from kd to $k+d$. In particular, the authors assumed that the spectral feature importance is independent from the spatial topography of each electrode. The MTL logistic regression model presented in this work is applicable to this assumption in the same way as the authors did for linear regression. Hence, instead of using (1) to predict the classes, we will use a bilinear hyperplane as the decision boundary that yields the model

$$h(X;\boldsymbol{w},\boldsymbol{a}) = \left(1+\exp\left(\boldsymbol{a}^T X \boldsymbol{w}\right)\right)^{-1} \in]0,1[\tag{10}$$

where $X \in \mathbb{R}^{k\times d}$ is an input feature, $\boldsymbol{w} \in \mathbb{R}^d$ is the parameter vector weighting the spectral features and $\boldsymbol{a} \in \mathbb{R}^k$ is the parameter vector weighting the spatial features.

Notice that we only changed the model from $h(\boldsymbol{x};\boldsymbol{w})$ to $h(X;\boldsymbol{w},\boldsymbol{a})$, hence, the MTL derivation is analogous to the previously presented non-FD case, except that we have decomposed the original weights into a spectral and a spatial part. We therefore incorporate two Gaussian priors $\mathcal{N}(\boldsymbol{\mu_a},\Sigma_{\boldsymbol{a}})$ and $\mathcal{N}(\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}})$ multiplicatively into the posterior in (5) with spectral task weights $\mathcal{W} = \{\boldsymbol{w}_t\}_{t=1}^m \subset \mathbb{R}^d$ and spatial task weights $\mathcal{A} = \{\boldsymbol{a}_t\}_{t=1}^m \subset \mathbb{R}^k$. The loss objective for the FD case then becomes

$$\begin{aligned} &L(\mathcal{W},\mathcal{A},\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}},\boldsymbol{\mu_a},\Sigma_{\boldsymbol{a}}) = \\ &-\sum_{t=1}^{m}\sum_{i=1}^{n_t} E_{ce}\left(\boldsymbol{w}^{(t)};X_i^{(t)},y_i^{(t)}\right) \\ &+\frac{1}{2}\sum_{t=1}^{m}\Omega\left(\boldsymbol{w}^{(t)},\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}}\right)+\Omega\left(\boldsymbol{a}^{(t)},\boldsymbol{\mu_a},\Sigma_{\boldsymbol{a}}\right) \end{aligned} \tag{11}$$

where E_{ce} and Ω are defined accordingly as in (6). Minimizing (11) is done equivalently to minimizing (6), i.e. we can use gradient based numerical optimization. The gradient w.r.t. a spectral task weight $\boldsymbol{w}^{(t)} \in \mathcal{W}$ reads

$$\begin{aligned} &\nabla L\left(\boldsymbol{w}^{(t)};\boldsymbol{a}^{(t)},\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}}\right) = \\ &\sum_{i=1}^{n_t}\left(h\left(X_i^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{a}^{(t)}\right)-y_i^{(t)}\right){X_i^{(t)}}^T\boldsymbol{a}^{(t)} \\ &+\Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu_w}\right) \end{aligned} \tag{12}$$

and similarly the gradient w.r.t. a spatial task weight $\boldsymbol{a}^{(t)} \in \mathcal{A}$

$$\begin{aligned} &\nabla L\left(\boldsymbol{a}^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{\mu_a},\Sigma_{\boldsymbol{a}}\right) = \\ &\sum_{i=1}^{n_t}\left(h\left(X_i^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{a}^{(t)}\right)-y_i^{(t)}\right)X_i^{(t)}\boldsymbol{w}^{(t)} \\ &+\Sigma_{\boldsymbol{a}}^{-1}\left(\boldsymbol{a}^{(t)}-\boldsymbol{\mu_a}\right). \end{aligned} \tag{13}$$

As the gradient for the spectral weights depend on the spatial parameters and vice versa we have to alternatingly fix one set of weights to compute the MAP estimate of the others. Apart from that, the learning algorithm is similar to the non-FD case but may be computationally a bit more expensive. In Fig.2 a FD learning procedure based on gradient descent is depicted.

Algorithm 2: FD Multi-task Logistic Regression

Data: Training sets $\mathcal{T}$ from m related tasks
Result: $\boldsymbol{\mu_w}$, $\Sigma_{\boldsymbol{w}}$, $\boldsymbol{\mu_a}$, $\Sigma_{\boldsymbol{a}}$

Initialize $\boldsymbol{\mu_w} = \mathbf{0}$ and $\Sigma_{\boldsymbol{w}} = I$;
Initialize $\boldsymbol{\mu_a} = \frac{1}{\sqrt{k}}\mathbf{1}$ and $\Sigma_{\boldsymbol{a}} = I$;
Arbitrary initialize $\mathcal{W} = \{\boldsymbol{w}^{(t)}\}_{t=1}^t$ and $\mathcal{A} = \{\boldsymbol{a}^{(t)}\}_{t=1}^m$;
while $\boldsymbol{\mu_w}$, $\Sigma_{\boldsymbol{w}}$, $\boldsymbol{\mu_a}$ and $\Sigma_{\boldsymbol{a}}$ not converged **do**
  for $\boldsymbol{w}^{(t)} \in \mathcal{W}$ and $\boldsymbol{a}^{(t)} \in \mathcal{A}$ **do**
    while $\boldsymbol{w}^{(t)}$ and $\boldsymbol{a}^{(t)}$ not converged **do**
      Choose some learning rate $\eta \in]0,\infty]$;
      Set $\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t)} - \eta\nabla L\left(\boldsymbol{w}^{(t)};\boldsymbol{a}^{(t)},\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}}\right)$;
      Set $\boldsymbol{a}^{(t)} = \boldsymbol{a}^{(t)} - \eta\nabla L\left(\boldsymbol{a}^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{\mu_a},\Sigma_{\boldsymbol{a}}\right)$;
  Update $\boldsymbol{\mu_w} = \mathrm{mean}(\mathcal{W})$ using (7);
  Update $\Sigma_{\boldsymbol{w}} = \mathrm{cov}(\mathcal{W};\boldsymbol{\mu_w})$ using (8);

Fig. 2. Gradient based MTL logistic regression algorithm to train a Gaussian prior based on a set of different task data sets in FD space.

D. Subject-specific Adaptation

Once we have trained the prior parameters using MTL we can immediately use the mean weight vector for prediction. When given a new feature $\boldsymbol{x}$, we only need to compute $h(\boldsymbol{x};\boldsymbol{\mu_w})$ for an out-of-the-box prediction of the class probability. When decoding brain states for a new subject, we are faced with a new task that has subject-specific variations. As more data for the new tasks becomes available (e.g. from a calibration phase), those variations will be captured by the task-specific data set and decoders trained from this set will eventually outperform the plain prior. The Bayesian framework naturally copes with this case; we have to just compute the MAP estimate of the adapted weights using (6) based on the new data set.

However, we do not know how much belief we should put into the prior to optimally trade-off between task-specific variations and shared task knowledge. This concept can be captured formally by introducing an additional regularization factor $\lambda \in [0,\infty]$ into the cross-entropy objective for subject adaptation. Given we denote our new task data with $\mathcal{D} = \{(\boldsymbol{x}_i,y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{C_1,C_2\}$ we obtain adapted weights by minimizing

$$L_\lambda(\boldsymbol{w}) = -\sum_{i=1}^{n_t} E_{ce}\left(\boldsymbol{w};\boldsymbol{x}_i,y_i\right)+\frac{\lambda}{2}\Omega\left(\boldsymbol{w},\boldsymbol{\mu_w},\Sigma_{\boldsymbol{w}}\right) \tag{14}$$

w.r.t. to the weights $\boldsymbol{w}$. This can be done again with numerical gradient based optimization techniques, where the gradient is analytically given by

$$\nabla L_\lambda(\boldsymbol{w}) = \sum_{i=1}^{n_t}\left(h\left(\boldsymbol{x}_i;\boldsymbol{w}\right)-y_i\right)\boldsymbol{x}_i+\lambda\Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w}-\boldsymbol{\mu_w}\right). \tag{15}$$

An regularization factor λ to find the optimal trade-off between prior and task adaptation can be obtained using model selection methods (e.g. cross-validation).

E. Experimental Setup

Evaluation of the model was conducted on real EEG signals recorded from 10 healthy subjects (two female, eight male, 22-28 years old, nine subjects were naïve to BCIs and one participated twice in BCI experiments) using a two-class sensorimotor rhythm paradigm, namely haptic motor imagery of left or right hand movements.

Each subject sat in a comfortable chair in front of a screen and performed 300 trials in the experiment (150 per condition of the binary paradigm, stimuli were presented in pseudorandom order and no feedback on the performance was provided). Each trial consisted of an initial pause of three second, followed by an imagery phase lasting seven seconds in which a centrally displayed arrow pointing to the left or right informed the subject to perform haptic left or right hand motor imagery, respectively.

Brain activity during the experiment was recorded using EEG with 128 electrodes positioned according to the extended 10-20 system (referenced at Cz). The signals were sampled at 500Hz using BrainAmp amplifiers[1] and a temporal analog high-pass filter with 10 seconds time constant.

After the experiment was conducted, data preprocessing solely consisted of spatially filtering the signals with a surface Laplace [24] to keep the results unbiased for evaluation. FD feature matrices were formed by applying the discrete Fourier transform with a Hann window to the motor imagery phase of each trial in order to extract equidistant log-band power features of 2Hz width within the frequency range from 7Hz to 31Hz from all electrodes. Hence, the FD space was spanned by 128×12-dimensional features.

III. Experimental Results

A. Classification Performance

The performance of MTL logistic regression on the real-world BCI paradigm was evaluated by comparing classification accuracies when different amount of subject-specific data is available. In particular, one out of the ten data sets corresponding to each subject was taken out to be regarded as subject-specific calibration data. Three models where used for comparison: FD MTL Logistic Regression with Gaussian prior trained using the algorithm shown in Fig.2, standard FD Logistic Regression with L2 regularization (i.e. uninformed prior) and finally FD MTL Linear Regression with a Gaussian prior trained as presented in the previous framework (with maximum-likelihood estimates for the variance hyperparameter). Two out of the ten subjects were performing near chance and were taken out from prior training (i.e. priors were finally trained from seven task data sets).

After obtaining priors for the models, the 300 samples from the subject-specific data were randomly divided into a training set (200 samples) and a test set (100 samples). Each model was successively trained on an increasing subset of the training set using a step size of 50 and 5-fold cross-validation to select a regularization hyperparameter from $\{\exp(-10), \exp(-9), \ldots, \exp(9), \exp(10)\}$. The trained models were then evaluated on the test set to compute their accuracy. The whole procedure was conducted for each subject using 100 runs where the data was randomly split into training and test set. The mean accuracy development over all subjects and runs is shown in Fig.3. Further, the development of the prior bias (mean regularization factor obtained from cross-validation) over the runs to trade off between subject adaptation and prior is visualized in Fig.4.

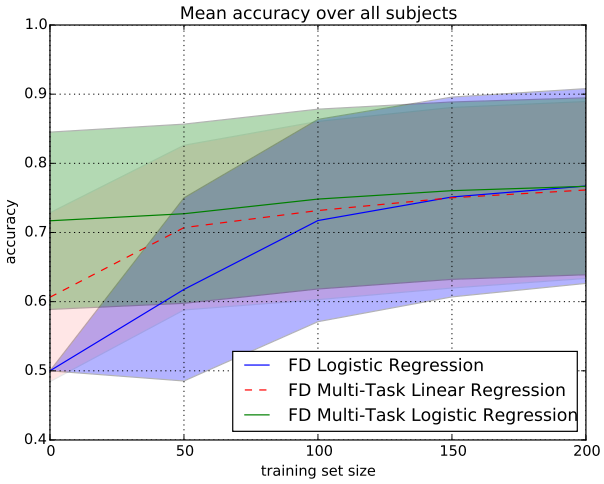

Fig. 3. Mean accuracy with shaded standard deviation on the test set over an increasing amount of subject-specific calibration data provided for training. The mean was taken over all 10 subjects using 100 runs with random splits of 200 training and 100 test samples. All models improve with increasing amount of calibration data and eventually converge to the same performance level. The accuracy development suggest that MTL logistic regression outperforms MTL linear regression and both MTL algorithms outperform uninformed logistic regression.

The results show that the performance gap between the models keeps getting smaller with increasing size of the training set and finally converge to a common performance level. However, MTL logistic regression with prior information outperforms the MTL model from the previous framework as well as standard logistic regression with uniformed prior. Remarkably, without using any subject-specific data for calibration the out-of-the box classification rate of the model prior from this work reaches the 70% mark (considered as the minimum requirement for reliable communication in BCIs). Development of the regularization factor over the runs and subjects shows a decreasing trend over increasing amount of calibration data from the new subjects. This indicates that the decoder is deviating from the prior in order to learn more task-specific structure, which is in fact a plausible statistical behavior; we expect that with more data from a problem the underlying structures will emerge stronger which can be captured by the model to improve on the subject-specific variations.

[1]BrainProducts GmbH, Gilching, Germany

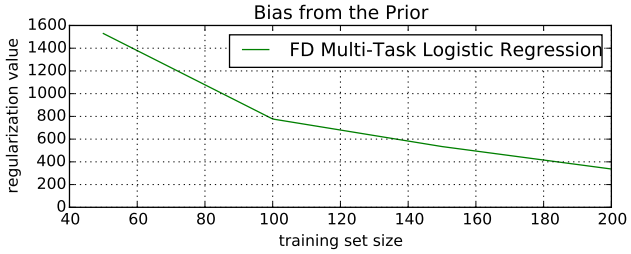

Fig. 4. Mean regularization value of MTL logistic regression (gathered through 5-fold cross-validation) for subject adaptation over all subjects and 100 runs (see Fig.3). Initial high regularization towards the prior for few calibration trials demonstrates that the model is relying stronger on the prior instead of the calibration data. As more subject-specific data is added to the training set, regularization drops and incorporates more exposed structure specific to the new subject.

B. Spatial and Spectral Prior

Trained models using the FD features have weight vectors for the spatial topography and the spectral frequency bins that indicate relevance of individual dimensions that are used by the model for prediction. In order to compare findings of the MTL algorithm with domain knowledge of the SMR paradigm, Gaussian prior parameters for the FD space were trained from all eight subjects (leaving out the two near chance performing subjects) using the algorithm in Fig.2. A visualization of the resulting priors is shown in Fig.5.

The trained prior identifies spatial relevance on electrodes placed above the left and right sensorimotor cortex. Those topographic features agree with domain knowledge of the neurophysiological characteristics known for this paradigm and indicate that indeed neural activity is used to predict the corresponding brain condition instead of artifacts. Furthermore, the model puts highest priority on the frequency bin for 11-13Hz in its prior mean, which corresponds to the μ-rhythm and agrees with the band-power modulation characteristics of the paradigm, too. Implicit feature selection used by the model for subject adaptation is likewise consistent as shown in the spectral covariance where we can see high covariates between the α-rhythm (9-13Hz) and β-rhythm (19-23Hz and 27-29Hz).

C. Null Hypothesis Pairwise Permutation Test

In III-A we compared the mean accuracies over each subject with increasing amount of calibration data in 100 runs. Here, we will perform a statistical test to examine if there is a significant difference in performance of the three tested models. In particular, a pairwise permutation test [25] between two models was conducted under the null hypothesis that their true mean performance is equal. The test was setup as follows: The mean accuracy over the 100 runs of each of the 10 subjects and three models were took for one calibration set size. Let $\mathcal{P}_a = \{a_1, a_2, \ldots, a_{10}\}$ be the performance samples from FD MTL logistic regression, $\mathcal{P}_b = \{b_1, b_2, \ldots, b_{10}\}$ the samples from FD mutli-task linear regression and $\mathcal{P}_c = \{c_1, c_2, \ldots, c_{10}\}$ from standard L2 logistic regression in FD space. Further, let μ_a, μ_b and μ_c denote the true mean

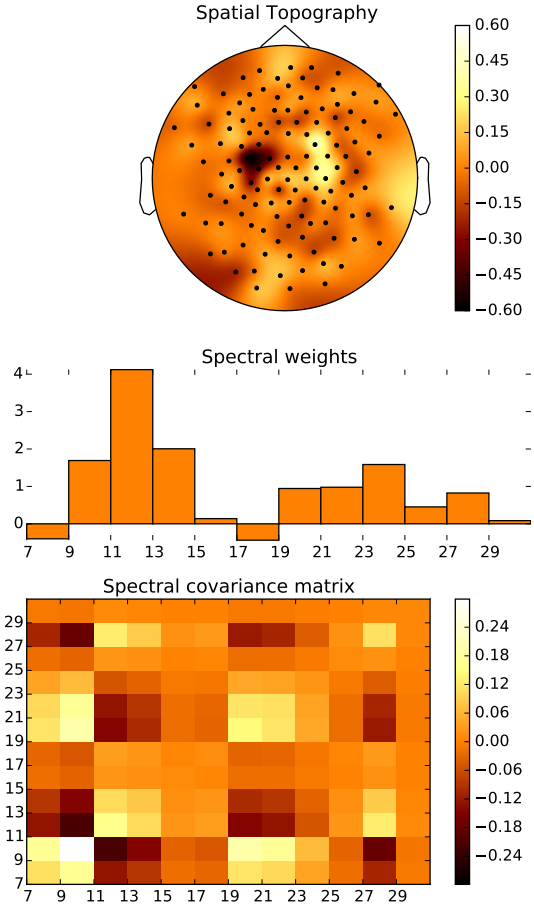

Fig. 5. The first plot shows the topography of the spatial weight prior trained by MTL logistic regression over eight subjects. The prior indicates relevant activity above the left and right sensorimotor cortex on electrodes C3 and C4. The second plot shows a bar chart of the trained spectral weight prior. They show high relevance on the frequency bin for 11-13Hz and its surrounding bins corresponding to the μ-rhythm, as well as moderate relevance within the β-range in 19-29Hz. The final third plot visualizes the spectral covariance prior trained by the algorithm. High positive covariates for spectral weights can be found within the α- and β-frequencies while negative covariates show up between those bands.

performance from which the samples $\mathcal{P}_a$, $\mathcal{P}_b$ and $\mathcal{P}_c$ were drawn, respectively. We tested two null hypotheses, the first was $H_0^* : \boldsymbol{\mu}_a = \boldsymbol{\mu}_b$ (i.e. MTL logistic and linear regression have the same performance) and the second $H_0^{**} : \boldsymbol{\mu}_a = \boldsymbol{\mu}_c$ (i.e. logistic regression with MTL prior and standard logistic regression with uninformed prior perform equally well). Using the test statistic $T(\mathcal{P}_x, \mathcal{P}_y) = \text{mean}(\mathcal{P}_x) - \text{mean}(\mathcal{P}_y)$ where x and y are substitutes for a, b or c the ρ-value was computed by

$$\rho = \frac{\sum_{i=1}^{n} \left[T\left(\mathcal{P}_x^{(i)}, \mathcal{P}_y^{(i)}\right) \geq T\left(\mathcal{P}_x, \mathcal{P}_y\right) \right]}{n}$$

where $\mathcal{P}_x^{(i)}$ and $\mathcal{P}_y^{(i)}$ are pseudorandomly generated pairwise permutations of $\mathcal{P}_x$ and $\mathcal{P}_y$. This means that each pair of samples (x_t, y_t) for the same subject t appears in $\mathcal{P}_x^{(i)}$ and

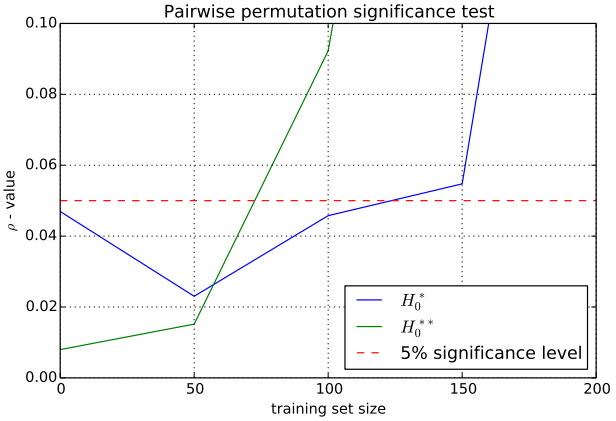

Fig. 6. ρ-value over an increasing amount of subject-specific calibration data using a pairwise permutation test for the null hypotheses H_0^* (MTL logistic and linear regression have same performance) and H_0^{**} (MTL logistic and standard logistic regression have same performance). For up to 50 calibration trials there is strong evidence against H_0^* and H_0^{**} at a 5% significance level. For up to 100 trials only H_0^* can be rejected. For up to 150 trials the ρ-value for H_0^* is marginal and for more trials no hypotheses can be rejected.

$\mathcal{P}_y^{(i)}$ again, but with a chance of 50% that the positions have switched to (y_t, x_t). The results for the ρ-value using $n = 100000$ permutations over different amounts of calibration data are shown in Fig.6.

The results show that without any or few subject-specific calibration data both H_0^* and H_0^{**} are rejected at a 5% significance level. Hence, together with the classification results from Fig.5, we can indeed observe a statistically significant improvement of MTL logistic regression over the MTL model from the previous framework and uninformed logistic regression in case of out-of-the-box decoding and short calibration phases (up to 50 trials). For moderate calibration sessions (up to 100 trials), we can only reject H_0^* and observe a significant improvement of MTL logistic regression over MTL linear regression.

IV. Discussion

This work extended a general framework from previous work used for MTL of linear regression models in BCIs by a logistic regression model with more suitable assumptions on the distribution of the dependent variable in case of binary classification. We demonstrated a significant improvement in classification accuracy of the new model over comparable models for calibration-free decoding and subject-specific adaptation with few calibration trials. The new model was able to learn spatially important locations on the scalp as well es relevant spectral frequency bands, both consistent with expert knowledge of the paradigm.

Besides an improved performance over the model used in the previous framework, logistic regression has the advantage of predicting class probabilities instead of direct classes, thus naturally incorporating mathematical uncertainty about the prediction. Adjusting the threshold at which we assign classes from the probability renders the model further optimizable for usage of well established methods like Receiver Operating Characteristic curves and other statistics. However, other models used throughout BCI research that turned out to work well may be examined for the generalized MTL approach and compared to each other. Prominent examples on which current work is in progress are derivations of Support Vector Machines and Linear Discriminant Analysis within the framework. Further approaches by using different prior structures or loss functions, as well as hybrid techniques of MTL together with advanced spatial filters are to be investigated and may further improve performance.

References

[1] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An eeg-based brain-computer interface for cursor control," Electroencephalography and Clinical Neurophysiology, vol. 78, no. 3, pp. 252 – 259, 1991.

[2] J. R. Wolpaw and D. J. McFarland, "Multichannel eeg-based brain-computer communication," Electroencephalography and Clinical Neurophysiology, vol. 90, no. 6, pp. 444 – 449, 1994.

[3] B. N., N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kubler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," Nature, vol. 398, no. 6725, pp. 297–298, 1999.

[4] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," National Academy of Sciences, vol. 101, no. 51, pp. 17 849–17 854, 2004.

[5] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying single trial eeg: Towards brain computer interfacing," in Advances in Neural Information Processing Systems 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 157–164.

[6] B. Schölkopf, J. Platt, and T. Hofmann, Adaptive Spatial Filters with predefined Region of Interest for EEG based Brain-Computer-Interfaces. MIT Press, 2007, pp. 537–544.

[7] N. J. Hill, T. N. Lal, M. Schröder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Schölkopf, A. Kübler, and N. Birbaumer, "Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 14, no. 2, pp. 183–186, 2006.

[8] M. Grosse-Wentrup and B. Schölkopf, A Review of Performance Variations in SMR-Based Brain–Computer Interfaces (BCIs), ser. SpringerBriefs in Electrical and Computer Engineering. Springer, 2013, ch. 4, pp. 39–51.

[9] T. Dickhaus, C. Sannelli, K.-R. Müller, G. Curio, and B. Blankertz, "Predicting BCI performance to study BCI illiteracy," in BMC Neuroscience 2009, vol. 10, 2009, p. (Suppl 1):P84.

[10] U. Hoffmann, J. M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain-computer interface for disabled subjects," J. Neurosci. Methods, vol. 167, no. 1, pp. 115–125, 2008.

[11] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for eeg-based motor imagery classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 16, no. 4, pp. 317–326, 2008.

[12] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones, "Brain-computer interfaces based on the steady-state visual-evoked response," IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pp. 211–214, 2000.

[13] T. Hinterberger, S. Schmidt, N. Neumann, J. Mellinger, B. Blankertz, G. Curio, and N. Birbaumer, "Brain-computer communication and slow cortical potentials," IEEE Trans Biomed Eng, vol. 51, no. 6, pp. 1011–1018, 2004.

[14] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multi-subject learning for common spatial patterns in motor-imagery bci," Computational Intelligence and Neuroscience, vol. 2011, no. 217987, pp. 1–9, Aug. 2011.

[15] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms," IEEE Transactions on Biomedical Engineering, vol. 58, no. 2, pp. 355–362, Feb 2011.

[16] W. Samek, F. C. Meinecke, and K. R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," IEEE Transactions on Biomedical Engineering, vol. 60, no. 8, pp. 2289–2298, 2013.
[17] S. Fazli, S. Dähne, W. Samek, F. Bießmann, and K. R. Müller, "Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based brain–computer interfaces," Proceedings of the IEEE, vol. 103, no. 6, pp. 891–906, June 2015.
[18] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010, Max-Planck-Gesellschaft. Cambridge, MA, USA: JMLR, May 2010, pp. 17–24.
[19] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 20–31, 2016.
[20] V. Jayaram and M. Grosse-Wentrup, "A transfer learning approach for adaptive classification in p300 paradigms," in Proceedings of the Sixth International BCI Meeting, 2016.
[21] C. M. Bishop, Pattern Recognition and Machine Learning. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
[22] J. Nocedal and S. J. Wright, Numerical Optimization, ser. Springer Series in Operations Research and Financial Engineering. Berlin: Springer, 2006.
[23] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," Journal of research of the National Bureau of Standards, vol. 49, pp. 409–436, 1952.
[24] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for eeg-based communication," Electroencephalography and Clinical Neurophysiology, vol. 103, no. 3, pp. 386 – 394, 1997.
[25] W. J. Welch, "Construction of permutation tests," Journal of the American Statistical Association, vol. 85, no. 411, pp. 693–698, 1990.