Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment

Isabela Albuquerque[†], João Monteiro[†], Olivier Rosanne and Tiago H. Falk*

Institut National de la Recherche Scientifique, Université du Québec, Montréal, QC, Canada

Assessment of mental workload in real-world conditions is key to ensuring the performance of workers executing tasks that demand sustained attention. Previous literature has employed electroencephalography (EEG) to this end despite having observed that EEG correlates of mental workload vary across subjects and physical strain, thus making it difficult to devise models capable of simultaneously presenting reliable performance across users. Domain adaptation consists of a set of strategies that aim at allowing for improving machine learning systems performance on unseen data at training time. Such methods, however, might rely on assumptions over the considered data distributions, which typically do not hold for applications of EEG data. Motivated by this observation, in this work we propose a strategy to estimate two types of discrepancies between multiple data distributions, namely marginal and conditional shifts, observed on data collected from different subjects. Besides shedding light on the assumptions that hold for a particular dataset, the estimates of statistical shifts obtained with the proposed approach can be used for investigating other aspects of a machine learning pipeline, such as quantitatively assessing the effectiveness of domain adaptation strategies. In particular, we consider EEG data collected from individuals performing mental tasks while running on a treadmill and pedaling on a stationary bike and explore the effects of different normalization strategies commonly used to mitigate cross-subject variability. We show the effects that different normalization schemes have on statistical shifts and their relationship with the accuracy of mental workload prediction as assessed on unseen participants at training time.

KEYWORDS

mental workload assessment, distribution shifts, electroencephalography, cross-subject generalization, domain adaptation

1. Introduction

Monitoring mental workload in a fast and accurate manner is critical in scenarios where the full attention of an individual is fundamental for the security of others. Firefighters, air traffic controllers, and first responders, for instance, are constantly exposed to such work conditions. In many cases, in addition to demanding mental tasks, individuals are also under varying levels of physical strain. Measuring mental workload under such scenarios is challenging, especially when relying on wearable sensors (Albuquerque et al., 2018).

Passive brain-computer interfaces (BCIs) have been widely used in the past for mental workload monitoring (e.g., Zhang et al., 2014, 2017; Wang et al., 2015). Existing models, however, exhibit high cross-subject variability, hence hindering their applicability in real-world scenarios. As pointed out in Yin and Zhang (2017), models are usually subject-specific and present poor generalization when training and testing conditions are distinct in terms of the represented individuals. Anatomic and environmental factors have been attributed as the main causes of the cross-subject variability (Wu et al., 2015a,b; Wei et al., 2018). Additionally, shifts between training and testing conditions could occur due to different data collection equipment, as well as changes in the electrodes positioning during an experimental session or even in the performance of each individual for the same task.

A standard way of compensating for the high cross-subject variability with EEG-based passive BCIs is to *calibrate* the model prior to applying it to an unseen individual. This is achieved by collecting a (usually small) number of labeled examples from this particular user and retraining or pruning the model to fine-tune it to the new user (Lotte, 2015). Recent work, however, has highlighted that this calibration step can be too costly and time-consuming, hence not very practical (Aricò et al., 2018; Lotte et al., 2018). Improving the cross-subject generalization of current BCIs is therefore critical for real-world applications, such as mental workload monitoring.

An alternative strategy to calibrating BCIs to unseen subjects/conditions is to develop methods that reduce the variability between training and testing conditions. To this end, methods such as domain adaptation (DA) have been proposed (Daume III and Marcu, 2006; Sun et al., 2017). A standard DA strategy corresponds to augmenting the learning objective of an algorithm with a term that accounts for how *invariant* the current model is with respect to data from different distributions (Ganin et al., 2016; Tzeng et al., 2017). The goal of this regularization term is to enforce the learned model to ignore domain-specific cues. It is important to emphasize that throughout the remainder of this paper, the terms *domain* and *distribution* will be used interchangeably.

Previous work on domain adaptation has shown that different techniques rely on distinct assumptions over the training and testing distributions (Ben-David et al., 2010; Zhao et al., 2019). For example, a common requirement is *covariate shift* assumption, which considers that the distributions of labels y conditioned on data x, $p(y|x)$, do not shift across training and testing conditions and only the marginal distributions $p(x)$ shift (Ben-David et al., 2010). In the case of EEG-based passive BCI applications, however, previous work has argued that $p(y|x)$ is likely to shift between different subjects (Wu et al., 2015a,b; Wu, 2016; Wei et al., 2018). Therefore, the covariate shift assumption cannot be taken for granted since, given feature vectors x_1 and x_2, respectively, acquired from two distinct subjects and represented in a shared feature space, $p_1(y|x_1) \neq p_2(y|x_2)$ even in the case where $x_1 = x_2$. As discussed in Zhao et al. (2019) and Johansson et al. (2019), when the covariate shift assumption does not hold, there is a trade-off between learning domain-invariant representations and obtaining a small prediction error across different domains that needs to be optimized.

Verifying whether the underlying assumptions of a particular approach hold in practice is a frequently overlooked step by domain adaptation approaches (Akuzawa et al., 2019). In this work, we claim that it is necessary to evaluate the underlying structure of a particular dataset in order to verify which types of distribution shifts exist and which assumptions could be safely considered (or not), when utilizing domain adaptation strategies. To this end, our main contributions are: (i) we introduce a method to estimate the cross-subject mismatch between the conditional label distributions; (ii) we apply a notion of divergence introduced in Kifer et al. (2004) to estimate the mismatch between marginal distributions of pairs of subjects; and (iii) we investigate whether common practices in the EEG literature to mitigate cross-subject variability, such as normalizing spectral features, are able to mitigate both conditional and marginal distributional shifts. Given the relevance of mitigating cross-subject variability on EEG-based mental workload assessment, we empirically validate our proposed method on the WAUC dataset (Albuquerque et al., 2020). The dataset is composed by EEG data collected during a mental workload modulation task with subjects performing different activity levels and activity types. In this contribution, we extend our first efforts toward quantifying cross-subjects statistical shifts as presented in Albuquerque et al. (2019), by considering a larger number of subjects in our analysis (total of 18 subjects), and, more importantly, we investigate how different ways to modulate physical activity affect the cross-subject statistical shifts on EEG correlates of mental workload.

The remainder of this paper is organized as follows: in Section 2 we provide an overview of domain adaptation and formalize the problem of generalizing across subjects under this setting. In Section 3, the proposed strategies to estimate conditional and marginal shifts are presented. In Sections 4, 5, we describe the experimental setup and present the

results, respectively. Finally, we outline the main conclusions in Section 6.

2. Domain adaptation and cross-subject generalization

Consider d-dimensional feature vectors $x \in \mathbb{R}^d$, computed from data through a deterministic mapping, such as power spectral density computations from EEG signals. We denote the feature space as $\mathcal{X}$. Further consider a labeling function $f : \mathcal{X} \rightarrow \mathcal{Y}$, where the label space is represented by $\mathcal{Y}$. For example, $\mathcal{Y}$ would be $\{0, 1\}$ for a binary classification case. A domain $\mathcal{D}$ is defined as a distribution over $\mathcal{X}$.

Moreover, let a hypothesis h be a mapping $h : \mathcal{X} \rightarrow \mathcal{Y}$, such that $h \in \mathcal{H}$, where $\mathcal{H}$ is a set of candidate hypothesis, or a hypothesis class. Finally, we define the risk R associated with a given hypothesis h on domain $\mathcal{D}$ as:

$$R[h] = \mathbb{E}_{x \sim \mathcal{D}} \ell[h(x), f(x)], \qquad (1)$$

where the loss $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ quantifies how different h is from the true labeling function f on $\mathcal{D}$. Supervised learning can be defined as searching the minimum risk hypothesis h^* within $\mathcal{H}$, i.e.,:

$$h^* = \arg\min_{h \in \mathcal{H}} R[h]. \qquad (2)$$

However, computing $R[h]$ is generally intractable since one does not usually have access to $\mathcal{D}$, but instead just observed samples from the domain.

2.1. Empirical risk minimization

Given the intractability of the risk minimization setting described above, empirical risk minimization is a common practical alternative framework for supervised learning. In such case, a sample X of size N is observed from $\mathcal{D}$, i.e., $X = \{x_1, x_2, \ldots, x_N\}$, where all x_n are assumed to be independently sampled from the domain $\mathcal{D}$ (i.e., the i.i.d. assumption holds). The empirical risk is thus defined as:

$$\hat{R}_X[h_X] = \frac{1}{n} \sum_{i=1}^{n} \ell[h_X(x_i), f(x_i)], \qquad (3)$$

and the generalization error (or generalization gap) will be the difference between the true and empirical risks, i.e., $\epsilon = |R[h_X] - \hat{R}_X[h_X]|$. Ideally, $\hat{R}_X[h_X] \approx 0$ and $\epsilon \approx 0$, in which case h_X is able to attain a low risk across new samples of $\mathcal{D}$, not observed at training time.

2.2. Domain adaptation

We now analyze the case such that the i.i.d. assumption, which considers x_n in X are all sampled according to a fixed domain $\mathcal{D}$, does not hold. More specifically, we assume that a set of M different domains exist. In the following, we describe two recent results and formally define the statistical shifts that might be observed when different domains are considered.

Since most relevant results and theoretical guarantees were proven specifically for the case in which $M = 2$, we consider such setting and define two domains, referred as the source and target domains $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. A bound for the risk of a given hypothesis on the target domain $R_T[h]$ was introduced in Ben-David et al. (2007). This result shows that $R_T[h]$ depends on $R_S[h]$, the risk of h on the source domain, a notion of divergence between both domains, as well as the minimum risk that can be achieved by some $h \in \mathcal{H}$ on both $\mathcal{D}_S$ and $\mathcal{D}_T$. We restate this result in the following Corollary.

Corollary 1 (Ben-David et al., 2007, Theorem 1): Consider two domains $\mathcal{D}_S$ and $\mathcal{D}_T$ over a shared feature space. The risk of a given hypothesis h on the target domain will be thus bounded by:

$$R_T[h] \leq R_S[h] + d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] + \lambda, \qquad (4)$$

where λ accounts for how "adaptable" the class $\mathcal{H}$ is and it is defined as the minimal total risk over both domains that can be achieved by some $h \in \mathcal{H}$:

$$\lambda = \min_{h \in \mathcal{H}} [R_S[h] + R_T[h]]. \qquad (5)$$

The term $d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ corresponds to the $\mathcal{H}\Delta\mathcal{H}$-divergence introduced in Kifer et al. (2004) for a hypothesis class $\mathcal{H}\Delta\mathcal{H} = \{h(x) \oplus h'(x) | h, h' \in \mathcal{H}\}$, where $\oplus$ is the XOR operation.

An extension of that result was introduced in Zhao et al. (2019) in order to replace λ by a term that explicitly accounts for a possible mismatch between the labeling rules of source and target domains, denoted as f_S and f_T, respectively. For that, the divergence between source and target is computed over a hypothesis class $\tilde{\mathcal{H}}$ defined as $\tilde{\mathcal{H}} = \{sign(|h(x) - h'(x)| - t) | h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. We state this result in the following Corollary:

Corollary 2 (Zhao et al., 2019, Theorem 4.1):

$$R_T[h] \leq R_S[h] + d_{\tilde{\mathcal{H}}}[\mathcal{D}_S, \mathcal{D}_T] + \\ \min\{\mathbb{E}_{x \sim \mathcal{D}_S} \mathbb{1}[f_S(x) \neq f_T(x)], \mathbb{E}_{x \sim \mathcal{D}_T} \mathbb{1}[f_S(x) \neq f_T(x)]\}, \qquad (6)$$

where $\min\{\mathbb{E}_{x \sim \mathcal{D}_S} \mathbb{1}[f_S(x) \neq f_T(x)], \mathbb{E}_{x \sim \mathcal{D}_T} \mathbb{1}[f_S(x) \neq f_T(x)]\}$ accounts the mismatch between the labeling functions.

In light of Corollaries 1 and 2, it is possible to point out the two main aspects that determine how well a hypothesis h generalizes from the source to the target domain. For that, the input space $\mathcal{X}$ must be such that the divergence $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ between the marginal distributions is low, while

the mismatch between labeling functions accounted by the term $\min\{\mathbb{E}_{x\sim\mathcal{D}_S}\mathbb{1}[f_S(x)\neq f_T(x)],\mathbb{E}_{x\sim\mathcal{D}_T}\mathbb{1}[f_S(x)\neq f_T(x)]\}$ is also small. Previous work on domain adaptation (e.g., Ganin et al., 2016) has mostly focused on mitigating the mismatch between marginal distribution and assumed that labeling functions were the same across domains. However, when this is not case, decreasing the discrepancy between marginal distributions (Zhao et al., 2019) or adding more data (Crammer et al., 2008) might actually hurt the performance of a model on the target domain.

2.3. Cross-subject generalization as domain adaptation

We formalize the problem of learning passive BCIs that generalize across subjects under the domain adaptation setting. For that, consider a dataset with a total of M subjects and that each subject is associated with domain $\mathcal{D}_i$ and labeling function f_i, $\forall i = \{1,\cdots,M\}$. Without loss of generality, assume that recordings from the first $M-1$ subjects are available at training time and we are interested in predicting how well a hypothesis $h \in \mathcal{H}$ would perform in the M-th subject, which was not considered at training time. Let $\mathcal{D}_S = \sum_{i=1}^{M}\pi_i\mathcal{D}_i$ be the source domain defined as a mixture of the domains corresponding to the training subjects with weights π_i equal to $\frac{1}{M}$ for all components, i.e., sampling from $\mathcal{D}_S$ yields data points from one of the training domains with equal probability. Taking into consideration Equation (6), we can bound the risk on the M-th unseen subject, $R_M[h]$ as

$$R_M[h] \leq R_S[h] + d_{\tilde{\mathcal{H}}}[\mathcal{D}_S,\mathcal{D}_M]+$$
$$\min\{\mathbb{E}_{x\sim\mathcal{D}_S}\mathbb{1}[f_S(x)\neq f_M(x)],\mathbb{E}_{x\sim\mathcal{D}_M}\mathbb{1}[f_S(x)\neq f_M(x)]\}. \tag{7}$$

In practice, we aggregate the available test samples from all the training subjects to estimate the risk of h in the source domain $R_S[h] = \sum_{k=1}^{M-1}R_i[h]$. However, there is no such straightforward way of estimating the two remaining terms of the bound. In the next Section, a strategy to compute these two terms is proposed.

While in this work we focus on formalizing the cross-subject generalization problem on EEG-based BCIs as domain adaptation and leverage the previous literature that set-up a theoretical framework for this topic to better understanding distribution shifts, past contributions applied domain adaptation techniques with the aim of mitigating the harms of domain shifts on BCIs. Under the assumption that only marginal shift is present, Ma et al. (2019) proposed an adversarial approach to enable improve generalization on unseen subjects via learning neural networks with components specific to each training subject's data, as well as parameters shared across all subjects. While the empirical performance of the proposed approach is promising in comparison to

previously proposed techniques, it is important to mention that, as we showed in this work, EEG dataset might also present conditional shifts and those must also be accounted for by domain adversarial methods (Zhao et al., 2019). Similarly, Li et al. (2018) proposed an adversarial approach based on mitigating marginal shifts via minimizing maximum mean discrepancies between distributions corresponding to training and test subjects by assuming unlabeled data from test subjects is available at training time.

3. Estimating shifts across multiple distributions

In this section, we provide practical strategies to estimate both conditional and marginal shifts for a case where multiple domains (subjects) are available. Quantifying such mismatch will enable us to:

- Shed light on which domain adaptation strategies should be used for a given scenario by verifying whether, for example, the covariate shift assumption holds.
- As these quantities are related to how well a particular hypothesis will perform on unseen subjects, we can use their estimates computed considering different feature spaces and infer which one would achieve better performance on unseen subjects.

3.1. Conditional shift

A conditional shift is observed across subjects when the labeling function (or, in the stochastic case, the conditional distribution of the labels given the input features) differ among the subjects, i.e., for M subjects, we have $f_i(x) \neq f_j(x)$, $\forall i,j = \{1,\cdots,M\}$. In order to characterize the cross-subject conditional shift of a dataset of M subjects, we consider the following quantity on the generalization bound presented in Corollary 2 for all pairs of subjects:

$$\min\{\mathbb{E}_{\mathcal{D}_i}[|f_i-f_j|],\mathbb{E}_{\mathcal{D}_j}[|f_j-f_i|]\}, \tag{8}$$

where $i,j = \{1,\cdots,M\}$. In practice, it is not possible to compute such quantity as one does not have access to the true labeling functions and computing the expectations in Equation (8) is intractable.

We thus propose to estimate such values by learning a labeling rule for each one of the domains, and account for how well it classifies examples from the other domain. Assuming that we are able to learn a good predictor for the labels of each domain, such approach is capable of accounting for how "close" the true labeling functions of different domains are. In practice, we consider that two labeled samples of size N from domains i

and j are available and compute the following estimator $\mu_{i,j}$ for the quantity $\mathbb{E}_{\mathcal{D}_i}[|f_i - f_j|]$:

$$\mu_{i,j} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}[f_i(x_n^i) \neq \tilde{f}_j(x_n^i)], \tag{9}$$

where $(x_n^i, y_n^i) \sim \mathcal{D}_i$, and $\tilde{f}_j$ is an approximated labeling function for the j-th subject. Beforehand, there are no particular constraints with respect to the hypothesis class and algorithm to approximate $\tilde{f}_j$. In light of that, in this work $\tilde{f}_j$ is set to be a decision procedure based on the Euclidean distance between data points in a fixed feature space. For that, we use a k-nearest neighbor (k-NN) labeling function, i.e., a k-NN binary classifier trained on $\mathcal{D}_j$ to classify as low or high mental workload condition data sampled from $\mathcal{D}_i$. Based on $\mu_{i,j}$ and $\mu_{j,i}$ we estimate the value $d_{i,j} = d_{j,i} = \min\{\mu_{i,j}, \mu_{j,i}\}$ and compose a Hermitian (elements symmetric with respect to the main diagonal are equal) disparity matrix D defined as:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,M} \\ d_{2,1} & d_{2,2} & \dots & d_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \dots & d_{M,M} \end{bmatrix}. \tag{10}$$

Notice that in the case we obtain optimal approximate labeling functions, i.e., $f_i(x_n^i) = \tilde{f}_j(x_n^i)$, $\forall i = j$, the trace of D is equal to 0. Finally, in order to obtain a single value representing the conditional shift of all subjects in a dataset, we aggregate the values of pairwise conditional shifts. For that, we compute the Frobenius norm $||.||_F$ of the disparity matrix D:

$$||D||_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} |d_{i,j}|^2}. \tag{11}$$

The resulting $||D||_F$ is then rescaled to the $[0,1]$ interval to allow for easier comparison across feature spaces.

3.2. Marginal shift

The $\mathcal{H}$-divergence between two distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ is defined as:

$$d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] = 2 \sup_{\eta \in \mathcal{H}} |\Pr_{x \sim \mathcal{D}_S}[\eta(x) = 1] - \Pr_{x \sim \mathcal{D}_T}[\eta(x) = 1]|. \tag{12}$$

As discussed by Ben-David et al. (2007), $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ can be estimated from the error ϵ of a binary classifier trained to distinguish samples from $\mathcal{D}_S$ and $\mathcal{D}_T$. The lower ϵ is, the highest the estimate of $d_{\mathcal{H}}$ will be, since in this case, there is a hypothesis η capable of distinguishing between $\mathcal{D}_S$ and $\mathcal{D}_T$ with high accuracy. Notice that the $\mathcal{H}$-divergence only

accounts for discrepancies between the marginal distributions of the domains, not accounting for how each data point is labeled. Therefore, it is not necessary to have access to labeled samples from the considered domains to estimate its value.

Our proposed approach to estimate the cross-subject marginal shift from a group of M domains (subjects) relies on estimating pair-wise domain divergences, i.e., we compute $d_{\mathcal{H}}[\mathcal{D}_i, \mathcal{D}_j]$ $\forall i,j = \{1, \cdots, M\}$. In the case of scenarios where EEG datasets are taken into account, estimating cross-domain marginal shifts consists in obtaining models capable of performing pair-wise discrimination of features extracted from recordings of different subjects. Similarly to the proposed strategy to estimating cross-subject conditional shift values, we introduce a Hermitian matrix H that accounts for marginal shifts between all subjects. Each entry of H corresponds to the average error rate of pair-wise subject classification. In practice, we use 5-fold cross validation to estimate the error rates. An aggregate value of marginal shift can also be obtained *via* the rescaled Frobenius norm of H.

4. Experimental setup

In this section we provide an overview of WAUC dataset, as well as introduce the features, normalization approaches, and the mental workload classification scheme utilized in the experiments. Moreover, we describe the implementation details in order to allow reproducibility of our experiments.

4.1. WAUC dataset

We consider the EEG recordings of the Workload Assessment Under physical aCtivity (WAUC) dataset (Albuquerque et al., 2020) for our experiments. This dataset was collected when subjects had cognitive and physical workload simultaneously modulated. Mental workload was modulated *via* the MATB-II task while physical activity consisted of running on a treadmill at 5 km/h or pedaling on a stationary bike at 70 rpm. EEG data was recorded using a Neurolectrics Enobio 8-channel wearable headset with a sampling rate of 500 Hz. Electrodes were placed following the 10-20 system at the frontal area in the positions AF7, FP1, FP2, and AF8. References were placed at FPz and Nz. The WAUC dataset also contains recordings from baseline periods during the data collection. There are two different types of baseline recordings: (1) EEG was recorded when no mental or physical effort was demanded from the participant (eyes-closed, no movement), and (2) Data was acquired when only physical effort was taken into account, i.e., subjects were running on the treadmill or pedaling at the specified speed while executing no mental task. Subjects performed two experimental sessions, each with an approximate duration of 10 min and under a different mental workload level.

For our experiments, we considered a total of 18 subjects from the dataset, whom half performed physical activity with the treadmill and the other half with the bike. From the selected subjects, 9 self-declared as male and 9 as female. The average age across subjects was 27 years old.

4.2. Feature extraction

Our preprocessing and feature extraction pipeline consisted in downsampling the EEG recording to 250 Hz, band-pass filtering from 0.5–45 Hz, and computing features over 4 s epochs with 3 s of overlap between consecutive windows. Considering a 10-min experimental session, after downsampling and epoching the data, we obtained an approximate total of 600 points per subject×session. As the literature has shown that increases in mental workload incur in changes in alpha, beta, and theta bands in the frontal cortex (Borghini et al., 2014; Hogervorst et al., 2014), we considered power spectral density (PSD) features in standard EEG frequency bands, namely: delta (0.1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (2–30 Hz).

4.3. Normalization

Feature normalization is a common practice used to minimize the effects of cross-subject variability for EEG-based classification tasks. Task-based features are typically normalized with respect to the statistics extracted from baseline periods (Bai et al., 2016; Bogaarts et al., 2016; Shedeed and Issa, 2016; Pati et al., 2020). The main goal of this strategy is to emphasize changes in the features that correspond to factors that were modulated during the experimental task. In the case of the WAUC dataset, normalizing the features with respect to the first baseline period (baseline 1) highlights changes on the PSD due to both mental and physical stimuli. In turn, normalization with respect to the statistics of recording collected during the second baseline highlights modifications stemming only from mental workload changes, as only physical strain was modulated during this step.

While commonly believed to improve classification accuracy, it is not clear from a statistical learning perspective whether and why these different normalization strategies work. Here, we quantitatively assess the impact that normalization has on mental workload performance under the lens of conditional and marginal shifts, as well as of cross-subject classification performance. As such, we perform a subject-wise normalization of each feature according to[1]:

$$x'_n = \frac{x_n - \beta}{\gamma},\qquad(13)$$

---

[1] Here, the operands are arrays and the subtraction and division symbols denote element-wise operations.

where β corresponds to the average feature vector and γ the standard deviation considering the data recorded for the respective subject during the baseline periods. Notice that for each subject, x_n corresponds to a data instance of size F, i.e., an array of size F, where F denotes the number of features considered for training the predictors for estimating each type of distribution shift. γ and β are arrays of size F with entries corresponding to the average and standard deviation of each feature across all data points, respectively.

In addition to the aforementioned normalization strategies, we also perform experiments with features obtained after per-subject whitening of the data, i.e., β is the sample average and γ the standard deviation for a given subject. This procedure is commonly referred to as z-score normalization. Lastly, we considered features without any normalization. As such, a total of four feature spaces are considered across our experiments: no normalization, whitening, and baselines 1/2 normalization.

4.4. Cross-subject mental workload classification

In addition to analyzing the estimated cross-subject conditional and marginal shift for a mental workload assessment task, we also evaluate the cross-subject classification performance in this scenario. For that, we consider a leave-one-subject-out (LOSO) evaluation scheme and train a different classifier per subject not included in the training set. Using this approach, we set our problem as a single-source single-target domain adaptation, where the source domain corresponds to the data of the all subjects pooled together, and the target domain corresponding to the subject left out as the test set. Although this is the setting considered in the experiments, we did not apply any domain adaptation scheme when learning classifiers since our objective in this work is to investigate distributional shifts and their relationship with out-of-distribution generalization.

4.5. Implementation details

We implemented all classifiers, normalization, and cross-validation schemes using Scikit-learn (Pedregosa et al., 2011). For all experiments, we performed 30 independent repetitions considering slightly different partitions of the available data examples by randomly selecting 300 data points per subject×session. To enforce reproducibility, the random seed for all experiments was set to 10. A Random Forest with 20 estimators is used as the subject classifier to estimate $d_{\mathcal{H}}$ for computing the marginal shift. For predicting mental workload using LOSO cross-validation, we also use a Random Forest classifier, but in this case with 30 estimators.

5. Results and discussion

In this section, we aim at answering the following main questions: (i) Do different feature normalization schemes yield different values of distributional shifts? (ii) Can the estimation of distributional shifts indicate how difficult it is to learn BCIs that generalize well on unseen subjects? (iii) For a fixed feature space, are our findings consistent across two partitions of the WAUC containing subjects that had physical activity levels modulated by either bike or treadmill?

5.1. Statistical shifts estimation

Figures 1, 2 show the boxplots with 30 independent estimates of the conditional shift for subjects corresponding to treadmill and bike, respectively. Considering the results obtained with the non-normalized version of the features as reference, it is possible to observe that whitening the features significantly improved the estimated aggregate conditional shift values (Equation 11) for both treadmill and bike cases. As expected, this type of normalization is widely used in machine learning and known to improve overall classification performance in different applications of EEG data (Sulaiman et al., 2010; Zhang et al., 2013; Cruz et al., 2021).

In the case of normalizing the features with respect to the baseline periods, our findings show large differences when comparing the treadmill and bike conditions. For the bike case, normalizing the features yielded only a slight decrease in the observed conditional shift for both baseline 1 and 2 periods. For the treadmill condition, on the other hand, normalizing relative to baseline 1 (no physical activity) resulted in an increase of the aggregated conditional shift, thus potentially negatively affecting the performance of the mental workload assessment model to unseen subjects. Baseline 2 normalization, in turn, reduced the estimated conditional shift to levels closer to that achieved with per-subject whitening.

In addition to investigating the aggregated conditional shift values, an in-depth analysis is also performed for the conditional shift values across all pairs of subjects in order to better understand the effects of feature normalization and the dependency on activity type. For that, Figures 3, 4 display the disparity matrices D computed considering features without normalization and whitening for both activity types, respectively. Notice that the entries at the main diagonal (i.e., within-subject disparity) were computed by having disjoint training and test sets, thus these values provide information about how good the employed labeling function approximation was. Also, these results correspond to a single estimate, thus do not show the variability of the reported quantities as it is the case in Figures 1, 2.

It can be observed that the cross-subject conditional shift for the bike condition is much higher in comparison to the

FIGURE 1
Boxplots with 30 independent estimates of the aggregate cross-subject conditional shift across different normalization strategies for participants which performed physical activity using a **treadmill**. Lower values represent smaller estimated conditional shifts.

FIGURE 2
Boxplots with 30 independent estimates of the aggregate cross-subject conditional shift across different normalization strategies for participants which performed physical activity using a **bike**. Lower values represent smaller estimated conditional shifts.

treadmill condition. This observation agrees with the findings of Ladouce et al. (2019) and Albuquerque et al. (2020), which observed that different methods for inducing physical activity generate different EEG responses. Our results indicate that in the case of PSD features, this difference can be observed in practice by EEG responses which are more subject-specific, resulting in lower classification performance for the case of performing

FIGURE 3
Pair-wise cross-subject conditional shift with non-normalized and whitened features computed from subjects that performed physical activity on the **treadmill**. **(A)** No normalization. **(B)** Z-score normalization.

FIGURE 4
Pair-wise cross-subject conditional shift with non-normalized and whitened features computed from subjects that performed physical activity on the **bike**. **(A)** No normalization. **(B)** Z-score normalization.

activity with a stationary bike, as reported by Albuquerque et al. (2020).

Similarly to the conditional shift analysis, we show in Figures 5, 6 boxplots for the estimated aggregated marginal shift computed 30 times for all the considered normalization procedures, for treadmill and bike conditions, respectively. It is important to highlight that higher values of marginal shift (i.e., high $d_{\mathcal{H}}$) indicate a higher accuracy on pair-wise cross-subject classification. As such, discriminating data from two subjects in the PSD feature space consists in an easier task, and this contributes to higher cross-subject variability.

We observe that for both treadmill and bike cases, subject-wise feature whitening decreased the estimated marginal shift, while baseline 1 and 2 normalization increased it. Intuitively, we expected z-score normalization to decrease the marginal shift, as the normalized features for all subjects have equal first and second order statistics. On the other hand, according to previous results on baseline normalization for EEG features, we expected that both baseline 1 and baseline 2 methods would make it more difficult for the classifier to discriminate subjects in the PSD feature space.
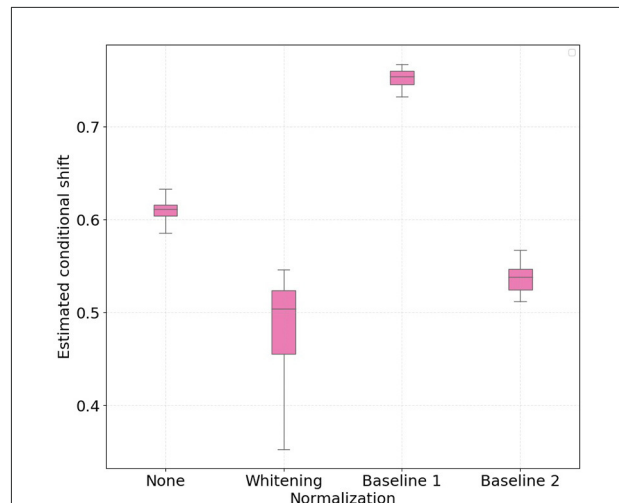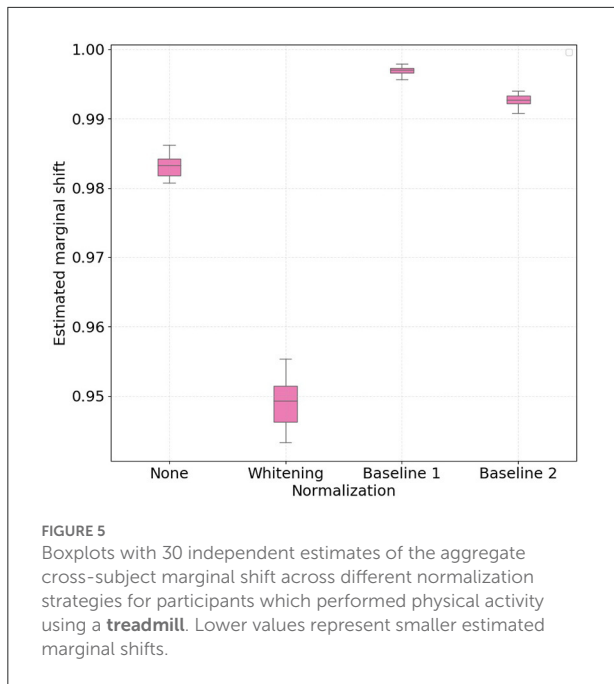
FIGURE 5
Boxplots with 30 independent estimates of the aggregate cross-subject marginal shift across different normalization strategies for participants which performed physical activity using a **treadmill**. Lower values represent smaller estimated marginal shifts.
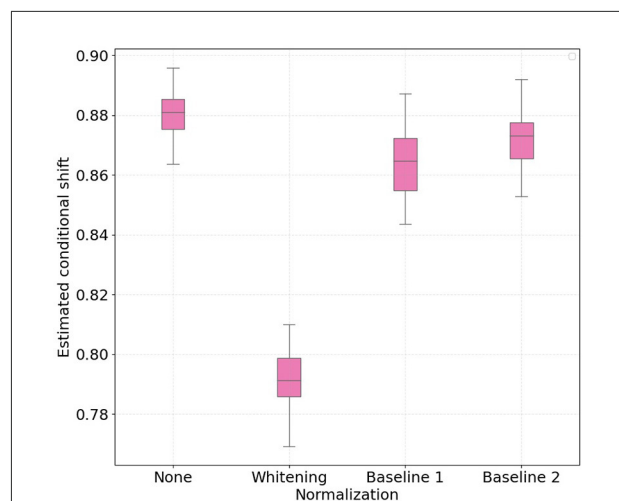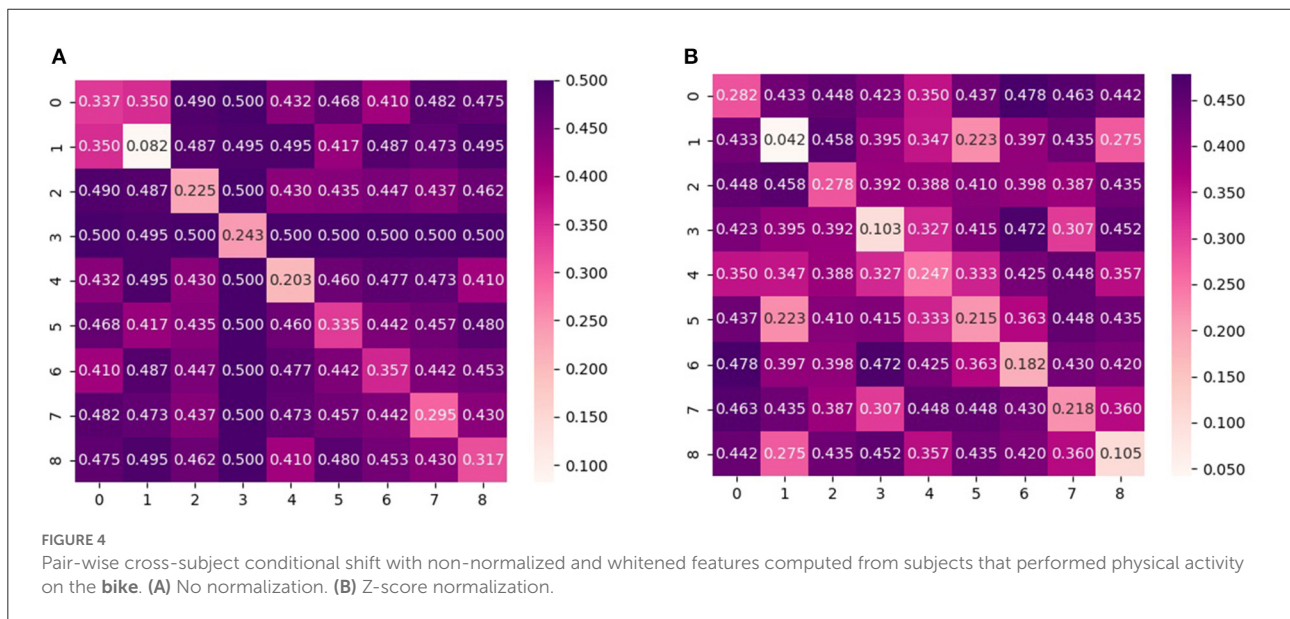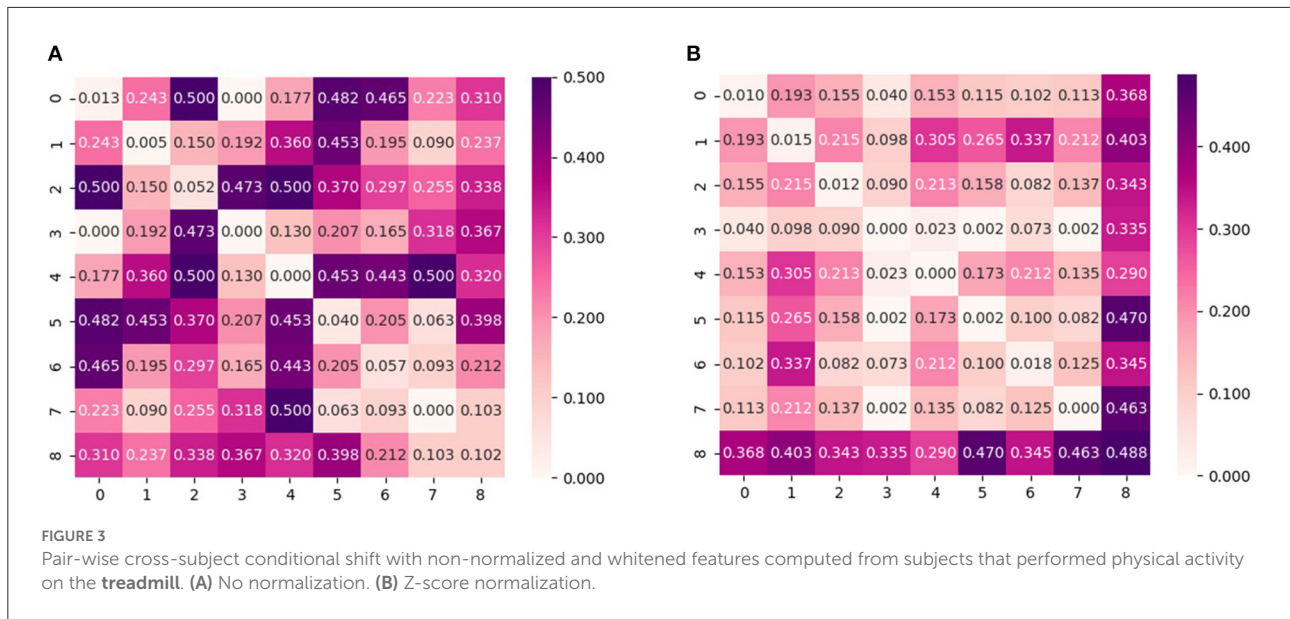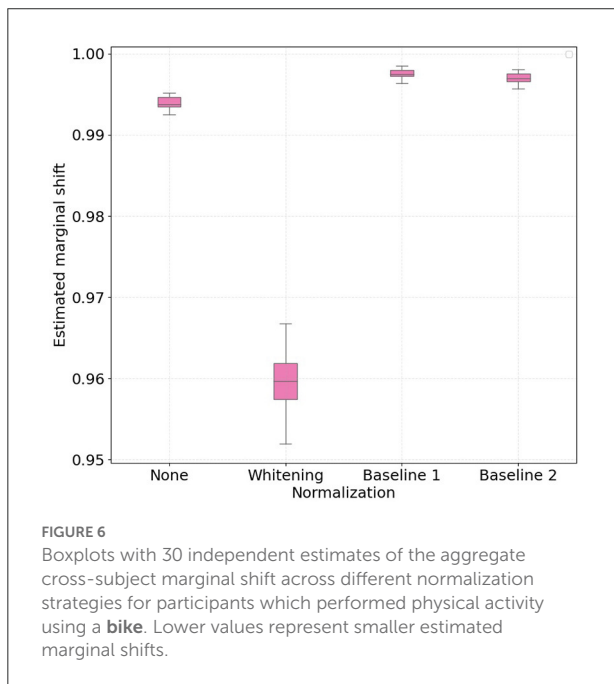

FIGURE 6
Boxplots with 30 independent estimates of the aggregate cross-subject marginal shift across different normalization strategies for participants which performed physical activity using a **bike**. Lower values represent smaller estimated marginal shifts.

5.2. Generalization gap

Lastly, target domain accuracy (i.e., test set or left-out subject) is reported for low/high mental workload classification using a LOSO cross-validation scheme. In addition to the test accuracy calculated on data from the subject left out, we also compute the classifier performance on the source domain by taking out from the training data 200 data points per subject.

Based on the bound shown in Equation (6), our goal is to verify whether the estimated conditional and marginal shift values provide a way to assess the generalization gap between source and target domains. We use the training accuracy to compute the empirical risk, as it is equal to $1 - \hat{R}_X[h_X]$ calculated with a 0-1 loss. Likewise, the true risk $R_X[h_X]$ was estimated as the accuracy on the test set. We calculated training and test average accuracy and the corresponding standard deviation across 30 independent runs. These values are shown per subject left out during training and averaged across all subjects. We also report average and standard deviation values of the generalization gap for each subject, calculated as the absolute difference between training and test accuracy. Tables 1, 2 present these quantities for the treadmill and the bike conditions, respectively.

According to the results presented in Table 1, we observe that results indicate z-score normalization, i.e., the features with lower conditional and marginal shifts, to yield the smallest average approximated generalization gap between source and target domains in most of the subjects. This finding is similarly observed in the case of the group of subjects that performed the experiment with the stationary bike, as shown by the results reported in Table 2. An overall comparison between treadmill and bike subjects also reveals that inter-subject generalization, as measured by the estimate of the risk on the source domain (training subjects), is considerably lower for the bike condition. This aspect could also have been predicted by the diagonal values of the disparity matrix (Figures 3, 4) which show that for the majority of the subjects the approximated labeling function seems to be better estimated for the treadmill condition.

Moreover, in the case of the treadmill group, we observe that baseline 1 normalization yielded a slightly smaller average generalization gap in comparison to baseline 2, even though it presented a considerably higher conditional shift. As both normalization strategies obtained close values of average marginal shift, we believe this indicates that the two analyzed statistical shifts might differ in their contribution to the generalization bound. Furthermore, considering the average results across all subjects, z-score normalization presented the best performance in terms of accuracy, being able to correctly classify roughly 70% of points from subjects not considered during training. It is important to highlight that as opposed to normalizing with respect to baseline recordings, which requires a calibration step to collect data prior to the actual task, z-score normalization does not need any extra information other than the features extracted from data corresponding to the task. On the other hand, despite better mitigating cross-subject variability and being more efficient in terms of data collection time, the intra-subject classification performance of models trained on z-score normalized features is worse in comparison with other strategies, indicating there might be a trade-off between improving cross-subject performance and maintaining good accuracy on the source domains.

TABLE 1 Results of binary mental workload classification with leave-one-subject-out cross validation for subjects that performed physical activity on the treadmill.

Subject	None	Whitening	Baseline 1	Baseline 2
S0	0.974 ± 0.004	0.936 ± 0.007	0.985 ± 0.003	0.982 ± 0.004
	0.764 ± 0.055	0.588 ± 0.018	0.889 ± 0.044	0.704 ± 0.028
	0.210 ± 0.055	0.348 ± 0.018	0.096 ± 0.044	0.279 ± 0.029
S1	0.974 ± 0.005	0.939 ± 0.010	0.985 ± 0.003	0.976 ± 0.003
	0.543 ± 0.043	0.628 ± 0.042	0.550 ± 0.037	0.560 ± 0.051
	0.431 ± 0.045	0.311 ± 0.044	0.435 ± 0.037	0.416 ± 0.050
S2	0.974 ± 0.004	0.941 ± 0.007	0.985 ± 0.003	0.979 ± 0.005
	0.575 ± 0.046	0.602 ± 0.058	0.524 ± 0.015	0.630 ± 0.052
	0.399 ± 0.046	0.340 ± 0.060	0.461 ± 0.016	0.349 ± 0.051
S3	0.974 ± 0.005	0.934 ± 0.008	0.984 ± 0.003	0.978 ± 0.004
	0.700 ± 0.079	0.968 ± 0.060	0.643 ± 0.082	0.603 ± 0.055
	0.249 ± 0.063	0.054 ± 0.065	0.292 ± 0.104	0.281 ± 0.093
S4	0.977 ± 0.003	0.939 ± 0.008	0.985 ± 0.003	0.983 ± 0.004
	0.662 ± 0.032	0.771 ± 0.056	0.540 ± 0.022	0.541 ± 0.024
	0.315 ± 0.032	0.168 ± 0.056	0.445 ± 0.022	0.441 ± 0.024
S5	0.973 ± 0.003	0.942 ± 0.009	0.989 ± 0.004	0.979 ± 0.004
	0.601 ± 0.044	0.851 ± 0.067	0.530 ± 0.030	0.554 ± 0.030
	0.372 ± 0.044	0.092 ± 0.067	0.454 ± 0.030	0.425 ± 0.028
S6	0.980 ± 0.005	0.945 ± 0.009	0.987 ± 0.003	0.978 ± 0.005
	0.751 ± 0.042	0.595 ± 0.037	0.588 ± 0.074	0.567 ± 0.049
	0.229 ± 0.043	0.350 ± 0.039	0.399 ± 0.074	0.411 ± 0.049
S7	0.973 ± 0.004	0.935 ± 0.007	0.985 ± 0.003	0.975 ± 0.004
	0.613 ± 0.088	0.862 ± 0.044	0.821 ± 0.074	0.565 ± 0.047
	0.360 ± 0.089	0.073 ± 0.047	0.164 ± 0.075	0.410 ± 0.047
S8	0.984 ± 0.003	0.959 ± 0.006	0.989 ± 0.003	0.982 ± 0.003
	0.608 ± 0.058	0.508 ± 0.004	0.597 ± 0.054	0.584 ± 0.061
	0.375 ± 0.059	0.451 ± 0.006	0.392 ± 0.055	0.398 ± 0.060
All	0.976 ± 0.005	0.941 ± 0.011	0.985 ± 0.004	0.979 ± 0.005
	0.649 ± 0.093	0.708 ± 0.155	0.637 ± 0.150	0.600 ± 0.078
	0.327 ± 0.093	0.242 ± 0.147	0.348 ± 0.140	0.379 ± 0.078
Cond. shift	0.608 ± 0.013	0.482 ± 0.060	0.753 ± 0.009	0.537 ± 0.014
Marg. shift.	0.981 ± 0.002	0.949 ± 0.003	0.997 ± 0.001	0.993 ± 0.001

For each subject, top and middle rows represent training and test accuracy, respectively. The estimated generalization gap is shown below the dotted line. Average and standard deviation across 30 independent runs are reported.

To provide further empirical evidence that the analysis of the statistical shifts as employed in this work can be used to select a feature normalization that yields better cross-domain (i.e., cross-subject) generalization, we show in Figure 7 boxplots of 30 independent generalization gap estimates for each subject within the treadmill group. In addition, we provide in Figure 8, a bar plot with average values of cross-subject disparity for all subjects that had physical workload modulated by the treadmill. These values were computed using the columns of the average disparity matrix resulting from the 30 repetitions executed to generate Figure 1. Notice that within this analysis we are not taking into account the marginal

shift. By comparing Figure 7 and Figure 8 we observe that for subjects 2, 3, 4, 5, 7, and 8 the normalization method with lower average conditional shift, yielded a smaller median estimated generalization gap. Importantly, we observe that subject 8 did not benefit from z-score normalization, as the conditional shift increased, along with an increase in the generalization and a decrease in the accuracy as shown in Table 1. While these results suggest the existence of correlation between our proposed estimator for the conditional shift and the generalization gap between unseen and training domains, a thorough investigation considering multiple datasets and a larger number of subjects should be carried out as part of

TABLE 2 Results of binary mental workload classification with leave-one-subject-out cross validation for subjects that performed physical activity on the bike.

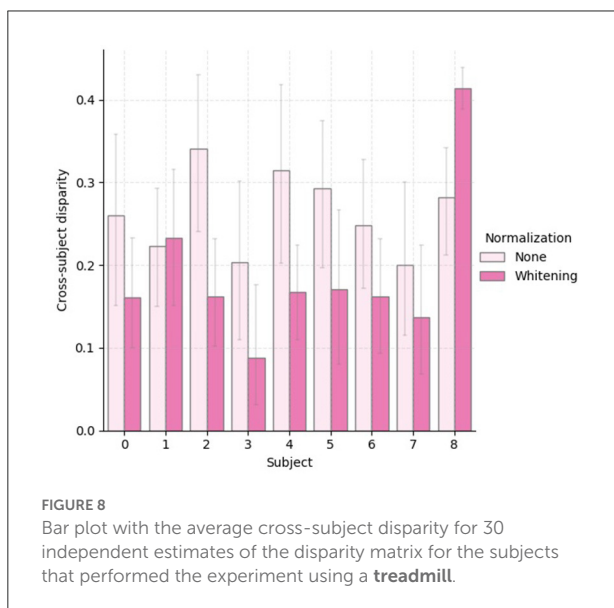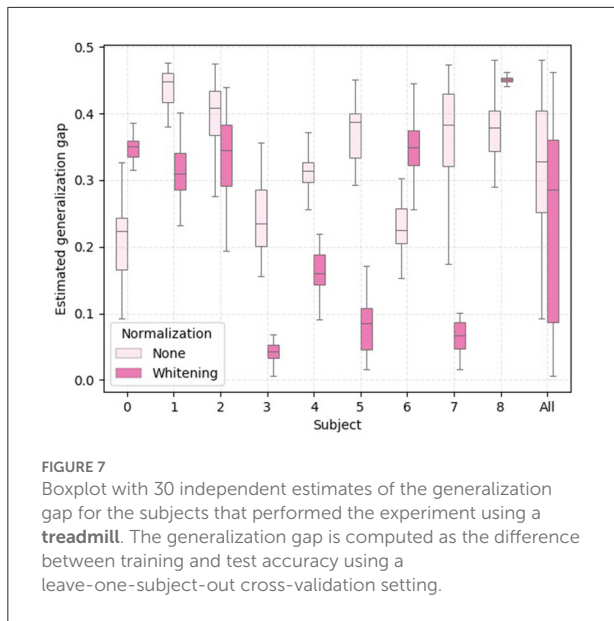| Subject	None	Whitening	Baseline 1	Baseline 2
S0	0.921 ± 0.008	0.845 ± 0.009	0.923 ± 0.007	0.920 ± 0.006
	0.534 ± 0.026	0.536 ± 0.023	0.525 ± 0.016	0.558 ± 0.022
	0.388 ± 0.028	0.309 ± 0.026	0.398 ± 0.016	0.363 ± 0.024
S1	0.893 ± 0.009	0.826 ± 0.015	0.899 ± 0.008	0.892 ± 0.007
	0.545 ± 0.041	0.579 ± 0.027	0.550 ± 0.046	0.568 ± 0.055
	0.348 ± 0.043	0.246 ± 0.030	0.348 ± 0.047	0.324 ± 0.053
S2	0.906 ± 0.007	0.829 ± 0.011	0.909 ± 0.008	0.904 ± 0.009
	0.545 ± 0.037	0.550 ± 0.026	0.507 ± 0.007	0.519 ± 0.014
	0.361 ± 0.038	0.279 ± 0.025	0.402 ± 0.012	0.385 ± 0.018
S3	0.892 ± 0.009	0.814 ± 0.012	0.896 ± 0.009	0.895 ± 0.009
	0.541 ± 0.039	0.681 ± 0.067	0.613 ± 0.078	0.578 ± 0.063
	0.351 ± 0.038	0.133 ± 0.067	0.284 ± 0.079	0.317 ± 0.063
S4	0.903 ± 0.008	0.836 ± 0.013	0.907 ± 0.008	0.900 ± 0.007
	0.549 ± 0.027	0.541 ± 0.024	0.575 ± 0.054	0.542 ± 0.034
	0.354 ± 0.029	0.295 ± 0.028	0.331 ± 0.051	0.358 ± 0.036
S5	0.910 ± 0.009	0.837 ± 0.012	0.914 ± 0.007	0.909 ± 0.008
	0.529 ± 0.020	0.555 ± 0.037	0.531 ± 0.019	0.522 ± 0.019
	0.380 ± 0.023	0.283 ± 0.040	0.383 ± 0.021	0.387 ± 0.020
S6	0.914 ± 0.008	0.847 ± 0.013	0.918 ± 0.008	0.914 ± 0.007
	0.529 ± 0.022	0.520 ± 0.015	0.535 ± 0.026	0.536 ± 0.025
	0.385 ± 0.022	0.327 ± 0.020	0.383 ± 0.027	0.378 ± 0.027
S7	0.900 ± 0.007	0.841 ± 0.009	0.905 ± 0.007	0.898 ± 0.010
	0.549 ± 0.033	0.547 ± 0.026	0.553 ± 0.033	0.557 ± 0.043
	0.350 ± 0.032	0.294 ± 0.027	0.352 ± 0.033	0.341 ± 0.043
S8	0.896 ± 0.009	0.841 ± 0.012	0.904 ± 0.008	0.900 ± 0.008
	0.551 ± 0.039	0.599 ± 0.030	0.546 ± 0.033	0.542 ± 0.028
	0.345 ± 0.040	0.242 ± 0.034	0.358 ± 0.033	0.358 ± 0.031
All	0.904 ± 0.012	0.835 ± 0.016	0.908 ± 0.011	0.904 ± 0.012
	0.541 ± 0.033	0.567 ± 0.057	0.548 ± 0.050	0.547 ± 0.042
	0.363 ± 0.037	0.268 ± 0.065	0.360 ± 0.054	0.357 ± 0.045
Cond. shift	0.880 ± 0.008	0.792 ± 0.010	0.864 ± 0.011	0.872 ± 0.010
Marg. shift.	0.994 ± 0.001	0.959 ± 0.004	0.998 ± 0.001	0.997 ± 0.001

For each subject, top and middle rows represent training and test accuracy, respectively. The estimated generalization gap is shown below the dotted line. Average and standard deviation across 30 independent runs are reported.

follow up work in order to verify this observation *via* rigorous statistical tools.

5.3. Main takeaways

In light of our results and discussion, we highlight the observations we found most relevant to be considered by future research. In case the goal is to improve out-of-distribution performance, normalization procedures that decrease the overall cross-subject conditional shift should be prioritized since they yield smaller generalization gaps. To devise passive BCIs with the aim of monitoring mental workload under physical activity, our analysis showed that z-score normalization provided the best strategy for normalizing EEG power spectral density features. Moreover, such normalized feature spaces should be considered in case representation learning based on domain adaptation are used to learn domain-invariant classifiers. Notice there is a caveat that should also be taken into account: the results shown in Tables 1, 2 consistently indicate (i.e., across equipment for modulating physical activity and normalization procedures) that improving out-of-distribution performance *via* normalizing the features leads to a decrease on the model accuracy computed on unseen data from the training subjects.

FIGURE 7
Boxplot with 30 independent estimates of the generalization gap for the subjects that performed the experiment using a **treadmill**. The generalization gap is computed as the difference between training and test accuracy using a leave-one-subject-out cross-validation setting.

FIGURE 8
Bar plot with the average cross-subject disparity for 30 independent estimates of the disparity matrix for the subjects that performed the experiment using a **treadmill**.

6. Conclusions

In this work, we present the first steps toward better understanding the cross-subject variability phenomena seen with passive EEG-based BCIs from a statistical learning perspective. We looked at this problem through the lens of domain adaptation and proposed strategies to estimate distributional shifts between conditional and marginal distributions corresponding to the data generating process of features and labels from different subjects. To evaluate the proposed approach, the WAUC dataset was used and binary mental workload assessment from EEG power spectral features

was performed. Our analysis showed that feature normalization, as well as data collection conditions such as the equipment used to induce physical workload, had a relevant impact in the estimated values of conditional shift. Importantly, our results showed that whitening the features (i.e., performing z-score normalization) mitigated both conditional and marginal shifts and improved mental workload assessment on unseen subjects at training time. Future work consists on further exploring the proposed estimators on datasets comprising different data modalities and tasks. Investigating the interplay between the choice of the class approximating the labeling functions would also be a relevant direction in order to obtain better estimators of conditional shift. In the case of continuing exploring the WAUC database, we believe a fine-grained analysis of distribution shifts considering the different subgroups with respect to the demographic attributes would shed light on how distribution shifts affect classification fairness. Finally, future work can also employ the developed strategies to estimate distributional shifts in order to better inform the development of domain adaptation methods for EEG applications.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: http://musaelab.ca/resources/.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Review Boards of INRS, Université Laval and the PERFORM Centre (Concordia University). The patients/participants provided their written informed consent to participate in this study.

Author contributions

IA: conceived and implemented the experimental analysis and writing. IA and JM: conceived the main ideas in the work. OR: feature extraction. TF: reviewing, funding, and supervision. All authors read, revised, and approved the manuscript.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akuzawa, K., Iwasawa, Y., and Matsuo, Y. (2019). *Domain Generalization via Invariant Representation Under Domain-Class Dependency.* Available online at: https://openreview.net/forum?id=HJx38iC5KX

Albuquerque, I., Monteiro, J., Rosanne, O., Tiwari, A., Gagnon, J.-F., and Falk, T. H. (2019). "Cross-subject statistical shift estimation for generalized electroencephalography-based mental workload assessment," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari), 3647–3653. doi: 10.1109/SMC.2019.8914469

Albuquerque, I., Tiwari, A., Gagnon, J.-F., Lafond, D., Parent, M., Tremblay, S., et al. (2018). "On the analysis of EEG features for mental workload assessment during physical activity," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (Miyazaki), 538–543. doi: 10.1109/SMC.2018.00101

Albuquerque, I., Tiwari, A., Parent, M., Cassani, R., Gagnon, J.-F., Lafond, D., Tremblay, S., and Falk, T. H. (2020). WAUC: a multi-modal database for mental workload assessment under physical activity. *Front. Neurosci.* 14:549524. doi: 10.3389/fnins.2020.549524

Aricò, P., Borghini, G., Di Flumeri, G., Sciaraffa, N., and Babiloni, F. (2018). Passive bci beyond the lab: current trends and future directions. *Physiol. Measure.* 39:08TR02. doi: 10.1088/1361-6579/aad57e

Bai, Y., Huang, G., Tu, Y., Tan, A., Hung, Y. S., and Zhang, Z. (2016). Normalization of pain-evoked neural responses using spontaneous EEG improves the performance of EEG-based cross-individual pain prediction. *Front. Comput. Neurosci.* 10:31. doi: 10.3389/fncom.2016.00031

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 137–144.

Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010). "Impossibility theorems for domain adaptation," in *International Conference on Artificial Intelligence and Statistics* (Sardinia), 129–136.

Bogaarts, J., Hilkman, D., Gommer, E. D., van Kranen-Mastenbroek, V., and Reulen, J. P. (2016). Improved epileptic seizure detection combining dynamic feature normalization with EEG novelty detection. *Med. Biol. Eng. Comput.* 54, 1883–1892. doi: 10.1007/s11517-016-1479-8

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Crammer, K., Kearns, M., and Wortman, J. (2008). Learning from multiple sources. *J. Mach. Learn. Res.* 9, 1757–1774.

Cruz, A., Pires, G., Lopes, A., Carona, C., and Nunes, U. J. (2021). A self-paced BCI with a collaborative controller for highly reliable wheelchair driving: experimental tests with physically disabled individuals. *IEEE Trans. Hum. Mach. Syst.* 51, 109–119. doi: 10.1109/THMS.2020.3047597

Daume, H. III, and Marcu, D. (2006). Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.* 26, 101–126. doi: 10.1613/jair.1872

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.1007/978-3-319-58347-1_10

Hogervorst, M., Brouwer, A.-M., and van Erp, J. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* 8:322. doi: 10.3389/fnins.2014.00322

Johansson, F. D., Sontag, D., and Ranganath, R. (2019). "Support and invertibility in domain-invariant representations," in *The 22nd International Conference on Artificial Intelligence and Statistics* (Naha), 527–536.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). "Detecting change in data streams," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases, Vol. 30* (Toronto, ON), 180–191. doi: 10.1016/B978-012088469-8.50019-X

Ladouce, S., Donaldson, D. I., Dudchenko, P. A., and Ietswaart, M. (2019). Mobile EEG identifies the re-allocation of attention during real-world activity. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-51996-y

Li, H., Jin, Y.-M., Zheng, W.-L., and Lu, B.-L. (2018). "Cross-subject emotion recognition using deep adaptation networks," in *International Conference on Neural Information Processing* (Siem Reap: Springer), 403–413. doi: 10.1007/978-3-030-04221-9_36

Lotte, F. (2015). Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces. *Proc. IEEE* 103, 871–890. doi: 10.1109/JPROC.2015.2404941

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Ma, B.-Q., Li, H., Zheng, W.-L., and Lu, B.-L. (2019). "Reducing the subject variability of EEG signals with adversarial domain generalization," in *International Conference on Neural Information Processing* (Sanur: Springer), 30–42. doi: 10.1007/978-3-030-36708-4_3

Pati, S., Toth, E., and Chaitanya, G. (2020). Quantitative EEG markers to prognosticate critically ill patients with COVID-19: a retrospective cohort study. *Clin. Neurophysiol.* 131:1824. doi: 10.1016/j.clinph.2020.06.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Shedeed, H. A., and Issa, M. F. (2016). Brain-EEG signal classification based on data normalization for controlling a robotic arm. *Int. J. Tomogr. Simul.* 29, 72–85.

Sulaiman, N., Taib, M. N., Aris, S. A. M., Hamid, N. H. A., Lias, S., and Murat, Z. H. (2010). "Stress features identification from EEG signals using EEG asymmetry & spectral centroids techniques," in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (Kuala Lumpur), 417–421. doi: 10.1109/IECBES.2010.5742273

Sun, B., Feng, J., and Saenko, K. (2017). "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*, ed G. Csurka (Springer), 153–171. doi: 10.1007/978-3-319-58347-1_8

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Hawaii). doi: 10.1109/CVPR.2017.316

Wang, S., Gwizdka, J., and Chaovalitwongse, W. A. (2015). Using wireless EEG signals to assess memory workload in the *n*-back task. *IEEE Trans. Hum. Mach. Syst.* 46, 424–435. doi: 10.1109/THMS.2015.2476818

Wei, C.-S., Lin, Y.-P., Wang, Y.-T., Lin, C.-T., and Jung, T.-P. (2018). A subject-transfer framework for obviating inter-and intra-subject variability in EEG-based drowsiness detection. *NeuroImage* 174, 407–419. doi: 10.1016/j.neuroimage.2018.03.032

Wu, D. (2016). Online and offline domain adaptation for reducing BCI calibration effort. *IEEE Trans. Hum. Mach. Syst.* 47, 550–563. doi: 10.1109/THMS.2016.2608931

Wu, D., Chuang, C.-H., and Lin, C.-T. (2015a). "Online driver's drowsiness estimation using domain adaptation with model fusion," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an), 904–910. doi: 10.1109/ACII.2015.7344682

Wu, D., Lawhern, V. J., and Lance, B. J. (2015b). "Reducing BCI calibration effort in rsvp tasks using online weighted adaptation regularization with source domain selection," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an), 567–573. doi: 10.1109/ACII.2015.7344626

Yin, Z., and Zhang, J. (2017). Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive

weights. *Neurocomputing* 260, 349–366. doi: 10.1016/j.neucom.2017.
05.002

Zhang, J., Yin, Z., and Wang, R. (2014). Recognition of mental workload levels
under complex human-machine collaboration by using physiological features and
adaptive support vector machines. *IEEE Trans. Hum. Mach. Syst.* 45, 200–214.
doi: 10.1109/THMS.2014.2366914

Zhang, J., Yin, Z., and Wang, R. (2017). Nonlinear dynamic classification of
momentary mental workload using physiological features and NARX-model-based

least-squares support vector machines. *IEEE Trans. Hum. Mach. Syst.* 47, 536–549.
doi: 10.1109/THMS.2017.2700631

Zhang, R., Xu, P., Guo, L., Zhang, Y., Li, P., and Yao, D. (2013). Z-score linear
discriminant analysis for eeg based brain-computer interfaces. *PLoS ONE* 8:e74433.
doi: 10.1371/journal.pone.0074433

Zhao, H., Combes, R. T. D., Zhang, K., and Gordon, G. J. (2019). On learning
invariant representation for domain adaptation. *arXiv preprint arXiv:1901.
09453*.