MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

AutoTransfer: Subject Transfer Learning with Censored Representations on Biosignals Data

Smedemark-Margulies, Niklas; Wang, Ye; Koike-Akino, Toshiaki; Erdogmus, Deniz

TR2022-098 September 17, 2022

Abstract

We investigate a regularization framework for subject transfer learning in which we train an encoder and classifier to minimize classification loss, subject to a penalty measuring independence between the latent representation and the subject label. We introduce three notions of independence and corresponding penalty terms using mutual information or divergence as a proxy for independence. For each penalty term, we provide several concrete estimation algorithms, using analytic methods as well as neural critic functions. We propose a hands-off strategy for applying this diverse family of regular- ization schemes to a new dataset, which we call "AutoTransfer". We evaluate the performance of these individual regularization strategies under our AutoTransfer framework on EEG, EMG, and ECoG datasets, showing that these approaches can improve subject transfer learning for challenging real-world datasets

International Conference of the IEEE Engineering in Medicine & Biology Society (EMBS) 2022

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

AutoTransfer: Subject Transfer Learning with Censored Representations on Biosignals Data

Niklas Smedemark-Margulies[1],
Northeastern University
Boston, MA, USA
smedemark-margulie.n@northeastern.edu

Ye Wang, Toshiaki Koike-Akino,
Mitsubishi Electric Research Labs. (MERL)
Cambridge, MA, USA
{yewang, koike}@merl.com

Deniz Erdoğmuş
Northeastern University
Boston, MA, USA
d.erdogmus@northeastern.edu

***Abstract*—We investigate a regularization framework for subject transfer learning in which we train an encoder and classifier to minimize classification loss, subject to a penalty measuring independence between the latent representation and the subject label. We introduce three notions of independence and corresponding penalty terms using mutual information or divergence as a proxy for independence. For each penalty term, we provide several concrete estimation algorithms, using analytic methods as well as neural critic functions. We propose a hands-off strategy for applying this diverse family of regularization schemes to a new dataset, which we call "AutoTransfer". We evaluate the performance of these individual regularization strategies under our AutoTransfer framework on EEG, EMG, and ECoG datasets, showing that these approaches can improve subject transfer learning for challenging real-world datasets.**

***Index Terms*—Transfer Learning, Deep Learning, Regularized Representation Learning, EEG, EMG, ECoG, AutoML**

I. Introduction

In this work, we investigate methods for transfer learning in the classification of biosignals data. Previous work has established the difficulty of transfer learning for biosignals and even the issue of so-called "negative transfer" [1], in which naïve attempts to combine datasets from multiple subjects or sessions can paradoxically decrease model performance, due to differences in response statistics. We address the problem of subject transfer by training models to be invariant to changes in a nuisance variable representing subject identifier (ID). We examine previously established approaches and develop several new approaches based on recent work in mutual information estimation and generative modeling. We evaluate these methods on a variety of electroencephalography (EEG), electromyography (EMG), and electrocorticography (ECoG) datasets, to demonstrate that these methods can improve generalization to unseen test subjects. We also provide an automated hyperparameter search procedure for applying these methods to new datasets, which we call "AutoTransfer".

Our basic approach to the transfer learning problem is to censor an encoder model, such that it learns a representation that is useful for the task while containing minimal information about changes in a nuisance variable (i.e., subject ID). The motivation behind our approach is related to the information bottleneck method [2], though with a key difference. Whereas the information bottleneck and related methods seek to learn a useful and compressed representation from a supervised dataset without any additional information about nuisance variation, we explicitly use additional nuisance labels in order to draw conclusions about the types of variation in the data that should not affect our model's output. Many transfer learning settings will have such nuisance labels readily available, and intuitively, the model should benefit from this additional source of supervision. Our method ranked first place in the subject-transfer task of the NeurIPS BEETL challenge [3].

The key contributions of this paper are three-fold:

- We introduce a framework for subject transfer learning with nuisance-censored representations.
- We derive regularization penalties to enforce censoring via mutual information or divergence measures, and provide concrete estimation algorithms for these penalties using techniques including neural critic functions and analytic divergence estimates.
- We thoroughly evaluate these methods on challenging real-world subject-transfer datasets, showing that these methods improve generalization to unseen subject data.

II. Learning Framework

Consider a dataset $\{(x, y, s)\}_1^N$ consisting of N triples of data $x \in \mathbb{R}^D$, discrete task labels $y \in \{1, \dots, C\}$, and discrete nuisance labels $s \in \{1, \dots, M\}$. Let the generative model for the data distribution be defined as:

$$p_{\text{data}}(x, y, s) = p(s)p(y|s)p(x|y, s). \tag{1}$$

Our transfer learning model consists of a parametric encoder $f_\theta(.) : \mathbb{R}^D \to \mathbb{R}^K$ producing a K-dimensional latent representation $z = f_\theta(x)$, as well as a parametric task classifier $g_\phi(.) : \mathbb{R}^K \to \mathbb{R}^C$. We train this model to approximate the posterior distribution over labels given an input item x: $g_\phi(f_\theta(x)) \approx p_{\text{data}}(y|x)$. For a given choice of parameters θ, our encoder model implies a conditional distribution $q_\theta(z|x)$ on features z given the data x. We can then define other conditional distributions of the features as:

$$q_\theta(z|y, s) = \int p(x|y, s)q_\theta(z|x)dx, \tag{2}$$

$$q_\theta(z|y) = \int p(x|y)q_\theta(z|x)dx, \tag{3}$$

$$p(x|y) = \int p(x|y, s)p(s|y)ds. \tag{4}$$

[1] Work done while NSM was on internship at Mitsubishi Electric Research Labs

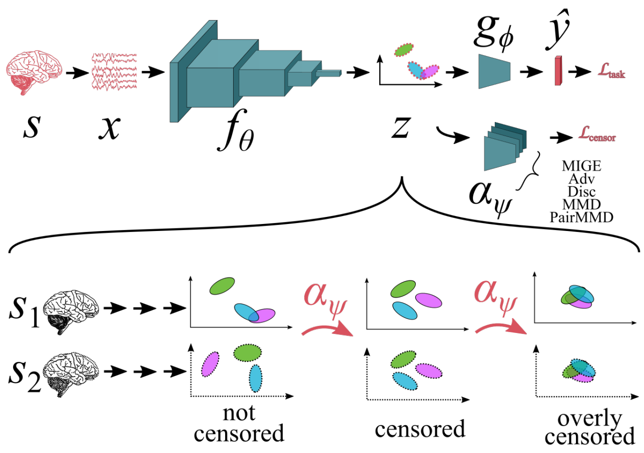

Fig. 1: AutoTransfer pipeline for subject-invariant feature censoring in transfer learning. Top: Model architecture. Subject s produces biosignals data x, which is mapped by encoder f_θ to latent code z. Classifier g_ϕ gives class probabilities $\hat{y}$, resulting in task loss $\mathcal{L}_{\text{task}}$. Censoring models α_ψ compute regularization penalty $\mathcal{L}_{\text{censor}}$ to enforce independence. Bottom: Regularization strategy. Subjects s_1, s_2 are encoded, and their latent feature distributions are regularized.

A. *Empirical Risk Minimization (ERM)*

In the standard ERM framework, we would seek to jointly learn parameters θ, ϕ that minimize the expected classification risk, which we would approximate using an empirical average over our training data. For some specified loss function $\mathcal{L}$ and data distribution $p_{\text{data}}(x, y, s)$, the risk is defined as the expected classification loss:

$$R(\theta, \phi) = \mathop{\mathbb{E}}_{p_{\text{data}}(x,y,s)} \left[\mathcal{L}(g_\phi(f_\theta(x)), y)\right]. \tag{5}$$

B. *Regularized ERM using Censoring*

We consider a regularized form of the ERM framework, with an added penalty to enforce independence between the learned representation z and the nuisance variable s, so that the classifier model g_ϕ can achieve similar performance across different subjects. This regularized learning framework is outlined in Fig. 1.

We consider three overall notions of independence between latent representation and nuisance variable, which we refer to as "censoring modes":

1) In "marginal censoring", we try to make representation z *marginally independent* of nuisance variable s: $z \perp s$.
2) In "conditional censoring", we try to make z *conditionally independent* of s, given the task label y: $z \perp s | y$.
3) In "complementary censoring", we partition the latent space into two halves $z = (z^{(1)}, z^{(2)})$, such that the first half $z^{(1)}$ is marginally independent of s, while maximizing the mutual information between the second half $z^{(2)}$ and s.

Marginal censoring captures the simplest notion of a "subject-independent representation". When the distribution of labels does not depend on the nuisance variable $p(y|s = s_1) = p(y|s = s_2)$, and the nuisance variable s is therefore not useful for the downstream task, this marginal censoring approach will not conflict with the task objective. However, there may naturally exist some correlation between y and s (i.e., subjects may perform the task differently); thus a representation z that is trained to be useful for predicting the task labels y may necessarily also be informative of s. Conditional censoring accounts for this conflict between the task objective and censoring objective by allowing that z contains some information about s, but no more than the amount already implied by the task label y. Complementary censoring accounts for this conflict by requiring that one part of the representation z is independent of the nuisance variable s, while allowing the other part to depend strongly on the nuisance variable.

We capture these three censoring modes in a regularized ERM objective:

$$(\theta^*, \phi^*) = \mathop{\arg\min}_{\theta,\phi} R(\theta, \phi) + \lambda\, \mathcal{L}_{\text{censor}}, \tag{6}$$

where $\mathcal{L}_{\text{censor}}$ is a penalty enforcing the desired independence. Table I details the different forms for this penalty that we consider, and the estimation methods used for each.

III. Estimation Techniques

In this section, we derive several concrete methods for estimating the mutual information and divergence penalties used in the regularized objective functions outlined above. Further details including pseudocode for the marginal, conditional, and complementary versions of each technique can be found starting in Appendix A1 of [8].

A. *Mutual Information Estimation Methods*

We consider two ways to estimate the mutual information penalties required for the censoring objectives given above. First, we use an adversarial nuisance classifier, whose cross entropy loss provides a surrogate for the mutual information between s and z (see Section III-A1). Second, we use Mutual Information Gradient Estimation (MIGE) [4], which uses score function estimators to compute the gradient of mutual information. We consider several kernel-based score function estimators (see Section III-A2).

1) Adversarial Censoring (Adv): We consider minimizing the conditional mutual information between z and s given y using an adversarial nuisance classifier model α with parameters ψ that maps latent representations z to a probability distribution over the nuisance variable s: $\alpha_\psi(.) : \mathbb{R}^K \to \mathbb{R}^M$. Previous research [5], [9] has established the technique of learning subject-invariant representations by training models in the presence of an adversarial subject classifier model.

Recall that for a given choice of encoder parameters θ, we obtain representations $z = f_\theta(x)$ for each data point. For a given choice of adversary parameters ψ and encoder parameters θ, computing the adversary's cross entropy loss $\mathcal{L}_{\text{CE, ADV}}(\theta, \psi) = \mathbb{E}_{p_{\text{data}}(x,y,s)}[-\log \alpha_\psi(s|z)]$ gives an upper bound on the conditional entropy $H(s|z)$. Noting that the mutual information can be decomposed as $I(z; s) = H(s) - H(s|z)$ and that the marginal entropy $H(s)$ is constant

TABLE I: Conceptual forms for regularization penalty $\mathcal{L}_{\text{censor}}$, and concrete estimation methods used for each penalty

Censoring Mode	Desired Effect	Mutual Information form of $\mathcal{L}_{\text{censor}}$	Divergence form of $\mathcal{L}_{\text{censor}}$
Marginal	$z \perp s$	$I(z; s)$	$\mathcal{D}(q_\theta(z)\|\|q_\theta(z\|s))$
Conditional	$z \perp s\|y$	$I(z; s\|y)$	$\mathcal{D}(q_\theta(z\|y)\|\|q_\theta(z\|s, y))$
Complementary	$z^{(1)} \perp s, \max I(z^{(2)}, s)$	$I(z^{(1)}; s) - I(z^{(2)}; s)$	$\mathcal{D}(q_\theta(z^{(1)})\|\|q_\theta(z^{(1)}\|s)) - \mathcal{D}(q_\theta(z^{(2)})\|\|q_\theta(z^{(2)}\|s))$
Estimation Methods	-	MIGE [4], Adversary [5]	MMD/Pairwise MMD [6], BEGAN Disc [7]

with respect to all model parameters, this gives a bound on the mutual information, which can be used as a surrogate objective for minimizing the mutual information:

$$I(z; s) \geq H(s) - \mathcal{L}_{\text{CE, ADV}}(\theta, \psi). \tag{7}$$

This bound will be tight for an adversary whose predicted distribution over subjects is close to the true posterior distribution; thus we can improve the quality of this surrogate objective by using a strong adversary model that is trained to convergence. See Appendix A1 of [8] for further details.

2) Mutual Information Gradient Estimation (MIGE) Censoring: Given the difficulty of estimating mutual information in high dimensions, Wen, Zhou, He, *et al.* [4] provide a method to estimate the *gradient* of mutual information directly. This suffices for cases like ours, in which an objective function containing a mutual information term will be minimized by gradient descent. Appendix Section C of Shi, Sun, and Zhu [10] derives a method for using a score function estimator to approximate the gradient of an entropy term. Appendix Sections A and B of Wen, Zhou, He, *et al.* [4] show how to apply this idea for estimating the gradient of entropy terms to estimating the gradient of mutual information.

a) Score Function Estimation: The score function terms $\nabla_z \log q_\theta(z)$ and $\nabla_z \log q_\theta(z|s)$ required for MIGE penalties can be computed using any score function estimation method available. The original implementation by Wen, Zhou, He, *et al.* [4] used the Spectral Stein Gradient Estimator (SSGE) [10]. We explore other kernel-based score function estimation methods based on the work of Zhou, Shi, and Zhu [11], who frame the problem of score function estimation as a regularized vector regression problem. See Appendix A2d of [8] for further details about the estimators used and how their hyperparameters were set.

B. Divergence Estimation Methods

As outlined in Table I, we also consider regularization penalties based on the divergence between two distributions. For marginal censoring, the definition of conditional probability tells us that the desired independence $z \perp s$ also implies that the distributions $q_\theta(z)$ and $q_\theta(z|s)$ are equivalent, or alternatively that the distributions $q_\theta(z|s_i)$ and $q_\theta(z|s_{j\neq i})$ are equivalent. Analogous divergences can be used for conditional or complementary censoring. We provide three methods for censoring using divergence estimates; the first two are closely related, while the third is quite distinct.

The first two methods rely on a kernel-based estimate of the Maximum Mean Discrepancy (MMD) [6], which provides a numerical measure of distance between two distributions. The MMD between two distributions is 0 when the distributions are equivalent. In Section III-B1, we use MMD as a surrogate for the divergence between $q_\theta(z)$ and $q_\theta(z|s)$, which we refer to as simply the "MMD" censoring approach. In Section III-B2, we use MMD as a surrogate for the divergence between $q_\theta(z|s_i)$ and $q_\theta(z|s_{j\neq i})$, which we refer to as the "Pairwise MMD" censoring approach.

In the third method, we use a neural discriminator model based on Boundary Equilibrium Generative Adversarial Networks (BEGAN) [7]. In the original work, this discriminator provides a surrogate measure of the divergence between real and generated data distributions. In our work, we use the discriminator to provide a measure of the divergence between $q_\theta(s)$ and $q_\theta(z|s)$, which allows us to reduce the dependence of z on s. See Section III-B3 for further details.

1) MMD Censoring: The MMD [6] provides a desirable measure of divergence between distributions because it makes no assumptions about the parametric form of the distributions being measured, and because it can be approximated efficiently with a kernel estimator, given a batch of samples from each distribution. The MMD is an integral probability metric, describing the divergence between two distributions $p(\cdot)$ and $q(\cdot)$ as the difference between the expected value of a test function $f \in \mathcal{F}$ under each distribution, for some worst-case f from a class of functions $\mathcal{F}$:

$$\mathsf{MMD}(\mathcal{F}, p, q) = \sup_{f\in\mathcal{F}} \left(\mathop{\mathbb{E}}_{x\sim p(x)} [f(x)] - \mathop{\mathbb{E}}_{y\sim p(y)} [f(y)] \right). \tag{8}$$

Gretton, Borgwardt, Rasch, *et al.* [6] derive a kernel estimate of the MMD using a radial basis function kernel (see details in Appendix B1 of [8]). We use their empirical estimate, with a kernel length scale set by the median heuristic [12] as we do for a subset of MIGE experiments.

2) Pairwise MMD (PairMMD) Censoring: Computing the MMD between $q_\theta(z)$ and $q_\theta(z|s)$ provides us a quantitative measure of the dependence between z and s, and by minimizing it we can enforce the indepences we desire. We can similarly measure the divergence between $q_\theta(z|s_i)$ and $q_\theta(z|s_{j\neq i})$ to enforce these independences. To compute an overall penalty using this "pairwise" approach, we consider all combinations of $\binom{M}{2}$ distinct values of the nuisance variable, and compute an average over these individual terms. Since computing the full quadratic set results in a potentially large overhead, we consider two approximations by selecting a random subset of terms to compute as below.

First, we consider using a parameter $b \in [0, 1]$ controlling a Bernoulli distribution to select a random subset of all possible pairs of s_i, s_j for $i \neq j$, which we call a "Bernoulli" subset selection. Second, we consider using an integer $d \in \{1, \ldots, M\}$ controlling the number of nuisance values included, and compute a term for all combinations within this subset, which we call a "clique" subset selection.

TABLE II: Censoring hyperparameters explored in AutoTransfer

Censoring Method	Parameter	Values Explored
Adv, MIGE, BEGAN Disc	Regularization Coefficient λ	$1, 0.3, 0.1, 0.03, 0.01$
MMD, PairMMD	Regularization Coefficient λ	$1, 3, 10, 30, 100$
MIGE	Score Function Estimator F_{score}	SSGE [10], MIGE, ν-Method, Tikhonov, Stein [11]
MIGE	Score Function Estimator Regularization γ	$0.01, 0.001, 0.0001$
MIGE	Adaptive Length-scale Method	Median, t-SNE-style

These two selection procedures are described in more detail in Appendix B2 of [8].

3) BEGAN Discriminator Censoring: BEGAN [7] uses an adversarial training scheme to learn a generative model. A generator network G tries to approximately map samples from a unit Gaussian distribution in its latent space to samples from the target data distribution, while a discriminator network D tries to distinguish real and fake data samples, as in a standard GAN setup [13]. This model uses an autoencoder as the discriminator to compute a lower bound on the Wasserstein-1 distance between the distribution of its autoencoder loss on real and generated data. In other words, the discriminator separates the two distributions by learning an autoencoder map that works well only for the "true" data distribution; the generator tries to produce data that matches the "true" data distribution and is well-preserved by this autoencoder map. The training is further stabilized by introducing a trade-off parameter to adaptively scale the magnitude of the discriminator's two loss terms.

The role of the discriminator in this original setup is to provide a surrogate objective so that the generator can bring two distributions (the true data distribution $p_{\text{data}}(x)$, and its own generated distribution $p_G(x)$) closer together. We use their method to provide a signal that allows our encoder model f_θ to approximately minimize the divergence terms.

IV. Experiments

In order to evaluate the proposed regularization approaches described in (6) and Table I, we perform experiments with several challenging real-world datasets. For each dataset, we explore all of the censoring estimation procedures described above. For detailed pseudocode, see Algorithms 1 through 17, detailed in Appendix B2 of [8]. We first search for promising hyperparameter ranges, then evaluate the most promising subset of hyperparameters using k-fold cross-validation and evaluate our AutoTransfer method on the resulting collection of models.

A. Datasets

We use a diverse set of physiological datasets: EEG (rapid serial visual presentation, RSVP [14]; error-related potentials, ErrP [15]), EMG (American Sign Language, ASL [16]), and ECoG (facial recognition, EcogFacesBasic [17]). To standardize the comparison across datasets, all data were preprocessed by z-scoring each channel of each trial. Additional feature engineering could improve absolute performance, though such techniques are orthogonal to the focus of the present study. For further dataset details, see Appendix C of [8].

B. Network Architectures and Training

The neural network architectures for our feature extractor network f_θ is based on EEGNet [18]. The classifier network g_ϕ consists of a single linear layer with softmax activation.

Models are trained to maximize balanced accuracy by using weighted cross entropy; for class k with N_k examples in the training set, the unnormalized weight $\tilde{w}_k$ is set as the inverse of the class proportion $\tilde{w}_k = \sum_i N_i / N_k$, and then the sum of weights is normalized to one, $w_k = \tilde{w}_k / \sum_i \tilde{w}_i$.

In each training fold, we designate one held-out subject for validation and one for test. The validation subject is used for model selection, as well as for early stopping. Models are trained for a maximum 500 epochs using the AdamW optimizer [19]. We begin with a learning rate of $\alpha_1 = 10^{-3}$, and decay the learning rate by the inverse square-root of the epoch number, such that at epoch t, we use a learning rate of $\alpha_t = \alpha_1 / \sqrt{t}$. The epoch of minimum validation loss is then evaluated on the test subject.

C. Hyperparameter Tuning

For each dataset under consideration, we tune several key hyperparameters for each of our estimation methods. We use the same data split for all settings explored; one subject is kept for validation, and another is kept for testing. We then select the best 3 settings for each of the five methods discussed above according to balanced validation accuracy. For each method, we vary the Lagrange multiplier coefficient λ. For the MIGE censoring, we also vary the score function estimator between the default SSGE, and three alternative kernel-based estimators discussed in Section III-A2a. For these three alternative score function estimators, we vary their own internal regularization parameter γ, and vary the method of setting the kernel length scale as discussed in Appendix A2e of [8]. Table II summarizes the range of hyperparameters explored for each method. Note that this hyperparameter search for both finite descrete values and contineous values can be fully automated by, e.g., Bayesian optimization in an AutoML framework [20].

D. Cross-Validation

We select the best 3 combinations of hyperparameter settings according to validation accuracy for each method for further examination by k-fold cross-validation. Specifically, for a dataset with M subjects, the cross-subject validation gives us a collection of M test accuracies, which we can visualize as a distribution of model performance. We then apply our AutoTransfer procedure; for a given dataset, we select the method whose 25th-percentile validation accuracy is highest. We select methods based on lower quartile performance as a way to avoid overfitting to the validation subjects.

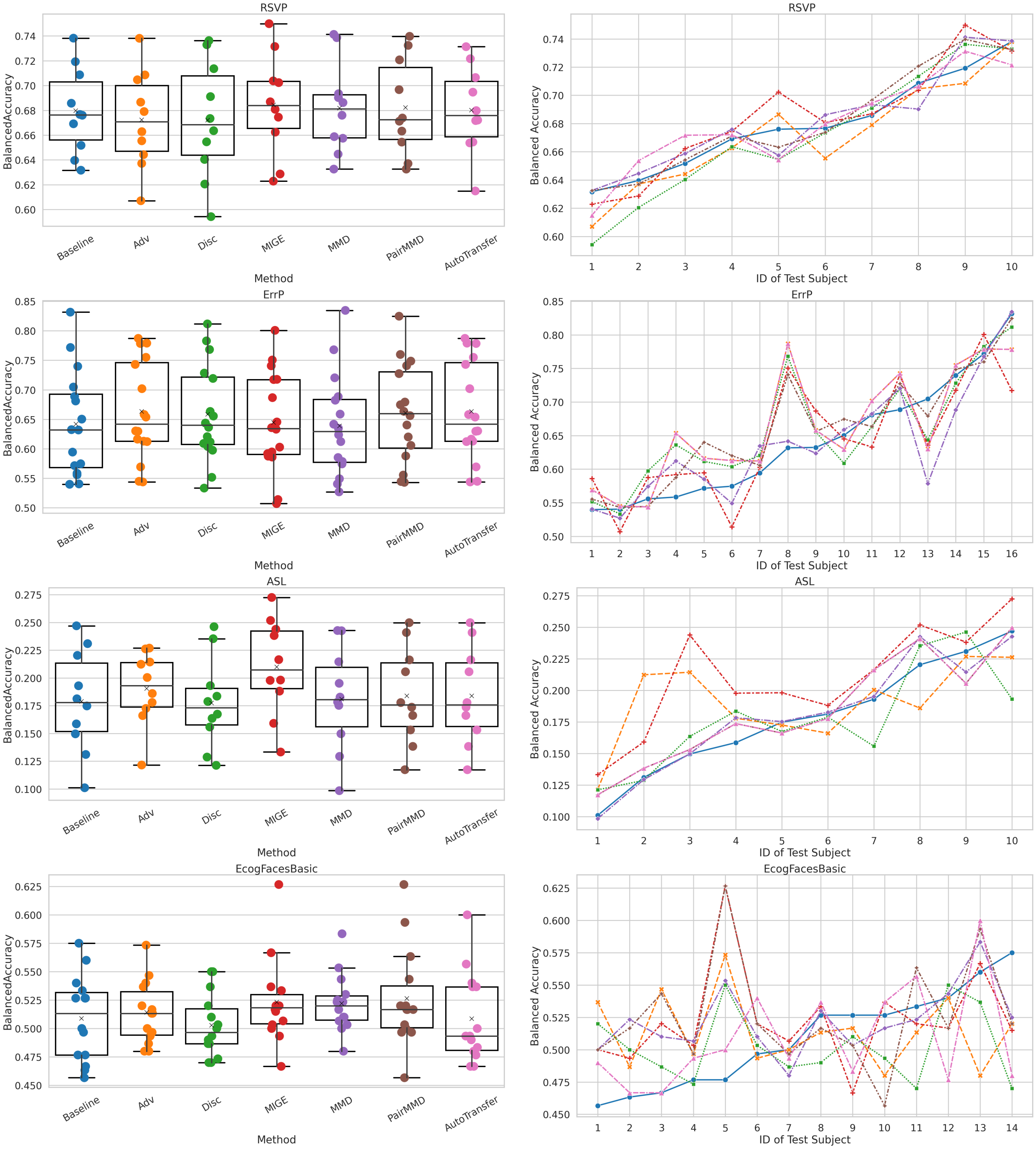

Fig. 2: Subject transfer balanced accuracy. Left: Test score from each fold, black ‘x’ indicates mean. Right: Accuracy vs test subject, sorted by baseline performance. Same colors for left and right. Censoring improves subject transfer, especially in subjects whose transfer performance is initially low. The best single censoring method is dataset dependent, while AutoTransfer consistently performs well and often matches the best method. See Sec. IV-A for dataset information.

Fig. 2 shows the results of these cross-validation experiments. On the left, each dot represents the transfer performance when a certain subject is used as the test set. The most striking observation is the wide variation in subject transfer performance; this is the heart of the difficulty in subject transfer learning. For each dataset, we observe that at least one of the estimation methods provides an improvement to the interquartile range, despite the presence of outlier subjects. Although the most popular censoring method based on adversarial training works well for most datasets, other censoring approach such as MIGE censoring can outperform it for some datasets. This suggests us that exploring different censoring methods is of great importance depending on datasets. On the right, each x-axis position represents a single test subject, and they are sorted according to their baseline transfer performance. Here we can see that our regularization penalties offer a strong benefit for some subjects, especially those whose baseline transfer accuracy is relatively worse.

V. Discussion

We addressed the problem of subject transfer learning for biosignals datasets by using a regularized learning framework to enforce one of several possible notions of independence, which we refer to as censoring modes. We derived estimation algorithms for each of the proposed regularization penalties. We evaluated these estimation algorithms on a variety of challenging real-world datasets including EEG, EMG, and ECoG, and found that these methods can offer significant improvement, especially for subjects originally near the lower quartile of transfer performance. Finally, we provided an automated end-to-end procedure for exploring and selecting a censoring method on a new dataset, which we call AutoTransfer. In our cross-validation experiments, AutoTransfer consistently offers an improvement over baseline transfer performance, though it may be below the maximal single-method performance due to the inherent inter-subject variability in these tasks. Note that we considered the case of discrete nuisance variables (subject ID), though most techniques are readily applicable to the case of continuous-valued nuisance variables.

Future Work: Our proposed approach is designed to improve subject transfer performance without making use of test-time adaptation or information about the statistics of the transfer subject's data. In order to construct a holistic transfer learning system, future work may therefore combine our training-time regularization with other test-time adaptation strategies such as few-shot learning, parameter prediction [21]–[23], data augmentation [24], [25], and self-supervision [26]–[29].

Furthermore, our methods are compatible with a number of other standard machine learning techniques that may improve absolute model performance. For example, feature engineering and hand-tuned feature extraction provide a way to strictly enforce prior knowledge about signal characteristics; when collecting biosignals datasets, this might include information such as sensor artifacts and signal frequency ranges. Since we propose a large family of estimation algorithms, our work may naturally benefit from model ensembling techniques, though the challenge in this context is to handle the overfitting issues that are inherent to these challenging subject transfer datasets. This large family of estimation algorithms also comes with a large set of model hyperparameters to consider; it is likely that further hyperparameter search may offer additional improvement in the final model performance. Finally, we note that it may be possible to use a decoder model with a reconstruction loss term as part of a method for enforcing the regularization strategies.

References

[1] Y.-P. Lin and T.-P. Jung, "Improving EEG-based emotion classification using conditional transfer learning," *Frontiers in human neuroscience*, vol. 11, p. 334, 2017.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[3] *NeurIPS 2021 BEETL Competition: Benchmarks for EEG Transfer Learning*, https://beetl.ai.

[4] L. Wen, Y. Zhou, L. He, M. Zhou, and Z. Xu, "Mutual information gradient estimation for representation learning," *arXiv preprint arXiv:2005.01123*, 2020.

[5] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Learning invariant representations from EEG via adversarial inference," *IEEE access*, vol. 8, pp. 27 074–27 085, 2020. DOI: 10.1109/ACCESS.2020.2971600.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[7] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[8] N. Smedemark-Margulies, Y. Wang, T. Koike-Akino, and D. Erdogmus, "AutoTransfer: Subject transfer learning with censored representations on biosignals data," *arXiv preprint arXiv:2112.09796*, 2021.

[9] M. Han, O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Disentangled adversarial transfer learning for physiological biosignals," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 422–425.

[10] J. Shi, S. Sun, and J. Zhu, "A spectral approach to gradient estimation for implicit distributions," in *International Conference on Machine Learning*, PMLR, 2018, pp. 4644–4653.

[11] Y. Zhou, J. Shi, and J. Zhu, "Nonparametric score estimators," in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 513–11 522.

[12] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for RKHS embeddings of probability distributions.," in *NIPS*, vol. 22, 2009, pp. 1750–1758.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] *RSVP EEG dataset*, https://repository.library.northeastern.edu/collections/neu:gm80jm78x.

[15] P. Margaux, M. Emmanuel, D. Sébastien, B. Olivier, and M. Jérémie, "Objective and subjective evaluation of online error correction during P300-based spelling,"

Advances in Human-Computer Interaction, vol. 2012, 2012.

[16] S. Y. Günay, M. Yarossi, D. H. Brooks, E. Tunik, and D. Erdoğmuş, "Transfer learning using low-dimensional subspaces for EMG-based classification of hand posture," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, IEEE, 2019, pp. 1097–1100.

[17] K. J. Miller, "A library of human electrocorticographic data and analyses," *Nature human behaviour*, vol. 3, no. 11, pp. 1225–1235, 2019.

[18] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[21] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in neural information processing systems*, 2016, pp. 523–531.

[22] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," *Advances in Neural Information Processing Systems*, vol. 32, pp. 7959–7970, 2019.

[23] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[24] K. Sohn, D. Berthelot, C.-L. Li, *et al.*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[25] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

[28] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[29] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[30] Y. Li and R. E. Turner, "Gradient estimators for implicit models," *arXiv preprint arXiv:1705.07107*, 2017.

[31] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*. Springer Science & Business Media, 1996, vol. 375.

[32] Y. Song and D. P. Kingma, "How to train your energy-based models," *arXiv preprint arXiv:2101.03288*, 2021.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.