Convolutional Monge Mapping between EEG Datasets to Support Independent Component Labeling

Austin Meek
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware
ajmeek@udel.edu

Carlos H. Mendoza-Cardenas
Twitch Interactive Inc.
San Francisco, California

Austin J. Brockmeier
Department of Electrical and Computer Engineering
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware

Abstract

EEG recordings contain rich information about neural activity and are used in the diagnosis and monitoring of multiple neuropathologies, including epilepsy and psychosis, but are subject to artifacts and noise. While EEG analysis can benefit from automating artifact removal through independent component analysis and automatic labeling of independent components (ICs), differences in recording equipment and context (the presence of noise from electrical wiring and other devices) may impact the performance of IC classifiers. Here we investigate how these differences can be minimized by appropriate spectral normalization through filtering using Convolutional Monge Mapping Normalization (CMMN), which was previously shown to improve deep neural network approaches for sleep staging. We propose a novel extension of the CMMN method with two alternative approaches to computing the source reference spectrum the target signals are mapped to: (1) channel-averaged and l_1-normalized barycenter, and (2) a subject-to-subject mapping that finds the source subject with the closest spectrum to the target subject. Notably, our extension yields space-time separable filters that can be used to map between datasets with different numbers of EEG channels. We apply these filters in an IC classification task, and show significant improvement in recognizing brain versus non-brain ICs.

In the context of EEG, differences in set-up (electrodes, amplifiers, analog and digital filters, power line noise) can create stark changes in the spectral content of recordings, which can deteriorate the performance of models, especially when trained on limited data. A solution to this problem is the Convolutional Monge Mapping Normalization (CMMN) approach, a method for spectral normalization based on optimal transport [7]. The CMMN method optimizes a unique linear filter for each subject and channel. After filtering, the spectra for each channel across all subjects are aligned to a common spectrum, which is the barycenter of the training subjects' spectra for that channel. Originally, CMMN was applied to sleep staging with deep neural networks. In this work, we extend it to the case of automatically labeling independent components, such as the popular ICLabel classifier [14], which can support automatic EEG artifact removal and denoising. In comparison to the ICLabel classifier, which uses a neural network model with scalp map and spectral features (including the autocorrelation sequence), and was trained on an extensive dataset [15], we consider custom

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time-Series for Health.

classifiers trained on a much smaller dataset collected in the US [13] using only time-series features extracted from each independent component [11]—that is, without any spatial information. We then test on an another independent dataset [8], collected in Europe. These two datasets were used in a prior work on IC classification [5], and are now both available on OpenNeuro [12, 9]. Because these datasets have expert-labeled ICs, we avoid filtering each channel with a different filter, which would change each IC differently; instead, we find a single CMMN filter for each subject that is applied to all channels. This channel-averaged CMMN filter is based on matching the channel-averaged spectrum of each test subject to either the barycenter of the channel-averaged spectra from the training set or the nearest spectrum in the training set in terms of the optimal transport distance. With a shared single temporal filter across channels, the filtering can be applied before or after the spatial filtering achieved by ICA's demixing matrix. Crucially, the use of the channel-averaging enables the application of CMMN to target domains with a differing number of channels. Here, the source (train) and target (test) datasets have 134–235 channels and 64 channels, respectively. This extension of the CMMN methodology achieves domain adaptation between two datasets with the pre-trained classifiers such that their performance (the subject averaged F1 score for the brain class is 0.91) exceeds the performance of ICLabel (0.88).

1 Methodology

The CMMN methodology enables domain adaptation between different neural signals [7] by mapping the spectra of the target subjects to the barycenter of the source subjects via a linear filter that achieves the optimal transport with respect to an ℓ_2-metric, assuming the signals are zero-mean stationary Gaussian discrete-time processes (see Appendix A for more details). We extend this methodology in two key ways: we propose to use channel-averaged PSDs yielding a single filter per subject and a subject-to-subject mapping scheme. A key benefit of using a single filter per subject is that it enables adapting between EEG datasets with a variable number of channels. We also propose to use ℓ_1-normalization of spectra for invariance to signal scale, which may occur due to impedance differences due to electrodes or scalp contact or differences in the electronic systems (amplifiers and filtering) or digital filtering.

We assume access to data (multichannel EEG) from I subjects in the source domain and one or more target subjects. We apply CMMN to transform the target EEGs to the source domain and apply classifiers trained on the source domain to the transformed target data. Let $x_c[n]$ denote the c-th channel of the target signal, $c \in \{1, \ldots, C^{\mathrm{T}}\}$. The CMMN transformation is a linear filtering of $x_c[n]$, with output $y_c[n] = (h * x_c)[n], \quad c \in \{1, \ldots, C^{\mathrm{T}}\}$, where $h[n]$ is the normalizing CMMN filter, and $y_c[n]$ is the transformed signal that will have a source-like spectrum (the channel-averaged PSDs will match). The estimation of the normalizing CMMN filter is a three-step process that uses a reference spectrum (computed in the second step) that is either a **Barycenter** from the source or the normalized PSD of a source subject via the subject-to-subject (**Subj-to-subj**) assignment: 1) power spectral density calculation, 2a) barycenter calculation, 2b) minimization of the Hellinger distance between target signals and source signals for the **Subj-to-subj** mapping scheme, and 3) calculation of the normalizing filters.

Step 1 PSD calculation. As in [7], we use the Welch periodogram method [17], which takes averages of the squared FFT from possibly overlapping windows, to calculate PSDs for each subject, as implemented by the SciPy library [16]. For real-valued signals, the spectra will be conjugate symmetric for positive and negative frequencies. In this case, $\boldsymbol{p} \in \mathbb{R}^P_{\geq 0}$ denotes the PSD for P non-negative frequencies (for zero-mean signals the DC component should be zero but is retained). For K equal-sized windows of length M, $\{\boldsymbol{x}_k\}_{k=1}^K = \mathcal{X} \subset \mathbb{R}^M$ with windowing function $\boldsymbol{w}$, the power spectral density estimate is $\boldsymbol{p} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} |\mathrm{RFFT}(\boldsymbol{w} \odot \boldsymbol{x})|^{\odot 2} \in \mathbb{R}^P_{\geq 0}$, where RFFT denotes the FFT operation for real-valued signals and returns the density estimate only for non-negative frequencies, $P = \texttt{nfft}/2 + 1$ is the number of non-negative frequencies, and $N = \texttt{nfft}$ is the length of the windows after zero-padding, assumed to be even. For a subject with C channels, the matrix of PSDs is $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_C]^\top \in \mathbb{R}^{C \times P}$. Unlike the original CMMN, which used different filters for each channel, we propose to compute a channel-averaged PSD as $\bar{\boldsymbol{p}} = \frac{1}{C_i} \sum_{c=1}^{C} \boldsymbol{p}_c$.[1]

[1] One possible concern is whether channel-averaging is a sufficient characterization of multi-channel EEG given the spectral differences in signals across the scalp. Of course, if the location of the electrodes on the

Step 2a Barycenter calculation. We compute a barycenter for the source subjects only, this serves as the template for mapping the target subjects. For I source subjects with channel-averaged PSDs $\bar{\boldsymbol{p}}_1^{\mathrm{S}}, \ldots, \bar{\boldsymbol{p}}_I^{\mathrm{S}}$, the barycenter, $\bar{\boldsymbol{p}}_{\mathrm{S}}$, and the ℓ_1-normalized barycenter, $\tilde{\boldsymbol{p}}_{\mathrm{S}}$, are defined as

$$\bar{\boldsymbol{p}}_{\mathrm{S}} = \frac{1}{I}\sum_{i=1}^{I} \bar{\boldsymbol{p}}_i^{\mathrm{S}}, \quad \tilde{\boldsymbol{p}}_{\mathrm{S}} = \frac{1}{I}\sum_{i=1}^{I} \frac{\bar{\boldsymbol{p}}_i^{\mathrm{S}}}{\|\bar{\boldsymbol{p}}_i^{\mathrm{S}}\|_1} = \frac{1}{I}\sum_{i=1}^{I} \tilde{\boldsymbol{p}}_i^{\mathrm{S}}, \quad \tilde{\boldsymbol{p}}_i^{\mathrm{S}} = \frac{\bar{\boldsymbol{p}}_i^{\mathrm{S}}}{\|\bar{\boldsymbol{p}}_i^{\mathrm{S}}\|_1}, \tag{1}$$

The l_1 normalization ensures that each subject contributes equally to the average. As the units of PSD are squared, outliers without normalization can influence the average.

Step 2b: Subject-to-Subject Mapping. As an alternative to the barycenter scheme, we map a target subject to the nearest source subject. For a target subject with an ℓ_1-normalized PSD $\tilde{\boldsymbol{p}}^{\mathrm{T}} = \frac{\bar{\boldsymbol{p}}^{\mathrm{T}}}{\|\bar{\boldsymbol{p}}^{\mathrm{T}}\|_1}$, we find the index i^* of the source subject that minimizes the Hellinger distance, which is applicable since the ℓ_1-normalized PSD is a probability mass function (PMF)—non-negative and sums to 1. The channel-averaged PSD of the best-matched source is then $\bar{\boldsymbol{p}}_*^{\mathrm{S}} = \bar{\boldsymbol{p}}_{i^*}^{\mathrm{S}}$, where

$$i^* = \underset{i\in\{1,\ldots,I\}}{\arg\min}\, d_{\mathrm{He}}(\tilde{\boldsymbol{p}}_i^{\mathrm{S}}, \tilde{\boldsymbol{p}}^{\mathrm{T}}) = \frac{1}{\sqrt{2}}\left\|(\tilde{\boldsymbol{p}}_i^{\mathrm{S}})^{\odot\frac{1}{2}} - (\tilde{\boldsymbol{p}}^{\mathrm{T}})^{\odot\frac{1}{2}}\right\|_2 = \sqrt{\frac{1}{2}\sum_{n=0}^{N-1}\left(\sqrt{\tilde{p}_i^{\mathrm{S}}[n]} - \sqrt{\tilde{p}^{\mathrm{T}}[n]}\right)^2}, \tag{2}$$

where d_{He} is the Hellinger distance. Assuming that $\tilde{\boldsymbol{p}}_i^{\mathrm{S}}, \tilde{\boldsymbol{p}}^{\mathrm{T}}$ are the eigenvalues of the trace normalized circulant matrices $\boldsymbol{\rho}_i^{\mathrm{S}} = \frac{\boldsymbol{\Sigma}_i^{\mathrm{S}}}{\mathrm{tr}(\boldsymbol{\Sigma}_i^{\mathrm{S}})}$ and $\boldsymbol{\rho}^{\mathrm{T}} = \frac{\boldsymbol{\Sigma}^{\mathrm{T}}}{\mathrm{tr}(\boldsymbol{\Sigma}^{\mathrm{T}})}$, where $\boldsymbol{\Sigma}_i^{\mathrm{S}}$ and $\boldsymbol{\Sigma}^{\mathrm{T}}$ are the covariance matrices of the discrete-time zero-mean stationary Gaussian random processes, then the Wasserstein-2 distance between the two variance-normalized processes is

$$W_2(\mathcal{N}(\mathbf{0}, \boldsymbol{\rho}_i^{\mathrm{S}}), \mathcal{N}(\mathbf{0}, \boldsymbol{\rho}^{\mathrm{T}})) = \sqrt{2} d_{\mathrm{He}}(\tilde{\boldsymbol{p}}_i^{\mathrm{S}}, \tilde{\boldsymbol{p}}^{\mathrm{T}}). \tag{3}$$

The use of Hellinger distance compared to Bures-Wasserstein distance means that signals are compared in terms of their spectral shape, ignoring difference in their variance.

Step 3 Normalizing filter. We define a real-valued, zero-phase linear filter with impulse response $h[n]$ and frequency response $\boldsymbol{H}$ given by the square root of the ratio of the channel-averaged PSDs [4]:

$$\boldsymbol{h} = \mathrm{IRFFT}_M(\boldsymbol{H}), \quad \boldsymbol{H} = \bar{\boldsymbol{p}}^{\mathrm{T}\odot-\frac{1}{2}} \odot \bar{\boldsymbol{p}}^{\mathrm{S}\odot\frac{1}{2}}, \quad H[n] = \frac{\sqrt{\bar{p}^{\mathrm{S}}[n]}}{\sqrt{\bar{p}^{\mathrm{T}}[n]}}, n \in \{0, \ldots, P-1\}, \tag{4}$$

where IRFFT_M is the inverse FFT for real-valued signals that operates on the non-negative frequencies. As discussed in Appendix A, by equalizing the channel-average PSD, this filter solves the optimal transport problem between zero-mean Gaussian distributions with circulant covariance matrices that commute and are diagonalized by the discrete-time Fourier transform. The only difference between the barycenter and subject-to-subject formulations is the choice of Gaussian distribution for the source. That is, the source PSD $\bar{\boldsymbol{p}}^{\mathrm{S}}$ can be either the barycenter of the source dataset $\bar{\boldsymbol{p}}_{\mathrm{S}}$ as in (1), the ℓ_1-normalized barycenter $\tilde{\boldsymbol{p}}_{\mathrm{S}}$ as in (1), or the closest matching source signal $\bar{\boldsymbol{p}}_*^{\mathrm{S}}$ as in (2).

1.1 Application of CMMN to pre-trained IC classifiers

While the proposed methodology is general and could be applied to normalize any target dataset for classifiers pre-trained (without CMMN) on a source dataset, we focus on classifiers for automatically labeling independent components of EEG into brain and other classes [18, 5, 14]. In particular, we build on recent work for labeling ICs without spatial information [11], applying the CMMN approach to random forest classifiers pre-trained on two different feature vectors: the PSD [2] and autocorrelation sequence of the independent components, as in the MNE-ICALabel implementation [10], a Python port of ICLabel [14].

scalp is vastly different such that the spectral content is not comparable, then channel-averaged PSD may not be meaningful, but for EEG montages with whole scalp this should not be an issue.

[2] In contrast to the original ICLabel implementation, we use a min-max normalization of the log-scale PSD instead of dividing by its max-absolute value because it is more robust to the scale of the EEG signals and provided a better classification performance.

2 Experiments and Results

We use two datasets for this domain adaptation study, the *Imagined Emotion* (*Emotion* for short) dataset from [12] as the source dataset and the *Cue* dataset from [8] as the target dataset. Both datasets include manually labeled independent components (ICs). The *Emotion* dataset has data from 32 subjects (13 male and 19 female, with an age mean and standard deviation of 25.5 $\pm$ 5 years). The EEG data has 180-232 channels and sampled at 256 Hz with durations ranging from 54 to 136 minutes. The dataset includes 935 expert-labeled ICs in three categories: brain (570), muscle (306), and eye (59). We used 27 subjects for training and 7 for testing. The training set contains 5,786 total ICs. The *Cue* dataset has data from 12 subjects (10 male, 2 female, with an age range of 21 to 25 years), using 64 channels during data collection. This dataset's recordings were 56 to 66 minutes long and were collected at 500 Hz, which we down-sampled to 256 Hz before performing domain adaptation. Expert annotations are available for 389 ICs: brain (261), muscle (102), eye (22), and heart (4). This dataset serves as the target for cross-dataset evaluation, testing generalization of the classifiers trained on the *Emotion* training data. Note that in comparison to the experiments in the original CMMN paper [7], our data contains different numbers of channels (134–235 vs. 64). Therefore, the use of channel-averaged PSDs for CMMN is essential. For further details on both datasets, please refer to [12, 8].

We show that the normalizing filters learned display intuitive mapping behaviors. The *Emotion* dataset [12] was collected in the US, and displays a characteristic 60 Hz line noise as seen in Fig. 2 in the appendix. The *Cue* dataset [8] was collected in Europe, and displays a characteristic 50 Hz line noise as seen in Fig. 3 in the appendix. When mapping from the Europe dataset to the US dataset. We can see in Fig. 1 that the filters learned to minimize the 50 Hz spike and introduce a spike at 60 Hz.

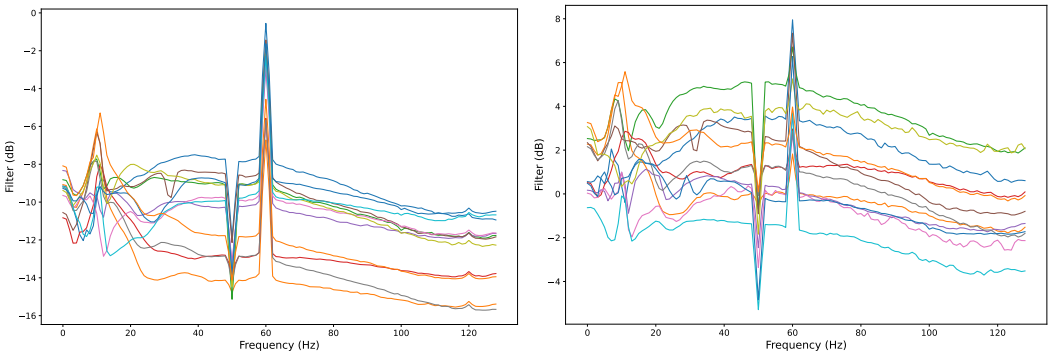

Figure 1: Frequency response of filters learned for mapping the *Cue* dataset to the *Emotion* dataset. (Left) **Barycenter** mapping scheme with a ℓ_1-normalized barycenter. Notice that the noise at 50 Hz is reduced and the noise at 60 Hz is relatively amplified. Overall, the filter attenuates since the normalized barycenter has lower magnitudes.(Right) **Subj-to-subj** mapping scheme. Even though the mapping scheme is different, the line noises are visibly still being 'swapped'.

We trained multi-class classifiers using both PSD and Autocorrelation features on the subset of 27 subjects from *Emotion* dataset, with hyperparameters selected by leave-one-subject-out (LOO) cross-validation. The classifier is trained on segments of length l_{train} (a hyperparameter), and validated (using the F1 score on the brain class) on segments of length $l_{\text{val}} \in \{5, 50\}$ minutes, with $l_{\text{train}} \leq l_{\text{val}}$. Given the best LOO hyperparameters, a new model is trained on the entire training set of 27 subjects. Here we report test results on segments of length equal to the validation segment length, from a disjoint test set of 7 source *Emotion* subjects and all the target subjects in the *Cue* dataset. We benchmark our method against ICLabel [14]. As a model trained on an extensive dataset of expert- and crowd-labeled ICs, ICLabel serves as a strong baseline that can be applied to both the source and target datasets without our proposed CMMN. We did not apply ICLabel to the CMMN normalized signals, as it was trained over a wide variety of datasets and should be already invariant. The source code for our implementation and experiments is publicly available at https://github.com/cniel-ud/ICWaves.

The performance on held-out subjects of the source *Emotion* dataset is shown in Table 1. PSD/Autocorr has the best performance at both time lengths.

Table 1: F1 score of brain-labeled independent components of classifiers within domain on held-out subjects.

Classifier	5 minutes	50 minutes
PSD/Autocorr.	**0.93**±0.05	**0.96**±0.05
ICLabel	0.88±0.05	0.89±0.07

Table 2: F1 score of brain-labeled independent components for different test segment lengths and CMMN schemes. Best performance for a classifier at a segment length is underlined and best classifier for a segment length is bolded. The last column contains the p-value for a one-sided Wilcoxon rank-sign test between no filtering and CMMN filtered (normalized barycenter for PSD/Autocorr.).

Classifier	Len.	No filtering	Barycenter	ℓ_1-norm. Bary.	Subj-to-subj	p-value
PSD/Autocorr.	5	0.77 ± 0.09	0.78±0.12	$\underline{0.84} \pm 0.07$	0.79 ± 0.17	0.0046
ICLabel	5	$\mathbf{0.88} \pm 0.06$				
PSD/Autocorr.	50	0.83 ± 0.09	0.86±0.09	$\underline{0.86} \pm 0.08$	0.85 ± 0.17	0.1696
ICLabel	50	$\mathbf{0.89} \pm 0.05$				

As seen in Table 2, our results show that, when compared to no filtering, using the appropriate CMMN scheme improves the performance of *Emotion*-trained classifiers on the *Cue* dataset. The subject-to-subject mapping scheme improves the brain class F1-score from 0.77 to 0.79 at 5 minutes and from 0.83 to 0.85 at 50 minutes. With the ℓ_1-normalized barycenter, the PSD/Autocorr classifiers' F1-score improved from 0.77 to 0.84 at 5 minutes, and from 0.83 to 0.86 at 50 minutes. We use a Wilcoxon rank sign test ($n = 12$, one-sided alternative) to test whether the filtering (normalized barycenter) has a significant improvement in F1 score for the brain-labeled ICs, and find significant improvement at a level of 0.05 for the 5 minute length.

3 Conclusion

In this work we made key extensions to the recently introduced CMMN methodology [7] to enable domain adaptation between EEG datasets for independent component classification for artifact removal. We introduce filters defined by channel-averaged PSDs along with a subject-to-subject mapping scheme, and show that our CMMN method results in improvements when classifying brain versus non-brain independent components, achieving domain adaptation between two different datasets with significant differences. Our method advances the work on EEG artifact-removal across distinct datasets, an area that is crucial for increasing the clinical utility of EEG recordings.

Acknowledgments and Disclosure of Funding

This work was supported in part by the University of Delaware General University Research fund. This research was supported in part through the use of DARWIN computing system: DARWIN – A Resource for Computational and Data-intensive Research at the University of Delaware and in the Delaware Region, which is supported by NSF under Grant Number: 1919839, Rudolf Eigenmann, Benjamin E. Bagozzi, Arthi Jayaraman, William Totten, and Cathy H. Wu, University of Delaware, 2021, URL: https://udspace.udel.edu/handle/19716/29071.

The authors thank Laura Frølich, Tobias Andersen, Klaus Gramann, and Morten Mørup for graciously providing access to the *Cue* dataset used for our experiments, and Prof. Gramann for allowing us to post it on OpenNeuro.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

[3] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[4] Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation, December 2020.

[5] Laura Frølich, Tobias S. Andersen, and Morten Mørup. Classification of independent components of EEG into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015. ISSN 1469-8986. doi: 10.1111/psyp.12290.

[6] Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[7] Theo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution Monge Mapping Normalization for learning on sleep data. *Advances in Neural Information Processing Systems*, 36:10457–10476, 2023.

[8] Klaus Gramann, Thomas Töllner, and Hermann J. Müller. Dimension-based attention modulates early visual processing. *Psychophysiology*, 47(5):968–978, 2010. ISSN 1469-8986. doi: 10.1111/j.1469-8986.2010.00998.x.

[9] Hermann J. Müller Klaus Gramann, Thomas Töllner. "dimension-based attention modulates early visual processing", 2025.

[10] Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. Mne-icalabel: Automatically annotating ica components with iclabel in python. *Journal of Open Source Software*, 7(76):4484, August 2022. ISSN 2475-9066. doi: 10.21105/joss.04484. URL `http://dx.doi.org/10.21105/joss.04484`.

[11] Carlos H. Mendoza-Cardenas, Austin Meek, and Austin J. Brockmeier. Labeling EEG Components with a Bag of Waveforms from Learned Dictionaries. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*, 2023.

[12] Julie Onton and Scott Makeig. *Imagined Emotion Study*. OpenNeuro, 2022. doi: 10.18112/openneuro.ds003004.v1.1.1.

[13] Julie A Onton and Scott Makeig. High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in human neuroscience*, 3:560, 2009.

[14] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage*, 198: 181–197, 2019. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.05.026.

[15] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. The iclabel dataset of electroencephalographic (eeg) independent component (ic) features. *Data in brief*, 25:104101, 2019.

[16] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[17] P. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, June 1967. ISSN 1558-2582. doi: 10.1109/TAU.1967.1161901.

[18] Irene Winkler, Stefan Haufe, and Michael Tangermann. Automatic classification of artifactual ica-components for artifact removal in eeg signals. *Behavioral and Brain Functions*, 7(1):30, 2011.

A Background on Convolutional Monge Mapping Normalization

The Wasserstein-2 distance between Gaussian distributions has a closed form [3, 6] known as the Bures-Wasserstein distance or Fréchet distance, expressed in terms of the means $\boldsymbol{m}^{\mathrm{S}}, \boldsymbol{m}^{\mathrm{T}}$ and covariance matrices $\boldsymbol{\Sigma}^{\mathrm{S}}, \boldsymbol{\Sigma}^{\mathrm{T}}$, as

$$\begin{aligned} &W_2(\mathcal{N}(\boldsymbol{m}^{\mathrm{S}}, \boldsymbol{\Sigma}^{\mathrm{S}}), \mathcal{N}(\boldsymbol{m}^{\mathrm{T}}, \boldsymbol{\Sigma}^{\mathrm{T}})) \\ &\quad = \sqrt{\|\boldsymbol{m}^{\mathrm{S}} - \boldsymbol{m}^{\mathrm{T}}\|_2^2 + \operatorname{tr}(\boldsymbol{\Sigma}^{\mathrm{S}} + \boldsymbol{\Sigma}^{\mathrm{T}} - (\boldsymbol{\Sigma}^{\mathrm{S}\frac{1}{2}} \boldsymbol{\Sigma}^{\mathrm{T}} \boldsymbol{\Sigma}^{\mathrm{S}\frac{1}{2}})^{\frac{1}{2}})}, \end{aligned} \tag{5}$$

where the second term in the square root is the squared Bures distance. Under the stationarity assumption, a discrete-time zero-mean Gaussian process is completely described by its auto-covariance matrix, which is a symmetric and Toeplitz matrix formed from the auto-correlation sequence $r[k] = \mathbb{E}[x[n]x[n+k]]$. Assuming a sufficiently long truncation of $r[\tau]$ that yields a $N \times N$ circulant matrix that is positive-semidefinite, then the discrete Fourier transform (DFT) matrix $\boldsymbol{F}$ diagonalizes it, $\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^H = \operatorname{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{F}^H$ denotes the Hermitian transpose of $\boldsymbol{F}$. The resulting spectrum $\boldsymbol{\lambda} = \operatorname{diag}(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^H)$ corresponds to an statistical estimate of the power spectral density (PSD). The squared Wasserstein-2 distance between two discrete-time, zero-mean stationary Gaussian random processes with auto-covariance $\boldsymbol{\Sigma}^{\mathrm{S}}$ and $\boldsymbol{\Sigma}^{\mathrm{T}}$, which are circulant, simplifies to

$$W_2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\mathrm{S}}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\mathrm{T}})) = \|\boldsymbol{\lambda}^{\mathrm{S}\odot\frac{1}{2}} - \boldsymbol{\lambda}^{\mathrm{T}\odot\frac{1}{2}}\|_2 = \sqrt{\sum_{n=0}^{N-1} \left(\sqrt{\lambda^{\mathrm{S}}[n]} - \sqrt{\lambda^{\mathrm{T}}[n]}\right)^2}. \tag{6}$$

The Bures-Wasserstein barycenter, which is the Gaussian special case of the Wasserstein-2 barycenter [1, 2], is the covariance matrix $\boldsymbol{\Sigma}_{\mathrm{S}}$ (in the set of positive semidefinite matrices $\mathcal{S}_N$) that minimizes the sum of squared Wasserstein-2 distances to a set of zero-mean Gaussian distributions described by covariance matrices $\{\boldsymbol{\Sigma}_i^{\mathrm{S}}\}_{i=1}^I$:

$$\boldsymbol{\Sigma}_{\mathrm{S}} = \underset{\boldsymbol{\Sigma} \in \mathcal{S}_N}{\arg\min} \sum_{i=1}^{I} \frac{1}{I} W_2^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i^{\mathrm{S}}), \tag{7}$$

Assuming all covariance matrices are circulant and positive semidefinite, it is straightforward to show that $\boldsymbol{\Sigma}_{\mathrm{S}} = \boldsymbol{F}^H \operatorname{diag}(\bar{\boldsymbol{\lambda}}_{\mathrm{S}})\boldsymbol{F}$, where $\bar{\boldsymbol{\lambda}}_{\mathrm{S}} = (\frac{1}{I}\sum_i \boldsymbol{\lambda}_i^{\mathrm{S}\odot\frac{1}{2}})^{\odot 2}$.

Consider a stationary target signal $x[n]$ with PSD $\boldsymbol{\lambda}^{\mathrm{T}} = \mathbb{E}[|\boldsymbol{F}\boldsymbol{x}|^{\odot 2}]$ that is convolved with the filter impulse response $h[n]$ (frequency response $\boldsymbol{H}$), yielding the signal $y[n] = (h * x)[n]$. Then under the stationarity assumption, the PSD of $y[n]$ is

$$\boldsymbol{\lambda}^y = \mathbb{E}[|\boldsymbol{F}\boldsymbol{y}|^{\odot 2}] = \mathbb{E}[|(\boldsymbol{F}\boldsymbol{x}) \odot \boldsymbol{H}|^{\odot 2}] = \mathbb{E}[|\boldsymbol{F}\boldsymbol{x}|^{\odot 2} \odot |\boldsymbol{H}|^{\odot 2}] = |\boldsymbol{H}|^{\odot 2} \odot \boldsymbol{\lambda}^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{H}$ is the discrete Fourier transform of the filter. Assuming $x[n]$ is also zero-mean and a Gaussian process $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\mathrm{T}})$, then $y[n]$ has a circulant auto-covariance matrix $\boldsymbol{\Sigma}^y = \boldsymbol{F}\operatorname{diag}(|\boldsymbol{H}|^{\odot 2} \odot \boldsymbol{\lambda}^{\mathrm{T}})\boldsymbol{F}^H$. The Wasserstein-2 distance between a source process (possibly the barycenter) and the filtered target process is then

$$W_2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\mathrm{S}}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^y)) = \|\boldsymbol{\lambda}^{\mathrm{S}\odot\frac{1}{2}} - |\boldsymbol{H}| \odot \boldsymbol{\lambda}^{\mathrm{T}\odot\frac{1}{2}}\|_2^2. \tag{9}$$

Clearly, the distance is zero for any filter $h[n]$ such that $|\boldsymbol{H}| = \boldsymbol{\lambda}^{\mathrm{S}\odot\frac{1}{2}} \odot \boldsymbol{\lambda}^{\mathrm{T}\odot-\frac{1}{2}} \in \mathbb{R}_{\geq 0}^N$, i.e., where the magnitude of the frequency response is the square-root of the power spectral densities $|H[n]| = \sqrt{\frac{\lambda^{\mathrm{S}}[n]}{\lambda^{\mathrm{T}}[n]}}, n \in \{0, \ldots, N-1\}$.

B Additional Results

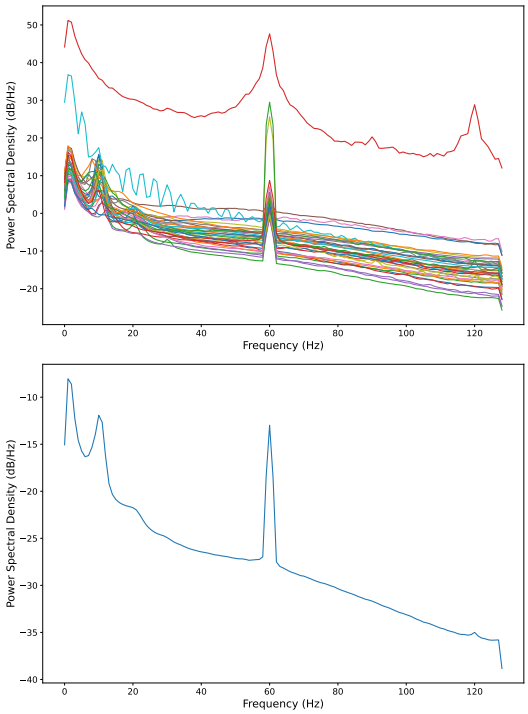

Figure 2: (Top) Channel-averaged power spectral densities for each subject in the *Emotion* dataset. Notice that some subjects are outliers in terms of overall amplitude. (Bottom) This is the ℓ_1-normalized barycenter computed from the source *Emotion* dataset. In the **Barycenter** mapping scheme, all target signals are filtered such that their robust channel-average PSD matches this. Notice the large spike at 60 Hz.

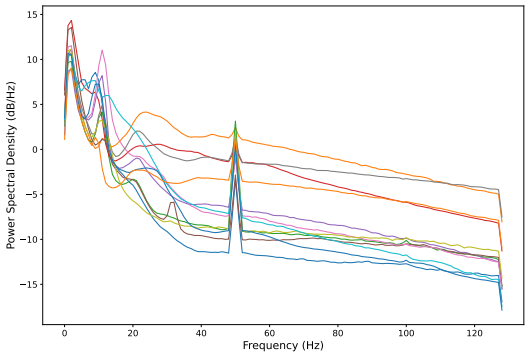

Figure 3: PSDs for different subjects in the target *Cue* dataset. Notice the spike at 50 Hz.