PROJENTROPY: USING ENTROPY TO OPTIMIZE SPATIAL PROJECTIONS

Austin J. Brockmeier, Eder Santanna, Luis G. Sanchez Giraldo, and Jose C. Principe

University of Florida
Electrical and Computer Engineering
Gainesville, Florida

ABSTRACT

Methods for hypothesis testing on zero-mean vector-valued signals often rely on a Gaussian assumption, where the second-order statistics of the observed sample are sufficient statistics of the conditional distribution. This yields fast and simple tests, but by using information-theoretic statistics one can relax the Gaussian assumption. We propose using Rényi's quadratic entropy as an alternative to the covariance and show how a linear projection can be optimized to maximize the difference between the conditional entropies. In addition, if the observed sample is actually a window of a multivariate time-series, then the temporal structure can be exploited using the generalized auto-correlation function, correntropy, of the projected sample. This both reduces the computational complexity and increases the performance. These tests can be applied for decoding the brain state from electroencephalogram (EEG) recordings. Preliminary results are demonstrated on a brain-computer interface competition dataset. On unfiltered signals, the projections optimized with the entropy-based statistic perform better than those of common spatial pattern (CSP) algorithm in terms of classification performance.

Index Terms— array signal processing, BCI, correntropy, EEG, entropy, feature extraction, hypothesis testing, projection pursuit

1. INTRODUCTION

Deciphering the brain state from electroencephalogram (EEG) recordings is a difficult challenge even in the two-condition setting common for brain computer interfaces (BCIs) [1]. One approach is to exploit spatial differences in the neural response across the electrode array. Essentially, this amounts to array signal processing with non-stationary signals in a low signal-to-noise setting. In this setting, any model may be hard to fit, but using a zero-mean multivariate Gaussian distribution as the model for the conditional distribution yields a simple test statistic for classification.

This work was partially supported in part by DARPA Contract N66001-10-C-2008.

This test statistic can be used to optimize a spatial projection matrix for EEG recordings, using only the estimates of the covariance in both classes and eigendecompositions. In the pattern recognition literature this is known as the Fukunaga-Koontz transform [2, 3], in the EEG analysis literature it has became popular for extracting features for motor imagery classification and is known as common spatial patterns (CSP) [4]. For both its speed, effectiveness, and ease-of-understanding this approach has been used extensively in the BCI research community, and there has been a wide literature of extensions to increase performance [5, 6].

Here we revisit this problem within a non-Gaussian framework using an information theoretic learning perspective [7]. We implicitly form a non-parametric model using kernel functions and use Renyi's quadratic entropy [8] as a test statistic. We optimize spatial projection that maximize the entropy of the data under one condition while minimizing it under the other condition. This use of an implicit model allows the projection to be optimized on unfiltered data.

As opposed to kernelizing the projection [9], which obfuscates a spatial interpretation, we use kernel functions to compute the test statistic, but a linear projection of the multichannel signal is still used. In terms of information theory, previous work [10] has proposed using approximations of negentropy to optimize projections as an alternative to CSP; however, the test statistic was still based on the variance of the projected signal.

While the proposed approach improves performance, it along with previous approaches often ignore any temporal information within a sample. Thus, we propose to use *a priori* information to select certain lags of the correntropy function, which is a generalized measure of the correlation function [11]. Not only does this approach improve performance, but it also decreases the computational complexity versus the temporally uninformed measure.

We test the approach using a benchmark BCI dataset. We show that—for unfiltered data—optimizing the spatial projection using the information-theoretic objective improves the performance further, and using correntropy over a subset of the lags yields the best performance.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. COVARIANCE-BASED TEST STATISTICS

Let X denote a random variable with domain $\mathbb{R}^d$. The distribution of X is conditioned on the value of a binary variable C, and the conditional probability density functions for $C = 0$ and $C = 1$ are denoted $f_X(x|0)$ and $f_X(x|1)$, respectively. We consider the problem of testing which value of C is more likely based on the logarithm of the likelihood ratio:

$$\mathcal{L}(x) = \ln\left(\frac{f_X(x|1)}{f_X(x|0)}\right). \tag{1}$$

When $f_X(x|c)$ is a zero-mean Gaussian distribution with covariance matrix Σ_c,

$$\mathcal{L}(x) \propto \operatorname{tr}\left[\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right) x x^{\mathrm{T}}\right]. \tag{2}$$

Given a sample of vectors $\{x\}$, a test statistic can be formed as $T(\{x\}) = \operatorname{tr}(AB) \propto \sum_i \mathcal{L}(x_i)$ where the matrix $A = \Sigma_0^{-1} - \Sigma_1^{-1}$ is symmetric, and $B = \sum_i x_i x_i^{\mathrm{T}}$ is scatter matrix of the sample.

This test uses the covariance of the sample in all dimensions. In practice it is sufficient to test the variance in a linear subspace in essence using a lower-rank approximation, $A \approx PP^{\mathrm{T}}$, where the matrix P defines the projection. The log-likelihood ratio is replaced with the new test statistic:

$$T_P(\{x\}) = \operatorname{tr}(PP^{\mathrm{T}}B) = \operatorname{tr}(P^{\mathrm{T}} \sum_i x_i x_i^{\mathrm{T}} P) \tag{3}$$

$$= \operatorname{tr}\left[\sum_i (P^{\mathrm{T}}x_i)(P^{\mathrm{T}}x_i)^{\mathrm{T}}\right] = \sum_i \|y_i\|_2^2, \tag{4}$$

where $y_i = P^{\mathrm{T}}x_i$. To improve the power of the test, the matrix P should be to chosen to maximize the divergence between $T_P(X|C = 1)$ and $T_P(X|C = 0)$. One approach is to constrain P to be orthonormal, $P^{\mathrm{T}}P = I$, and maximize $\mathsf{E}[T_P(X|C = 1)]$ while minimizing $\mathsf{E}[T_P(X|C = 0)]$. In terms of the covariance matrices,

$$\mathsf{E}[T_P(X|c)] = \operatorname{tr}\left(P^{\mathrm{T}}\mathsf{E}[XX^{\mathrm{T}}|c]P\right) = \operatorname{tr}\left(P^{\mathrm{T}}\Sigma_c P\right). \tag{5}$$

This leads to the trace ratio problem:

$$\underset{P^{\mathrm{T}}P=I}{\text{maximize}} \left\{\frac{T_P(X|1)}{T_P(X|0)} = \frac{\operatorname{tr}\left(P^{\mathrm{T}}\Sigma_1 P\right)}{\operatorname{tr}\left(P^{\mathrm{T}}\Sigma_0 P\right)}\right\}. \tag{6}$$

There is no closed-form optimal solution to the trace-ratio problem [12], but the ratio of the determinants requires no constraint on the orthonormality of P,

$$\text{maximize } \frac{\left|P^{\mathrm{T}}\Sigma_1 P\right|}{\left|P^{\mathrm{T}}\Sigma_0 P\right|} \tag{7}$$

and can be solved as a generalized eigenvalue problem.

Practically, the test statistic $T_P(\{x\}) = \sum_i \|y_i\|_2^2$ can be treated as a feature for the binary classification problem. Instead of using only a projection that maximizes the objective, projections that minimize the objective are also used. When a sample is a window of a time series, then the test statistic is a short-term measure of the projected signal's power. To normalize the conditional power distribution a log transform is typically used [4]. In addition, the variance in each direction of the projection can be normalized by the total variance of the projected sample. Explicitly, if P has m columns then each feature vector has m elements and the jth element is $\ln\{\sum_i y_i^2(j)/[\sum_k \sum_i y_i^2(k)]\}$ [4]. However, this normalization eliminates one degree of freedom of the features and can decrease performance.

3. KERNEL-BASED QUADRATIC ENTROPY

Again let X denote the random variable with domain $\mathfrak{X} = \mathbb{R}^d$ and $\{x\}$ denote a sample. Now consider a possibly nonlinear embedding of this sample into a reproducing kernel Hilbert space (RKHS) defined by the positive definite kernel $\kappa : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$. Associated with κ there is an implicit mapping $\phi : \mathfrak{X} \to \mathcal{H}$ that maps any element $x \in \mathfrak{X}$ to an element in the Hilbert space $\phi(x) \in \mathcal{H}$ such that the kernel evaluation corresponds to an inner- or dot-product: $\kappa(x, x') = \langle\phi(x), \phi(x')\rangle = \phi(x)^{\mathrm{T}}\phi(x')$. Here we abuse notation the notation for matrix transpose for elements in the RKHS. Given a sample, the kernel matrix is formed as $K_{ij} = \kappa(x_i, x_j)$; the resulting matrix K is positive definite.

Using the RKHS embedding, we propose to compute the test statistic (3) using the the operator $\Psi\Psi^{\mathrm{T}}$. The new test statistic is computed as $S(x) = \operatorname{tr}\left[\Psi\Psi^{\mathrm{T}}\phi(x)\phi(x)^{\mathrm{T}}\right]$.

Again, Ψ can be chosen to maximize the power of the test. Instead of directly optimizing the operator, we propose to adapt ϕ and fix $\Psi\Psi^{\mathrm{T}}$ as $\mathsf{E}_{X'}[\phi(X')\phi(X')^{\mathrm{T}}] = \mathsf{E}_{X'}[\phi(X')]\mathsf{E}_{X'}[\phi(X')]^{\mathrm{T}}]$, where X' is an independent random variable with the same distribution as X.

We write the test statistic using properties of the trace:

$$S(x) = \operatorname{tr}\left[\mathsf{E}_{X'}[\phi(X')\phi(X')^{\mathrm{T}}]\phi(x)\phi(x)^{\mathrm{T}}\right] \tag{8}$$

$$= \mathsf{E}_{X'}[\phi(X')^{\mathrm{T}}\phi(x)\phi(x)^{\mathrm{T}}\phi(X')] \tag{9}$$

$$= \mathsf{E}_{X'}[\kappa(X', x)\kappa(x, X')] = \mathsf{E}_{X'}[\kappa^2(x, X')] \tag{10}$$

The mean of the test statistic,

$$\mathsf{E}_X S(X) = \mathsf{E}_X \mathsf{E}_{X'}[\kappa^2(X, X')] = V(X, X'), \tag{11}$$

is known as the cross-information potential [7] between X and X'. Since X' is simply an independent but identically distributed version of X we denote $V(X, X')$ as $V(X)$.

Now, we consider optimizing ϕ to separate $\mathsf{E}_{X|0}S(X)$ and $\mathsf{E}_{X|1}S(X)$. Since each point in a sample corresponds to a vector, we use a multivariate Gaussian kernel parametrized by a projection matrix P:

$$\kappa_P(x, x') = \exp\{-(x - x')^{\mathrm{T}}PP^{\mathrm{T}}(x - x')\}. \tag{12}$$

With this kernel, the test statistic is proportional to the density of an unweighted mixture of Gaussians with equal covariance

$$S_P(x) = \mathsf{E}_{X'}\left[\exp\{-(x-X')^{\mathrm{T}}PP^{\mathrm{T}}(x-X')\}\right]. \quad (13)$$

In terms of kernel density estimation, κ_P^2 is a Parzen window and $S_P(x) = \mathsf{E}_{X'}[\kappa_P^2(x,X')] \propto p_X(x)$. Taking the expected value, $\mathsf{E}[S_P(x)] = V_P(X) \propto \mathsf{E}_X[p_X(X)]$; the negative logarithm of the right hand side is Rényi's quadratic entropy [8].

For the conditional distributions,

$$V_P(X|c) = \mathsf{E}[S_P(X|c)] = \mathsf{E}_{X|c}\mathsf{E}_{X'|c}[\kappa_P^2(X,X')]. \quad (14)$$

Under the convention above, we would like to choose P to maximize $\mathsf{E}[S_P(X|1)]$ while minimizing $\mathsf{E}[S_P(X|0)]$. This leads to an unconstrained optimization problem

$$\underset{P}{\text{maximize}}\ g(P) = \frac{\mathsf{E}[S_P(X|1)]}{\mathsf{E}[S_P(X|0)]} = \frac{V_P(X|1)}{V_P(X|0)}. \quad (15)$$

Taking the logarithm of the objective function,

$$\ln V_P(X|1) - \ln V_P(X|0)) \propto -H_2(X|1) + H_2(X|0). \quad (16)$$

Thus, it is clear that optimizing with this objective yields a projection that minimizes the entropy of $X|1$ while maximizing the entropy of $X|0$.

Given a single sample $\mathcal{X} = \{x\}$ and the corresponding kernel matrix K_P, a biased estimator[1] of $V_P(X)$ is

$$\tilde{V}_P(\mathcal{X}) = \sum_{i,j}\kappa_P^2(x_i,x_j) = \sum_{i,j}[K_P]_{i,j}^2 = \mathrm{tr}(K_P^2). \quad (17)$$

$$\text{Then } h_P(\mathcal{X}) = -\ln \tilde{V}_P(\mathcal{X}) = -\ln \sum_{i,j}[K_P]_{i,j}^2 \quad (18)$$

is a test statistic that is proportional to Rényi's quadratic entropy [13] under the model defined by P. Thus, for succinctness, we refer to Eq. (18) as *projentropy*.

In an information-theoretic setting, it would seem more efficient to exploit the divergences between the classes [6]. For instance if the distribution of X' was fixed to the distribution of either $X|1$ or $X|0$, then Eq. (13) would utilize any difference in the distributions. However, this would require to choose and keep a dictionary of samples to represent X'. With the proposed approach, the test statistic does not require access to any previous samples; this eliminates the need to store samples between trials, or have to store every sample a problem that can plague kernel-based learning machines.

3.1. Gradients and optimization

Using the entropy ratio (16) as a surrogate for log-likelihood ratio (2), we consider optimizing P using the following cost function

$$\begin{aligned}\mathcal{F}(P) &= \mathsf{E}_{\mathcal{X}|1}[h_P(\mathcal{X})] - \mathsf{E}_{\mathcal{X}|0}[h_P(\mathcal{X})] \\ &= -\mathsf{E}_{\mathcal{X}|1}[\ln \mathrm{tr}(K_P^2)] + \mathsf{E}_{\mathcal{X}|0}[\ln \mathrm{tr}(K_P^2)]. \end{aligned} \quad (19)$$

[1] An unbiased estimator simply ignores the diagonal of the kernel matrix.

The projection that minimizes the entropy of one class is not necessarily orthogonal to the one that minimizes the entropy of the other class. Thus, for feature extraction both projections should be found. In addition, for the multi-class setting each class can be substituted for class 1, and the rest of the classes can be pooled into class 0.

The gradient of the cost function, with respect to P, is

$$\nabla_P\mathcal{F}(P) = -\mathsf{E}_{\mathcal{X}|1}[4X^{\mathrm{T}}DXP] + \mathsf{E}_{\mathcal{X}|0}[4X^{\mathrm{T}}DXP] \quad (20)$$

$$D = (G \circ K_P) - \mathrm{diag}(\mathbf{1}^{\mathrm{T}}(G \circ K_P)) \quad (21)$$

$$G = \nabla_{K_P}\ln \mathrm{tr}(K_P^2) = \frac{2K_P}{\mathrm{tr}(K_P^2)} \quad (22)$$

where here X is a $n \times d$ matrix where each row corresponds to a point in the sample, $\circ$ is the Hadamard element-wise product, and $\mathbf{1}$ is a vector of ones. We use limited-memory BFGS to optimize P in MATLAB using `minFunc` [14]. The initial solution is chosen to have normally distributed entries with unit variance, and the optimization is stopped after 50 iterations.

The computational complexity of the statistic for a sample of n points with a $n \times n$ kernel matrix is $\mathcal{O}(n^2)$; whereas, the covariance statistic has complexity $\mathcal{O}(m^2 n)$ for a m dimensional projection. With a set of N samples, the computational complexity of the derivative is $\mathcal{O}(Nn^2)$, and the convergence time is dependent on m—this is much greater than the overall complexity of CSP that is $\mathcal{O}(m^3)$.

4. TEMPORALLY-INFORMED PROJENTROPY

The covariance and entropy-based tests ignore any temporal structure within a sample, treating the set of points as independently distributed. When each sample is a window from a time-series, we can consider using statistics based on the auto-correntropy function [11], which is a generalized version of the auto-correlation function. Correntropy has been used to analyze the periodicity of signals [15] or extract information via the kernel-based representation [16]. Specifically, the auto-correntropy function is defined for the process $\mathbf{x}_t$ as $v(\tau) = \mathsf{E}[k(\mathbf{x}_t, \mathbf{x}_{t+\tau})]$, where k is chosen as a shift-invariant kernel function.

We assume we have *a priori* information on the specific temporal structure. Let $\mathcal{T} = \{\tau_1, \tau_2, \ldots \tau_m\}$ denote the set of lags of interest. Then a simple test statistic can be formed as an unweighted combination across lags

$$V_{\mathcal{T}}(\mathbf{x}_t) = \sum_{\tau\in\mathcal{T}} v(\tau) = \sum_{\tau\in\mathcal{T}} \mathsf{E}[k(\mathbf{x}_t, \mathbf{x}_{t+\tau})]. \quad (23)$$

Returning to multivariate time series, we set $k = \kappa_P^2$ which uses a linear projections across channels. Then $V_{\mathcal{T}}(\mathbf{x}_t)$ is a combination of the statistic (13) with the distribution of X' fixed at $\mathbf{x}_{t+\tau}, \tau \in \mathcal{T}$.

Given a sample $\mathcal{X}$ that corresponds to a window of a discretely sampled time series, i.e., $\mathcal{X} = [x]_{i=1}^n$, the

sample-version of correntropy uses the time average: $\tilde{v}(\tau) = \sum_{i=1}^{n-\tau} \frac{1}{n-\tau} k(x_i, x_{i+\tau})$. The test statistic for a sample is

$$\tilde{V}_{P,\mathcal{T}}(\mathcal{X}) = \sum_{\tau \in \mathcal{T}} \tilde{v}(\tau) = \sum_{\tau \in \mathcal{T}} \sum_{i=1}^{n-\tau} \frac{1}{n-\tau} \kappa_P^2(x_i, x_{i+\tau}). \quad (24)$$

Then $h_{P,\mathcal{T}}(\mathcal{X}) = -\ln \tilde{V}_{P,\mathcal{T}}(\mathcal{X})$ is a *temporally informed* version of projentropy (18), which uses only a part of the kernel matrix (17). Beyond the ability to restrict attention to certain periodicities, the calculation of $h_{P,\mathcal{T}}(\mathcal{X})$ requires $\mathcal{O}(mn)$ operations where $m = |\mathcal{T}|$. Since m can be chosen to be much smaller than n, this requires far fewer calculations than the entropy-based statistic. We leave the problem of optimizing temporally-informed projentropy for future work.

5. EXPERIMENTAL SETUP AND RESULTS

For preliminary testing we used the BCI competition III dataset IV(a), [17], provided by Fraunhofer FIRST, Intelligent Data Analysis Group (Klaus-Robert Müller, Benjamin Blankertz), and Campus Benjamin Franklin of the Charité–University Medicine Berlin, Department of Neurology, Neurophysics Group (Gabriel Curio) [18]. Five healthy subjects performed cued segments of right hand and right foot motor imagery. For each subject, 140 trials of each class, 280 trials in total were provided. Classification was performed using 5×5 cross validation, where the labels for the test trials set apart in the competition are also used indiscriminately.

We used the framework and implementations provided in the NFEA toolbox [19, 20] to analyze the dataset in MATLAB. For each trial the 3.5 s window during the visual cue was extracted, and only the window from .5 s to 2.5 s was used (201 time points). *No filtering was applied*.

A simple setup for features extraction and classification was used to compare the covariance-based projection with the projentropy-based projection. We used the traces from 13 electrodes over the left motor area: (C5, C1, CCP7, CCP3, CCP1, CP5, CP3, CP1, CPz, PCP5, PCP3, PCP1). Two spatial projections, one for each class, were optimized using only samples in the training set. Each projection itself was 1-dimensional. Thus, only 2 features were used to classify each trial.

We used either the variance, entropy, or correntropy at specific lags of the projected signal as the features. The variance required no free parameter. In addition, there is no free parameter when the projentropy statistic is used with a projection optimized using projentropy. When the projentropy statistics are computed on the CSP projection, the projection is scaled such that median distance between all projected time points is 1.

For the temporally-informed projentropy, lags were chosen *a priori* to correspond to the fundamental and second harmonic of signals at around 6.7 Hz. This is close to the high-pass cutoff when filtering is applied. A number of lags surrounding the fundamental period of 15 samples at 100 Hz, were used: $\mathcal{T} = \{0, 1, 13, 14, 15, 16, 17, 27, 28, 29, 30, 31, 32, 33\}$.

Table 1. Motor imagery classification accuracy (% correct). The average and standard deviation across 5×5 cross-validation (224 training and 58 testing samples). Projections are found via common spatial patterns (CSP) or projentropy ($\mathcal{P}$) and the variance (σ^2), Rényi entropy (H_2), or correntropy at certain lags ($V_{\mathcal{T}}$) are used as features. Only two one-dimensional projections are found for each method. Linear discriminant analysis is used for classification. (**Bold** indicates methods with significantly better performance for a given subject or average as determined by a paired sign test with significance of $\alpha = 0.1$.)

Subject	CSP$\|\sigma^2$	CSP$\|H_2$	CSP$\|V_{\mathcal{T}}$	$\mathcal{P}\|H_2$
aa	63±4.0	63±4.0	64±4.1	**74±1.3**
al	88±1.2	88±0.8	**90±0.4**	87±0.5
av	56±1.5	56±2.0	**63±1.4**	**63±3.5**
aw	**75±2.5**	**75±3.0**	**77±0.9**	**76±0.3**
ay	73±1.0	76±1.6	**87±1.1**	74±1.2
Across all	71±12	71±12	**76±12**	75±8.5
Run-time (s)	0.1±0.0	1.7±0.1	1.3±0.1	89±3.1

After extraction, the two-dimensional feature vectors were classified using linear discriminant analysis [21]; the results are shown in Table 1. A significant increase in classification performance was seen when using the alternative statistics as features. Optimizing the two projections using projentropy-based cost function yielded a increase of 4 percentage points over CSP. Even with the CSP projection, the temporally informed projentropy-based features yielded the highest accuracy with a increase of 5 percentage points. As stated in Section 4, future work should address optimizing the projection based on temporally informed projentropy $\mathcal{P}|V_{\mathcal{T}}$.

6. CONCLUSION

We proposed *projentropy* as a method for training projections that discriminative based on the entropy of a sample instead of the variance. For unfiltered signals, projentropy improves the classification accuracy over using CSP. The main drawback of projentropy is the increased computational complexity. It is noteworthy that the temporally informed measure, which has the lowest computational complexity, yielded the best results. This motivates future work on the study of the effect of temporal information to substitute for linear filtering.

7. REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer

interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] K. Fukunaga and W. L. Koontz, "Application of the Karhunen-Loève expansion to feature selection and ordering," *Computers, IEEE Transactions on*, vol. 100, no. 4, pp. 311–318, 1970.

[3] S. Zhang and T. Sim, "Discriminant subspace analysis: A fukunaga-koontz approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, pp. 1732–1745, 2007.

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

[5] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K.-R. Müller, "Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., pp. 113–120. MIT Press, Cambridge, MA, 2008.

[6] W. Samek, M. Kawanabe, and K. Muller, "Divergence-based framework for common spatial patterns algorithms," *Biomedical Engineering, IEEE Reviews in*, vol. PP, no. 99, 2013, Pre-print.

[7] J. C. Principe, "Information theoretic learning: Renyi's entropy and kernel perspectives," 2010.

[8] A. Rényi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1960, pp. 547–561.

[9] Q. Zhao, T. M. Rutkowski, L. Zhang, and A. Cichocki, "Generalized optimal spatial filtering using a kernel approach with application to eeg classification," *Cognitive Neurodynamics*, vol. 4, no. 4, pp. 355–358, 2010.

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

[11] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *Signal Processing, IEEE Transactions on*, vol. 55, no. 11, pp. 5286–5298, 2007.

[12] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[13] L. G. Sanchez Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Transactions on Information Theory*, 2013, *submitted*, `arXiv:1211.2459 [cs.LG]`.

[14] Mark Schmidt, "minFunc," 2012, Software available at `http://www.di.ens.fr/~mschmidt/Software/minFunc.html`.

[15] P. Huijse, P. A. Estevez, P. Protopapas, P. Zegers, and J. C. Principe, "An information theoretic algorithm for finding periodicities in stellar light curves," *Signal Processing, IEEE Transactions on*, vol. 60, no. 10, pp. 5135–5145, 2012.

[16] E. Santana, J. C. Principe, E.E. Santana, R.C.S. Freire, and A.K. Barros, "Extraction of signals with specific temporal structure using kernel methods," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5141–5150, 2010.

[17] B. Blankertz, K.-R. Muller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, Jd R. Millan, M. Schroder, and N. Birbaumer, "The BCI competition III: Validating alternative approaches to actual BCI problems," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, no. 2, pp. 153–159, 2006.

[18] G. Dornhege, B. Blankertz, G. Curio, and K. Muller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 993–1002, 2004.

[19] A. H. Phan, "NFEA: Tensor toolbox for feature extraction and applications," 2011, Technical Report, Lab for Advanced Brain Signal Processing, BSI, RIKEN. Software available at `http://www.bsp.brain.riken.jp/~phan/nfea/nfea.html`.

[20] A. H. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets," *Nonlinear Theory and Its Applications, IEICE*, vol. 1, no. 1, pp. 37–68, 2010.

[21] M. Kiefte, "Discriminant analysis toolbox," 1999, Software available from `http://www.mathworks.com/matlabcentral`.