

Discriminant Multitaper Component Analysis of EEG

Mads Dyrholm* and Paul Sajda†

*Center for Visual Cognition, Department of Psychology, University of Copenhagen, Denmark

†Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, USA

Abstract. This work extends Bilinear Discriminant Component Analysis to the case of oscillatory activity with allowed phase-variability across trials. The proposed method learns a spatial profile together with a multitaper basis which can integrate oscillatory power in a band-limited fashion. We demonstrate the method for predicting the handedness of a subject's button press given multivariate EEG data. We show that our method learns multitapers sensitive to oscillatory activity in the 8-12Hz range with spatial filters selective for lateralized motor cortex. This finding is consistent with the well-known mu-rhythm, whose power is known to modulate as a function of which hand a subject plans to move, and thus is expected to be discriminative (predictive) of the subject's response.

Keywords: Single-trial, EEG, oscillatory, signal detection

PACS: 87.19.le; 87.19.lt; 87.85.D-

INTRODUCTION

In this paper we propose a Component Analysis method for single-trial EEG. By 'Component Analysis' we refer generally to a method that characterizes each individual EEG trial in terms of learned spatial and temporal profiles. Component Analysis methods can be either unsupervised: as for instance ICA [1] or PARAFAC [2, 3]; or supervised: as for instance BDCA [4, 5, 6]. Supervised methods are appealing in situations where the number of trials is limited. Other supervised methods incorporate power spectral estimation [7, 8, 9] and hence enable detection of oscillatory activity which is not phase locked across trials. However, these methods do not learn the temporal envelope associated with the oscillatory activity. The method that we propose here learns (for each 'component') a spatial profile and an associated taper basis for power spectral discrimination. The taper basis can thus represent a temporally enveloped and filtered subspace to optimally tradeoff bandwidth, temporal support, and sidelobe power [10, 11].

The proposed method evolves from the BDCA model (Logistic Regression), where we model the expected class label y_n for trial n by the sigmoid function

$$E[y_n] = \pi[\mathbf{X}_n] = \frac{1}{1 + e^{-(w_0 + \psi[\mathbf{X}_n])}} \quad (1)$$

where $\mathbf{X}_n \in \mathbb{R}^{D,T}$ are trial epochs, trial numbers are $n \in [1, N]$, $\psi[\mathbf{X}_n]$ is a discriminating function of the data and w_0 models an offset in the argument to the sigmoid. D is the number of electrodes and T is the number of samples per epoch. In BDCA the

discriminating projection, with R 'components', is given by

$$\psi_{\text{BDCA}}[\mathbf{X}_n] = \sum_{r,i,j=1}^{R,D,T} (\mathbf{u}_r)_i (\mathbf{X}_n)_{i,j} (\mathbf{v}_r)_j \quad (2)$$

and it enables the detection of evoked activity with (for 'component' r) a spatial profile \mathbf{u}_r and a temporal profile \mathbf{v}_r [5]. The modification that we propose here is to replace the component-wise temporal profiles \mathbf{v}_r with a temporal taper matrix \mathbf{V}_r , and integrate projected power, by redefining the discriminating function

$$\psi[\mathbf{X}_n] \equiv \sum_{r,i,b=1}^{R,D,B} (\mathbf{u}_r)_i \left(\sum_{j=1}^T (\mathbf{X}_n)_{i,j} (\mathbf{V}_r)_{j,b} \right)^2 \quad (3)$$

where R is the number of components, \mathbf{u}_r defines the component spatial power integral, B is a 'bandwidth' parameter defining the number of tapers (column vectors) in \mathbf{V}_r . Our method can in principle learn specialized tapers such as Discrete Prolate Spheroidal Sequences (DPSS) often used in multi-taper spectral analysis. See [10, 11] on the relationship between bandwidth, temporal support, and number of tapers for the case of DPSS.

LIKELIHOOD AND GRADIENT

The log likelihood per sample is given by

$$l_n = y_n(w_0 + \psi[\mathbf{X}_n]) - \log[1 + e^{w_0 + \psi[\mathbf{X}_n]}] \quad (4)$$

see e.g. [12]. The gradient expressions are given by

$$\frac{\partial l_n}{\partial w_0} = y_n - \pi[\mathbf{X}_n] \quad (5)$$

$$\frac{\partial l_n}{\partial (\mathbf{u}_r)_i} = (y_n - \pi[\mathbf{X}_n]) \frac{\partial \psi[\mathbf{X}_n]}{\partial (\mathbf{u}_r)_i} \quad (6)$$

$$\frac{\partial l_n}{\partial (\mathbf{V}_r)_{j,b}} = (y_n - \pi[\mathbf{X}_n]) \frac{\partial \psi[\mathbf{X}_n]}{\partial (\mathbf{V}_r)_{j,b}} \quad (7)$$

where

$$\frac{\partial \psi[\mathbf{X}_n]}{\partial (\mathbf{u}_r)_i} = \sum_{b=1}^B \left(\sum_{j=1}^T (\mathbf{X}_n)_{i,j} (\mathbf{V}_r)_{j,b} \right)^2 \quad (8)$$

$$\frac{\partial \psi[\mathbf{X}_n]}{\partial (\mathbf{V}_r)_{j,b}} = 2 \sum_{i=1}^D (\mathbf{X}_n)_{i,j} (\mathbf{u}_r)_i \times \left(\sum_{j'=1}^T (\mathbf{X}_n)_{i,j'} (\mathbf{V}_r)_{j',b} \right) \quad (9)$$

REGULARIZATION

As in [4, 5] we declare Gaussian Process priors to regularize the parameter estimation for smoothness. Assume \mathbf{u}_r drawn from $\mathcal{N}(\mathbf{0}, \mathbf{K}_u)$ and the columns of \mathbf{V}_r drawn from $\mathcal{N}(\mathbf{0}, \mathbf{K}_v)$, where the covariance matrices \mathbf{K}_u and \mathbf{K}_v define the degree and form of smoothness of \mathbf{u}_r and \mathbf{V}_k respectively. We define these (positive definite) covariance matrices by evaluating the covariance function

$$(\mathbf{K}_u)_{i,i'} = \sigma^2 k_{\text{Matérn}}(r_{i,i'}) \quad (10)$$

where $r_{i,i'}$ is the distance between electrode- i and electrode- i' and similarly

$$(\mathbf{K}_v)_{j,j'} = \sigma^2 k_{\text{Matérn}}(r_{j,j'}) \quad (11)$$

where $r_{j,j'}$ is the difference between sample- j and sample- j' . In both cases the hyper-parameter σ^2 defines the overall parameter scale, and the Matérn covariance function is given by

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{l} \right)^\nu \text{B} \left(\frac{\sqrt{2\nu}r}{l} \right) \quad (12)$$

where $\text{B}(\cdot)$ is a modified Bessel function, l is a length-scale hyper-parameter, and ν is a shape hyper-parameter. The parameter l can roughly be thought of as the distance within which points are significantly correlated [13]. The parameter ν defines the degree of ripple. The regularized cost function is obtained by subtracting the log priors from the sample negative log likelihood (4)

$$E = \sum_r \frac{1}{2} \mathbf{u}_r^T \mathbf{K}_u^{-1} \mathbf{u}_r + \frac{1}{2} \text{Trace}[\mathbf{V}_r^T \mathbf{K}_v^{-1} \mathbf{V}_r] - \sum_n l_n \quad (13)$$

We learn $\{\mathbf{u}_r, \mathbf{V}_r\}$ by optimizing E using BFGS Quasi-Newton optimization, with soft line search and trust region monitoring as implemented by H. B. Nielsen [14]. For that purpose the gradient of the cost function is given by

$$\frac{\partial E}{\partial w_0} = - \sum_n \frac{\partial l_n}{\partial w_0} \quad (14)$$

$$\frac{\partial E}{\partial \mathbf{u}_r} = \mathbf{K}_u^{-1/2} \mathbf{u}_r - \sum_n \frac{\partial l_n}{\partial \mathbf{u}_r} \quad (15)$$

$$\frac{\partial E}{\partial \mathbf{V}_r} = \mathbf{K}_v^{-1/2} \mathbf{V}_r - \sum_n \frac{\partial l_n}{\partial \mathbf{V}_r} \quad (16)$$

EXPERIMENT

Data

As a way to test our method, we consider the benchmark EEG data first published as ‘Dataset IV’ as part of the ‘BCI Competition 2003’ [15]. A single subject performed a

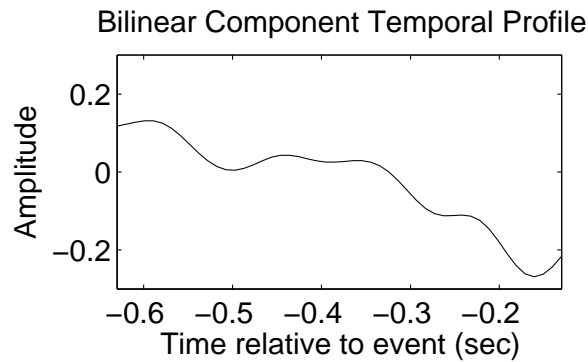


FIGURE 1. From BDCA: Temporal profile represents the Bereitschaftspotential (negative).

Bilinear Component Spatial Profile

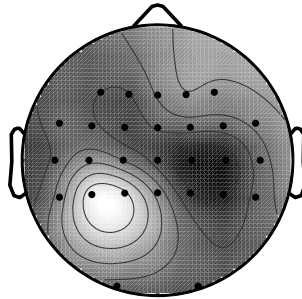


FIGURE 2. From BDCA: Spatial profile showing the discriminative lateralization of the Bereitschaftspotential profile.

self-paced key typing task, where they pressed keys with their index and little fingers using a self-chosen order, timing and handedness. The goal in the BCI Competition was to predict the laterality (handedness) of the upcoming finger movement. Typing was done at an average speed of 1 key per second. 28 channels were recorded at 1000 Hz with a pass-band between 0.05 and 200 Hz, then downsampled to 100Hz sampling rate. Trial matrices of length 500ms were extracted by epoching the data starting 630ms before each key-press. 316 epochs were to be used for classifier training.

BDCA

First, we trained a BDCA with one component using the method of [5]. Cross validated Area Under the ROC Curve was $AUC = 0.95$. The temporal profile is shown in FIGURE 1, and the spatial profile is shown in FIGURE 2. The profiles correspond well with previously reported potentials for similar paradigms, see e.g. [16, 9].

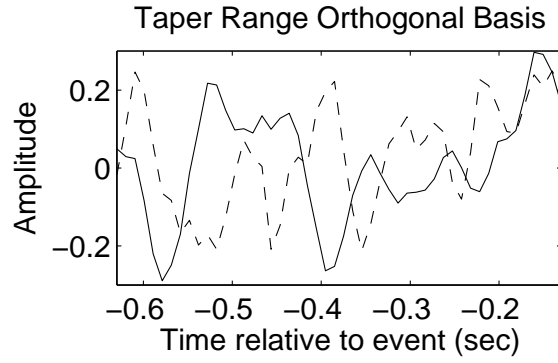


FIGURE 3. From new taper method: An orthogonal basis illustrating the range of the learned taper matrix.

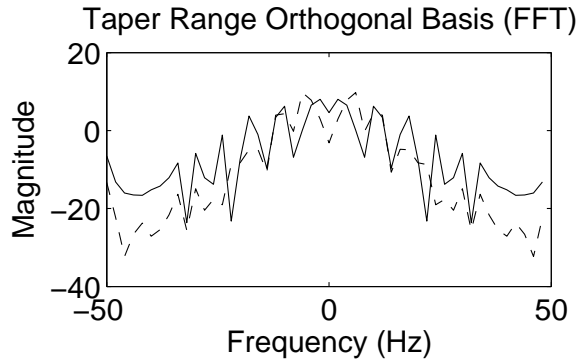


FIGURE 4. From new taper method: Both orthogonal tapers show a harmonic pattern with a mu-band fundamental frequency. This finding supports the hypothesis of discriminant mu-rhythm activity.

New taper method

We then used the new taper method incrementally on top of the BDCA result. First, we subtracted all activity in the BDCA discriminating subspace from each trial, i.e. $\mathbf{X}_n \leftarrow \mathbf{X}_n - \mathbf{u}_1 \mathbf{u}_1^T \mathbf{X}_n \mathbf{v}_1 \mathbf{v}_1^T / (\|\mathbf{u}_1\|_2^2 \|\mathbf{v}_1\|_2^2)$. By doing so we ensured that the following taper method analysis would not learn from those evoked (non-oscillatory) potentials. We then trained the new taper method on the modified data. The BFGS Quasi Newton learning was sensitive to scaling, and we found that scaling the data by 10^{-1} was useful. We obtained cross validated $\text{AUC} = 0.75$ which indicated that the new method had indeed learned something that the BDCA had previously ignored. An orthogonal basis of the range of the learned tapers are shown in FIGURE 3 and the basis is clearly oscillatory. The taper range FFT magnitudes are shown in FIGURE 4 and the two basis vectors show harmonic structure with a fundamental frequency in the mu-band. The spatial profile is shown in FIGURE 5 and exhibits a lateralization in correspondence with [16], i.e. a lateralized power decrease in the mu band due to desynchronization.

Spatial Map For Taper Method

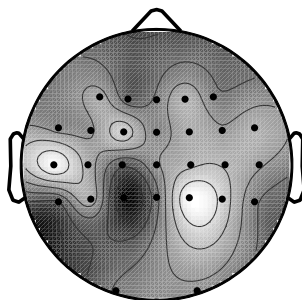


FIGURE 5. Spatial map from new taper method. The spatial map supports the hypothesis of a lateralized decrease in activity (mu-rhythm c.f. FIGURE 4).

CONCLUSION

Discriminant Multitaper Component Analysis exploits power changes in oscillatory activity that is discriminating of trial type, and complements our previously described Bilinear Discriminant Component Analysis which utilizes linear changes in activity. Both approaches are well-suited for analysis of neural data, particularly EEG, where both evoked response potentials and power changes in specific frequency bands hold information predictive of subject response and behavior.

REFERENCES

1. S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent Component Analysis of Electroencephalographic Data," in *Advances in Neural Information Processing Systems*, edited by M. Mozer, and M. Hasselmo, 1996, pp. 145–151.
2. F. Miwakeichi, E. Martinez-Montes, P. A. Valdes-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, *Neuroimage* **22**, 1035–45 (2004).
3. M. Mørup, L. K. Hansen, C. S. Hermann, J. Parnas, and S. M. Arnfred, *NeuroImage* **29**, 938–947 (2006).
4. M. Dyrholm, and L. C. Parra, "Smooth bilinear classification of EEG," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006.
5. M. Dyrholm, C. Christoforou, and L. C. Parra, *Journal of Machine Learning Research* **8**, 1097–1111 (2007).
6. L. Parra, C. Christoforou, A. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. Philiastides, and P. Sajda, *IEEE Signal Processing Magazine* **25**, 95–115 (2008).
7. S. Lemm, B. Blankertz, G. Curio, and K. Müller, *IEEE Trans Biomed Eng* **52**, 1541–1548 (2005).
8. G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, *IEEE Trans Biomed Eng* (2006).
9. C. Christoforou, P. Sajda, and L. C. Parra, "Second Order Bilinear Discriminant Analysis for single trial EEG," in *Advances in Neural Information Processing Systems*, 2007, vol. 21.
10. D. Slepian, *The Bell Systems Tech. J.* **57**, 1371–1430 (1978).
11. D. J. Thomson, *Proc. IEEE* **70**, 1055–1096 (1982).
12. C. E. McCulloch, and S. R. Searle, *Generalized, Linear, and Mixed Models*, Wiley, 2001, ISBN 0-471-19364-X.

13. C. E. Rasmussen, and C. K. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, The MIT Press, Cambridge, MA, USA, 2006.
14. H. B. Nielsen, UCMINF - an algorithm for unconstrained, nonlinear optimization, Tech. Rep. Report IMM-REP-2000-19, Technical University of Denmark (2000).
15. B. Blankertz, K.-R. Müller, G. Curio, T. Vaughan, G. Schalk, J. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroder, and N. Birbaumer, *Biomedical Engineering, IEEE Transactions on* **51**, 1044–1051 (2004).
16. G. Pfurtscheller, C. Brunner, A. Schlögl, and d. Lopes, *Neuroimage* **31**, 153–9 (2006).