Mixed-norm regularization for Event-Related Potential based Brain-Computer Interfaces

Rémi Flamary
Joint work with: Alain Rakotomamonjy

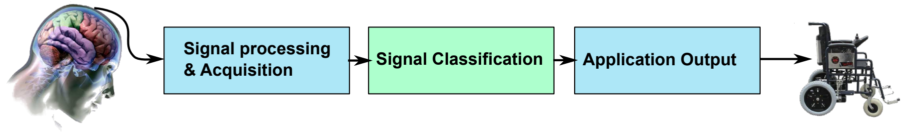
LITIS EA 4108, INSA-Université de Rouen
76800 Saint Etienne du Rouvray, France

April 19, 2011

Table of contents

Section

Brain-Computer Interfaces



Aim

Providing a direct communication channel between the human brain and an external device.

Challenges

- Providing robust classifiers.
- Learning quickly (time and learning examples).

BCI Types

- Motor Imagery.
- *Event-Related Potential*.

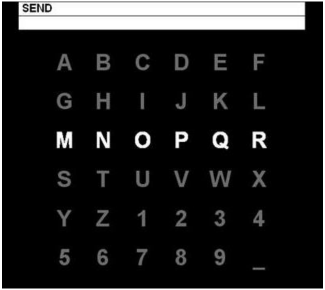
Event Related Potential (ERP)

ERP-based BCI [Luck, 2005]

- ERP: signal emitted by the brain after a given event occurs.
- Recording done with ElectroEncephalograms: noisy signal.
- Usually linear classifiers are sufficient.

P300 Speller

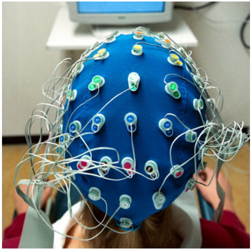
- P300 ERP occurs 300 ms after a rare event.
- The subject focuses on a letter.
- The columns and lines of the keyboard are flashed randomly.
- P300 appears when the column/line is flashed.
- The classifier output for all colmuns/lines are added in order to find the selected letter.



Sensor selection

Why ?

- All sensors are not relevant.
- Reduce implementation cost (short setup time, smaller EEG cap).



How is it done?

- Prior knowledge (discriminant areas of the brain).
- Recursive Feature Elimination (RFE) maximizing performances through Cross-Validation [Rakotomamonjy and Guigue, 2008].
- RFE using a relevance criterion (SSNR) [Rivet et al., 2010].
- Discriminant framework with sparsity inducing regularization [Tomioka and Müller, 2010].

Multi-task learning

Why?

- In BCI, learning a classifier for one subject is one task.
- A way to transfer knowledge between subjects (transfer learning).
- Good results obtained for Motor Imagery in BCI [Alamgir et al., 2010].
- Better performances for a small number of training samples.

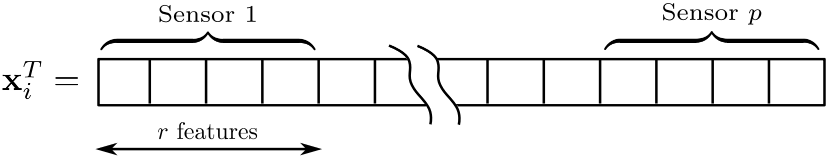
How is it done?

Learning jointly all the tasks and promoting similarity between them by:

- Minimizing the variance of the classifiers [Evgeniou and Pontil, 2004].
- Forcing the classifier to lie on a low dimensional space [Argyriou et al., 2008].
- Selecting jointly the relevant features [Rakotomamonjy et al., tted].

Section

Definitions for sensor selection



Learning set

- $\{\mathbf{x}_i, \mathbf{y}_i\}_{i\in\{1...n\}}$ the n training examples.
- $\mathbf{x}_i \in \mathbb{R}^d$ with $d = r \times p$ (r temporal features for each of the p sensors)

Linear classifier

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b \tag{1}$$

with $\mathbf{w} \in \mathbb{R}^d$ the separating hyperplane and $b \in \mathbb{R}$ the bias term.

Optimization framework

Discriminative framework

$$\min_{\mathbf{w},b} \quad \sum_{i}^{n} L(\mathbf{y}_i, \mathbf{x}_i^T\mathbf{w} + b) + \lambda\Omega(\mathbf{w}) \qquad (2)$$

where:

- $L(\cdot,\cdot)$ is a loss function measuring the discrepancy between actual and predicted labels.
- In this work, $L(y,\hat{y}) = max(0, 1 - y\hat{y})^2$ is the squared hinge loss.
- $\Omega(\cdot)$ is the regularization term.
- Regularization controlled by λ.

Regularization term

- Avoid over-fitting.
- Select relevant sensors through sparsity.

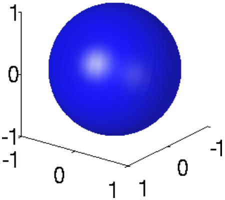
Regularization terms I

ℓ_2 – norm

$$\Omega_2(\mathbf{w}) = ||\mathbf{w}||_2^2$$

Where $||\cdot||_2$ is the euclidean norm.

- Not sparse.
- All components are regularized independently.

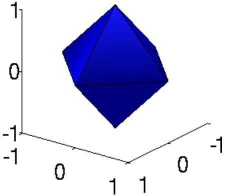


Feasible region for $\Omega_2(\mathbf{w}) \leq 1$

ℓ_1 – norm

$$\Omega_1(\mathbf{w}) = \sum_{i=1}^{d} |\mathbf{w}_i|$$

- Sparsity on the features of $\mathbf{w}$.
- All components are regularized independently.



Feasible region for $\Omega_1(\mathbf{w}) \leq 1$

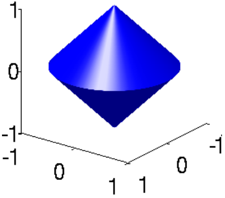
Regularization terms II

$\ell_1 - \ell_p$ mixed norm

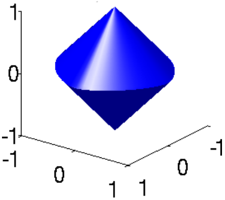
$$\Omega_{1-p}(\mathbf{w}) = \sum_{g \in \mathcal{G}} ||\mathbf{w}_g||_p$$

where $\mathcal{G}$ contains non-overlapping groups of $\{1..d\}$, $1 \leq p \leq 2$ and $||\mathbf{x}||_p = \left(\sum_i \mathbf{x}_i^p\right)^{1/p}$.

- ℓ_1 norm on the vector containing the ℓ_p norm of each group.
- p controls regularization between $\ell_1 - \ell_1 = \ell_1$ and $\ell_1 - \ell_2$ also known as group-lasso.
- We group the features by sensor.



Feasible region for $\Omega_{1-2}(\mathbf{w}) \leq 1$



Feasible region for $\Omega_{1-p}(\mathbf{w}) \leq 1$

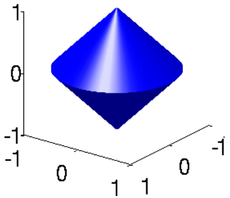
Regularization terms III

Adaptive $\ell_1 - \ell_2$ mixed norm

$$\Omega_{a:1-2}(\mathbf{w}) = \sum_{g \in \mathcal{G}} \beta_g ||\mathbf{w}_g||_2$$

where the weights β_g are selected to enhance sparsity.

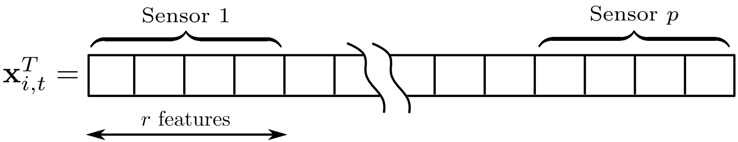
- Problem solved with $\beta_g = 1$.
- Then problem is solved iteratively with $\beta_g = 1/||\mathbf{w}_g^*||_2$, $\mathbf{w}^*$ being the optimal classifier from last iteration.
- Stop when convergence or after max number of iterations.
- Sparser results as groups with small norms are more penalized.
- Better theoretical properties [Bach, 2008].



Feasible region for $\Omega_{1-2}(\mathbf{w}) \leq 1$

Similar for $\Omega_{a:1-2}(\mathbf{w})$ with a scaling β_i on each dimension.

Definitions for multi-task learning



Learning set

- $\{\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\}_{i\in\{1...n\}}$ for each task $t \in 1 \ldots m$.
- $\mathbf{x}_{i,t} \in \mathbb{R}^d$ with $d = r \times p$ (r temporal features for each of the p sensors)

Tasks

- One task per subject.
- We learn jointly $(\mathbf{w}_t, \mathbf{b}_t)$ for each task t.

Optimization framework for MTL

Discriminative framework for MTL

$$\min_{\mathbf{W},\mathbf{b}} \quad \sum_{t}^{m}\sum_{i}^{n} L(\mathbf{y}_{i,t}, \mathbf{x}_{i,t}^T\mathbf{w}_t + \mathbf{b}_t) + \Omega_{\text{mtl}}(\mathbf{W}) \tag{3}$$

where:

- $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_m] \in \mathbb{R}^{d\times m}$ is a matrix concatenating all the classifiers.
- $\Omega_{\text{mtl}}(\mathbf{W})$ is the regularization term.

Regularization term

- Avoid over-fitting.
- Select relevant sensors through sparsity.
- Promote similarity between tasks.

MTL Regularization

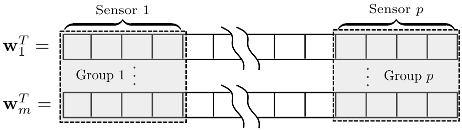
Regularization term

$$\Omega_{\text{mtl}}(\mathbf{W}) = \lambda_r \underbrace{\sum_{g \in \mathcal{G}'} ||\mathbf{W}_g||_2}_{\text{Mixed norm}} + \lambda_s \underbrace{\sum_{t=1}^{m} ||\mathbf{w}_t - \hat{\mathbf{w}}||_2^2}_{\text{Similarity}} \quad (4)$$

where λ_r and λ_s weight the mixed norms and similarity regularization.

Mixed norm

$\mathcal{G}'$ contains groups of sensors in $\mathbf{W}$:



Similarity

- $\hat{\mathbf{w}} = \frac{1}{m}\sum_t \mathbf{w}_t$ is the average classifier across tasks
- Minimize the variance of the classifiers [Evgeniou and Pontil, 2004].

Algorithm

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [Beck and Teboulle, 2009]

Can be used whenever the objective function can be expressed as:

$$f_1(\mathbf{w}) + f_2(\mathbf{w})$$

with:

- $f_1(\cdot)$ a gradient Lipschitz continuous term.
- $f_2(\cdot)$ a non-differentiable term having a closed form proximal operator:

$$Prox(\mathbf{v}) = \text{argmin}_{\mathbf{w}} \|\mathbf{v} - \mathbf{w}\|^2 + f_2(\mathbf{w})$$

Advantages

- Simple and efficient algorithm.
- Convergence properties.
- Fast regularization path thanks to warm-start.

Algorithmic implementation

Sensor selection problem

- $f_1(\mathbf{w}) = \sum_i^n L(\mathbf{y}_i, \mathbf{x}_i^T \mathbf{w} + b)$, with the squared hinge loss is gradient Lipschitz continuous term.
- $f_2(\mathbf{w}) = \Omega(\mathbf{w})$, has a closed form proximal for the proposed regularization terms. Example for the ℓ_1 norm:

$$Prox_{\Omega_1}(\mathbf{v})_i = \begin{cases} 0 & if \quad |\mathbf{v}_i| \leq \lambda \\ \mathbf{v}_i - \lambda sign(\mathbf{v}_i) & if \quad |\mathbf{v}_i| > \lambda \end{cases}$$

Multi-task problem

- $f_1(\mathbf{w}) = \sum_{t,i}^{m,n} L(\mathbf{y}_{i,t}, \mathbf{x}_{i,t}^T \mathbf{w}_t + \mathbf{b}_t) + \sum_t^m ||\mathbf{w}_t - \hat{\mathbf{w}}||_2^2$, that is provably gradient Lipschitz continuous.
- $f_2(\mathbf{w}) = \Omega_{1-2}(\mathbf{W})$, has a closed form proximal for the proposed regularization terms. Example for the $\ell_1 - \ell_2$ norm:

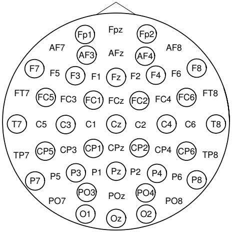
$$Prox_{\Omega_{1-2}}(\mathbf{v})_g = \begin{cases} \mathbf{0} & if \quad ||\mathbf{v}_g||_2 \leq \lambda \\ \mathbf{v}_g(1 - \frac{\lambda}{||\mathbf{v}_g||_2}) & if \quad ||\mathbf{v}_g||_2 > \lambda \end{cases}$$

Section

P300 datasets

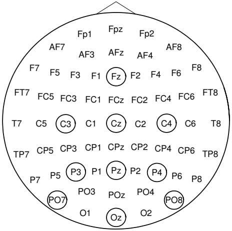
EPFL Dataset [Hoffmann et al., 2008]

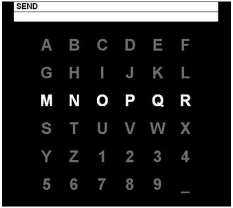
- P300 with 3 × 2 image selection.
- 8 subjects.
- 32 electrodes.
- 3000 examples, 1000 for training/validation.



UAM Dataset [Ledesma Ramirez et al., 2010]

- P300 Speller with standard 6 × 6 virtual keyboard.
- 30 subjects.
- 10 electrodes.
- 3000 examples, 1000 for training/validation.





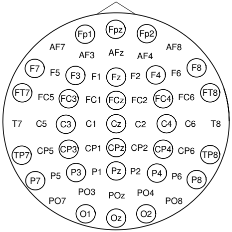
Error Related Potential dataset

Experimental setup

- Subjects asked to memorize the position of 2 to 9 digits.
- They had to recall the position of one of these digits.
- Signal recorded after the visualization of the result (correct/error) .

Dataset

- ErrP Event Related Potential.
- 8 subjects.
- 31 electrodes.
- 72 examples, 57 for training/validation.



Methods evaluation

Sensor selection methods

Method	Reg.	Groups
SVM	ℓ_2	-
SVM-1	ℓ_1	feature
GSVM-2	$\ell_1 - \ell_2$	sensor
GSVM-p	$\ell_1 - \ell_p$	sensor
GSVM-a	Adapt. $\ell_1 - \ell_2$	sensor

- Classification performance measured with Area Under the ROC Curve.
- Groups correspond to sensors.
- Dataset randomly split ($10\times$).
- λ selected through Cross-Validation.

Multi-task methods

Method	Reg.	Groups
SVM-Full	ℓ_2	-
MGSVM-2	$\ell_1 - \ell_2$	sensor
MGSVM-2s	$\ell_1 - \ell_2$ and Sim.	sensor

- Classification performance measured with Area Under the ROC Curve.
- Groups correspond to sensors (across tasks).
- Use a small number of examples .
- λ_r and λ_s selected through Cross-Validation.

Classification performances for P300

Datasets	EPFL Dataset (8 Sub., 32 Ch.)			UAM Dataset (30 Sub., 10 Ch.)		
Methods	Avg AUC	Avg Sel	p-value	Avg AUC	Avg Sel	p-value
SVM	80.35	100.00	-	84.47	100.00	-
SVM-1	79.88	87.66	0.15	84.45	96.27	0.5577
GSVM-2	80.53	78.24	0.31	84.94	88.77	0.0001
GSVM-p	80.38	77.81	0.74	84.94	90.80	0.0001
GSVM-a	79.01	26.60	0.01	84.12	45.07	0.1109

Performance Results

- AUC, percent of selected sensors and Signrank Wilcoxon test p-value.
- GSVM-2 gives the best performance but uses 80-90% of the sensors.
- GSVM-a provides the best selection with a slight performance loss.
- Some subjects in UAM dataset perform poorly for all methods ($< 60\%$ AUC).

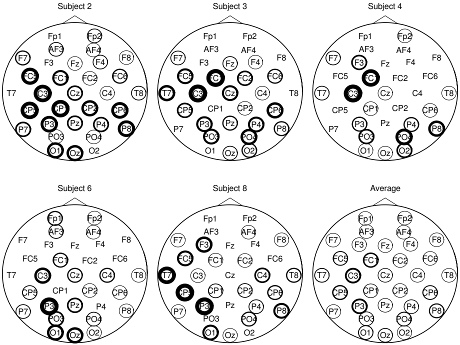
Classification performances for Error Related Potential

Datasets	ErrP Dataset (8 Sub., 32 Ch)		
Methods	Avg AUC	Avg Sel	p-value
SVM	76.96	100.00	-
SVM-1	68.84	45.85	0.3125
GSVM-2	77.29	29.84	0.5469
GSVM-p	76.84	37.18	0.7422
GSVM-a	67.25	7.14	0.1484

Performance Results

- AUC, percent of selected sensors and Signrank Wilcowon test p-value.
- GSVM-2 gives the best performance with 30% of the sensors.
- GSVM-a is statistically equivalent to SVM but loses 10% AUC.
- Difficult to select the regularization parameter on 57 examples!

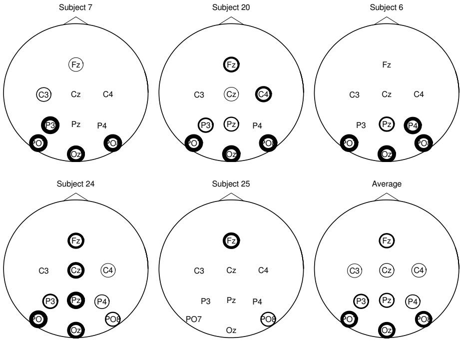
Selected sensors for EPFL Dataset



Results for GSVM-a

- Selected sensors are highly dependent on the subject.
- Sensors from the occipital area [Krusienski et al., 2008].
- And other areas such as T7 and C3 [Rivet et al., 2010, Rakotomamonjy and Guigue, 2008].

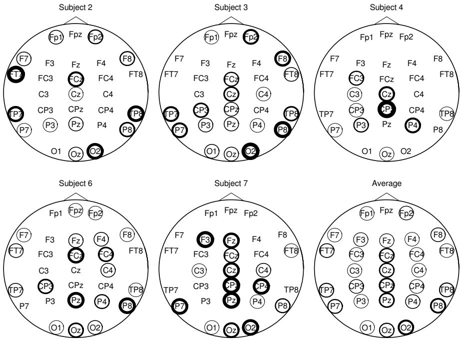
Selected sensors for UAM dataset



Results for GSVM-a

- Classical P300 experimental setup.
- Less sensors selected.
- Sensors from the occipital area [Krusienski et al., 2008].

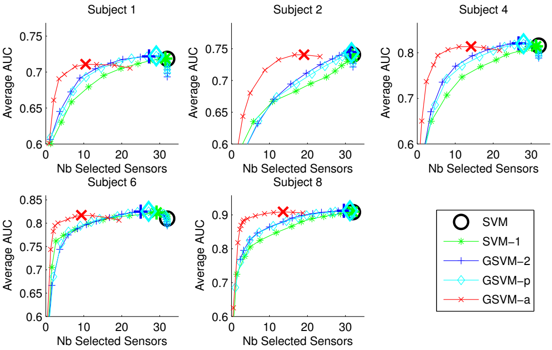
Selected sensors for ErrP dataset



Results for GSVM-2

- Important variances across subjects.
- Sensors in the central area selected in average [Dehaene et al., 1994].
- Small dataset.

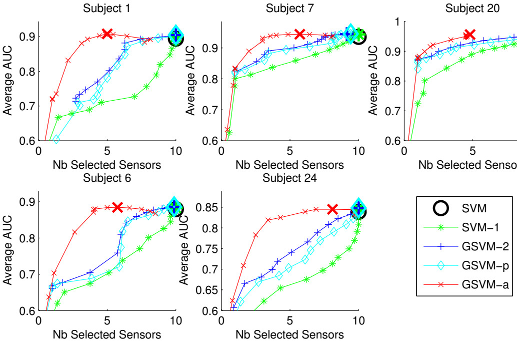
Sensor selection performance for EPFL Dataset



Results

- Performance vs sparsity plots.
- GSVM-a clearly outperforms the other methods for sensor selection.

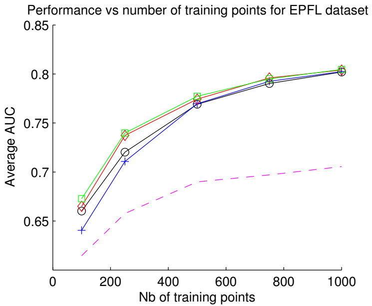
Sensor selection performance for UAM Dataset

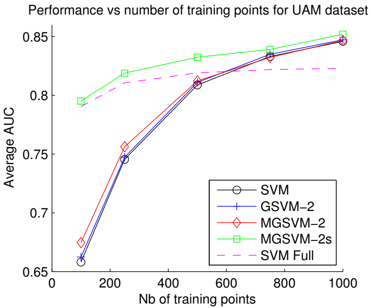


Results

- Performance vs sparsity plots (10 sensors).
- GSVM-a clearly outperforms the other methods for sensor selection.

Multi-task learning Results





Results

- Average performances for different number of training examples.
- MTL regularization leads to the best results.
- Promoting similarity drastically improves performances for UAM.

MTL results for difficult subjects

Method	Sub. 28	Sub. 25	Sub. 4	Sub. 8
SVM	0.5492	0.5643	0.6559	0.7198
MGSVM-2s	0.6417	0.6507	0.7144	0.7725

Results

- Average AUC for the most difficult subjects of the UAM dataset.
- 500 training/validation examples.
- Performance gain up to 15 % AUC.
- Ability to handle better "BCI illiteracy".

Section

Conclusion

This work

- Discriminative optimization framework for sensor selection and multi-task learning.
- Comparison on several Datasets.
- Group-lasso for classification performances.
- Adaptive Group-lasso for sensor selection.
- Multi-task learning when small number of training examples available.

Future works

- Investigate different groups for MTL.
- Automatically perform pre-processing through sparsity.

Section

References I

[Alamgir et al., 2010] Alamgir, M., Grosse-Wentrup, M., and Altun, Y. (2010).
Multi-task Learning for Brain-Computer Interfaces.
In *AI & Statistics*.

[Argyriou et al., 2008] Argyriou, A., Evgeniou, T., and Pontil, M. (2008).
Convex multi-task feature learning.
Machine Learning, 73(3):243–272.

[Bach, 2008] Bach, F. (2008).
Consistency of the group Lasso and multiple kernel learning.
Journal of Machine Learning Research, 9:1179–1225.

[Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009).
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2:183–202.

[Dehaene et al., 1994] Dehaene, S., Posner, M., and Tucker, D. (1994).
Localization of a neural system for error detection and compensation.
Psychological Science, 5(5):303–305.

References II

[Evgeniou and Pontil, 2004] Evgeniou, T. and Pontil, M. (2004).
Regularized multi-task learning.
In *Proceedings of the tenth Conference on Knowledge Discovery andData minig*.

[Hoffmann et al., 2008] Hoffmann, U., Vesin, J., Ebrahimi, T., and Diserens, K. (2008).
An efficient p300-based brain-computer interface for disabled subjects.
Journal of Neuroscience Methods, 167(1):115–125.

[Krusienski et al., 2008] Krusienski, D., Sellers, E., McFarland, D., Vaughan, T., and Wolpaw, J. (2008).
Toward enhanced P300 speller performance.
Journal of neuroscience methods, 167(1):15–21.

[Ledesma Ramirez et al., 2010] Ledesma Ramirez, C., Bojorges Valdez, E., Yáñez Suarez, O., Saavedra, C., Bougrain, L., and Gentiletti, G. G. (2010).
An Open-Access P300 Speller Database.
In *Fourth International Brain-Computer Interface Meeting*.

[Luck, 2005] Luck, S. (2005).
An introduction to the event-related potential technique.
MIT Press.

References III

[Rakotomamonjy et al., tted] Rakotomamonjy, A., Flamary, R., Gasso, G., and Canu, S. (Submitted).
$\ell_p - \ell_q$ penalty for sparse linear and sparse multiple kernel multi-task learning.
IEEE Trans. Neural Networks.

[Rakotomamonjy and Guigue, 2008] Rakotomamonjy, A. and Guigue, V. (2008).
BCI competition III: Dataset II - ensemble of SVMs for BCI P300 speller.
IEEE Trans. Biomedical Engineering, 55(3):1147–1154.

[Rivet et al., 2010] Rivet, B., Cecotti, H., Phlypo, R., Bertrand, O., Maby, E., and Mattout, J. (2010).
EEG sensor selection by sparse spatial filtering in P300 speller brain-computer interface.
In *Proc. EMBC nt. Conf. IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, pages –.

[Tomioka and Müller, 2010] Tomioka, R. and Müller, K. (2010).
A regularized discriminative framework for EEG analysis with application to brain-computer interface.
NeuroImage, 49(1):415–432.