Channel Selection for Brain Signal Classification by Penalized Automatic Relevance Determination

Reo Togashi and Yoshikazu Washizawa,
The University of Electro-Communications, Tokyo, Japan.
E-mail: t1431074@edu.cc.uec.ac.jp and washizawa@uec.ac.jp

***Abstract*—Channel selection or reduction in Brain computer interface (BCI) is important to reduce the cost and improve the generalized accuracy. A channel selection method using group automatic relevance determination (GARD) for P300 based BCI has been reported. In this paper, we apply the penalized ARD (PARD) which is an extension of ARD, and compare with GARD in our auditory BCI. Experimental results show that PARD provides more sparse solution than GARD while PARD shows almost the same classification accuracy as GARD.**

I. Introduction

Brain computer interface (BCI) enables disabled patients to communicate with people or a device [1], [2]. P300 which is a component of the event-related potential (ERP) and evoked about 300 ms after the presentation of a low frequent stimulus, has been widely used for BCI [3], [4]. Fig. 1 is averaged target and non-target responses measured by electroencephalography (EEG). P300 is a peak around 500 ms. P300-based BCI detects this peak to estimate a subject's intention.

Brain signal is measured by multi-channel EEG. P300 is mainly observed around Pz. However, since its amplitude is small compared to background EEG or artifact, multiple electrodes are used to capture P300 [4], [5]. Additionally, the position of P300 depends on the subject's age, condition and the degree of concentration for the stimulus. Therefore we need to find the optimal number and the position of the electrodes for each subject [4], [6], [7], [8].

Previously the least absolute shrinkage and selection operator (Lasso) and Group Lasso (GL), which is group version of Lasso, have been proposed to find a sparse solution [9], [10]. However, these methods have a problem that the regularization parameter should be tuned. In general, the regularization parameter is selected by the cross-validation, which is wasteful in computation time. On the other hand, the automatic relevance determination (ARD), relevance vector machine (RVM), and grouped ARD (GARD) have been proposed to obtain a sparse solution from Bayesian framework [11], [12], [13]. Unlike Lasso or GL, these Bayesian-based approaches can obtain the sparse solution without tuning the regularization parameter by assuming the uniform prior distribution for the variance of the weight. Penalized ARD (PARD) obtains the sparse solution by using non-uniform prior for the variance of the weight [14]. PARD requires to tune a hyper-parameter as well as Lasso and GL.

In this paper, we compare GARD and PARD with respect to the classification accuracy and sparsity in four command auditory BCI. We found that PARD obtained more sparse weight than GARD while keeping classification accuracy. In other words, PARD achieves almost the same classification performance using smaller number of electrodes.

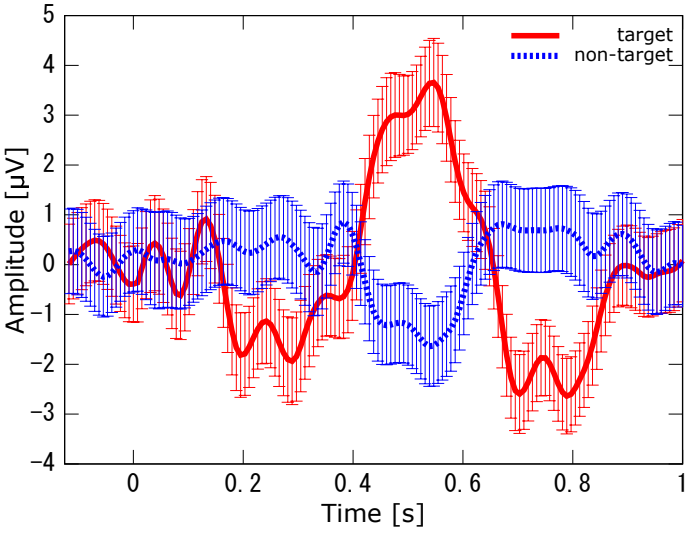

Fig. 1. Averaged event related potential signals of the target and non-target response. P300 is observed around 500 ms latency after the onset of the desired stimulus.

II. Algorithm of Penalized Automatic Relevance Determination

Let $\{(\boldsymbol{x}_n, t_n)\}_{n=1}^N$ be a data set, where $\boldsymbol{x}_n \in \mathcal{R}^d$ is the nth input vector and t_n is its output value, N is the number of samples and d is the number of dimensions. To consider sparse estimation by group, Both $\boldsymbol{x}$ and a weight vector $\boldsymbol{w} \in \mathcal{R}^d$ are partitioned into G groups,

$$\boldsymbol{x} = ({\boldsymbol{x}^{(1)}}^\top, \ldots, {\boldsymbol{x}^{(g)}}^\top, \ldots, {\boldsymbol{x}^{(G)}}^\top)^\top \quad (1)$$

$$\boldsymbol{w} = ({\boldsymbol{w}^{(1)}}^\top, \ldots, {\boldsymbol{w}^{(g)}}^\top, \ldots, {\boldsymbol{w}^{(G)}}^\top)^\top, \quad (2)$$

where G is the number of groups. Let us consider the linear regression model,

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{x} = \sum_{g=1}^{G} {\boldsymbol{w}^{(g)}}^\top \boldsymbol{x}^{(g)}. \quad (3)$$

We assume that t consists of $y(\boldsymbol{x}, \boldsymbol{w})$ and noise η,

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \eta = \sum_{g=1}^{G} \boldsymbol{w}^{(g)\top} \boldsymbol{x}^{(g)} + \eta, \tag{4}$$

and η follows the Gaussian distribution,

$$\eta_n \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where σ^2 is a variance of η. The probability density function (PDF) for t_n $(n = 1, \ldots, N)$ is given by

$$p(t_n|\boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2), \tag{6}$$

and the joint PDF for $\boldsymbol{t} = (t_1, \ldots, t_N)^\top$ is then given by

$$p(\boldsymbol{t}|\boldsymbol{w}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp(-\frac{1}{2\sigma^2}||\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}||^2), \tag{7}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top \in \mathcal{R}^{N \times d}$.

Next, let us introduce the prior distribution for $\boldsymbol{w}$. Assume that $\boldsymbol{w}_g$ also follows the zero-mean Gaussian distribution. The prior distribution for $\boldsymbol{w}_g$ $(g = 1, \ldots, G)$ is given by

$$p(\boldsymbol{w}^{(g)}|\lambda_g) = \mathcal{N}(w^{(g)}|\boldsymbol{0}, \lambda_g \boldsymbol{I}_{d_g}), \tag{8}$$

where λ_g is a variance of $\boldsymbol{w}^{(g)}$, $\boldsymbol{I}_a$ is $a \times a$ the identity matrix and d_g is the number of dimensions of $\boldsymbol{w}^{(g)}$ which satisfies $d = \sum_{g=1}^{G} d_g$.

In ARD and GARD, the prior distribution of λ is assumed to be uniform. On the other hand, PARD supposes that λ_g follows the gamma distribution whose shape parameter equals to one. The prior distribution of λ_g is

$$p_\gamma(\lambda_g) = \begin{cases} \gamma \exp(-\gamma\lambda_g) & (\lambda_g \geq 0) \\ 0 & (\lambda_g < 0) \end{cases} \tag{9}$$

where $\frac{1}{\gamma}$ is the scale parameter.

$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_G)^\top$ and σ^2 are estimated by the following point estimation problem,

$$\arg\max_{\boldsymbol{\lambda}, \sigma^2} \; J(\boldsymbol{\lambda}, \sigma^2) = p(\boldsymbol{\lambda}, \sigma^2|\boldsymbol{t}) \tag{10}$$

By using Bayes's theorem, the objective function is reduced to

$$\begin{aligned} J(\boldsymbol{\lambda}, \sigma^2) &= \int p(\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2|\boldsymbol{t}) d\boldsymbol{w} \\ &= \frac{p(\boldsymbol{\lambda}, \sigma^2)}{p(\boldsymbol{t})} \int p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2) d\boldsymbol{w}. \end{aligned} \tag{11}$$

We assume that $\boldsymbol{\lambda}$ and σ^2 are independent, that is

$$p(\boldsymbol{\lambda}, \sigma^2) = p(\boldsymbol{\lambda}) p(\sigma^2). \tag{12}$$

Then we have

$$J(\boldsymbol{\lambda}, \sigma^2) = \frac{p(\boldsymbol{\lambda}) p(\sigma^2)}{p(\boldsymbol{t})} \int p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2) d\boldsymbol{w}. \tag{13}$$

Since $p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) = p(\boldsymbol{t}|\boldsymbol{w}, \sigma^2)$ and $p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2) = p(\boldsymbol{w}|\boldsymbol{\lambda})$, the natural logarithm of the integral part is computed as follows,

$$\begin{aligned} &\log \int p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2) d\boldsymbol{w} \\ &= \frac{d - N}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t}, \end{aligned} \tag{14}$$

where

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_N + \boldsymbol{X} \boldsymbol{\Lambda} \boldsymbol{X}^\top \tag{15}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}([\lambda_1, \ldots, \lambda_1, \lambda_2, \ldots, \lambda_2, \lambda_G, \ldots, \lambda_G]) \tag{16}$$

Finally, the optimization problem is reduced to

$$\arg\min_{\boldsymbol{\lambda}, \sigma^2} \; -\log J(\boldsymbol{\lambda}, \sigma^2) \tag{17}$$

$$= \arg\min_{\boldsymbol{\lambda}, \sigma^2} \; \frac{1}{2} \log|\boldsymbol{\Sigma}| + \frac{1}{2} \boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t} + \gamma \sum_{g=1}^{G} \lambda_g. \tag{18}$$

The optimization problem (18) is reduced to GARD if $\gamma = 0$. $\boldsymbol{\lambda}$ and σ^2 are optimized by the gradient method.

After $\boldsymbol{\lambda}$ and σ^2 are obtained, $\boldsymbol{w}$ is obtained using Bayes's theorem,

$$\max_{\boldsymbol{w}} \; p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2, \boldsymbol{t}) = \max_{\boldsymbol{w}} \; \frac{p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2)}{p(\boldsymbol{t}|\boldsymbol{\lambda}, \sigma^2)}. \tag{19}$$

Since $p(\boldsymbol{t}|\boldsymbol{\lambda}, \sigma^2)$ is a constant for $\boldsymbol{w}$, the optimization problem (19) is simplified to

$$\max_{\boldsymbol{w}} \; p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\lambda}, \sigma^2) \tag{20}$$

$$= \max_{\boldsymbol{w}} \; \exp(-\frac{1}{2}\{\boldsymbol{w}^\top (\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Lambda}^{-1}) \boldsymbol{w} - 2\sigma^{-2} \boldsymbol{t}^\top \boldsymbol{X}^\top \boldsymbol{w}). \tag{21}$$

The optimal $\boldsymbol{w}$ is derived from (21) by maximizing the index part, that is

$$\begin{aligned} \hat{\boldsymbol{w}} &= \sigma^{-2} (\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{X}^\top \boldsymbol{t} \\ &= \boldsymbol{\Lambda} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t} \end{aligned} \tag{22}$$

III. Application for BCI

Let $x_{n,c}(i)$ $(n = 1, ..., N, c = 1, \ldots, N_{ch}, i = 1, \ldots, T)$ and $t_n \in \{-1, 1\}$ be an observed signal and label respectively, where N is the number of trials, N_{ch} is the number of channels and T is the number of sampling points. The feature vector is defined by

$$\boldsymbol{z}_n = (\boldsymbol{x}_{n,1}^\top, \ldots, \boldsymbol{x}_{n,c}^\top, \ldots, \boldsymbol{x}_{n,N_{ch}}^\top)^\top, \tag{23}$$

where $\boldsymbol{x}_{n,c} = [x_{n,c}(0), \ldots, x_{n,c}(T-1)]^\top$. To perform the channel selection/reduction with PARD and GARD, G is set to the number of channels ($G = N_{ch}$). The estimated label $\hat{t}$ is obtained by classifying the feature vector $\boldsymbol{z}$ using $\boldsymbol{w}$ as follows,

$$\begin{cases} \hat{t} = 1 & (\boldsymbol{w}^\top \boldsymbol{z} > 0) \\ \hat{t} = -1 & (\boldsymbol{w}^\top \boldsymbol{z} < 0) \end{cases} \tag{24}$$

The algorithm procedure is summarized in Fig. 2.

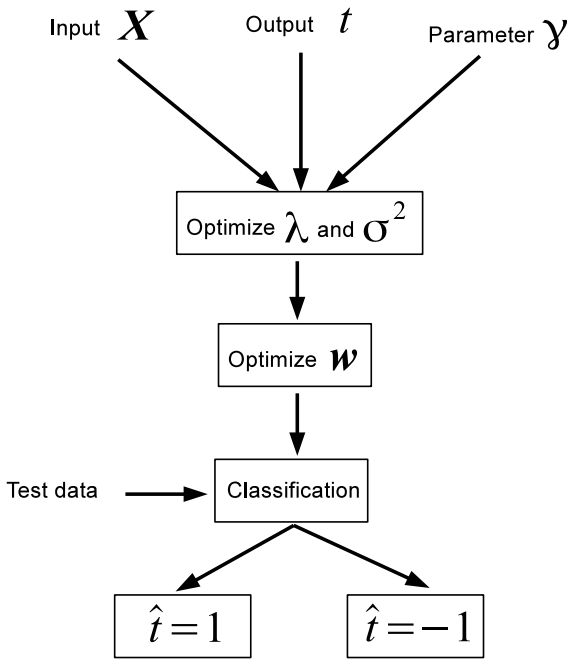

Fig. 2. Summary of the PARD.

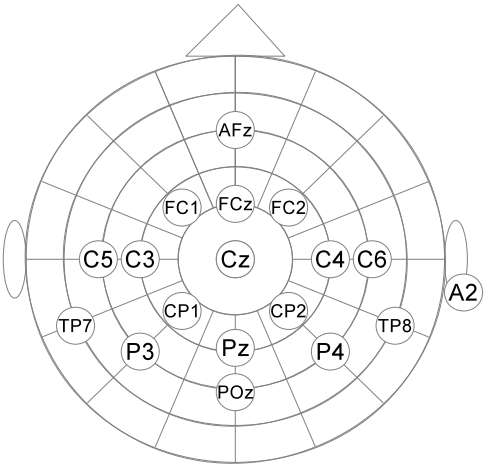

Fig. 3. Location of electrode.

IV. Experiment

The brain signal was recorded by an active 16ch EEG system (g.GAMMAcap2, g.LADYbird (active), g.GAMMAbox) produced by Guger technologies, and amplified by a biological signal amplifier (BA 1008, Digitex). The stimulus trigger was generated from the audio interface, and electrically recorded by an AD converter. The electrodes were located in FCz, FC2, FC1, Cz, CP1, CP2, Pz, POz, P3, P4, TP8, TP7, C3, C4, C5 and C6, and the ground was AFz and the reference was A2 (Fig. 3). The electrodes were arranged to capture ERP around Pz and the area of the temporal lobe related to cognition.

Four speech stimuli, “jou,” “ge,” “sa,” and “yu” were used. They respectively mean “up,” “down,” “left,” and “right” in Japanese. These stimuli were randomly given from one of four

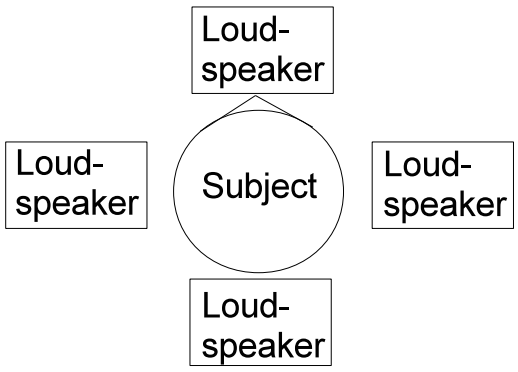

Fig. 4. Location of loud speaker.

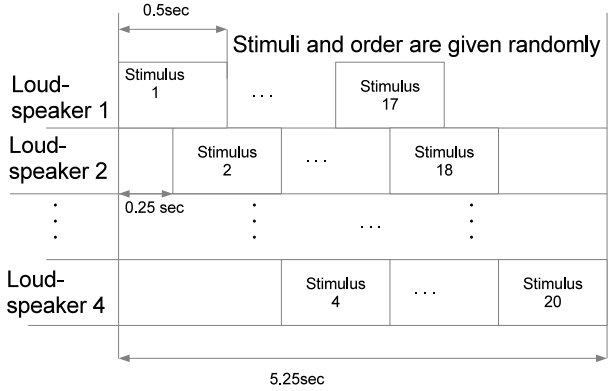

Fig. 5. Stimulus presentation of the experiment.

loudspeakers located back, front, left, and right (Fig. 4). 20 stimuli were presented to a subject in one trial. Each speech stimulus was presented five times. The order of stimulus presentation and the kind of stimulus were random. However, the same stimulus was not continuously presented, and the loudspeaker did not sequentially give the stimulus. Fig. 5 shows the presentation scheme. Each subject conducted 50 trials. Five healthy subjects agreed to take part in the experiment and signed the consent form approved by the research ethics committee of The University of Electro-Communications. The participants were received instructions which stimulus to focus on by a monitor in each trial. During the experiment, they were required to attend the target stimulus, close their eyes, and count the number of the target stimulus when the desired stimulus was presented.

A high-pass analog filter with cutoff frequency of 0.5 Hz and a low-pass analog filter with cutoff frequency of 100 Hz were applied by amplifier. The sampling frequency was 512 Hz. A digital band-pass filter with cutoff frequency of 1 Hz and 12 Hz was applied for the recorded EEG. EEG signal from 0 ms to 750 ms after the onset of the stimulus was extracted to classify.

V. Result

Five-fold cross validation procedure was applied to evaluate the BCI classification accuracy and sparsity. Fig. 6 shows

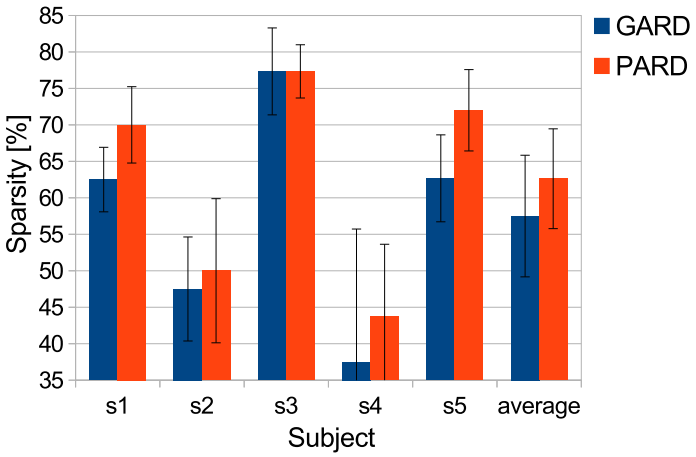

Fig. 6. The result of sparsity between GARD and PARD.

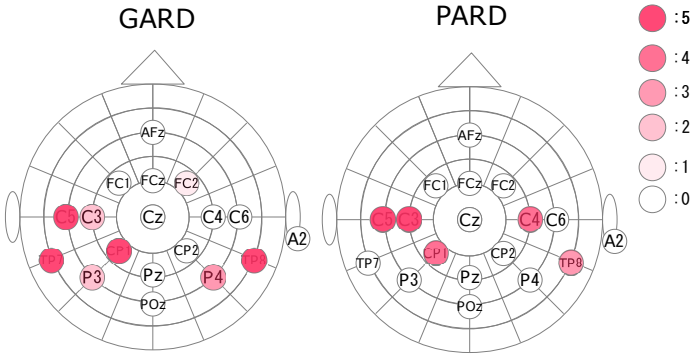

Fig. 7. Channel selection of subject 5 between GARD and PARD. FC1 was not used in this subject because it didn't work well.

result of the sparsity analysis of both GARD and PARD methods. The sparsity is measured by the ratio of zero components in weight $\boldsymbol{w}$. PARD shows higher sparsity than that of GARD for all subjects except for subject 3.

Fig. 7 shows a visualization of the channel selection/reduction results of the sparsity improvement best scoring subject number 5. The number that the electrode has a non-zero coefficient in the cross validation is represented by the depth of the color. The deeper color stands for frequently selected electrode. We can see that the frequently selected electrodes are similar in both GARD and PARD.

Fig. 8 shows the classification accuracies of both GARD and PARD. Although the number of electrodes for the classification was small in case of PARD method application, the final BCI classification accuracy of this method was as good as in the GARD case. From the above discussion, it is concluded that PARD method resulted with the similar classification accuracies using the smaller number of EEG electrodes as compared with GARD technique.

VI. Conclusion and Future Work

PARD was applied to auditory BCI and compared with GARD with respect to the classification accuracy and sparsity. We found that PARD shows more sparse solution and almost the same classification accuracy compared with GARD.

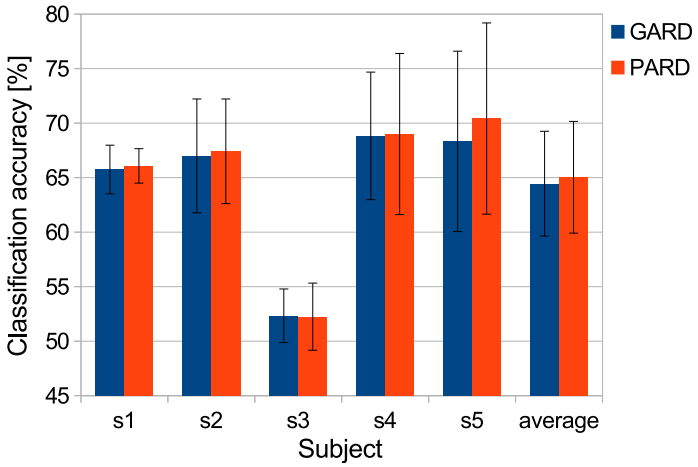

Fig. 8. The result of classification accuracy between GARD and PARD.

As a future work, we will obtain more sparsity and classification accuracy by changing the prior of the variance for weight. The variances are assumed to be independent in the model of our prior. However this assumption is unreasonable since brain signals measured from close electrodes are expected to have larger correlation [15]. Therefore the performance will be improved by introducing additional information related to electrode's position as a correlated prior information.

Acknowledgment

The auditory experiment environment was provided by Shuho Yoshimoto. This work was supported by Grant-in-Aid for Scientific research (B) 26280054.

References

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: communication and restoration of movement in paralysis," *The Jounal of Physiology*, vol. 579, no. 3, pp. 621–636, 2007.

[3] O. I. Khan, S.-H. Kim, T. Rasheed, A. Khan, and T.-S. Kim, "Extraction of P300 using constrained independent component analysis," *31st Annual International Conference of the IEEE EMBS*, pp. 4031–4034, 2009.

[4] T. Hruby and P. Marsalek, "Event-related potentials - the P3 wave," *Acta Neurobiol. Exp.*, pp. 55–63, 2003.

[5] A. Mourauxa and G. D. Iannetti, "Across-trial averaging of event-related eeg responses and beyond," *Magnetic Resonance Imaging*, vol. 26, no. 7, pp. 1041–1054, 2008.

[6] B. M, R. Bouguerra, and T. Choufa, "Estimation of amplitude and latency changes of P300 response in real-time," *EUROCON 2005*, vol. 1, pp. 21–24, 2005.

[7] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, 2004.

[8] D. Jarchi, B. Makkiabadi, and S. Sanei, "Estimation of trial to trial of P300 subcomponents by coupled rao-blackwellised particle filtering," *IEEE/SP 15th Workshop on Statistical Signal Processing*, pp. 17–20, 2009.

[9] R. Tobshirani, "Regression shirinkage and selection via the LASSO," *Jounal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[10] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Jounal of the Royal Statistical Society. Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[11] D. J. MacKay, "Bayesian non linear modeling for the prediction competion," *ASHRAE Trans*, pp. 3704–3716, 1994.

[12] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Jounal of Machine Learning Research*, pp. 211–244, 2001.
[13] T. Yu, Z. Yu, Z. Gu, and Y. Li, "Grouped automatic relevance determination and its application in channel selection for P300 BCIs," *Neural Systems and Rehabilitation Engineering*, pp. 1–10, 2015.
[14] A. Aravkin, J. V. Burke, and A. C. G. Pillonetto, "Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ARD and Glasso," *Journal of Machine Learning Research*, pp. 217–252, 2014.
[15] H. Higashi and T. Tanaka, "Regularization using similarities of signals observed in nearby sensors for feature extraction of brain signals," *35th Annual International Conference of the IEEE EMBS*, pp. 7420–7423, 2013.