# Outlier Detection for Single-Trial EEG Signal Analysis

Boyu Wang, Feng Wan, Peng Un Mak, Pui In Mak, and Mang I Vai

*Abstract*—**The performance of a brain computer interface (BCI) system is usually degraded due to the outliers in electroencephalography (EEG) samples. This paper presents a novel outlier detection method based on robust learning of Gaussian mixture models (GMMs). We apply the proposed method to the single-trial EEG classification task. After trial-pruning, feature extraction and classification are performed on the subset of training data, and experimental results demonstrate that the proposed method can successfully detect the outliers and therefore achieve more reliable result.**

## I. Introduction

A brain computer interface (BCI) is a system forming a direct connection between brain and machine, which enables individuals with severe motor disabilities to have effective control over external devices without using the traditional pathways as peripheral muscle or nerves [1-2]. The brain activities are often recorded noninvasively by electroencephalogram (EEG), which has excellent temporal resolution and usability, and the EEG signal is therefore a popular choice for BCI research. In order to control an EEG-based BCI, the user must produce different brain activity patterns, which are recorded by electrodes on the scalp, and then features are extracted from EEG signals and translated into the control commands. In most existing BCIs, this translation relies on a classification algorithm [3].

Gaussian mixture models (GMMs) [4] have been applied to model the feature extracted from EEG signal analysis in BCI system. In [5] and [6], the mixture of Gaussian was introduced as the online classifier and the parameters were updated in a simulated online scenario. In [7] a GMM-based classifier was used to separate the signal into different classes of mental task, where adaptation is concerned by using a supervised method. Similarly, [8] and [9] also proposed an online GMM classifier via the decorrelated least mean square (DLMS) algorithm. On the other hand, GMMs can be also applied to model the features extracted from EEG data in which the rest or active state of brain signals are modeled so that the changes in EEG signal can be detected rather than classified [10], [11].

Noise is ubiquitous in EEG signals due to the factors such as measurement inaccuracies, physiological variations in background EEG, muscle and eyes blink artifacts. Therefore, contaminated samples in EEG data should be pruned to achieve a reliable classification result. Unfortunately, the estimation of the parameters of GMMs by the traditional algorithm (i.e., expectation-maximization, EM) is usually sensitive to the atypical samples, but none of the GMM-based EEG analysis algorithm, to our best knowledge, considered the negative effects of the outliers in EEG data.

The motivation of this work is twofold. First, we develop a robust learning algorithm for mixture models. On the other hand, it has been demonstrated that by proper feature extraction algorithms (e.g., common spatial pattern, CSP), the features of EEG signals are similar to normal distribution [16]. Therefore, we can apply the proposed outlier detection method for single-trial EEG analysis. In particular, we propose a deterministic annealing learning approach for robust fitting of GMMs. The GMMs are learned based on maximum trimmed likelihood (MTL) [13] via EM algorithm and the learning process is controlled by annealing temperatures, leading gradual optimization of the objective function, so that the outliers can be automatically detected and the estimation of parameters of GMMs is more robust and reliable. We apply the proposed method to single-trial EEG signal analysis task to prune the trials contaminated by artifacts, and the experimental results demonstrate robust performances of the proposed algorithm.

## II. Methods

### *A. Mixture Models, Trimmed Likelihood Estimator and FAST-TLE*

A mixture model with $K$ Gaussian densities can be defined as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p\left(\mathbf{x} \mid \boldsymbol{\theta}_k\right), \text{ with } 0 \le \pi_k \le 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \quad (1)$$

where $\pi_k$ is the mixing coefficient, and $\boldsymbol{\theta}_k$ is the set of parameters for the $k$th component.

Define $\boldsymbol{\Theta} \equiv \{\pi_1, ..., \pi_K, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ as the complete set of the parameters needed to specify the mixture model. The ML estimate of the optimal set of the parameters is defined as a maximum of the log-likelihood function:

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\theta}_k\right) \quad (2)$$

It is well known that the ML solution via EM algorithm is sensitive to the atypical observations in the data. To resolve this problem, robust estimation based on trimmed likelihood has been developed [13], and the trimmed log-likelihood function for mixture models is defined as

The authors are with the Department of Electrical and Electronics Engineering, Faculty of Science and Technology, University of Macau, Av. Padre Tomás Pereira,Taipa, Macau.

$$\log p_{TL}(\mathbf{X}\,|\,\boldsymbol{\Theta}) = \sum_{n=1}^{N}\omega_n \log\sum_{k=1}^{K}\pi_k p\left(\mathbf{x}_n\,|\,\boldsymbol{\theta}_k\right) \tag{3}$$

where $\omega_n \in \{0,1\}\ \forall\ n=1,...,N,$ and $\sum_{n=1}^{N}\omega_n = M.$ When $\omega_n = 1,$ it indicates that $\mathbf{x}_n$ is a typical sample. Otherwise, $\omega_n = 0$ and $\mathbf{x}_n$ is detected as a trimmed off samples.

The basic idea of MTL is to maximize the log-likelihood function of a selected subset consisting of *N-M* samples. It is based on the assumption that the outliers usually have lower likelihood than the typical observations. Hence, the main challenge for MTL is how to simultaneously select the subset of the data with maximum sum of likelihood values and to evaluate the parameters of the mixtures which maximize the corresponding log-likelihood. Although FAST-TLE has been proposed in [13] to handle this problem, this method is a local algorithm, and therefore highly depends on the initialization.

### *B. Deterministic Annealing Based Robust Learning*

The ultimate solution of $\omega_n$ is a "hard" solution in the sense that $\omega_n \in \{0,1\}, \forall\ n=1,...,N$. Such a constraint makes the objective function non-differentiable. To overcome this problem, we can resort the deterministic annealing (DA) approach, and consider the following objective function:

$$F(\boldsymbol{\Theta},\omega) = -\sum_{n=1}^{N}\omega_n \log p\left(\mathbf{x}_n\,|\,\boldsymbol{\Theta}\right) - TH_{\omega} \tag{4}$$

under the constraints of $\sum_{n=1}^{N}\omega_n = M$ , $\sum_{k=1}^{K}\pi_k = 1$ , and

$$H_{\omega} = -\sum_{n=1}^{N}\omega_n \log\omega_n \tag{5}$$

where *T* is the Lagrange multiplier, which is analogous to the temperature in statistics physics, $H_{\omega}$ is the Shannon entropy, which represents a specified level of randomness. At high value of *T*, the objective function is very smooth and we mainly maximize entropy, with $\sum_{n=1}^{N}\omega_n = M,$ yielding $\omega_n = M/N.$ As *T* is gradually lowered, the influence of log-likelihood function is increasing, which makes the solution of $\omega_n$ harder and harder. Finally, as *T* approaches to zero, the optimization is carried out directly on the trimmed log-likelihood function, forcing $\omega_n$ to either zero or one, which yields the MTL.

The motivation of DA learning procedure is that there is no guarantee that the selected subset in the early stage of learning is near the true one. Therefore, all of the samples should be equally treated at first, and the constraint is gradually relaxed during the learning process to increase the effect of the selection of subset, so that the global (at least a better local) optimal solution could be achieved.

To implement the proposed algorithm, an annealing loop is imposed outside the conventional EM loop, consisting of the E-step, and the modified M-step, which not only re-estimates the parameter set $\boldsymbol{\Theta}$ , but also updates the indicators $\{\omega_n\}$ . For GMMs, given the fixed $\{\omega_n\}$ , setting the derivatives of the objective function with respect to $\pi_k$ , $\boldsymbol{\mu}_k$ , and $\boldsymbol{\Sigma}_k$ to zero, under the constraints $\sum_{k=1}^{K}\pi_k = 1$ and $\sum_{n=1}^{N}\omega_n = M$ , we have

$$\pi_k = \frac{1}{M}\sum_{n=1}^{N}\omega_n p\left(k\,|\,\mathbf{x}_n\right), \tag{6}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\omega_n p\left(k\,|\,\mathbf{x}_n\right)\mathbf{x}_n}{\sum_{n=1}^{N}\omega_n p\left(k\,|\,\mathbf{x}_n\right)} \tag{7}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N}\omega_n p\left(k\,|\,\mathbf{x}_n\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^T}{\sum_{n=1}^{N}\omega_n p\left(k\,|\,\mathbf{x}_n\right)} \tag{8}$$

where

$$p\left(k\,|\,\mathbf{x}_n\right) = \frac{\pi_k N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \tag{9}$$

It is clear that the E-step of the proposed algorithm is the same as conventional EM algorithm. The main difference lies on the M-step where all of the parameters are updated via a weighted posterior, and the mixing coefficients are obtained by being divided by *M* rather than *N*. It can be observed that the larger the value of $\omega_n$ is, the more the corresponding sample contributes to the estimation of parameters. As *T* approaches to zero, $\omega_n$ skews either to one or zero, indicating whether the sample is viewed as typical or eliminated as an outlier.

For the parameters $\{\omega_n\}$, taking account of the constraint $\sum_{n=1}^{N}\omega_n = M$ , we minimize the following quantity

$$F + \beta\left(\sum_{n=1}^{N}w_n - M\right) \tag{10}$$

where $\beta$ is the Lagrange multiplier. Setting the derivative of (10) with respect to $\omega_n$ to zero and eliminate $\beta$, finally we obtain

$$w_n = \frac{Mp\left(x_n\,|\,\boldsymbol{\Theta}\right)^{\frac{1}{T}}}{\sum_{n=1}^{N}p\left(x_n\,|\,\boldsymbol{\Theta}\right)^{\frac{1}{T}}} \tag{11}$$

(11) does not take account of the constraint $\omega_n \subseteq [0,1]$ . For simplicity, we can just force $w_m = 1$ when $w_m > 1$, and then remove $\mathbf{x}_m$ from the data set and re-calculate the $\{w_n\}$, $\forall\ n=1,...,m-1,m+1...,N$ until all of the indicators satisfy the constraint $\omega_n \subseteq [0,1]$ and $\sum_{n=1}^{N}\omega_n = M.$ Although it may introduce some errors in the estimation of $\{\omega_n\}$ , the proposed algorithm still shows good performance in

the experimental result, as shown in the following section.

Clearly, the learning process consists of repeated E-step and modified M-step while lowering the temperature, and its monotonicity in objective function is obvious. When the temperature approaches to zero, the method degenerates to FAST-TLE algorithm, of which the monotonicity has been proved in [14]. The convergence property of deterministic annealing has also been discussed in [15].

## III. Experiments

In this section we apply our method on both synthetic data and EEG data. Before the experiments, some parameters of the proposed method should be specified carefully. Specifically, we choose the initial temperature as 100, the temperature scaling factor as 0.9, and the final temperature as 0.01.

### A. Synthetic Data

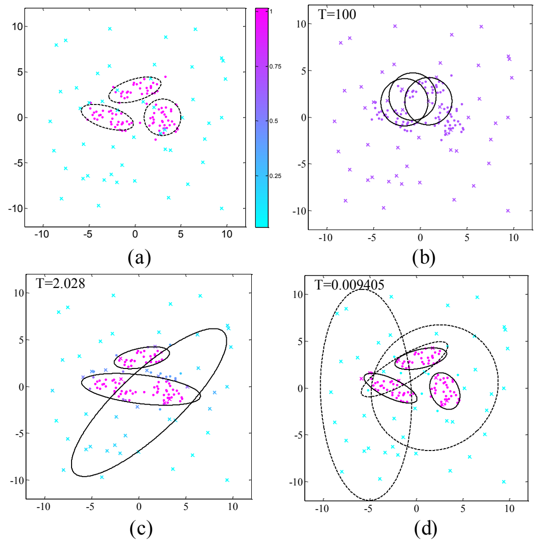
The first example is a synthetic dataset which consists of 100 samples from three two-dimensional Gaussian components with equal mixing coefficients, and the parameter set of each component is given by

$$\boldsymbol{\mu}_1 = [0,\ 3]^T,\ \boldsymbol{\mu}_2 = [3,\ 0]^T,\ \boldsymbol{\mu}_3 = [-3,\ 0]^T$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}$$

In addition, 50 noise points generated from a uniform distribution within [-10, 10] on each dimension are added to the typical samples, which is similar to the data set discussed in [13]. Thus, the total number of samples $N$ = 150, and the number of typical samples $M$ = 100. The obtained the samples, as well as the Gaussian components are shown in Fig. 1(a). The typical samples are marked by magenta dots, whereas the outliers are marked by cyan crosses. On the other hand, the colors of the observations also indicate the values of weights $\omega_n \in [0,1]$ which are represented by cyan when $\omega_n = 0$ (the corresponding sample is viewed as outliers) and by magenta when $\omega_n = 1$ (the corresponding sample is viewed as typical observation). The colors of samples vary smoothly from cyan to magenta as the values of $\{\omega_n\}$ approach from zero to one.

Fig. 1(b)-(c) demonstrate the learning process of proposed method. At the beginning, three components are randomly initialized among the samples, and the values $\{\omega_n\}$ at high temperature are almost same (equal to $M/N$ = 2/3, marked by purple dots or crosses). As T is lowered, the components converge to the true model, and the atypical observations are gradually detected and eliminated (depicted by the smooth varying of the colors from purple to cyan). Fig. 1(d) also shows the result of conventional EM, which is marked by dashed line. It can be observed that it fails to fit the samples due to the existing of outliers.



(a)

(b)

(c)

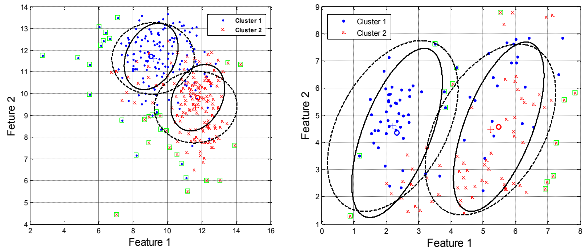
(d)

Fig. 1. Fitting a three-components Gaussian mixture with noise (typical samples are marked by magenta dots, and the outliers are marked by cyan crosses): (a) the true model shown by dashed line; (b)-(c) the learning process of the proposed method; (d) the results of the proposed method (solid lines) and the conventional EM algorithm (dashed lines).

### B. EEG Data

We consider the classification task of EEG signals, a non-invasive measurement of brain activities. In general, the EEG signals can be categorized as multi-channel and bi-channel EEG signals. For the multi-channel EEG data, features are usually extracted via spatial filtering techniques (e.g., CSP); for bi-channel signals, band power (BP) is usually used as the features. Because of the artifacts and other types of noise samples in EEG data, the classification result is usually unreliable. Fig. 2 [12] depicts the modeling result where EEG samples are clustered by GMMs with (solid lines) and without (dashed lines) outlier detection. The centers are marked with cross and circle respectively, and the outliers are indicated by the green panes. It can be observed that no matter which type of feature extraction algorithm is used, the covariances of the both clusters are enlarged due to the noisy data and outliers. In addition, the directions of the dispersions and the centers of each cluster are also drawn toward the noise samples. All of these factors may cause the negative effects on the model estimation and the further analysis, but such negative effects cannot be eliminated by feature extraction algorithms.

Now we apply our algorithm to the data set IIa from BCI competition IV [18], which consists of EEG data sets from 9 subjects. For each subject, two sessions were recorded, each of which consists of 288 trials with duration of 7s. In addition, the data set for each subject also contains some rejected trials, which are contaminated by noise or artifacts. For detailed description of this data set, see [18]. Before feature extraction, the EEG signals are filtered by 0.5-30Hz band pass filter. Then we apply CSP to extract the features from multichannel EEG



(a) (b)

Fig. 2. The clusters estimated by GMMs with and without outlier detection for features extracted by CSP (a) and BP (b) [12].

signals. Since the CSP is a data-driven feature extraction approach, after the elimination of noise samples, we re-train the CSP and GMM classifiers.

Table I. illustrates the average classification accuracies of the EEG signals of nine subjects obtained by conventional EM algorithm ($\epsilon$ = 0) and the proposed robust approach with different trimming levels. The classification accuracy of the EM algorithm is not improved significantly for the original EEG data. However, when the signals are contaminated by the rejected trials, the performance of conventional EM algorithm deteriorates obviously, whereas our proposed approach can detect and eliminate the outliers, so that more robust and reliable results can be obtained. In other words, the proposed algorithm can successfully reduce the negative effects of EEG signals contaminated by artifacts and noise.

## IV. Conclusions

In this paper, we proposed a novel DA based EM algorithm to detect outliers for single-trial EEG analysis. Experiments demonstrate that the performances of conventional learning approaches are significantly deteriorated due to the outliers while our method can successfully alleviate the negative effects of outliers. In addition, being applied on BCI system, the proposed method can automatically prune off the EEG signals contaminated by artifacts and noise without any additional channel rejection operation or visual inspection of an expert. Since the noise is ubiquitous in EEG signals, it is necessary to prune off a small account of samples to achieve reliable result even though the noise level is unknown.

The future work will focus on the reduction on the dependence on the prior knowledge of the trimming level. It should be noted that although our method is applied to Gaussian mixtures, it can be extended to non-Gaussian cases, which will be also considered in the future work.

## Acknowledgment

The authors gratefully acknowledge support from the Macau Science and Technology Department Fund (Grant FDCT/036/2009/A) and the University of Macau Research Fund (Grants RG059/08-09S/FW/FST and RG080/09-10S/WF/FST).

TABLE I
Classification Accuracies for the EEG Data Sets with and without Rejected Trials (%)

| Data Sets | Trimming Level $\epsilon$ (%) | | | |
|---|---|---|---|---|
| | 0 | 3 | 5 | 10 |
| Original Data | 73.3 | 74.1 | 75.2 | 74.8 |
| Data with Rejected Trials | 67.4 | 70.2 | 71.7 | 72.8 |

## References

[1] J. R. Wolpaw *et al.*, "Brain–computer interface technology: a review of the first international meeting," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 164–173, June 2000.

[2] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, pp. R32–R57, 2007.

[3] D. J. McFarland, C.W. Anderson, K.-R.Müller, A. Schlogl, and D. J. Krusienski, "BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 135-138, June 2006.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Press, New York, 2006.

[5] J. R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Brain-actuated interaction," *Artif. Intell.*, vol. 159, pp. 241-259, Nov. 2004.

[6] J. R. Millán, "On the need for on-line learning in brain–computer interfaces," in *Proc. Int. Joint Conf. Neural Networks*, Budapest, Hungary, 2004, pp. 2877–2882.

[7] A. Buttfield, and J. del R. Millán, "Online classifier adaptation in brain-computer interfaces," *Tech. Report*, IDIAP–RR 06-16, March 2006.

[8] S. Sun, C, Zhang, and N. Lu, "On the on-line learning algorithms for EEG signal classification in brain computer interfaces," *Lect Notes Comput Sci*, vol. 3614, pp. 638-647, 2005.

[9] S. Sun, and C, Zhang, "Learning on-line classification via decorrelated LMS algorithm: application to brain–computer interfaces," *Lect Notes Comput Sci*, vol. 3735, pp. 215-226, 2005.

[10] G. Schalk, P. Brunner, L. A. Gerhardt, H. Bischof, and J. R. Wolpaw, "Brain-computer interfaces (BCIs): Detection instead of classification," *Neurosci. Meth.*, vol. 167, pp. 51-62, Jan 2008.

[11] S. Fazli, M. Danóczy, F. Popescu, B. Blankertz, and K.-R.Müller, "Using rest class and control paradigms for brain computer interfacing," in *Proc. 10th Int. Work-Conf. Artificial Neural Networks*, Salamanca, Spain, 2009, pp. 651-665.

[12] B. Wang, *et al.*, "Gaussian Mixture Model Based on Genetic Algorithm for Brain-Computer Interfaces," in *Proc. 3rd Int. Cong. Image and Signal Processing*, Yantai, China, 2010, pp. 4079 – 4083.

[13] N. Neykov, P. Filzmoser, R. Dimova, P. Neytchev, "Robust fitting of mixtures using trimmed likelihood estimator," *Comput. Statist. Data. Anal.*, vol. 52, pp. 299–308, 2007.

[14] N. Neykov, and C. Müller, "Breakdown point and computation of trimmed likelihood estimators in generalized linear models," in *Developments in Robust Statistics*, Heidelberg: Physica-Verlag, 2003, pp. 277–286.

[15] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239

[16] B. Blankertz *et al.*, "Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 11, pp. 127–131, June 2003.

[17] BCI competition IV. Available: http://bbci.de/competition/iv/.