Computational Neuroscience

Validating rationale of group-level component analysis based on estimating number of sources in EEG through model order selection

Fengyu Cong\textsuperscript{a,}\textsuperscript{*}, Zhaoshui He\textsuperscript{b,1}, Jarmo Hämäläinen\textsuperscript{c}, Paavo H.T. Leppänen\textsuperscript{c}, Heikki Lyytinen\textsuperscript{c}, Andrzej Cichocki\textsuperscript{d}, Tapani Ristaniemi\textsuperscript{a}

\textsuperscript{a} Department of Mathematical Information Technology, University of Jyväskylä, Finland
\textsuperscript{b} Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China
\textsuperscript{c} Department of Psychology, University of Jyväskylä, Finland
\textsuperscript{d} Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan

HIGHLIGHTS

- Validating rationale of group component analysis (CA) to extract desired components of event-related potentials (ERPs) is critical, but largely ignored.
- Performing group CA on the temporally concatenated EEG datasets theoretically assumes the numbers of sources are invariable among individual datasets.
- Estimating the number of sources in individual EEG datasets via model order selection can logically test the rationale.

ARTICLE INFO

Article history:
Received 10 June 2012
Received in revised form 7 September 2012
Accepted 25 September 2012

Keywords:
EEG
Event-related potential
Filter
Group component analysis
Mismatch negativity
Model order selection
Number of sources

ABSTRACT

This study addresses how to validate the rationale of group component analysis (CA) for blind source separation through estimating the number of sources in each individual EEG dataset via model order selection. Control children, typically reading children with risk for reading disability (RD), and children with RD participated in the experiment. Passive oddball paradigm was used for eliciting mismatch negativity during EEG data collection. Data were cleaned by two digital filters with pass bands of 1–30 Hz and 1–15 Hz and a wavelet filter with the pass band narrower than 1–12 Hz. Three model order selection methods were used to estimate the number of sources in each filtered EEG dataset. Under the filter with the pass band of 1–30 Hz, the numbers of sources were very similar among different individual EEG datasets and the group ICA would be suggested; regarding the other two filters with much narrower pass bands, the numbers of sources were relatively diverse, and then, applying group ICA would not be appropriate. Hence, before group ICA is performed, its rationale can be logically validated by the estimated number of sources in EEG data through model order selection.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Event-related potentials (ERPs) have been acknowledged as an important research tool of human cognition. ERPs, such as mismatch negativity (MMN), are often processed and analyzed by the group-level methods (Näätänen et al., 2011). Also, blind source separation (BSS) methods such as principal/independent component analysis (PCA/ICA) across a group of individuals has been often used for extracting ERP components of interest (Calhoun et al., 2009; Kayser and Tenke, 2003; Vakorin et al., 2010; Vigario and Oja, 2008; Hyvärinen, 2011). In this study, we name group-level PCA and group ICA as group component analysis (CA) for simplicity. Basically, the intrinsic idea for group CA is to decompose data of multiple participants and conditions together to facilitate the further statistical analysis for extracted ERPs. For instance, the extensively used EEGLAB supports temporal concatenation of EEG data of different participants in different groups under different experimental conditions; then, ICA can be performed on the connected data for source separation (Delorme and Makeig, 2004). One extracted component may include one ERP activity related brain responses of those participants in those groups under those conditions. The smallest scale of group CA is to concatenate the EEG data of two experimental conditions for one participant. Kalyakin et al. (2008) demonstrated that ICA was performed on the concatenated...
EEG data of two deviant stimuli in a passive oddball paradigm, and one extracted component by the group ICA contained two MMN related peaks associated with the two deviant stimuli.

For BSS and ICA in neuroinformatics, there are two roles which are rejection of artifacts and extraction of desired brain activities (Vigario and Oja, 2008). Regarding the former application, eye blinks are one of key artifacts to be rejected. In contrast to the brain activities elicited by stimuli in an ERP experiment, variance of eye blinks can be much greater. Although ICA does not inherently rank variances of sources to be estimated, it tends to stably extract the sources with larger variances (Delorme and Makeig, 2004; Huovinen and Ristaniemi, 2006). Hence, ICA has been extensively used to reject eye blinks (Jung et al., 2000; Joyce et al., 2004; Hoffmann and Falkenstein, 2008). In this study, we focus on the application of ICA to extract the desired brain activities in an ERP experiment after the artifacts are rejected. Under such circumstance, we assume that there are a few sources of interest among all sources and the variance of each source of interest is comparable with other sources of interest.

Indeed, group CA for EEG data conforms to the linear transform model (LTM). In the model, the sources of electrical activities in the brain are mapped onto the points along the scalp through a mixing matrix and the mixtures (i.e., EEG data) of those sources are recorded by electrodes. Hence, the underlying basic assumption for group CA is that the mixing matrices for the EEG data of different participants under different experimental conditions are identical because only one mixing matrix is usually estimated for all participants and conditions by group CA. This also indicates that the orders of sources, at least, the most interesting sources, and the numbers of the sources in the models for the EEG data of different participants under different experimental conditions are invariable (we will show these analyses in next section in detail). Actually, the assumption is very strong in practice and it is crucial to investigate whether such an assumption is reasonable or not for group CA, although checking the validity of this assumption seems to be largely ignored in the EEG literature.

Regarding a LTM, theoretically, we cannot know the mixing matrix through PCA or ICA, but the multiplication of a source and its corresponding mixing coefficients (Cong et al., 2010, 2011b,c,d). Furthermore, an order of extracted components by ICA is inherently indeterminate (Hyvärinen et al., 2001), as a result, it is difficult to straightforwardly validate the rationale of the assumption for group CA regarding the mixing matrix. Nevertheless, when the number of sensors is greater than that of sources, the LTM is over-determined, and in this case, the model order selection methods (Cong et al., 2011a, 2012b; Knosche et al., 1998; Li et al., 2007) may be used to estimate the number of sources from EEG data and fMRI data. Regarding group CA, if numbers of sources in data of different participants are identical, we may at least partially enable the elementary assumption of group CA to be acceptable. The current study examines the model order selection methods for validating the group CA assumptions.

As for model order selection, there are many existing methods mainly based on the information criteria principles (Akaike, 1974; Rissanen, 1978; Stoica and Selen, 2004), Bayesian learning rules (Schwarz, 1978; Seghouane and Cichocki, 2007), and the gap in a sorted parameter sequence (Niesing, 1997; He et al., 2009, 2010; Cong et al., 2012b). This study is not targeted to develop a new model order selection method, but to examine whether the numbers of sources in EEG data of different participants in different groups are identical or not. Hence, the Akaike’s information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), and GAP (He et al., 2009, 2010) were chosen as the representatives of those three groups of methods to estimate the number of sources in EEG data. Furthermore, in order to improve signal to noise ratio (SNR) of EEG data, they are often filtered to remove the interference and noise by a digital filter or a wavelet filter (Burger et al., 2007; Cong et al., 2012a; Kalaykin et al., 2007; Quian Quiroga and Garcia, 2003). Consequently, the number of sources was estimated from the cleaned EEG by different filters in this study, which also provides the opportunity to analyze whether different filters affect the rationale to concatenate EEG data for group CA.

In the current study, the MMN responses were examined from the EEG data of children. Participants included three groups of children which were typical readers (denoted as CONT hereinafter), typical readers with risk of reading disability (RD) (denoted by CONT-Risk hereinafter), and the at-risk children with RD (denoted by RD hereinafter). Our interest was to show whether the numbers of sources in EEG data of different participants were identical or not.

2. Method

2.1. Data description

The children comprised a sub-sample of the Jyväskylä Longitudinal Study of Dyslexia (JLD) in which more than 200 children at risk for dyslexia and their controls have been monitored since birth (Lyttinen et al., 2004, 2005). The numbers of children in CONT, CONT-Risk and RD were 28, 40, and 27 in this study.

The goal of the experiment was originally to identify the ERPs to pitch and time rise in children with reading disabilities and typically reading children through a passive oddball paradigm (Hämäläinen et al., 2008). In this study, only responses to the pitch change were taken for analysis. The stimuli consisted of pairs of harmonic sinusoidal tones. In the middle of a tone pair, a silence gap of 255 ms separated the two tones. The durations of two tones in sequence were 100 ms and 150 ms. The fundamental frequencies of the first and second tones were 250 Hz and 500 Hz, respectively. Each tone was with three additional harmonics of their individual fundamental frequency. The inter stimulus interval (ISI) was 610 ms between the pairs in this study. In the oddball paradigm, the probability of one deviant was 0.1, and 3–5 standard sounds existed between the presentations of deviant sounds. For further details on the participants, stimuli and procedure, refer to the study of Hämäläinen et al. (2008).

During the experiment, the participants were instructed to pay their attention in watching a silenced video or playing a computer game; data was collected with 128 electrodes using the vertex as the reference; a high-pass filter of 0.1 Hz and a low-pass filter of 100 Hz were used to filter the collected data; the sampling rate was 500 Hz. After the artifact rejection (with the criteria of 200 μV for peak-to-peak amplitude and 115 μV for transients), at least 60 trials among 125 collected ones remained to any subject. EEG data was offline filtered with the band pass filter from 0.53 Hz to 35 Hz to further remove noise. Having been averaged over the kept single trials, the data was re-referenced to the average over all electrodes. The averaged trace last 900 ms, and the first 100 ms were the recordings of the pre-stimulus baseline. Then, the difference wave (DW: responses of deviant stimuli minus those of standard stimuli) was produced for further processing in MATLAB (Version R2010b, The Mathworks, Inc., Natick, MA).

2.2. Model and assumption of group component analysis

A LTM of EEG data of one participant under one experiment condition can be stated as

\[ X(i) = A(i)S(i) + E(i) \]  

(1)

where for the participant-\( i (i = 1, \ldots, I) \), a data matrix \( X(i) \in \mathbb{R}^{M \times T} \) denotes the recorded EEG data of \( M \) channels by \( T \) samples, a data
matrix $S(I) \in \mathbb{R}^{m \times T}$ represents the $N(i)$ sources of electrical activities in the brain and each source has $T$ samples, a data matrix $A(I) \in \mathbb{R}^{M \times N(I)}$ is the mixing matrix which maps the sources from the brain to the points along the scalp, and a data matrix $E(I) \in \mathbb{R}^{M \times T}$ expresses the sensor noise. Regarding group CA, the LTM of the concatenated EEG data of $I$ participants is

$$X = A S + E$$

(2)

where the connected data are $X = \{X(1), \ldots, X(I), \ldots, X(T)\} \in \mathbb{R}^{M \times N(I)}$, $S = \{S(1), \ldots, S(I), \ldots, S(T)\} \in \mathbb{R}^{N \times N(I)}$, and $E \in \mathbb{R}^{M \times T}$. Hence, the basic assumption for group CA is

$$A = A(1) = \cdots = A(I) = \cdots = A(L)$$

(3)

Eq. (3) means that the mixing matrices in Eq. (1) for different participants are identical, and the orders of the sources in Eq. (1) are identical, and the numbers sources in Eq. (1) are invariable too, i.e.,

$$N = N(1) = \cdots = N(I) = \cdots = N(L)$$

(4)

Consequently, Eq. (4) implies that if the numbers of sources in EEG data of different participants were not the same, the assumption for group CA would not be satisfied and it could not be reasonable to perform group CA on the concatenated EEG data of different participants any more. Indeed, this results in the importance to study the numbers of sources in EEG data of different participants. In other words, estimating the number of sources cannot sufficiently validate the rationale of group ICA, but may suggest whether group CA can be used or not.

Through ICA, independent components are extracted through a unmixing matrix $W$ as

$$Y = W X = W A S + W E = CS + W E$$

where, $C = W A$ is a global matrix (Cichocki and Amari, 2003),

$$Y = \{y_1, y_2, \ldots, y_n\}^T \in \mathbb{R}^{N \times N(I)}, S = \{s_1, s_2, \ldots, s_n\}^T \in \mathbb{R}^{N \times N(I)}.$$

For an ideal ICA decomposition, there is only one non-zero element in each column and each row of $C$ (Cong et al., 2011c,d; Cichocki and Amari, 2003). Consequently, $y_n = y_n(1), \ldots, y_n(I), \ldots, y_n(T)$, where

$$y_n = y_n(1), \ldots, y_n(I), \ldots, y_n(T)$$

(5)

where $y_n(1) \in \mathbb{R}^{1 \times T}$ is the $k$th column of $S(I)$, $y_n(I) \in \mathbb{R}^{1 \times T}$. Since $s_1(1), \ldots, s_n(1), \ldots, s_1(I), \ldots, s_n(I)$ reveal the temporal evolutions of the same type of brain activities they can be very similar with each other, and since they are all relevant to the same column of the mixing matrix in Eq. (2) they shall the same spatial map along the scalp, resulting in the same location of the brain activity in the brain. Hence, for the satisfactory decomposition of group ICA, $y_n(1), \ldots, y_n(I), \ldots, y_n(L)$ will be very similar with each other and they share the same spatial map which is represented by the $i$th column of $Y^{-1}$ (Cong et al., 2011c,d; Makeig et al., 1999, 1997). This can be a post criterion to check the feasibility of group CA. The high similarity among $y_n(1), \ldots, y_n(I), \ldots, y_n(T)$ can indicate a success of group CA, at least from the view of the model of concatenated EEG data. This study is devoted to estimating the number of sources in each individual dataset, and the similarity among $y_n(1), \ldots, y_n(I), \ldots, y_n(T)$ is not discussed.

### 2.3. Model order selection

Considering a multiple-input–multiple-output (MIMO) signal model, an array of $m$ electrodes sensing signals $x = \{x_1, \ldots, x_m\} \in \mathbb{R}^{m \times 1}$ are from $n$ sources $s = \{s_1, \ldots, s_n\} \in \mathbb{R}^{n \times 1}$ through a gain (or named as mixing) matrix $A \in \mathbb{R}^{m \times n}$, i.e.,

$$x = As + e$$

(5)

where $e = (e_1, \ldots, e_m)^T \in \mathbb{R}^{m \times 1}$ is the noise vector. In this problem (5), both $A$ and $s$, as well as the number of sources $n$, are unknown. The task in this study is to seek the number of sources from the observed data. To achieve this goal, first, we make three trivial assumptions as the following: (i) $A$ is a tall matrix ($m > n$) and full column rank, (ii) noise signals $e_1, \ldots, e_m$ are mutually independent and follow identical Gaussian distribution $N(0, \sigma^2)$; (iii) the noise is independent statistically with the sources $s_1, \ldots, s_n$ (He et al., 2010). Then, we can obtain

$$C_e = E[x(1)s(T)^T] = A \cdot C_s \cdot A^T + \sigma^2 I$$

(6)

where $E$ denotes the mathematical expectation, $I$ is the identity matrix, and $C_s = E[s(t)s(t)^T]$. Since the rank of $A$ is $n$, one can readily derive

$$\lambda_1 \geq \cdots \geq \lambda_n \geq \lambda_{n+1} = \cdots = \lambda_m = \sigma^2$$

(7)

where $(\lambda_i)_{m=1}^n$ are the eigenvalues of matrix $C_s$ in the descending order.

Usually, when the first $k$ principal components explain over 80% or 90% of variance of the observed $x$, those components are assumed to compose the signal subspace and then are selected for the further analysis (Jolliffe, 2002). However, this method is not adaptive to the data, which is the main obstacle for it to obtain optimal results. The model order selection may overcome this drawback, and can be used to detect the number of sources for separating the signal and the noise subspace in high-density EEG recordings (Cong et al., 2011a; Knosche et al., 1998). For completeness of the study, we briefly introduce the GAP, AIC and BIC algorithms as follows.

In order to identify the parameter $n$ by searching the gap between $\lambda_n$ and $\lambda_{n+1}$ in (7), a gap measure (He et al., 2010) has been defined

$$\text{GAP}(p) = \left\{ \begin{array}{ll} \frac{\text{var}(\lambda_{i+1}^{m-1})}{\text{var}(\lambda_i^{m-1})} & \text{for } \text{var}(\lambda_i^{m-1}) \neq 0 \\ +\infty & \text{for } \text{var}(\lambda_i^{m-1}) = 0 \end{array} \right.$$

(8)

where $p = 1, \ldots, m-2$ and

$$\text{var}(\lambda_i^{m-1}) = \frac{1}{m-p} \sum_{i=p}^{m-1} (\lambda_i - \frac{1}{m-p} \sum_{i=p}^{m-1} \lambda_i)^2$$

denotes the sample variance of the sequence $(\lambda_i^{m-1})_i$, and $\lambda_i = \lambda_i - \lambda_{i+1}, i = 1, \ldots, m-1$. Then, we determine the number of sources by the criterion (He et al., 2009, 2010):

$$n = \arg \min_{p=1,\ldots,m-2} \text{GAP}(p)$$

(10)

The AIC (Akaike, 1974) and BIC (Schwarz, 1978) can read as

$$\text{AIC}(k) = -2L(x(\Theta_k)) + 2G(\Theta_k)$$

(11)

$$\text{BIC}(k) = -2L(x(\Theta_k)) + G(\Theta_k) \log T$$

(12)

where

$$L(x(\Theta_k)) = (T/2) \log (\prod_{i=k+1}^{N+1} s_i^{-1} x_i^{-m-k}) / ((1/m-k) \sum_{i=k+1}^{N+1} s_i^{-1} x_i^{-m-k}),$$

and $G(\Theta_k) = 1 + mk - (1/2)k(k-1), T$ is the number of samples, $L(x(\Theta_k))$ is the maximum log-likelihood of the observation based on the model parameter set $\Theta_k$ of the $k$th order and $G(\Theta_k)$ is the penalty for model complexity given by the total number of the free parameters in $\Theta_k$ (Wax and Kailath, 1985).

### 2.4. Data processing

To show the effectiveness of the model order selection in estimating the number of sources from EEG data, a simulation
study was first carried out. Fourteen ICA sources extracted from a 128 channel the EEG data in our previous study were used (Cong et al., 2011a,e). The sources were mixed according to Eq. (5), and the number of sensors was 128, and the mixing matrix whose elements ranged from −1 to 1 was randomly generated in every run of simulations. When noise was zero, the noise free mixtures which are the signals in the definition of SNR were produced. White Gaussian noise with different SNR starting from 5 dB was added, and the noise was the same to each channel. For SNR, the mean of energy of all 128 channels’ signals was calculated to represent the energy of signals, and the energy of the noise at one channel was used to denote the energy of noise. PCA was performed on the generated mixtures, i.e., the simulated EEG data, to obtain the eigenvalues from which the number of sources was estimated by AIC, BIC and GAP.

2.4.2. Real data

The DW was firstly filtered by two band-pass digital filters and a wavelet filter. The two digital filters were designed based on Fast Fourier Transform (FFT) (Kalaykin et al., 2007). For the digital filters, the number of points for FFT was 5000, and within the pass-band, the transformed coefficients of DW in the frequency-domain were kept and others were set to be zero, and then, all the coefficients were transformed back to the time-domain to reconstruct the filtered DW. The pass bands of the two digital filters were from 1 to 30 Hz (digital filter-1) and 1 to 15 Hz (digital filter-2). Regarding the wavelet filter which had been used before for denoising MMN (Cong et al., 2011a,b,e), the reversal biorthogonal wavelet with the order of 6.8 (Daubechies, 1992) was used, and a DW was decomposed into nine levels, and the coefficients from the sixth level to eighth level were used to reconstruct the desired MMN. The reason to select the coefficients of those levels for the reconstruction is that the magnitude responses of such a wavelet filter under the configurations mentioned above should match the spectral properties of ERPs (Cong et al., 2012a). The magnitude responses of the filters are shown in Fig. 1. It is evident that the wavelet filter seems like a digital filter with the pass-band narrower than from 1 to about 12 Hz and the attenuation of the input by the wavelet filter started reducing from 7 Hz to about 12 Hz gradually.

After the filters were applied, PCA was performed on the filtered data to obtain the eigenvalues from which the number of sources was estimated by AIC, BIC and GAP.

2.4.3. Frequency response of a filter

For a digital filter, when the input is the unit impulse, the output is the impulse response, and then, the Fourier transform of the impulse response produces the frequency response of the filter (Cong et al., 2012a; Mitra, 2005). In this study, the magnitude responses of a filter were calculated according to the following four steps:

(1) The unit impulse with the duration from −900 ms to 900 ms was generated under the sampling frequency of 500 Hz. The duration was referenced to the length of one epoch in the MMN experiment in this study. Regarding the unit impulse, its value at zero time was 1 and at other timestamps, its values were zero.

(2) The unit impulse was filtered to produce the impulse response by the filters used in this study.

(3) The impulse response was transformed by FFT with the number of points 5000 to produce the frequency responses of the filter.

(4) The magnitude of the frequency responses was calculated to obtain the magnitude responses.

2.4.4. Clustering ratio: criterion to determine the rationale of group CA

Eq. (4) means that the numbers of sources in the individual EEG datasets should be identical if group CA is applied, which is indeed a theoretical assumption. In practice, if the numbers of sources in the majority of individual EEG datasets are similar, group CA may be suggested for data processing.

Hence, in order to investigate the similarity of the numbers of sources of all participants, the estimated numbers of sources in this study were clustered into two clusters under each model order selection method for each filter. Then, the ratio of sizes of two clusters (larger cluster to smaller cluster) was calculated for each model order selection method regarding each filter. Next, under each filtering method the ratios for three methods (GAP, AIC, BIC) were averaged to produce the ratio which is named as ‘Clustering ratio’ in this study. Finally, if the clustering ratio is larger than three, it means the majority of the participants has a similar number of sources and group ICA may be performed; otherwise, group ICA is not suggested. The two-step clustering method in IBM SPSS (V.19) was used in this study.

Fig. 1. Magnitude responses of three filters.
3. Result

3.1. Simulated data

Fig. 2.1 shows the 14 sources which were used in every run of simulation. The sources had very similar energy (the mean of power of 14 sources was 1.32 $\mu$V$^2$ and the standard variance was 0.20). For demonstration of the simulation, Figs. 2.2 and 2.3 describe the 13 channels' mixtures among 128 ones, and the eigenvalues estimated by PCA from noisy mixtures. Here, for clarity, not all 128 channels' mixtures are presented. In this example, the SNR was 8 dB, and AIC, BIC and GAP all estimated 14 sources. The cost computing time by GAP was about one third of any of AIC and BIC. Regarding the eigenvalues in Fig. 2.3, the one with the magnitude about 10 is the 14th among all 128 eigenvalues. After this one, all the next 114 eigenvalues are very close to each other since they correspond to noise subspace, and the first 14 ones are associated with the signal subspace.

Fig. 2.4 demonstrates the results of estimation by GAP, AIC and BIC when the SNR ranged from 5 to 28 dB. The results were the average over 30 runs of simulations. Such a range of SNR is often reported in the model order selection studies (He et al., 2009, 2010).

The simulation shows that when the EEG data meets a LTM and the assumptions of model order selection methods are satisfied, the number of sources can be correctly estimated when SNR is larger than 5 dB in our study.

3.2. Real data

The ERP data have been reported in previous studies for the purpose of effects of MMN (Hämäläinen et al., 2008). Hence, the
data are not presented hereinafter. Instead, we mainly report the results of model order selection.

Tables 1–3 describe the estimation by three methods for all participants from the filtered EEG data by the two digital filters with the pass bands from 1 to 30 Hz and from 1 to 15 Hz and the wavelet filter, respectively. The estimation is about the number of sources and the ratio between the corresponding number of participants under this number of sources and 95 (the number of all participants). Generally, under any filter, the estimation by AIC and BIC was entirely identical and the estimated number of sources by GAP was smaller than those by AIC and BIC.

After the estimated numbers of sources under each model order selection method were clustered, the clustering ratios defined above were 3.02, 2.74 and 1.87 respectively for the digital filter with the pass bands from 1 to 30 Hz, the digital filter with the pass bands from 1 to 15 Hz, and the used wavelet filter in this study. Hence, only the digital filter with the pass band of 1–30 Hz can be accepted as the preprocessing method before group CA in this study.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Estimation under the digital filter with the pass-band of 1–30Hz.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td></td>
<td>58 56 57 79% 20% 1%</td>
</tr>
<tr>
<td>AIC</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td></td>
<td>62 63 64 65 73% 25% 1% 1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Estimation under the digital filter with the pass-band of 1–15Hz.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td></td>
<td>32 30 29 81% 18% 1%</td>
</tr>
<tr>
<td>AIC</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td></td>
<td>34 35 36 66% 32% 2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Estimation under wavelet filter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td>14</td>
<td>57% 15% 3%</td>
</tr>
<tr>
<td>AIC</td>
<td>Estimated number of sources Number of participants/number of all participants</td>
</tr>
<tr>
<td>22</td>
<td>49% 21% 24%</td>
</tr>
</tbody>
</table>

4. Discussion

Group-level blind source separation methods or component analysis (CA) have been generally used to extract ERP components from the concatenated EEG data of different participants of different groups under different experimental conditions for group-level analysis in the cognitive research (Calhoun et al., 2009; Kayser and Tenke, 2003; Vakorin et al., 2010; Hyvärinen, 2011; Hämäläinen et al., 2008). This study examined the rationale of the basic assumption of group CA. The theoretical principle is that if the numbers of sources in EEG data of different participants were identical, it might be reasonable to concatenate the EEG data of different participants together for interesting ERP component extraction through group CA. In practice, the majority of participants should share a similar number of sources in individual EEG datasets; otherwise, logically the group CA would not be facilitated.
Based on the clustering ratio defined in this study, we found that the numbers of sources of most participants in the filtered EEG data by the digital filter with a pass band of 1–30 Hz were very similar when they were estimated by AIC, BIC and GAP model order selection methods. If the EEG data were filtered by the digital filter with a pass band of 1–15 Hz or the used wavelet filter whose pass band was even narrower than 1–12 Hz, the estimated numbers of sources became relatively diverse.

It should be noted that although three filters are used in this study, we do not target to suggest using all of them in other ERP studies. The selection of filters in this study is based on the previous publications (Cong et al., 2012a; Kalyakin et al., 2007). Filters used in this study possess the superposition rule. Hence, as long as the filters are appropriately designed according to the properties of the sources of interest, such filters do not change those sources (Cong et al., 2011e). In this study, the estimated number of sources is for ERP data of children. Usually, ERP data of children are noisier and are not as homogeneous as the data of adults. Hence, the number of sources in other ERP datasets can be different.

In our study, it is also interesting to observe that the number of sources estimated by AIC and BIC was greater than that by GAP in the real data although for the simulated data the estimates were entirely the same. Indeed, we do not know the estimation by which method is more accurate since it is impossible to know the true number of sources in the real EEG data modeled by a LTM. Hence, the clustering ratio is defined as the average over the three ratios of cluster sizes for the three methods. Moreover, the required conditions for model order selection might not be completely satisfied. For example, the GAP requires noise to be white Gaussian, which might not be the case in practice. Usually, when AIC is used for the model order selection, AIC assumes the samples in a time series are independent; hence, it is necessary to examine what the true number of independent samples is in a time series (Li et al., 2007). We repeated the methods introduced by Li et al. (2007) and Pyper and Peterman (1998), and the estimation for the number of sources by AIC was the same to the results as mentioned above. This might result from the fact the number of samples is just 450 in our study and most of the sources in our data are very sparse as shown in Fig. 2.

For the validation of group CA, one plausible option seems to estimate the number of sources in the concatenated data in contrast to the individual datasets. For different participants, the levels of noise can be different and the strength of sources of interest can be different either. If ERP data of different participants were concatenated, signal to noise ratio would become very complicated. Model order selection may be very sensitive and may not be accurate when SNR is low (Cong et al., 2012b), hence, concatenating data of different participants may result in difficulty and bias for model order selection. As a result, we do not suggest performing model order selection on the concatenated data of different datasets of different participants.

Nevertheless, since the model order, i.e., the number of underlying sources in data, is important for group CA (Abou-Elseoud et al., 2010; Allen et al., 2012), it would be interesting to examine the impact of the number of sources by analyzing separately data of participants with different numbers of sources. Furthermore, it has recently been shown that ‘Independent EEG Sources Are Dipolar’ (Delorme et al., 2012). Hence, our study can also be significant for the source localization of ERPs (Hämäläinen et al., 2011; Ortlz-Mantilla et al., 2012). The estimation of number of sources in this study helps in the estimation of number of dipole sources within the time range of one epoch in an ERP experiment. With ICA, the components with evident peaks can be extracted (Cong et al., 2011b,e), and then, the number of components within one short time range is available, which can be used as the reference of the number of dipoles for source localization.

Within the best knowledge of authors, validation of the rationale to perform group ICA on ERPs is seldom done theoretically or practically although group ICA has been often applied to study ERPs. Furthermore, the model including the number of sources in ERP data of a participant can be affected by many factors including the participant, experimental condition and design, type of stimuli, etc. So, before group ICA is applied, we think it is necessary to check the rationale to perform group ICA on the preprocessed data by any preprocessing method. We hope to draw the attention of the researchers who have used or will use group ICA to consider the rationale of using this method. Hence, in this study, we investigate the theoretical model of group ICA and propose one solution to test the rationale to perform group ICA through examining the numbers of sources of all participants. We do not think there is a method to confidently determine whether the group ICA can be used or not regardless of different preprocessing procedures since a preprocessing method might change the model of the original data.

In summary, it is beneficial to examine the number of sources in each individual EEG dataset for the rationale of group CA since different preprocessing methods, such as different filters, may differently impact this parameter. The MATLAB code for the simulation through GAP in this study is available via http://users.jyu.fi/~fecong/wICA.html.

Acknowledgements

This work was partly supported by TEKES (Finland) grant 40334/10 ‘Machine Learning for Future Music and Learning Technologies’ Cong F. thanks Professor Asoke K. Nandi for the discussion of model order selection, and the international mobility grant (2009, 2010 and 2011) of Research and Innovation Office of the University of Jyväskylä. This work was supported in part by National Natural Science Foundation of China under Grant 60974072, Natural Science Foundation of Guangdong Province under Grant 2011030002886 (team project), Program for New Century Excellent Talents in University under Grant NCTT-11-0911, and Special Scientific Funds approved in 2011 for the Recruited Talents by Guangdong Provincial universities.

References


Hämäläinen JA, Ortiz-Mantilla S, Benasich AA. Source localization of event-related potentials to pitch change mapped onto age-appropriate MRIs at 6 months of age 2011;54:1901–8.


Huovinen J, Jolliffe I. Independent component analysis using successive interference cancellation for oversampled data 2006;17:577–89.


