scientific reports

Check for updates

OPEN

CW_ICA: an efficient dimensionality determination method for independent component analysis

Yuyan Yi[1], Nedret Billor[1], Arne Ekstrom[2] & Jingyi Zheng[1]✉

Independent component analysis (ICA) is a widely used blind source separation method for signal pre-processing. The determination of the number of independent components (ICs) is crucial for achieving optimal performance, as an incorrect choice can result in either under-decomposition or over-decomposition. In this study, we propose a robust method to automatically determine the optimal number of ICs, named the column-wise independent component analysis (CW_ICA). CW_ICA divides the mixed signals into two blocks and applies ICA separately to each block. A quantitative measure, derived from the rank-based correlation matrix computed from the ICs of the two blocks, is utilized to determine the optimal number of ICs. The proposed method is validated and compared with the existing determination methods using simulation and scalp EEG data. The results demonstrate that CW_ICA is a reliable and robust approach for determining the optimal number of ICs. It offers computational efficiency and can be seamlessly integrated with different ICA methods.

Independent component analysis is a statistical tool to extract hidden information from the observed (mixed) signals. Assuming that these observed signals are linear combinations of mutually independent and non-Gaussian source signals, ICA seeks to discover the linear combination of these mixed signals to recover the original source signals. The performance of ICA is measured by the independence or non-Gaussianity of the estimated ICs.

ICA finds extensive applications in various disciplines. In the realm of biomedical science, it has been leveraged to investigate the brain function, primarily by extracting temporal and spatial information from functional magnetic resonance imaging (fMRI)[1,2] and electroencephalograph (EEG)[3]. In pharmaceutical fields, ICA enables the examination of the distribution of actives and major excipients within tablets by comparing the calculated signals with the pure spectra of the formulation compounds[4]. In chemistry, ICA is widely used for the separation of unknown sample mixtures[5] via peak detection and matching in high-performance liquid chromatography. Additionally, ICA is also extensively used in signal pre-processing for identifying and removing noise and contaminations, thereby extracting effective information.

However, a crucial challenge when utilizing ICA lies in determining the optimal number of ICs for the accurate decomposition. Both under-decomposition (too few ICs) and over-decomposition (too many ICs) can hamper effective source separation. Commonly used determination techniques include information criteria, eigenvalue spectrum (ES), bootstrap resampling (BS), and cross-validation (CV), among others. Nevertheless, these methods have their drawbacks. For instance, information criteria may suffer from overfitting when the sample size is small or strict model assumptions are made. Eigenvalue spectrum methods can be subjective in the choice of threshold and may be affected by noisy signals. Bootstrap resampling techniques, although comprehensive, can be computationally expensive. Cross-validation, while generally reliable, may introduce data partition bias and incur computational costs.

To address the issues with these determination methods, researchers have proposed several alternatives, that can be categorized into three classes. Firstly, methods leveraging original signal information have been introduced. Wang et al.[6] introduced Mean-field ICA (MF-ICA), a Bayesian-based approach that determines the optimal number of ICs by evaluating the square-root sum of the residual between original and reconstructed data. This method excels in separating complex mixtures, such as those encountered in chemistry. Other approaches in this class, such as those proposed by Monakhova et al.[7] and Kassouf et al.[8] leverage different metrics for the

[1]Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA. [2]Department of Psychology and Evelyn McKnight Brain Institute, University of Arizona, Tucson, AZ 85721, USA. ✉email: jingyi.zheng@auburn.edu

determination of the optimal number of ICs. Monakhova et al. introduced an index, known as the Amari Index, whereas Kassouf et al. employed a correlation method, referred to as ICA_corr_y. Both methods, however, necessitate prior knowledge of the ground truth or mixing matrix. Secondly, there are methods employing visual analysis. Bouveresse et al.[9] proposed two techniques, one of which is the ICA-by-Blocks method, employing a "signal-correlation" plot to determine the optimal number of ICs. The other method uses a heatmap generated by the Durbin–Watson criterion. Thirdly, there are methods that require specific data structures. Bach et al.[10] suggested a determination method based on a forest-structured graphical model, which is limited to dependencies among sources within a forest structure and may not apply to broader classes of dependencies. Kassouf et al.[8] presented a determination method using the Kaiser–Meyer–Olkin (KMO) index, a measure indicating the presence of a partial correlation among at least two residual signals. Nonetheless, if there is a small number of ICs, this method encounters a potential pitfall in cases of complete or high correlation among variables, which makes the correlation matrix non-invertible and further poses challenges to the analysis. While these methods offer computational efficiency and require fewer assumptions compared to other techniques, they do have certain drawbacks. For instance, they require enough mixed signals and structured signals (such as sparse, periodic, linear, etc.). Additionally, in some methods, the optimal number of ICs must be visually identified from a plot, which can be subjective. Furthermore, the existing methods for determining the number of ICs may not be universally applicable to all ICA methods, which introduce additional challenges, such as uncertainty and instability. It is also worth noting that the robustness of the determination method for signals with different characteristics plays an important role in determining the optimal number of ICs. However, this factor has not been extensively studied in the context of existing determination methods (see Table 1 for a summary of advantages and disadvantages of these techniques categorized into three groups).

Given the limitations of current determination methods, we propose a method called column-wise independent component analysis (CW_ICA) to automatically determine the optimal number of ICs. Inspired by the ICA-by-Blocks approach, the proposed method addresses challenges related to computational efficiency, consistency, and robustness. Instead of using Pearson correlations, CW_ICA employs Spearman correlations among the ICs obtained from the different blocks. This choice offers advantages in terms of capturing monotonic relationships between ICs, which can be valuable in various scenarios. Moreover, we introduce a novel metric based on the column-wise maximum rank-based correlations between the extracted ICs in CW_ICA. This metric serves as a criterion for determining the optimal number of ICs. Therefore, compared to the existing determination methods, the major advantages of CW_ICA are:

1. **Efficiency**: the computational cost is significantly less than existing methods.
2. **Consistency**: the optimal numbers of ICs obtained by CW_ICA are consistent when it is coupled with different ICA methods.
3. **Robustness**: CW_ICA is robust for signals with different characteristics.

The rest of the paper is organized as follows. In section "Background", we provide a brief introduction to the ICA method and review the current determination methods. Section "Proposed CW_ICA method" presents the CW_ICA method, using a simple example for illustration. We also compare the CW_ICA method with existing determination methods using both simulation and real data in order to evaluate its performance in section "Assessment of proposed method". A detailed discussion of our findings can be found in section "Discussion", and our conclusion is outlined in last section.

Background

Fundamentals of ICA

Suppose there are p mixed signals with length being n. Denote the observed signal matrix as $\boldsymbol{X}_{p\times n}$, where p is the number of mixed signals and n is the signal length. Assuming the observed signals are linear mixtures of q source signals. Then the ICA model is formulated as:

Category	Selected methods	Pros	Cons
Require original source signals information	MF-ICA	Effective for separating complex mixtures, particularly in Chemistry	Requires original source signals
	Amari index based	A quantitative measure with intuitive interpretation	Requires true mixing matrix
	ICA_corr_y	A data-driven approach with computational efficiency	Requires at least one original source signal
Visual determination	ICA-by-Blocks	Flexible in block size, easy interpretation via plot	High computational complexity
	DW criterion	Determine the number of IC based on signal/noise ratio	Fails if variance of DW values among mixed signals is large
Require specific data structure	FCA	Model both inter-cluster independence and intra-cluster dependence	Requires forest structured signals
	KMO index based	A quantitative measure with intuitive interpretation	Fails for cases where small number of ICs occurs
	DW criterion	Determine the number of IC based on signal/noise ratio	Requires structured signals

Table 1. Summary of existing determination methods.

$$\boldsymbol{X}_{p\times n} = \boldsymbol{A}_{p\times q}\boldsymbol{S}_{q\times n} \tag{1}$$

where $\boldsymbol{A}_{p\times q}$ is the "mixing matrix", which specifies the contributions of the source signals to each mixture, and $\boldsymbol{S}_{q\times n}$ is the matrix of source signals. ICA aims to determine both the mixing matrix and the source signal matrix. According to the number of observed signals (p) and source signals (q), ICA methods can be divided into two cases: (over)determined ICA (i.e., $p \geq q$) (e.g., FastICA[11], JADE[12], Infomax[13], etc.), and underdetermined ICA ($p < q$) (e.g., FastFCA[14], MAICA[15], OICD[16], etc.). In this paper, we focus on (over)determined ICA ($p \geq q$), where the mixing matrix $\boldsymbol{A}$ is invertible. The object of ICA can be achieved by estimating the de-mixing matrix $\boldsymbol{W} = \boldsymbol{A}^{-1}$, and the estimated source signals (ICs) can be further obtained by projecting the whitened data onto the matrix $\boldsymbol{W}$.

For comparison purposes, we employed the three most commonly used ICA methods—FastICA, Infomax and JADE when combined with the determination methods.

Overview of existing determination methods

The selection of determination methods is contingent upon the specific structure and characteristics of the data under consideration. In this section, we provide a concise summary of the current state-of-the-art in this field, highlighting key differences among these methods. For a detailed comparison of the pros and cons of the existing determination methods, please refer to Table 1.

Outlined below are algorithms for three different determination methods. These methods will be compared with the proposed technique in "Proposed CW_ICA method". In the algorithms provided, $\boldsymbol{X}_{p\times n}$ represents p mixed signals each of length n. The residual signal matrix is calculated by subtracting estimated signals from initial signals, $\boldsymbol{X}$, denoted by $\boldsymbol{R} = \boldsymbol{X} - \hat{\boldsymbol{X}}$. A_{max} represents the maximum number of ICs.

Durbin–Watson (DW) criterion

The DW statistic is a well-known test statistic used for detecting the presence of autocorrelation in the residuals from a regression analysis[17]. It also serves as a measure of the signal-to-noise ratio in signals, which provides a method for determining the number of ICs[18]. The value of DW criterion for the ith mixed signal, X_i, is defined as:

$$DW_i = \frac{\sum_{t=2}^{n}(r_{i,t} - r_{i,t-1})^2}{\sum_{t=1}^{n}(r_{i,t})^2}, \quad i = 1, \ldots, p \tag{2}$$

where $r_{i,t}$ denotes the value of ith residual signal at time point t. If the DW_i is close to 0, the signal is noise-free, implying that the extracted IC necessitates further decomposition, However, if it is near 2, the signal is inundated with noise, which indicates that the signal is over-decomposed. The average of DW_i values over all p signals is employed as a measurement for determining the number of ICs. Nevertheless, the variance of DW_i's tends to be large in real datasets due to the non-linear behavior exhibited in real-world signals, which contradicts the linearity assumption inherent to ICA methods. In practice, heatmaps are used to depict DW values for each mixed signal produced by models with varying numbers of ICs. From the heatmap plot, q_{opt} is determined to be the optimal number of ICs whenever a sudden increase occurs in the DW values of all mixed signals. The procedure that determines the optimal number of ICs based on the heatmap is summarized as below:
Algorithm based on the Durbin–Watson criterion.

```
**Input:** Observed signals $\boldsymbol{X}$, maximum number of ICs $A_{max}$
**Output:** Optimal number of ICs $q_{opt}$
**Initial:** Number of ICs, $q = 2$
**While** $2 \leq q \leq A_{max}$ **do**
    Perform ICA with presetting number of ICs, $q$, on mixing matrix, $\boldsymbol{X}$.
    Calculate residual matrices by subtracting estimated signals from initial signals, $\boldsymbol{R}_q \leftarrow \boldsymbol{X} - \hat{\boldsymbol{X}}$.
    Calculate DW criteria for each signal: $DW_{q,1}, \ldots, DW_{q,p}$.
    $q \leftarrow q + 1$.
**End While**
Generate heatmap for $\{DW_{i,j}\}, i = 1, \ldots, A_{max}, j = 1, \ldots, p$. The larger the value, the lighter the color.
**If** for most of the columns, blocks in row $q + 1$ is significantly lighter than the one in row $q$ **then**
 $q_{opt} \leftarrow q$.
**End If**
**Return** $q_{opt}$
```

Due to the large variance of DW values among mixed signals, especially for the large number of mixed signals, it becomes challenging to visually select the optimal number of ICs. Further, another limitation of Durbin–Watson criterion is that it can only be used for structured signals (i.e., sparse, periodic, linear, etc.).

ICA_corr_y

ICA_corr_y was proposed to select the optimal number of ICs but requires a known source signal (y)[8]. The main focus of this method lies in determining the correlation between the estimated source signals ($\hat{\boldsymbol{S}}$) and the known source signal (y). The highest correlation is expected to be observed when the optimal IC number is extracted,

despite potential experimental errors. In other words, among ICA models with different number of ICs, if the ICA model with q_{opt} ICs includes the IC with the highest correlation to the known signal y, q_{opt} is identified as the optimal number of ICs. Details of the algorithm is presented below:
Algorithm ICA_corr_y.

```
**Input:** Observed signals $\boldsymbol{X}$, maximum number of ICs $A_{max}$, a source signal $y$
**Output:** Optimal number of ICs $q_{opt}$
**Initial:** Number of ICs, $q = 2$
**While** $2 \leq q \leq A_{max}$ **do**
    Perform ICA with presetting number of ICs, $q$, on mixing matrix, $\boldsymbol{X}$.
    Calculate correlation coefficient between $y$ and each extracted ICs, respectively, $r_1$,..., $r_q$.
    Record $corr_q = \max_{1 \leq i \leq q} (r_i)$ for ICA model with $q$ ICs.
    $q \leftarrow q + 1$.
**End While**
**If** $corr_q = \max_{2 \leq i \leq A_{max}} (corr_i)$ **then**
    $q_{opt} \leftarrow q$.
**End If**
**Return** $q_{opt}$
```

However, having at least one source signal known is a strong requirement, and is not common in many scientific fields such as the scalp EEG signals. Due to this stringent requirement, despite its simplicity, this method is not widely adopted.

ICA-by-blocks

ICA-by-Blocks was proposed to determine the number of ICs based on the correlation of ICs between blocks[9]. The original data matrix is split into B blocks, which is decided in advance. Then, A_{max} ICA models with 1 to A_{max} ICs are computed for each of these predefined blocks. ICs corresponding to "true" source signals are expected to be found in all blocks. Such "true" ICs derived from different blocks should be highly correlated with each other. If all extracted ICs in each block are "true" ICs, the correlation between these ICs in different blocks should be close to 1. If too many ICs are extracted from the blocks, the extraneous ICs will contain a substantial portion related to noise, causing them to exhibit markedly lower correlations with all the ICs derived from other blocks. The ICA-by-Blocks algorithm is outlined below:
Algorithm ICA-by-blocks.

```
**Input:** Observed signals $\boldsymbol{X}$, maximum number of ICs $A_{max}$, number of blocks $B$
**Output:** Optimal number of ICs $q_{opt}$
**Initial:** Number of ICs, $q = 2$
Randomly separate observed signals $\boldsymbol{X}$ into $B$ blocks.
**While** $2 \leq q \leq A_{max}$ **do**
    Perform ICA with presetting number of ICs, $q$, on each block, respectively.
    Generate Pearson correlation coefficient matrix $\boldsymbol{P}_{qB \times qB}$ for all $q \times B$ ICs from all $B$ blocks.
    Vectorize correlation matrix $\boldsymbol{P}$ and sort elements from the largest to the smallest, $\boldsymbol{V}_{(qB)^2 \times 1}$.
    Extract values from $(q \times B + 1)^{th}$ to $(q \times B + q \times (B^2 - B))^{th}$ in the vector $\boldsymbol{V}$, denote as
    $\boldsymbol{L}_{q(B^2-B) \times 1}$.
    Generate the signal-correlation plot for visualization based on every second value in the vector $\boldsymbol{L}$.
    $q \leftarrow q + 1$.
**End While**
**If** most significant drop occurs at $q^{th}$ point in the signal-correlation plot, **then**
    $q_{opt} \leftarrow q$.
**End If**
**Return** $q_{opt}$
```

However, this method is constrained by the number of mixed signals, p, since both the choice of the number of blocks, B, and maximum number of ICs, A_{max}, depend on the sample size (i.e., $A_{max} \leq \frac{p}{B}$). While multiple blocks are desired to better measure the correlation between ICs extracted from different blocks, too many blocks will restrict the maximum number of ICs and increase the computational complexity. Further, this method does not use a quantitative measure to determine the optimal number of ICs, instead a signal-correlation plot is used visually to determine it. Thus, it is rather time-consuming and potentially prone to subjective errors.

Proposed CW_ICA method

Algorithm development

The CW_ICA method starts by randomly splitting the initial data matrix into two sample blocks ($\boldsymbol{B}_1$ and $\boldsymbol{B}_2$), each containing an approximately equal number of signals. The maximum number of computed ICs, denoted by A_{max}, is preset and assumed to be less than the number of signals in each block, ($\frac{p}{2}$). When p is an odd number, the mixed signals are randomly divided into two blocks, with one containing $\frac{p+1}{2}$ signals and the other containing $\frac{p-1}{2}$ signals. We then perform ICA models on each block respectively with the same number of ICs. The process is repeated with the number of ICs varying from 2 to A_{max}. ICs corresponding to the true source signals are expected to be in two blocks. We assume that for each true IC extracted from Block 1, there is a highly correlated IC extracted from the Block 2. In order to measure the correlation between ICs from different blocks, we use a rank-based correlation matrix,$\mathrm{P}_{2q\times 2q}$, that measures a monotonic relationship among extracted ICs, in other words, not restricted to linear relationship as the Pearson correlation matrix does. Further, due to the symmetry of the correlation matrix, we only need to perform further analysis on the off-diagonal block, $\mathbf{P}'\ _{q\times q}$ (see Fig. 1 for a detailed process).

For each model with q ICs, we record the maximum absolute value of each column in $\mathbf{P}'$, indicating the strongest correlation between each pair of ICs ($\rho_1,\ \rho_2, \ldots,\ \rho_q$). Then, we record the smallest of these q values to represent the least absolute correlation coefficient between a pair of ICs. This leads to a quantitative measurement for the ICA model with q ICs, defined as

$$R_q = \min_{1\leq i\leq q} \{\max\{|\rho_i|\}\} \tag{3}$$

where r_i is the ith column of the matrix $\mathbf{P}'$. When q is small, it is likely to underestimate the source signals, resulting in highly correlated extracted ICs, i.e., R_q is closed to 1. As q increases, it tends to overestimate, introducing noise signals and causing some extracted ICs to be uncorrelated, i.e., R_q is closed to 0. Therefore, to identify the optimal number of IC, we observe the changes of R_q as q grows. The q_{opt} is selected based on the pattern of R_q, where R_q is relatively high while R_{q+1} is significantly lower. Specifically, when the number of ICs exceeds q_{opt}, if R_q decreases significantly and remains consistently low as q increases, we claim that the optimal number of ICs is q_{opt}.

To quantify the "significant drop", we calculate the first-order difference of $R_2,\ \ldots,\ R_{A_{max}}$

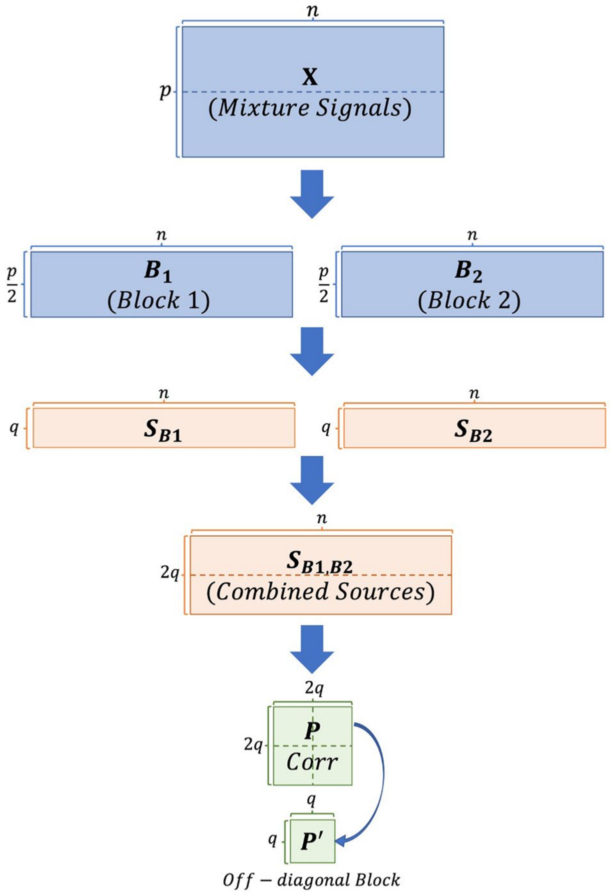

Figure 1. Stages of data structures in CW_ICA method.

$$D_{i,q} = R_{i,q} - R_{i,q-1}, \quad q = 3, \ldots, A_{max}, \quad i = 1, \ldots, Rep \tag{4}$$

where $R_{i,q}$ is the smallest column-wise maximum absolute correlation value, Rep is the number of repetitions. A negative value of $D_{i,q}$ signifies a decrease in the measurement, with a smaller value indicating a more substantial decrease. Thus, the optimal number of ICs is automatically selected out according to the index of the smallest first-order difference, that is $\min_{3 \le q \le A_{\max}} \{D_{i,q}\}$.

We repeat this procedure multiple times, record the detected optimal number of ICs each iteration. The optimal number of ICs is the one that occurs most frequently over all repetitions. The steps are summarized below:

Algorithm CW_ICA.

```
Input: Observed signals $\boldsymbol{X}$, maximum number of ICs $A_{max}$, number of repetitions $Rep$
Output: Optimal number of ICs $q_{opt}$
Initial: Number of ICs, $q = 2$
For $1 \le i \le Rep$ do
    Randomly and evenly split mixed signals into two blocks, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.
    If $q = 2$ then
        Perform ICA with presetting number of ICs, $q$, on each block, respectively.
        Generate Spearman correlation coefficient matrix $\boldsymbol{P}_{2q\times 2q}$ for all $2q$ ICs from two blocks.
        Calculate the smallest column-wise maximum value, $R_{i,q}$, in one off-diagonal block matrix $\boldsymbol{P'}_{q\times q}$.
    End If
    While $3 \le q \le A_{max}$ do
        Perform ICA with presetting number of ICs, $q$, on each block, respectively.
        Generate Spearman correlation coefficient matrix $\boldsymbol{P}_{2q\times 2q}$ for all $2q$ ICs from two blocks.
        Calculate the smallest column-wise maximum absolute value, $R_{i,q}$, in one off-diagonal block matrix $\boldsymbol{P'}_{q\times q}$.
        Calculate the difference between $R_{i,q}$ and $R_{i,q-1}$ and record as $D_{i,q} \leftarrow R_{i,q} - R_{i,q-1}$.
        $q \leftarrow q + 1$.
    End While
    If $D_{i,j} = \min_{3 \le q \le A_{max}} \{D_{i,q}\}$ then
        $q_{opt,i} \leftarrow j - 1$.
    End If
    Generate the signal-correlation plot for visualization based on recorded $R_{i,2},\ldots, R_{i,A_{max}}$.
    $i \leftarrow i + 1$.
$q_{opt} \leftarrow mode(q_{opt,1},\ldots,q_{opt,rep})$.
Return $q_{opt}$
```

The signal-correlation plot shown in Fig. 2C, which depicts the change in R_q as q increases, can also be used to visually identify the optimal number of ICs via detecting a significant drop. However, this method could be time-consuming and may introduce subjective bias.

Validation

CW_ICA starts by randomly partitioning mixed signals into two blocks. This is pivotal as excessively dividing the data into numerous blocks might result in each block containing only a limited number of ICs[8]. In this method, we suggest using a rank-based correlation coefficient, specifically, the Spearman correlation coefficient, as a measure for determining the correlation between ICs. The Spearman correlation coefficient is a statistical measure used to assess the strength and direction of the relationship (i.e., the monotonic relationship) between two variables. As a nonparametric measure, the Spearman correlation coefficient refrains from making assumptions about data distribution or homoscedasticity. Instead of relying on the actual values, the Spearman correlation is based on ranking the data points for both variables. Moreover, the Spearman correlation is robust against outliers because it relies on ranking rather than actual values, making it less susceptible to the influence of extreme data points. The determination of the optimal number of ICs in the ICA-by-Blocks method heavily relies on visual interpretation of a signal-correlation plot. However, this approach is inherently subjective and potentially time-consuming, particularly when the plot shows multiple significant drops at similar levels. In contrast, the proposed CW_ICA method automates this process by quantitatively defining what constitutes a "significant drop", thereby eliminating the need for manual interpretation or inspection.

Additionally, the ICA-by-Blocks method lacks repetitions, which increases the risk of obtaining results by chance. To avoid the risk of biased random splitting due to specific row distribution, we propose repeating the entire process—randomization, equitable division of mixed signals into two blocks, and the subsequent steps—multiple times, denoted as Rep. This systematic repetition within the CW_ICA methodology ensures a more robust and reliable determination of the optimal number of ICs.

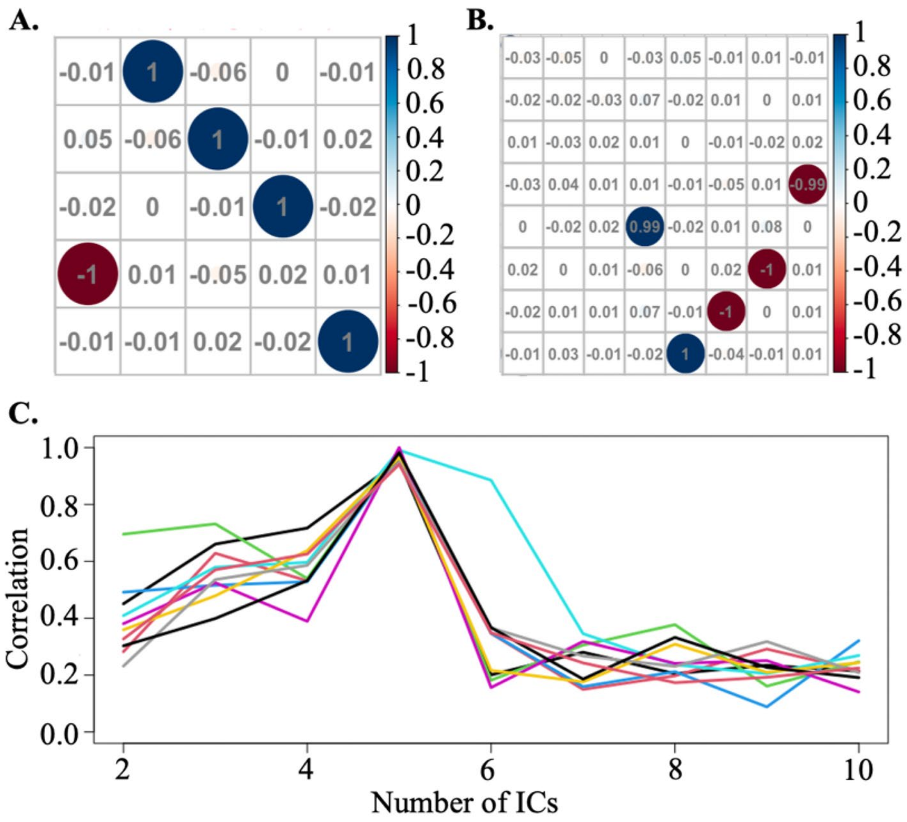

Figure 2. (**A**) Off-diagonal correlation plot for estimated ICs from two blocks (q = 5). (**B**) Off-diagonal correlation plot for estimated ICs from two blocks (q = 8). (**C**) Signal-correlation plot of one replicate. The significant drop at 5th indicates the optimal number of ICs is 5 at this replicate.

Illustrative example

To provide a clear explanation of the CW_ICA procedure, we consider a dataset comprising $p = 20$ mixed signals, each with a length of $n = 500$, which are generated by linear combination of 5 source signals ($q_{opt} = 5$). Our objective is to determine the optimal number of ICs, q_{opt}.

CW_ICA starts with randomly dividing the 20 mixed signals into two blocks: $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, with the dimension of each block being $\frac{p}{2} \times n$ (i.e., 10×500). With the maximum number of ICs being 10 ($A_{max} = 10$), we perform ICA on each block using a range of preset IC numbers $\left(q = 2, \ldots, 10\right)$. For instance, when $q = 8$, after applying ICA to both $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, we obtain two sets of ICs, one from each block, resulting in a combined matrix of $2q \times n$ (i.e., 16×500). Then we compute the Spearman correlation coefficient between each pair of ICs and the correlation matrix is denoted as $\boldsymbol{P}$ with dimension being $2q \times 2q$ (i.e., 16×16). $\boldsymbol{P}$ is composed of four distinct blocks with each block being a $q \times q$ (i.e., 8×8) matrix. The diagonal blocks are indeed identity matrices because they are the correlation between ICs from the same block, which are orthogonal. The two symmetric off-diagonal blocks, which depict correlations between the ICs in $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, contains the same information. therefore, we only need to consider one off-diagonal matrix, $\boldsymbol{P'}$, as shown in Fig. 2B,

For each column of $\boldsymbol{P'}$, we first record the highest absolute correlation value ($\rho_1, \ \rho_2, \ldots, \ \rho_8$), which indicates the strength of correlation between each IC in $\boldsymbol{B}_1$ and each IC in $\boldsymbol{B}_2$. Then, the quantitative measurement of this model is $R_8 = \min_{1 \leq i \leq 8} \{\rho_i\}$. If R_8 is close to 1, it suggests that all extracted ICs in $\boldsymbol{B}_1$ have strong correlations with ICs from $\boldsymbol{B}_2$. This indicates that the optimal number of ICs is greater than or equal to 8. Conversely, if R_8 is close to 0, it implies that some extracted ICs from $\boldsymbol{B}_1$ are not correlated with any ICs from $\boldsymbol{B}_2$ These redundant ICs indicate that the optimal number of ICs should be less than 8. In Fig. 2B, we observe that R_8 is close to 0, which indicates that q_{opt} is less than 8 in this iteration. Moreover, we also show the situation when q is 5, which is indeed the true number of signals, in Fig. 2A. R_5 is close to 1, suggesting that the q_{opt} is greater than or equal to 5.

To determine the q_{opt}, we look at $R_2, \ldots, R_{10}$ and locate the significant drop. Specifically, we calculate the first order differences between these values. For instance, we compute $D_6 = R_6 - R_5$, if D_6 is found to be the smallest value among all the differences, we conclude that the optimal number of ICs in the first iteration is $q_{opt,1} = 5$. Furthermore, we generate a signal-correlation plot to visually examine if there is a significant drop at the 5th point. A significant drop in the plot suggests $q_{opt,1}$.

In the simulation, we repeat the whole process 10 times, i.e., $Rep = 10$, and record the optimal number of ICs from each iteration as $q_{opt,i}, \quad i = 1, \ldots, 10$. Based on previous observations, the performance of CW_ICA, regardless of combining with various ICA methods, become stable in 10 repetitions. Nevertheless, it is essential to adjust the number of repetitions based on the number of input signals to avoid excessive computation time while maintaining accurate outcomes. By examining the 10 signal-correlation lines overlaid on the plot (Fig. 2C), we consistently observe the significant drop occurring at the 5th point from most iterations. Based on this frequent occurrence, we confidently conclude that the optimal number of ICs is $q_{opt} = 5$.

Assessment of proposed method

Artifacts in the collected scalp EEG signals are inevitable and can affect the subsequent analysis of brain activity. To address this issue, ICA techniques are widely utilized to "clean" scalp EEG signals by filtering out artifacts (e.g., eye movements, cardiac activity, muscle activity, etc.) from brain signals. If too few ICs are used, the resulting brain signal may still contain artifacts, reducing the effectiveness of artifact removal. On the other hand, using an excessive number of ICs can lead to over-separation of the brain signal, potentially causing the loss of important features and information. Therefore, determining the accurate number of source signals is crucial when applying ICA to scalp EEG signals. In this section, we carry out simulations and implement the proposed method on real EEG data to assess the proposed CW_ICA.

Simulation

Firstly, we assess the performance of the proposed method by applying CW_ICA in conjunction with different ICA methods on simulated EEG signal data. Since the true number of source signals is known in simulation data, the accuracy of CW_ICA along with three determination methods, the DW, ICA-by-Blocks and ICA_corr_y methods, will be compared. We select three widely used ICA methods, namely FastICA, Infomax, and JADE, to combine with each determination method. However, it is important to note that JADE have convergence issues if the preset number of ICs is greater than the number of source signals. Therefore, JADE is only performed on the real data, where the true number of signals is unknown.

Simulated data generation

According to characteristics of EEG components[19], eight analog EEG source signals (i.e., true $q = 8$), which consist of both periodic and non-periodic signals as shown in Fig. 3, are simulated. Then, $p = 30$ mixed signals, denoted as $\boldsymbol{X}$, are generated by linearly combining the source signals as $\boldsymbol{X} = \boldsymbol{AS}$, where $\boldsymbol{A}$ is a randomly generated mixing matrix whose elements are normally distributed.

Impact of correlation coefficients on determination methods

To assess the impact of different correlation coefficients on two determination methods, specifically ICA-by-Blocks and CW_ICA, we conduct this simulation study. Noting that the true optimal number of ICs is $q_{opt} = 8$ in this simulation.

We employ both CW_ICA and ICA-by-block coupled with Pearson and Spearman correlation to determine the optimal number of IC, q_{opt}. Figure 4 shows the signal-correlation plots. For ICA-by-Blocks, when it is coupled with Spearman correlation (Fig. 4A2,A4), the estimated q_{opt} is 8, which is the same as the true q. However, when coupled with Pearson correlation (Fig. 4A1,A3), ICA-by-block determines q_{opt} being 10, which leads to over-decomposition. This discrepancy suggests that the choice of correlation coefficient greatly influences the determination accuracy of ICA-by-Blocks. On the other hand, CW_ICA, clearly exhibits a sharp drop at the true number of ICs (i.e. $q_{opt} = 8$), which indicates that CW_ICA outperforms ICA-by-Blocks regardless of the type of correlation coefficient employed (Fig. 4B1–B4). Moreover, the result obtained by the Spearman correlation-based CW_ICA provides even much more compelling evidence for accurately identifying the true

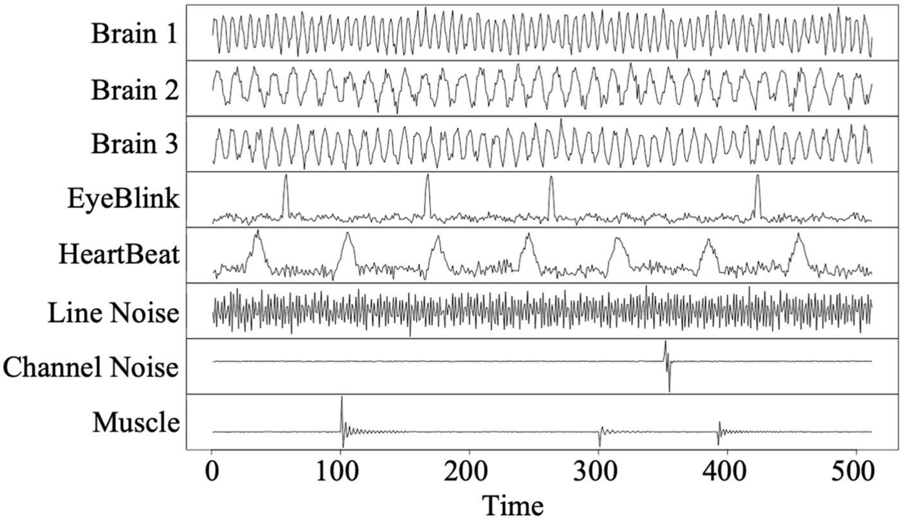

Figure 3. Plots of simulated EEG component signals. The signal length is set as 512.

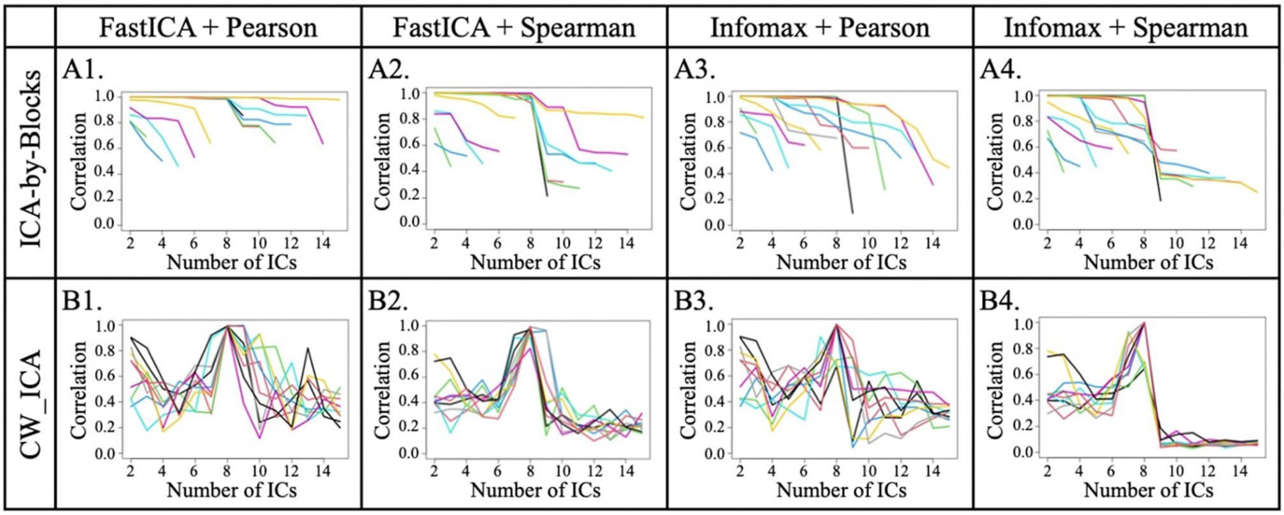

Figure 4. Signal-Correlation plots for (**A**) ICA-by-Blocks and (**B**) CW_ICA methods.

q_{opt} (Fig. 4B2,B4). This is because Spearman correlation coefficient captures monotonic relationships that exist among ICs, providing enhanced performance in determining the optimal number of ICs.

Accuracy
Since the true number of IC q is known in simulation, we can evaluate the accuracy of q_{opt} obtained by different determination methods including our method and ICA-by-Blocks, DW, and ICA_corr_y. In this simulation, multiple mixed-signal datasets, $\boldsymbol{X}$, are generated by repeatedly and randomly generating the mixing matrix, $\boldsymbol{A}$, while keeping the original source signal, $\boldsymbol{S}$, unchanged. Accuracy in this context refers to the percentage of simulation runs that that a determination method correctly identifies the true number of ICs:

$$Accuracy = \frac{m}{N} \times 100\% \tag{5}$$

where m is the number of simulations runs that correctly identify the true number of ICs and N denotes the total number of simulations runs. To estimate the optimal number of ICs, all four determination methods (i.e., CW_ICA, ICA-by-Blocks, DW and ICA_corr_y) are combined with FastICA and Infomax, respectively. Note that source 8 is chosen to be the known source signal required by ICA_corr_y. Limited number of simulations runs (i.e., $N = 5,\ 10,\ 25$), are carried out because both DW and ICA-by-Blocks methods are graph-based identification methods which are time consuming.

As shown in Fig. 5, CW_ICA and ICA_corr_y exhibits the highest accuracy among four determination methods, providing almost 100% accuracy when combined with either FastICA or Infomax. The accuracy of ICA-by-Blocks combined with FastICA is slightly higher than the result when combined with Infomax. However, the DW method performs the worst among the four methods, especially when combined with FastICA.

Robustness
To further understand the robustness of CW_ICA and compare with existing determination methods, especially when the observed data has varying characteristics, we simulate datasets, $\boldsymbol{X}$, with different properties, such as numbers of mixed signals, signal lengths, signal-to-noise ratio (SNR), and frequency ranges, by changing the parameter setting of simulated source signals, $\boldsymbol{S}$, and mixing matrix, $\boldsymbol{A}$. In this simulation, the true number of source signals is set to be 5. We compare four determination methods for accurately determining the number of ICs across datasets with diverse characteristics.

Figure 6 illustrates the estimated number of the source signals, q_{opt}, obtained by four methods as we vary the levels of mixed signals, signal lengths, signal-to-noise ratio, and frequency ranges. First, we conclude that DW coupled with FastICA is not suitable, since the Fig. 6A1–A4 show that the results obtained using the DW criteria is inconsistent as the signal parameters change. Second, we observe that if ICA-by-Blocks is used in conjunction with Infomax, it generates inconsistent results in certain cases, for instance as the number of mixed signals increases, or the signal length changes (i.e., Fig. 6B1,B3). Additionally, the results obtained by ICA_corr_y show variations with changes in the characteristics of mixed signals (i.e., Fig. 6B3,B4). Furthermore, when combined with FastICA, ICA_corr_y produces incorrect results under certain conditions (i.e., Fig. 6A2). In comparison, CW_ICA shows more consistent results regardless the characteristic of mixed signals and the ICA methods.

Scalp EEG data application

Researchers recruited a total of 19 adults (7 females, 12 males) from the University of Arizona in this study[20,21]. Participants were asked to monitor the distances travelled: short (100 virtual meters) vs long (200 virtual meters) distances while navigating in the virtual reality. Each task was repeated 24 trials, and each trial lasted 5.656 s.

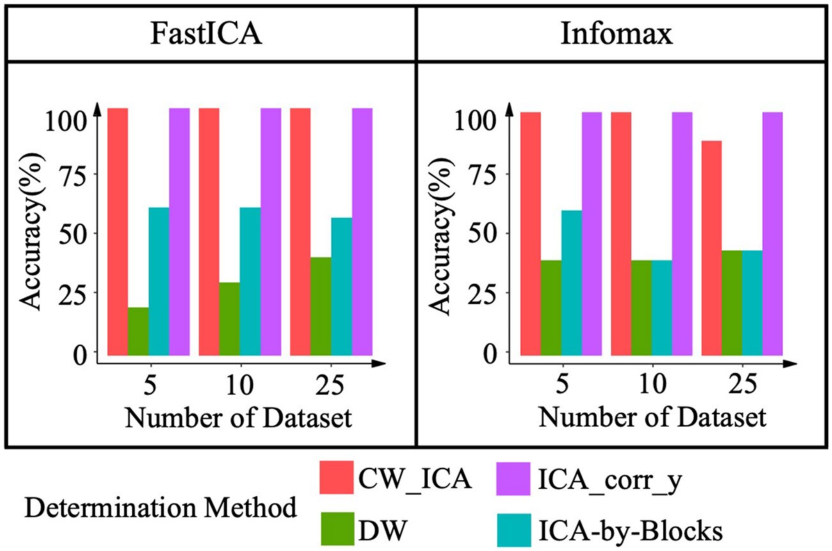

Figure 5. Accuracy of the three determination methods.

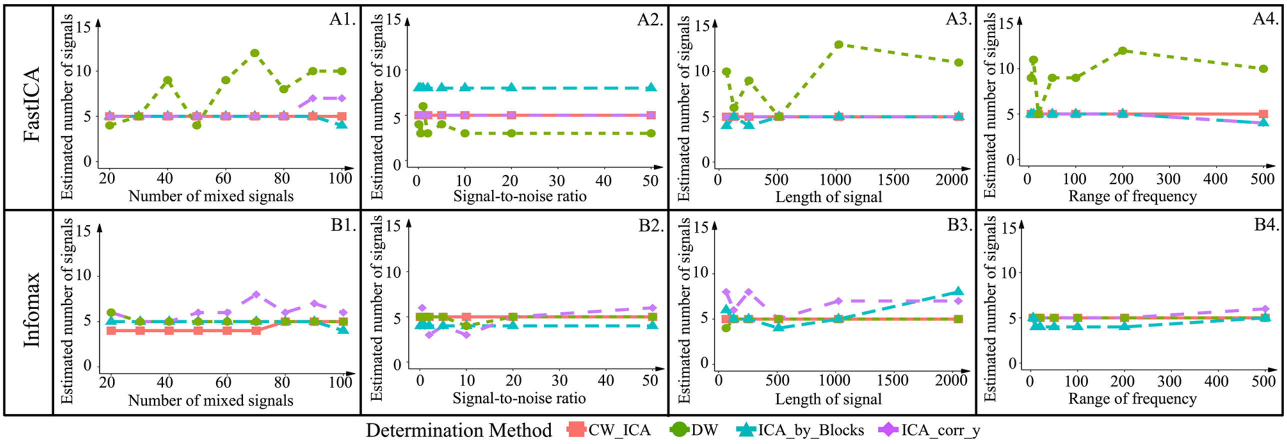

Figure 6. Estimated number of source signals with variety parameters in mixed signals.

Participants walked freely on an omnidirectional treadmill while wearing a wireless scalp EEG cap. The sampling rate was 500 Hz. Details of experiment design can be found in[20,21].

The raw scalp EEG data collected from this study is used to validate the effectiveness and robustness of the proposed CW_ICA, we also consider two other determination methods (i.e., DW, ICA-by-Blocks), with each coupled with three ICA methods (Fast ICA, Infomax, and JADE). In our analysis, we aim to determine the optimal number of ICs at two levels: subject-wise and channel-wise.

First, the optimal number of ICs is determined for each subject separately, considering mixed signals X with dimensions of 2828 × 48 (n) by 64 (p). The heatmap in Fig. 7 shows the optimal number of ICs for each subject, obtained using CW_ICA, DW, and ICA-by-Blocks in combination with FastICA, Infomax, and JADE. We observe that for each subject, the resulting ICs determined by the CW_ICA method exhibit more consistency across the three ICA methods, compared to DW and ICA-by-Blocks. The variation of the determined number of ICs, quantified by the standard deviation (Std), is also summarized in Fig. 8. This indicates that the CW_ICA method has the minimal variation when combining with different ICA methods. The optimal IC number obtained by the CW_ICA method for all subjects is around 7–11 which is close to the widely used IC numbers in neuroscience, while DW and ICA-by-Blocks give smaller IC number, around 5–8. This consistent and reliable performance of CW_ICA in determining the optimal number of ICs, irrespective of the specific ICA algorithm used, underscores its robustness and reliability.

In real data, there is no ground truth, therefore, we further investigate the corresponding ICs obtained using the optimal IC number determined by three determination methods. Let's consider subject 18 as an example. The optimal IC number obtained by the CW_ICA method is 11 while the q_{opt} determined by the other two determination methods is 5. Therefore, we preset the number of IC being 5 and 11 respectively when applying ICA and the obtained ICs are shown in Fig. 9. With the number of ICs being 5, some channel noise can be

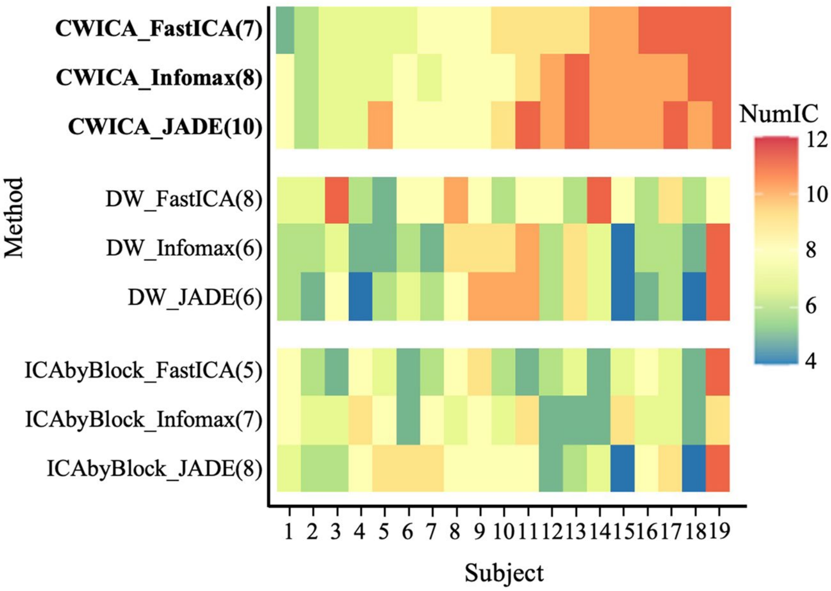

Figure 7. Heatmap of estimated number of ICs. Each column describes the estimated number of ICs by 9 methods from the same subject. Most estimated number of ICs are around 7 to 12, which indicates the number of sources contained in raw EEG signals. The value in parentheses represents the most frequency result of each method.

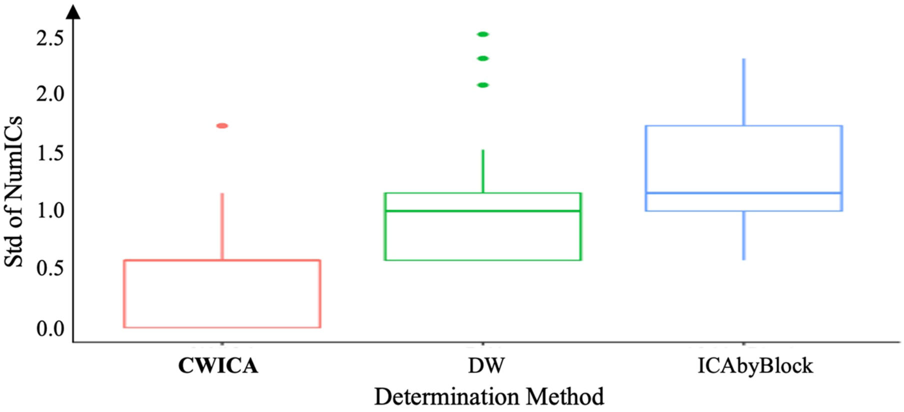

Figure 8. Precision comparison of the estimated number of ICs for three determination methods for every individual.

successfully separated, as evident in IC 1 and 2. However, the remaining three ICs still contain mixed signals and are not fully separated. some channel noise can be separated (e.g., IC 1 and 2). However, the rest three ICs are still mixed signals. With IC number being 11, sources are much better separated, for instance, we obtain the successful separation of channel noise in IC 1 and 2 and, the separation of muscle artifacts in IC 6 and 7.

Besides subject-wise analyses, we also consider channel-wise determination of the optimal IC number. This approach accounts for the possibility that the number of source signals may vary across different brain regions. For example, channels located near the eyes may have a higher number of sources compared to other regions. To determine the channel-wise optimal IC number we implement the nine algorithms on the mixed signals $\boldsymbol{X}$, with dimensions of 2828 (n) by 48 (p), for each subject and each channel. The average of obtained optimal IC number across all 19 participants is calculated and summarized in the heatmap given in Fig. 10. When examining the standard deviation (Std) of the determined IC numbers in Fig. 11, it is evident that the CW_ICA method exhibits more consistent outcomes when combined with different ICA methods, even though the optimal IC number may vary across different channels.

Topographical scalp maps (Fig. 12) represent the channel-wise optimal number of ICs for all 19 participants and provide valuable insights into the distribution of the optimal IC numbers across different brain regions. Overall, the optimal IC number obtained by the CW_ICA method is relatively lower in the frontal cortex, indicating potentially less interference from physical artifacts. Whereas a larger number of ICs may be required

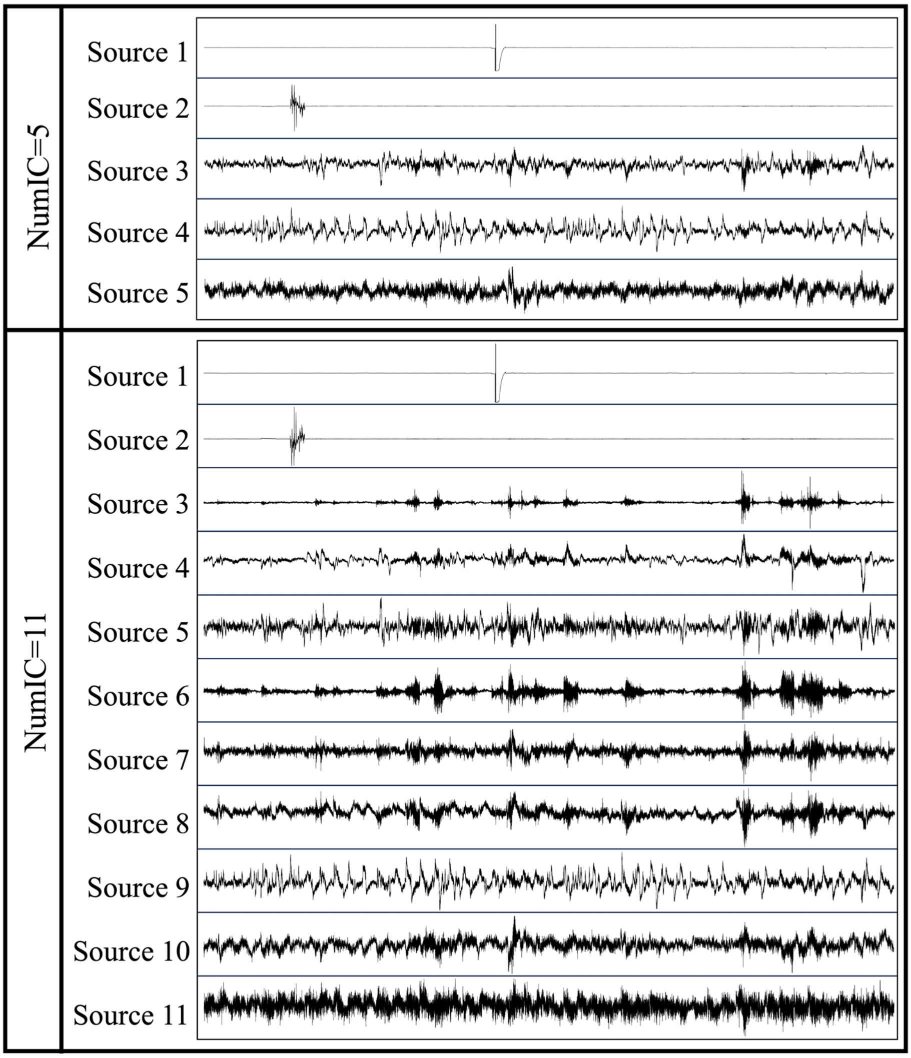

Figure 9. Extracted EEG source signals from subject 18 with different number of ICs.

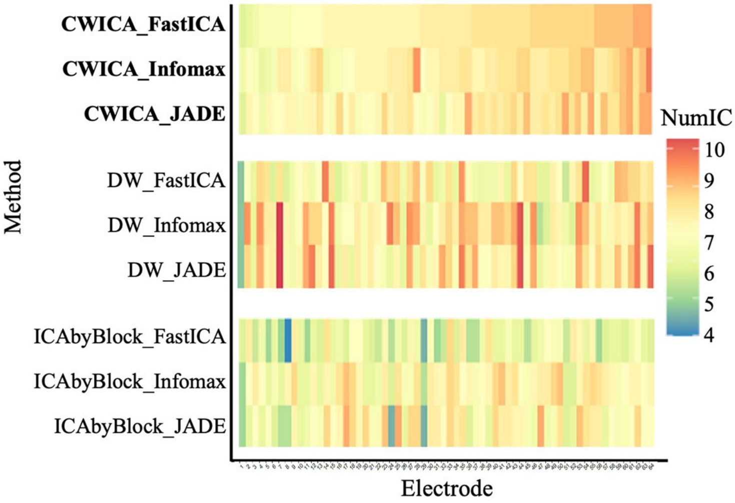

Figure 10. Heatmap of estimated number of ICs. Each column describes the estimated number of ICs by 9 methods at the same electrode.

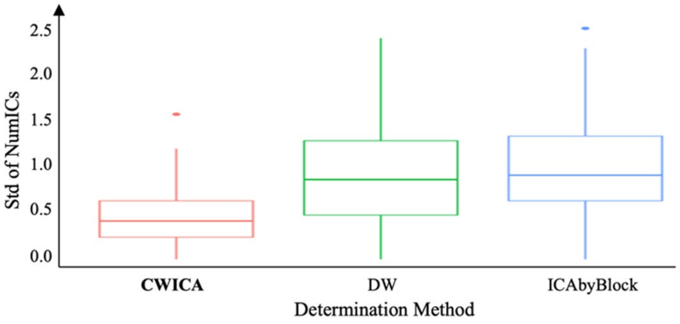

Figure 11. Precision comparison of the estimated number of ICs for three methods of determining the number of ICs for each channel over all 19 individuals.

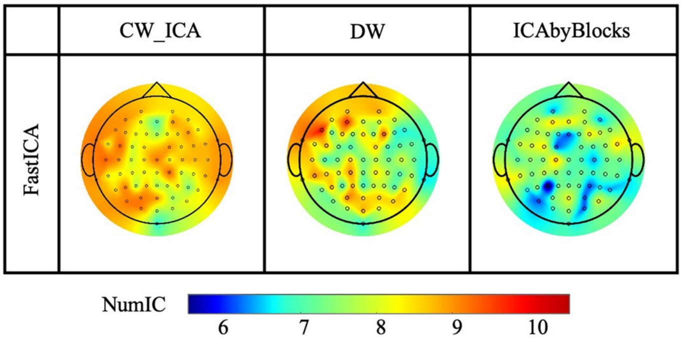

Figure 12. Topographical scalp maps. The number of ICs detected by CW_ICA is relatively larger at electrodes in the temporal and prefrontal cortex than in other brain regions.

to effectively separate the brain signals from the artifacts in prefrontal and temporal cortex. This pattern can be attributed to the fact that the prefrontal and temporal cortex regions are near facial muscles and signals are more susceptible to physical artifacts, such as muscle movements and eye blinking. In contrast, the other two determination methods, particularly ICA-by-Blocks, do not show significant variations in the optimal IC numbers across different electrodes.

Discussion

In this study, we propose a robust method called CW_ICA for determining the optimal number of ICs. Current determination methods visually determine the optimal number of ICs from a plot, they are typically used in conjunction with JADE[9], and applied to the real data in specific fields. However, there is a lack of information on whether these methods can be effectively combined with other widely used ICA methods, such as FastICA and Infomax, and their robustness in various scenarios. Therefore, to address these gaps and overcome the limitations of existing determination methods, we propose CW_ICA. This method can be combined with multiple ICA methods and automatically determine the optimal number of ICs.

We summarize the significant advantages of the proposed CW_ICA method over existing approaches here. First, its computational cost is significantly lower than the existing methods since it eliminates the need for source signal-related information. Moreover, it relies on a quantitative measurement instead of visual identification. Secondly, the CW_ICA method simplifies the process of determining the optimal number of ICs compared to the ICA-by-Blocks method. Rather than determining the number of blocks in ICA-by-Blocks, CW_ICA divides mixed signals into only two blocks. This simplification eliminates potential challenges in selecting the maximum number of ICs when dealing with a large number of blocks. In addition to the aforementioned advantages, CW_ICA offers several other notable benefits. Firstly, CW_ICA extracts only one value from each column of the off-diagonal matrix, whereas ICA-by-Blocks preserves all values in the off-diagonal matrix. This streamlined approach significantly reduces complexity and computational overhead while maintaining the accuracy and reliability of determining the optimal number of ICs. Secondly, CW_ICA can be coupled with multiple ICA methods, consistently yielding reliable results. Researchers can focus solely on selecting an appropriate ICA method based on the properties of the mixed signal, without concerns about the compatibility of the determination method and the ICA method. Thirdly, CW_ICA is a robust method because it uses rank-based correlation instead of Pearson correlation coefficient as used in ICA-by-Blocks. The rank-based correlation measures the relationship between ICs from different blocks based on ranks, avoiding reliance on assumptions and generating more robust results. Finally, CW_ICA automatically determines the optimal number of ICs. Unlike ICA-by-Blocks and DW, which require visual identification from plots, CW_ICA quantifies the identification process. This quantification allows for automated determination, eliminating the time-consuming manual analysis required by the other methods.

To better illustrate the advantages of CW_ICA, we conduct extensive comparisons using both simulated signals and collected raw EEG signals. First, we compare the performance of two correlation coefficients (Pearson and Spearman correlation coefficients) in combination with ICA-by-Blocks and CW_ICA. The results clearly show that CW_ICA with the Spearman correlation coefficient displays a significant drop at the true number of ICs, 8, providing more accurate determination (Fig. 4). Second, we evaluate the accuracy of CW_ICA by applying it to multiple datasets and comparing it with ICA-by-Blocks, ICA_corr_y and DW. The findings imply that CW_ICA consistently maintains an accuracy rate of nearly 100%, while the performance of the other two methods varies when they are combined with different ICA methods (Fig. 5). Next, to evaluate the robustness of CW_ICA, it is applied to multiple sets of mixed signals with different properties (i.e., length, quantity, signal-to-noise ratio, and range of frequency). We compare the results obtained by CW_ICA with those of ICA-by-Blocks, ICA_corr_y, and DW. The findings indicate that the correct number of ICs is identified by CW_ICA and not affected by changes in the characteristics of the mixed signals (Fig. 6). Further, we compare nine combinations of determination methods (CW_ICA, ICA-by-Blocks, DW) and different ICA methods (FastICA, Infomax, JADE) to determine the optimal number of ICs in raw EEG signals (Figs. 7, 8, 9, 10, 11, 12). Among these combinations, only the proposed CW_ICA provides the same number of ICs across each electrode, demonstrating its compatibility with different ICA algorithms and consistent determination capability. It is worth noting that our methods are adaptable and can be applied to other datasets for the determination of the number of source signals as well.

While CW_ICA demonstrates consistent results when combined with multiple ICA methods, the efficiency of ICA is inherently tied to the selection of specific ICA methods, which are determined based on the characteristics of the signal data. Additionally, the validity of the assumption in ICA that source signals are mutually statistically independent can impact the overall performance of CW_ICA. Moreover, the computational complexity escalates with an increase in the number of mixed signals and signal length, as the study exclusively focuses on classical ICA methods. Furthermore, the study imposes limitations on the choice of ICA methods by assuming that the number of source signals is less than or equal to half the number of mixed signals (i.e., $q \leq \frac{p}{2}$). To broaden the scope, future investigations will explore other over-determined ICA methods, aiming to unleash the potential for a greater number of source signals. Additionally, CW_ICA can be expanded to functional CW_ICA, considering the time-dependence of data, thereby mitigating the impact caused by the length and pattern of the signal.

Conclusion

The proposed CW_ICA method addresses limitations in current determination methods by introducing a quantitative measurement and a block splitting approach to reduce computational complexity. By focusing on the smallest column-wise maximum absolute value, CW_ICA offers a versatile solution that can be seamlessly integrated with various ICA methods. Moreover, it leverages the robustness of the Spearman correlation coefficient, which leads to reliable and consistent results in determining the optimal number of ICs automatically. To evaluate the performance of CW_ICA, it is compared with existing determination methods in combination with multiple ICA methods using extensive simulated data and real raw EEG signals. In conclusion, the proposed CW_ICA method offers a versatile and robust approach for automatically determining the number of ICs in signal analysis. Its compatibility with multiple ICA methods, reduction in computational complexity, utilization of Spearman correlation coefficient, and strong performance in comparative evaluations make it a valuable tool for researchers in various fields.

Data availability

The scalp EEG datasets are available at https://osf.io/3vxkn/. Written informed consent was obtained from participants and study was approved by the Institutional Review Board at the University of Arizona and is conducted in accordance with relevant guidelines and regulations. Additional simulated datasets and code are available at https://github.com/yzy0080/CW_ICA.git.

Received: 4 August 2023; Accepted: 7 December 2023
Published online: 02 January 2024

References

1. Adali, T., Anderson, M. & Fu, G.-S. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *Signal Process. Mag. IEEE* **31**, 18–33 (2014).
2. Hu, G. *et al.* Snowball ICA: A model order free independent component analysis strategy for functional magnetic resonance imaging data. *Front. Neurosci.* **14**, 569657 (2020).
3. Rejer, I. & Gorski, P. Benefits of ICA in the case of a few channel EEG. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2015**, 7434–7437 (2015).
4. Boiret, M., Rutledge, D. N., Gorretta, N., Ginot, Y.-M. & Roger, J.-M. Application of independent component analysis on Raman images of a pharmaceutical drug product: Pure spectra determination and spatial distribution of constituents. *J. Pharm. Biomed. Anal.* **90**, 78–84 (2014).
5. Debrus, B. *et al.* Application of new methodologies based on design of experiments, independent component analysis and design space for robust optimization in liquid chromatography. *Anal. Chim. Acta* **691**, 33–42 (2011).
6. Wang, G., Cai, W. & Shao, X. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemom. Intell. Lab. Syst.* **82**, 137–144 (2006).
7. Monakhova, Y. B., Astakhov, S. A., Kraskov, A. & Mushtakova, S. P. Independent components in spectroscopic analysis of complex mixtures. *Chemom. Intell. Lab. Syst.* **103**, 108–115 (2010).
8. Kassouf, A., Jouan-Rimbaud Bouveresse, D. & Rutledge, D. N. Determination of the optimal number of components in independent components analysis. *Talanta* **179**, 538–545 (2018).

9. Jouan-Rimbaud Bouveresse, D., Moya-González, A., Ammari, F. & Rutledge, D. N. Two novel methods for the determination of the number of components in independent components analysis models. *Chemom. Intell. Lab. Syst.* **112**, 24–32 (2012).
10. Bach, F. & Jordan, M. Finding clusters in independent component analysis. In *4th International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003* (2003).
11. Boppidi, P. K. R. *et al.* Implementation of fast ICA using memristor crossbar arrays for blind image source separations. *IET Circuits Devices Syst.* **14**, 484–489 (2020).
12. Rutledge, D. N. & Jouan-Rimbaud Bouveresse, D. Independent components analysis with the JADE algorithm. *TrAC Trends Anal. Chem.* **50**, 22–32 (2013).
13. Sahonero-Alvarez, G. & Calderon, H. A comparison of SOBI, FastICA, JADE and infomax algorithms (2017).
14. Ito, N., Ikeshita, R., Sawada, H. & Nakatani, T. A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1950–1965 (2021).
15. Rejer, I. & Górski, P. MAICA: An ICA-based method for source separation in a low-channel EEG recording. *J. Neural Eng.* **16**, 056025 (2019).
16. Hao, Y., Song, L., Wang, M., Cui, L. & Wang, H. Underdetermined source separation of bearing faults based on optimized intrinsic characteristic-scale decomposition and local non-negative matrix factorization. *IEEE Access* **7**, 11427–11435 (2019).
17. White, K. J. The Durbin–Watson test for autocorrelation in nonlinear models. *Rev. Econ. Stat.* **74**, 370–373 (1992).
18. Gómez-Carracedo, M. P., Andrade, J. M., Rutledge, D. N. & Faber, N. M. Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. *Anal. Chim. Acta* **585**, 253–265 (2007).
19. Pion-Tonachini, L., Kreutz-Delgado, K. & Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* **198**, 181–197 (2019).
20. Zheng, J. *et al.* Time-frequency analysis of scalp EEG with Hilbert-Huang transform and deep learning. *IEEE J. Biomed. Health Inform.* **26**, 1549–1559 (2022).
21. Liang, M., Zheng, J., Isham, E. & Ekstrom, A. Common and distinct roles of frontal midline theta and occipital alpha oscillations in coding temporal intervals and spatial distances. *J. Cogn. Neurosci.* **33**, 2311–2327 (2021).

Acknowledgements

The authors would like to thank anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper. This paper is based upon work supported by the National Science Foundation under Grant No. 2153492 and the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant No. UL1TR003096.

Author contributions

Y.Y., N.B., A.E. and J.Z designed the experiments. Y.Y. performed the method development, experiments and testing. Y.Y., N.B., A.E. and J.Z. contributed to writing and revising the manuscript. All authors have approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.