A methodology for synthesizing interdependent multichannel EEG data with a comparison among three blind source separation techniques

Ahmed Al-Ani[1], Ganesh R. Naik[1], and Hussein A. Abbass[2]

[1] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo NSW 2007, Australia
[2] University of New South Wales, School of Engineering and Information Technology, Canberra, ACT 2600, Australia.
{Ahmed.Al-Ani@uts.edu.au, Ganesh.Naik@uts.edu.au, H.Abbass@adfa.edu.au}

Abstract. In this paper, we introduce a novel method for constructing synthetic, but realistic, data of four Electroencephalography (EEG) channels. The data generation technique relies on imitating the relationships between real EEG data spatially distributed over a closed-circle. The constructed synthetic dataset establishes ground truth that can be used to test different source separation techniques. The work then evaluates three projection techniques – Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Canonical Component Analysis (CCA) – for source identification and noise removal on the constructed dataset. These techniques are commonly used within the EEG community. EEG data is known to be highly sensitive signals that get affected by many relevant and irrelevant sources including noise and artefacts.
Since we know ground truth in a synthetic dataset, we used differential evolution as a global optimisation method to approximate the "ideal" transform that need to be discovered by a source separation technique. We then compared this transformation with the findings of PCA, ICA and CCA. Results show that all three techniques do not provide optimal separation between the noisy and relevant components, and hence can lead to loss of useful information when the noisy components are removed.

Keywords: Synthetic multichannel EEG; artefact removal; PCA; ICA; CCA; optimal projection; differential evolution

1 Introduction

The Electroencephalography (EEG) is a well established modality for recording brain's electrical activities. In addition to the various signal sources arising from the cerebral, scalp EEG is also influenced by a number of noise and artefact sources. These include scalp muscles, eye movements and blinks, breathing, heart beat, and electrical line noise. Proper interpretation of EEG data is very important as artefact and noise may bias the neurological interpretation [1–4].

There has been attempts to extract relevant information from the EEG signals using classical signal processing techniques, including adaptive supervised filtering, parametric and nonparametric spectral estimation, time frequency analysis, and higher-order statistics. Unfortunately, all techniques face difficulties arising from the spectrum overlap of brain signals with artifacts. For instance, most of these techniques have been reported to fail in completely eliminating ocular artifacts [5], [6], [7].

Multivariate techniques, such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Canonical Component Analysis (CCA) have been widely used in identifying the relevant sources and denoising the EEG data. The aforementioned techniques are commonly used in processing EEG data [6, 8, 7]. However, some important information might be lost while applying the above procedures to EEG data. Identification and removal of exact location of noisy components are of great interest to both engineers and clinicians. However, the number of sources, whether of cerebral origin or artefacts, that contribute to the recorded scalp signal as well as their combination methodology are not known. Thus, one possible approach to overcome this limitation is the use of synthetic EEG data to model and project the sources.

This research reports on the ability of the three techniques – PCA, ICA and CCA – in re-constructing the relevant sources using synthetic data. A novel method for constructing synthetic data of four EEG channels is presented with the aim of imitating relationships between real EEG channels, spatially distributed with varying distances. Five measures are presented to evaluate the performance of these methods, where the first three evaluate the reconstruction of sources without eliminating noisy components, while the remaining two evaluate the information loss and noise elimination ability of these methods when removing noisy component(s). A differential evolution (DE) algorithm is utilized to search for the optimal transformation, as the synthetic sources are known a priori, to estimate the difference in performance of the three multivariate methods to that of the "best" projection for each of the five evaluation measures.

The paper is organized as follows. Section 2 describes the construction of synthetic data. Evaluation of the three projection techniques is presented in section 4, and a conclusion is given in section 5.

2 Construction of Synthetic Data

In the construction of synthetic data, we consider the case of limited number of EEG channels (four in particular), and presume that there are five brain sources, four of them are local, one for each channel, while the fifth one is global. We also presumed that there are two noise sources, similar to the synthetic data described in [4]. The seven synthesised sources are described as shown in Table 1. We also decided to change the frequency of the second component for each of the five EEG sources in the second half of the signal and make the EMG artefacts active during limited portions of the signal. Figure 1 shows the seven sources, which are sampled at 256 Hz.

In order to define relationships between the four channels, we studied a real EEG dataset that consists of 64 channels according to the montage shown in Fig. 2. The four synthesized channels are considered to form a rectangle shape with different lengths of its horizontal and vertical sides, as described in Table 2. The real EEG dataset was used

Table 1. Synthesised sources

Source	Equation	Description
1	$14\sin(2\pi \times 4t) + 52\sin(2\pi \times 22t)$	Delta and Beta
2	$23\sin(2\pi \times 7t) + 70\sin(2\pi \times 19t)$	Theta and Beta
3	$16\sin(2\pi \times 5t) + 43\sin(2\pi \times 11t)$	Theta and Alpha
4	$44\sin(2\pi \times 9t) + 56\sin(2\pi \times 47t)$	Alpha and Gamma
5	$34\sin(2\pi \times 6t) + 24\sin(2\pi \times 45t)$	Theta and Gamma
6	$144\sin(2\pi \times 31t) + 337\sin(2\pi \times 51t)$	EMG artefact
7	$282\sin(2\pi \times 28t) + 246\sin(2\pi \times 49t)$	EMG artefact

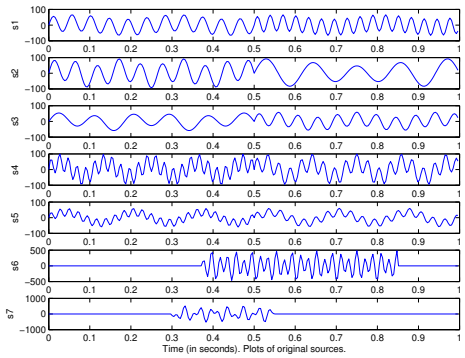

Fig. 1. The seven original signal sources of the synthetic data

to calculate the correlation values between the six pairs of the four channels, i.e., {ch1, ch2}, {ch1, ch3}, {ch2, ch3}, {ch1, ch4}, {ch2, ch4} and {ch3, ch4}.

The nine cases listed in the table have different number of channel combinations, which are: 38, 33, 26, 22, 18, 13, 10, 4 and 2 respectively, as the smaller the rectangle the larger the number of possible combinations that can be formed using the 64 channels. In general, the obtained results indicate that the longest the rectangle side the smaller the correlation between channels. We also noticed that vertical channels tend to have smaller correlation compared to horizontal ones with similar distance (i.e., larger inter-lobe differences compared to intra-lobe), and that the lower horizontal channels (parietal/occipital) usually have higher correlation than their corresponding higher ones (frontal/central).

We considered one of the two synthetic noise sources to be close to the upper left corner of the rectangle (e.g., close to the left frontal channels), while the other has a stronger effect on the lower right corner of the rectangle (e.g., close to the right parietal/occipital channels). The observed synthetic signals of the four channels are formed using a weighted sum of the seven sources. We fixed the weight of the local source of each channel to 1.0 and varied the weights of EMG artefacts based on their distances from each of the four channels. The objective was to search for an appropriate weight value for the other local and global sources to achieve correlation values that are close to the ones listed in Table 2. The weight search that was implemented using Differential

Table 2. Considered scenarios based on distances between the four channels

Case	D-h	D-v	Example {Ch1, Ch2, Ch3, Ch4}	Median Correlation between channels {1,2}, {1,3}, {2,3}, {1,4}, {2,4}, {3,4}
1	1	1	{F5, F3, FC5, FC3}	{0.84, 0.78, 0.73, 0.71, 0.76, 0.83}
2	2	1	{F5, F1, FC5, FC1}	{0.65, 0.79, 0.57, 0.52, 0.75, 0.65}
3	2	2	{F5, F1, C5, C1}	{0.64, 0.52, 0.36, 0.30, 0.42, 0.66}
4	3	2	{F5, FZ, C5, CZ}	{0.30, 0.55, 0.27, 0.23, 0.42, 0.38}
5	4	2	{F5, F2, C5, C2}	{0.20, 0.57, 0.19, 0.20, 0.47, 0.20}
6	4	3	{F5, F2, CP5, CP2}	{0.24, 0.24, 0.13, 0.22, 0.18, 0.21}
7	5	3	{F5, F4, CP5, CP4}	{0.22, 0.28, 0.18, 0.27, 0.19, 0.52}
8	6	4	{F5, F6, P5, P4}	{0.29, 0.10, 0.27, 0.37, 0.13, 0.59}
9	7	4	{F5, F8, P5, P8}	{0.15, 0.07, 0.25, 0.37, 0.07, 0.54}

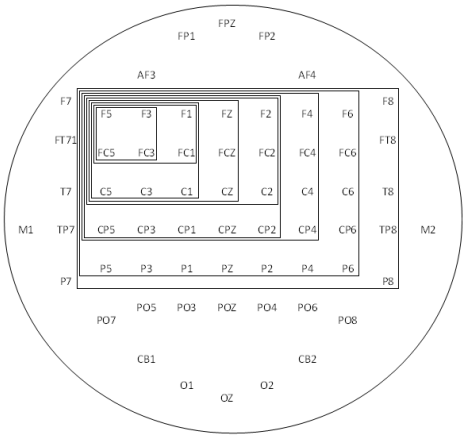

Fig. 2. EEG montage that shows distribution of 64 channels with examples of different distances between the four selected channels (corners of the rectangles)

Evolution (DE) [9, 10] was restricted by an upper limit. The obtained weight matrices are listed in Table 3. The nine obtained sets shown in Fig. 3 demonstrate that the varying level of correlation between the four signals with the first set (Fig. 3(a)) having the maximum correlation between the four signals that are noticeably affected by the first noise source. The effect of the first noise source was kept constant for the fist signal and gradually decreased for the remaining three signals, especially the fourth one, as the distance between the channels increased. In contrast, the effect of the second noise source gradually increased for the fourth signal.

3 Evaluation of The Projection Techniques

It is important to find the optimal reconstruction matrices as their projections will serve as a baseline to evaluate the performance of PCA, ICA and CCA. Finding these matrices will also be helpful for the development of future projection techniques. Two objectives will be defined and a search mechanism using DE will be utilized for this purpose. The first objective is to search for the weight reconstruction matrix that maximises correla-

Table 3. Weigh matrices for each of the nine cases. Weights that are in bold font were fixed, while the remaining ones were optimized given that they do not exceed an upper limit

	W1				**W2**				**W3**			
S1	**1.000**	0.500	-0.500	-0.500	**1.00**	-0.450	0.450	0.450	**1.000**	0.450	-0.450	-0.353
S2	0.224	**1.000**	0.494	0.112	0.450	**1.000**	-0.183	0.370	-0.056	**1.000**	-0.450	-0.450
S3	0.500	-0.500	**1.000**	-0.127	0.348	-0.070	**1.000**	-0.304	0.234	0.019	**1.000**	-0.410
S4	0.500	0.379	0.440	**1.000**	-0.450	0.224	0.450	**1.000**	-0.45	0.240	0.450	**1.000**
S5	-0.122	0.110	-0.092	0.074	0.143	-0.281	-0.066	0.047	-0.248	-0.250	0.087	-0.134
S6	**0.500**	**0.400**	**0.350**	**0.300**	**0.500**	**0.350**	**0.350**	**0.250**	**0.500**	**0.350**	**0.250**	**0.200**
S7	**0.020**	**0.030**	**0.040**	**0.050**	**0.020**	**0.040**	**0.04**	**0.070**	**0.020**	**0.040**	**0.070**	**0.150**
	W4				**W5**				**W6**			
S1	**1.000**	-0.400	-0.400	-0.400	**1.000**	-0.300	-0.300	-0.300	**1.000**	-0.250	-0.250	-0.250
S2	-0.400	**1.000**	-0.230	-0.400	-0.300	**1.000**	-0.300	-0.300	-0.250	**1.000**	-0.250	-0.250
S3	-0.400	-0.400	**1.000**	0.400	-0.280	-0.300	**1.000**	-0.172	-0.250	-0.250	**1.000**	0.250
S4	-0.400	0.198	-0.400	**1.000**	-0.300	0.300	-0.300	**1.000**	-0.250	0.016	-0.216	**1.000**
S5	-0.500	0.500	0.500	0.219	-0.750	0.750	0.418	0.070	-0.750	0.750	0.750	-0.266
S6	**0.500**	**0.300**	**0.250**	**0.170**	**0.500**	**0.250**	**0.250**	**0.130**	**0.500**	**0.250**	**0.150**	**0.100**
S7	**0.020**	**0.050**	**0.070**	**0.200**	**0.020**	**0.055**	**0.070**	**0.250**	**0.020**	**0.055**	**0.100**	**0.300**
	W7				**W8**				**W9**			
S1	**1.000**	-0.250	-0.250	0.250	**1.000**	-0.200	-0.200	0.200	**1.000**	-0.026	-0.150	0.150
S2	-0.250	**1.000**	-0.250	-0.250	-0.200	**1.000**	0.066	-0.200	-0.150	**1.000**	0.019	-0.150
S3	-0.250	-0.250	**1.000**	0.250	-0.200	-0.200	**1.000**	0.200	-0.150	-0.150	**1.000**	0.150
S4	-0.069	-0.059	0.250	**1.000**	0.200	-0.077	0.200	**1.000**	0.150	-0.150	0.150	**1.000**
S5	-0.750	0.672	0.551	0.094	-0.750	0.294	0.230	-0.603	-0.750	0.750	0.380	-0.732
S6	**0.500**	**0.200**	**0.150**	**0.070**	**0.500**	**0.150**	**0.050**	**0.020**	**0.500**	**0.100**	**0.050**	**0.020**
S7	**0.020**	**0.060**	**0.100**	**0.350**	**0.020**	**0.065**	**0.150**	**0.450**	**0.020**	**0.070**	**0.150**	**0.500**

tion with the original sources. We will first attempt to maximise the correlation with the four local sources only, and then with all five EEG sources. Those two implementations will be named DE1 and DE2. The second objective is to dedicate one component to the two noise sources, i.e, maximise correlation with them, while the remaining three components are to be dedicated to the EEG sources (one version (DE3) for the four local sources, and another one (DE4) for all five EEG sources). This will enable the removal of noisy component similar to how ICA, PCA and CCA are usually utilized in EEG processing. We used five measures for evaluating the performance of ICA, PCA and CCA. The first three for evaluating the transformation without removing any component, and the the other two evaluate the estimated noise free signals after removing noisy component(s). The five measures are:

1. Distance of reconstructed sources from the original local sources, i.e., distance between $Corr(S_r, S_l)$ and $Corr(S_l, S_l)$, where $Corr(S_r, S_l)$ is the cross correlation matrix between the reconstructed sources, S_r, and the original local sources, S_l, while $Corr(S_l, S_l)$ is the autocorrelation matrix of the original local sources. Note that we manually arranged the components of ICA, PCA and CCA to achieve maximum cross correlation between S_r and S_l.
2. Correlation between the reconstructed sources and the two noise sources, $Corr(S_r, S_n)$, where S_n represents the two noisy sources.
3. Correlation between the reconstructed sources and the original global source, $Corr(S_r, S_g)$, where S_g is the global source.

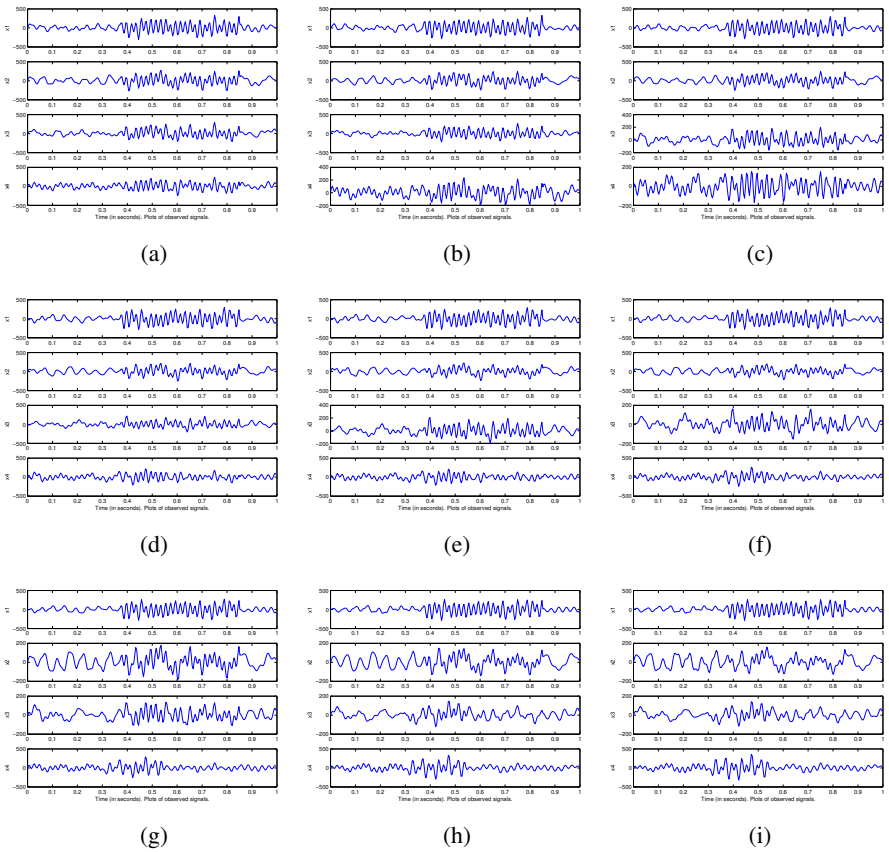

Fig. 3. Synthesised signals using the nine weight matrices.

4. Error remaining in the reconstructed signals after removing noisy component(s), which is calculated using $Corr(Sig_R, S_n)$, where Sig_R are the signals after removing noisy components, while S_n is the noisy sources.
5. Information loss due to removal of noisy component(s), which is estimated using $Corr(S_{rn}, [S_l, S_g])$, where S_{rn} is(are) the reconstructed noisy component(s) that is(are) removed to obtain an estimate of noisy-free signals.

Results of the ICA, PCA, CCA and the four DE optimized methods, which represent the upper limits to compare with, are shown in Table 4. The first three measures shown in the table indicate that the three methods of ICA, PCA and CCA could not reach the optimal performance in terms of correlation with the local and global EEG sources as well as reducing the influence of noise. The table shows that the performance of ICA was slightly better than CCA and noticeably better than PCA for the first two measures, but slightly worse than them for the third measure. On the other hand, the performance of CCA was slightly better than ICA and PCA in reconstructing the global source.

Table 4. Performance of ICA, PCA, CCA and the four DE-optimized benchmark methods

	Method	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Mean
Measure 1	ICA	0.1624	0.2305	0.1836	0.2448	-	0.2922	0.2564	0.2229	0.2171	0.2262
	PCA	0.2378	0.3341	0.2350	0.3201	0.2495	0.2562	0.2539	0.2179	0.2136	0.2576
	CCA	0.3087	0.2119	0.1929	0.2131	0.2251	0.2388	0.2569	0.2743	0.2589	0.2423
	DE1	0.0996	0.1710	0.1453	0.1854	0.1869	0.1886	0.1804	0.1208	0.1129	**0.1545**
	DE2	0.1131	0.1708	0.1446	0.1920	0.1994	0.1963	0.1852	0.1318	0.1347	0.1631
Measure 2	ICA	0.2804	0.3203	0.3488	0.3536	0.4433	0.3479	0.3337	0.3181	0.3252	0.3412
	PCA	0.2738	0.3144	0.3291	0.3533	0.3667	0.3829	0.4264	0.4273	0.4341	0.3676
	CCA	0.2729	0.2907	0.3301	0.3526	0.3695	0.3889	0.3673	0.3622	0.3701	0.3449
	DE1	0.2391	0.0712	0.1588	0.0964	0.1412	0.1476	0.1658	0.3419	0.3612	0.1915
	DE2	0.2075	0.0682	0.1605	0.0904	0.1170	0.1377	0.1702	0.3287	0.3365	**0.1796**
Measure 3	ICA	0.0946	0.1700	0.0965	0.2925	0.1710	0.3872	0.3288	0.2612	0.3692	0.2412
	ICA	0.0990	0.1713	0.1079	0.2613	0.3413	0.4189	0.3116	0.2450	0.3752	0.2591
	ICA	0.0954	0.1704	0.1130	0.2905	0.3628	0.4467	0.3743	0.2876	0.4024	0.2826
	ICA	0.0988	0.1714	0.1140	0.3378	0.4308	0.5435	0.4716	0.2646	0.3490	0.3091
	ICA	0.0981	0.1745	0.1137	0.3598	0.4467	0.5722	0.4981	0.2939	0.4296	**0.3319**
Measure 4	ICA	0.1268	0.1511	0.1591	0.1630	0.1397	0.1229	0.1202	0.0671	0.4641	0.1682
	PCA	0.0697	0.1145	0.1019	0.0922	0.1342	0.1661	0.2033	0.1192	0.3433	0.1494
	CCA	0.0702	0.0588	0.1119	0.1216	0.1505	0.1690	0.0335	0.0838	0.2148	0.1127
	DE3	0.0081	0.0303	0.0180	0.0071	0.0068	0.0113	0.0013	0.0047	0.0013	**0.0099**
	DE4	0.0089	0.0211	0.0091	0.0364	0.0139	0.0067	0.0089	0.0121	0.0070	0.0138
Measure 5	ICA	0.2837	0.2945	0.3051	0.3388	0.3352	0.3781	0.3478	0.3546	0.3537	0.3324
	PCA	0.3398	0.3742	0.3400	0.3835	0.3715	0.3855	0.3733	0.3332	0.3332	0.3594
	CCA	0.3531	0.3066	0.2908	0.3198	0.3271	0.3236	0.3741	0.3682	0.3627	0.3362
	DE3	0.1730	0.1258	0.0906	0.0294	0.0506	0.0754	0.1102	0.1887	0.1878	**0.1146**
	DE4	0.1836	0.1287	0.0917	0.0342	0.0489	0.0804	0.1104	0.1966	0.1938	0.1187

Please note that measure 1 could not be calculated for ICA in the fifth case, as it only managed to find two components and could not converge when calculating the third one. For the fourth and fifth measures that deal with removing noisy component(s), CCA tends to perform slightly better than ICA and PCA in terms of removing the noise sources and not losing relevant EEG information, however, all three methods are not optimal as removal of noisy sources led to noticeable loss of information.

4 Conclusion

In this paper, we proposed a new approach for the construction of synthetic EEG signals using limited number of channels. The construction aimed at mimicking relationships between real EEG channels. The distances were varied between channels, and hence, varying their cross-correlation. The projections obtained using PCA, ICA and CCA were evaluated using five measures and compared to a best approximate projection that was obtained using a differential evolution algorithm. Results reveal that all three projection techniques are not optimal in terms of reconstructing the original sources. Also, the removal of noisy component(s) led to a loss of relevant information for all three methods. These findings motivate the need for more research on the reconstruction of relevant EEG sources. The proposed synthetic dataset construction method can serve as one of the testbeds for evaluating EEG source separation techniques.

References

1. Platt, B., Riedel, G.: The cholinergic system, eeg and sleep. Behavioural brain research **221**(2) (2011) 499–504
2. Mirowski, P., Madhavan, D., LeCun, Y., Kuzniecky, R.: Classification of patterns of eeg synchronization for seizure prediction. Clinical neurophysiology **120**(11) (2009) 1927–1940
3. Simon, M., Schmidt, E.A., Kincses, W.E., Fritzsche, M., Bruns, A., Aufmuth, C., Bogdan, M., Rosenstiel, W., Schrauf, M.: Eeg alpha spindle measures as indicators of driver fatigue under real traffic conditions. Clinical Neurophysiology **122**(6) (2011) 1168–1178
4. Goh, S.K., Abbass, H.A., Tan, K.C., Al-Mamun, A.: Decompositional independent component analysis using multi-objective optimization. Soft Computing (2015) 1–16
5. Cichocki, A., Amari, S.i.: Adaptive blind signal and image processing: learning algorithms and applications. Volume 1. John Wiley & Sons (2002)
6. Sweeney, K.T., McLoone, S.F., Ward, T.E.: The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique. Biomedical Engineering, IEEE Transactions on **60**(1) (2013) 97–105
7. Urigüen, J.A., Garcia-Zapirain, B.: Eeg artifact removalstate-of-the-art and guidelines. Journal of neural engineering **12**(3) (2015) 031001
8. Mammone, N., Foresta, F.L., Morabito, F.C.: Automatic artifact rejection from multichannel scalp eeg by wavelet ica. Sensors Journal, IEEE **12**(3) (2012) 533–542
9. Sarker, R., Elsayed, S., Ray, T.: Differential evolution with dynamic parameters selection for optimization problems. Evolutionary Computation, IEEE Transactions on **18**(5) (Oct 2014) 689–707
10. Guo, S.M., Yang, C.C.: Enhancing differential evolution utilizing eigenvector-based crossover operator. Evolutionary Computation, IEEE Transactions on **19**(1) (Feb 2015) 31–49