34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

ICA Order Selection Based on Consistency: Application to Genotype Data

Jiayu Chen, Vince D. Calhoun, and Jingyu Liu

Abstract— **Independent component analysis (ICA), a blind source separation method, has been shown to be a useful approach to identify genetic components representing combined effects from multiple mutations. However, the ICA order selection for genotype data has been a challenge, since a genetic component usually accounts for a small amount of variance of the data, and makes it difficult to distinguish true signals from background. To address this issue, we propose to select ICA order based on consistency and implement three strategies in this study. Simulations demonstrate robust performances of all three strategies where the selected orders lead to optimal results regardless of ICA performances.**

I. Introduction

Independent component analysis (ICA) is a blind source separation method which has been widely used in many fields such as signal and image processing [1, 2]. A variety of algorithms were developed to achieve the independence among extracted components. Two often used algorithms are Infomax [3] and fast-ICA [4]. While the latter extracts one independent component at a time, the former requires selecting the order, or the component number, before data decomposition. For ICA algorithms that need order selection, information-theoretic criteria, such as Akaike information criterion (AIC) and minimal description length (MDL), have been employed [5-8]. In particular, a modified MDL criterion was specifically developed for ICA applied to functional magnetic resonance imaging (fMRI) data [9].

More recently, the application of ICA was extended to genotype data [10-12] and showed great promise due to its multivariate nature. For instance, applying ICA to single nucleotide polymorphism (SNP) data, we can identify components that represent combined effects from multiple SNPs and may further be associated with a given phenotype. Again, depending on the ICA algorithm, the order needs to be selected. The order selection is much more challenging for genotype data compared with fMRI data, since in general each genetic component accounts for a small amount of variance embedded in the genome (except for those accounting for the population structure), making it difficult to separate true signals from the background. In addition, a principal component analysis (PCA) data reduction is usually applied before Infomax-ICA to select out the same number of principal components accounting for the most variance of the data. This PCA reduction obviously does not guarantee the inclusion of information related to a genetic component carrying small variance. While using variance to identify the true component number works less effectively for genotype data, we observed that using consistency leads to relatively more accurate results. Thus, instead of using the information-theoretic criteria, we propose to select the order based on consistency for genotype data.

*This work was supported by National Institutes of Health grants R01EB005846 and R33DA027626.

J. Chen, J. Liu and V. D. Calhoun are with The Mind Research Network, Albuquerque, NM 87106 USA (corresponding author to provide phone: 505-504-0143; fax: 505-272-8002; e-mail: jchen@mrn.org).

J. Chen, J. Liu and V. D. Calhoun are with the Electrical Engineering Department, University of New Mexico, Albuquerque, NM 87131 USA.

II. Method

The proposed order selection procedure consists of three steps: ICA runs, consistency map construction and order selection.

ICA runs: We apply Infomax-ICA to a given dataset $X_{M1\times M2}$ with different orders (denoted as n), as shown in (1). S^n and A^n respectively represent the components and loadings extracted by ICA with an order of n. The maximal tested order is denoted as N.

$$X^n{}_{M1\times M2} = A^n_{M1\times n} \cdot S^n_{n\times M2}\,;\ (n = 2,3,\dots,N) \qquad (1)$$

Consistency map construction: Given the ICA results from different tested orders, two consistency maps are constructed, one for components (*CS*) and the other for loadings (*CA*). The consistency evaluates the overall components' or loadings' similarity measured by correlations within a range of tested orders. Specifically, for the k^{th} component extracted in an ICA run with order n (denoted as $S^n(k)$), we identify the most similar component extracted in the following ICA run with order $n+1$ (denoted as $S^{n+1}(k')$), and then record the absolute value of their correlation as an element $CS(k,n)$ in the component consistency map, as shown in (2). This procedure is repeated for each component extracted in each ICA run, and thus the component consistency map, *CS,* is constructed as the upper triangular part of an $N\times N$ matrix. In a similar way, we construct the loading consistency map *CA*. Within the consistency matrices *CS* and *CA*, each column of the upper triangle reflects the overall consistency across all components or loadings extracted in one ICA run, while each row depicts the consistency evolution of one specific component or one set of loadings across all the tested orders.

$$CS(k,n) = abs\left[corr\left(S^n(k)\,, S^{n+1}(k')\right)\right] \qquad (2)$$

Order selection: In this step, we locate the desired order which leads to, relatively speaking, the most accurate components and loadings. Three strategies can be applied: overall consistency, reference-blind consistency, and reference-specific consistency.

A. Selection based on the overall consistency (overall)

Within the component consistency map, we focus on its upper triangle and calculate the mean of each column to

obtain the overall component consistency CS_{ova} for each tested order n, as shown in (3). It is expected that the overall consistency remains stable with low orders and starts to decrease quickly when the increasing order results in a components over-splitting situation. Thus, the turning point provides a good guidance on the order selection. To avoid catching local oscillations, we search for a component order range, R_S, covering 10 consecutive tested orders, where the overall consistency exhibits the largest descending gradient (G). The above procedure is repeated for the loading consistency map and results in an order range R_A. Finally, to balance both component and loading consistencies, the median value of the overlapped range between R_S and R_A is selected as the final order, denoted as n_{sel}.

$$CS_{ova}(n) = \frac{1}{n}\sum_{k=1}^{n} CS(k,n) \quad (3)$$
$$G(\tilde{n}) = CS_{ova}(n) - CS_{ova}(n+9),\ \ \tilde{n} = \{n, \dots, n+9\}$$
$$R_s = \{\tilde{n} | max[G(\tilde{n})]\}$$
$$n_{sel} = median[intersect(R_s, R_A)]$$

B. Selection based on the consistency of a reference

Given a component of interest, S_r, as a reference, we select out from each ICA run one counterpart component S_c^n that exhibits the most similar pattern to the reference. Then to evaluate the reference's consistency across tested orders, we apply a sliding window covering 10 consecutive orders and calculate the overall consistency CS_c (average of all pairwise correlations) among counterpart components within that window, as shown in (4). To avoid overfitting, among the windows exhibiting relatively high consistencies ($>CS_{c,th}$, chosen empirically), we select the leftmost to be the component order range, denoted as R_S. The above procedure is also repeated for the loadings, resulting in the order range R_A. Finally, to balance component and loading consistencies, the median value of the overlapped range between R_S and R_A is selected as the final order n_{sel}. Depending on the purpose of the study, the reference selection can be guided by the consistency map or phenotypical information, as described below:

$$CS_c(\tilde{n}) = mean\{abs[corr_{pairwise}(S_c^n, \dots, S_c^{n+9})]\} \quad (4)$$
$$CS_{c,th} = 0.9 \cdot median[CS_{c(top10)}]$$
$$R_s = min\{\tilde{n} | CS_c(\tilde{n}) > CS_{c,th}\}$$
$$n_{sel} = median[intersect(R_s, R_A)]$$

- Reference selected based on the consistency map (reference-blind): In the consistency map, a segment in a single row exhibiting consecutively high correlations indicates a high regional stability. The corresponding component is likely to be true and can serve as a good reference.
- Reference selected based on phenotypical information (reference-specific): The selection of reference can also be guided by phenotypical information such as diagnoses or assessments of studied samples. For instance, in a schizophrenia study, we can select a component whose loadings differentiate patients from controls as a reference.

To assess this method's performance, we applied the order selection procedure to simulated datasets. ICA results derived from different orders were compared with the ground truth, and the average accuracies were calculated as a function of the tested order. Specifically, the component accuracy was evaluated by sensitivity, which is the ratio of correctly identified causal loci over the known true loci. The loading accuracy was reported as the absolute value of the correlation between the simulated case-control pattern and the extracted loadings. Based on the resulting accuracy, we examined whether the selected order would lead to the optimal results.

We conducted the primary test described above with a dataset consisting of 200 samples and 5000 SNP loci. 8 components were simulated using PLINK [13], each involving 150 causal loci and a different case-control pattern. The causal loci exhibited different levels of effect sizes, ranging from 1.77 to 18.86 with a median of 2.20. Furthermore, we investigated the robustness of the procedure under different conditions, including effect size of causal loci, number of samples, number of SNP loci and number of true components.

III. Results

In the primary test, we performed ICA runs with orders ranging from 2 to 100 and then constructed the component and loading consistency maps, as shown in Fig. 1, where the color map indicates the strength of correlation.

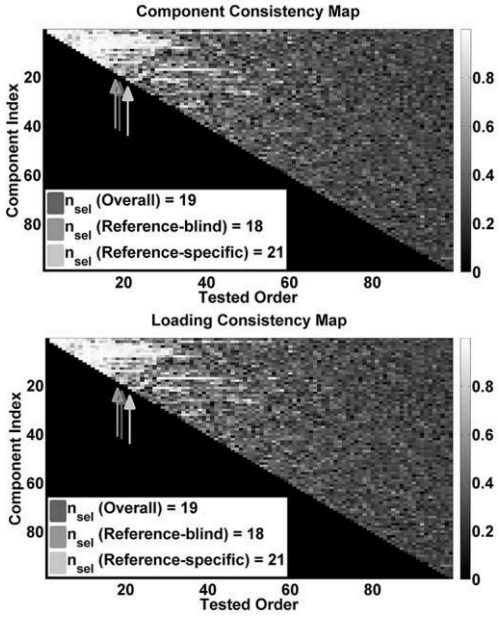

Fig. 1. Component and loading consistency maps.

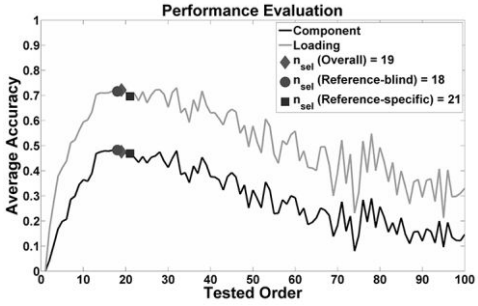

Fig. 2. Performance evaluation of the primary test (200 samples, 5000 SNP loci, 8 true components, median effect size of 2.20).

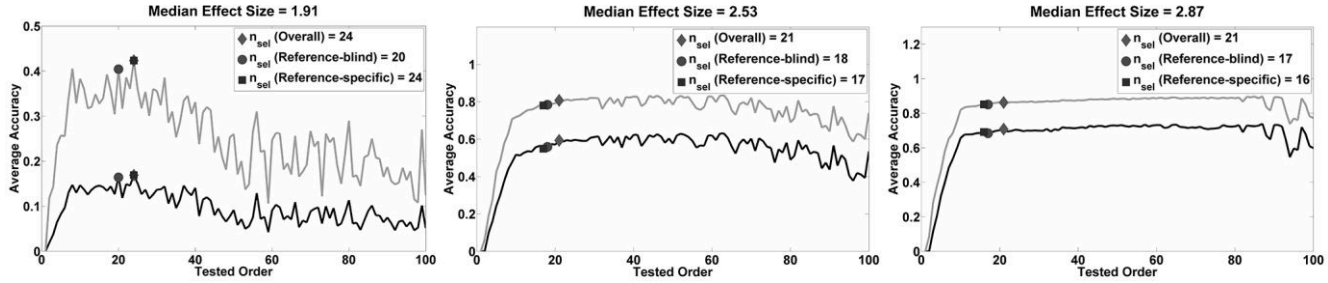

Fig. 3. Performance evaluations on datasets with causal loci of different effect sizes (200 samples, 5000 SNP loci, 8 true components). Black and gray lines represent component and loading accuracies respectively.

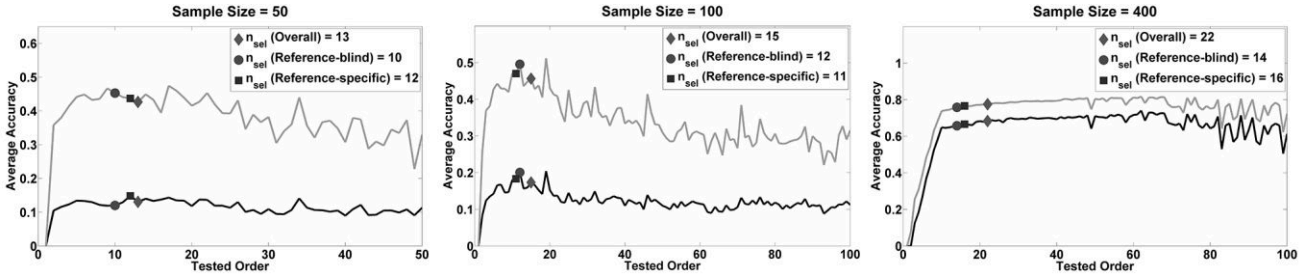

Fig. 4. Performance evaluations on datasets with different sample sizes (5000 SNP loci, 8 true components, median effect size of 1.99). Black and gray lines represent component and loading accuracies respectively.

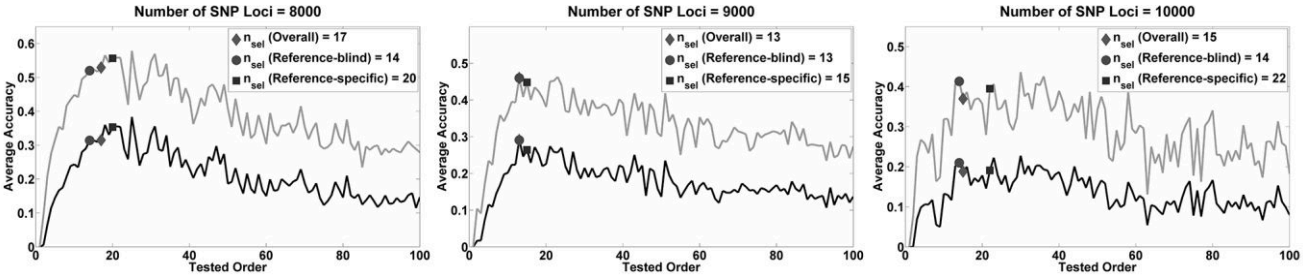

Fig. 5. Performance evaluations on datasets with different numbers of SNP loci (200 samples, 8 true components, median effect size of 2.04). Black and gray lines represent component and loading accuracies respectively.

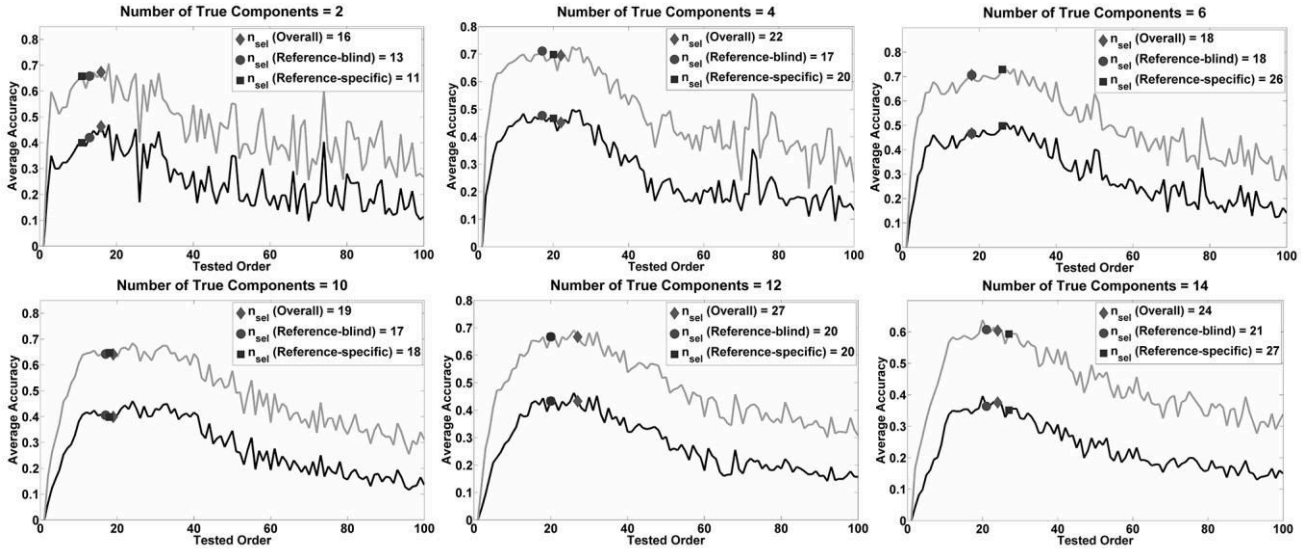

Fig. 6. Performance evaluations on datasets with different numbers of true components (200 samples, 5000 SNP loci, median effect size of 1.95). Black and gray lines represent component and loading accuracies respectively.

All three selection strategies were tested. Using the overall consistency, the order was selected to be 19. Using the 8th component extracted with the order 17 as a reference (reference-blind), the order was selected to be 18. Using the case-control pattern of the first simulated component as a reference (reference-specific), the order was selected to be

21. The selected orders are marked in Fig.1 and Fig. 2, where Fig 1 shows the positions and consistency values of the selected orders in the two consistency maps, and Fig. 2 provides a summary of the performance evaluation across tested orders, indicating that the selected orders lead to the optimal results.

The performances of the proposed procedure on datasets with different conditions are summarized in Fig. 3-6, where the selected orders are marked and compared with other tested orders in terms of the resulting accuracies. It can be seen that we are mainly identifying the leftmost sliding window exhibiting an optimal accuracy. In general, the selected orders lead to relatively accurate components and loadings regardless of the ICA performances.

IV. Discussions and Conclusions

The proposed order selection procedure employs consistency as a criterion to locate the optimal order that results in relatively accurate components and loadings. Given its robustness, we expect that ICA can consistently extract a true component within a range of varying orders. This consistent region can be captured with different strategies, either through evaluating the overall consistency across all components or evaluating the consistency of a specific component across different orders, which can be selected based on regional stability or phenotypical information. Simulations demonstrate robust performances of all three strategies under different conditions.

Effect size of causal loci, number of samples and number of SNP loci: These varying conditions result in components accounting for different amounts of variance of the data. With a larger effect size, more samples or less input SNP loci, the simulated components account for more variance of the data than those with a smaller effect size, less samples or more input loci. When the components carry an adequate amount of variance, they can be accurately identified by ICA. In cases where ICA performs well, the order selection procedure accurately pinpoints the optimal order providing the best results. In cases where components are extracted with low accuracies, the proposed procedure still captures the range where relatively accurate components and loadings can be obtained, as shown in Fig. 3-5

Number of true components: We also simulated datasets with different numbers of true components, ranging from 2 to 14. Fig. 6 summarizes the performance on these datasets. Overall, the proposed procedure exhibits robust performance where the selected order consistently leads to reasonable results regardless of varying numbers of true components. In addition, this evaluation clearly shows that, when a genetic component accounts for a small amount of variance, a true component number does not guarantee optimal results, since the component may be neglected in the PCA reduction applied before Infomax-ICA.

Among the three order selection strategies, the "overall" and the "reference-blind" methods are completely data-driven, while the "reference-specific" method involves phenotypical information. To investigate whether the selection of phenotypical information would affect the performance of the "reference-specific" method, we simulated components with different case-control patterns, yet always used the pattern of the first component to guide the reference selection. The simulation results indicate that the selected orders result in optimal average accuracies of all components and loadings regardless of the choice of phenotype. Thus we conclude that the reference selection can be guided by any phenotypical information and the performance of the procedure is not sensitive to this selection.

In summary, we design a procedure to select the ICA order based on consistency. The goal is to locate an order which allows ICA to extract relatively accurate, consistent components and loadings, while the components and background signal carry comparable variations. Three strategies have been implemented based on Infomax-ICA to achieve this goal. Simulation results indicate robust performances of all three strategies under different conditions and it is noteworthy that the procedure is able to select a reasonable order even when ICA operates less efficiently. While it awaits further evaluation with different ICA algorithms, we believe that there will be many applications for this procedure, not limited to genotype data, but any data with very low signal-to-noise ratio. Although the procedure proposed here is not mathematically 'hard' or 'novel', it will bring in great practical benefit for many researches.

References

[1] P. Comon, "Independent Component Analysis, a New Concept," *Signal Process*, Vol. 36, pp. 287-314, 1994.

[2] A. Hyverinen, J. Karhunen, and E. Oja, *Independent Component Analysis* 1ed, New York: Wiley, 2001.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput*, Vol. 7, pp. 1129-59, 1995.

[4] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput*, Vol. 9, pp. 1483-1492, 1997.

[5] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. of 2nd International Symposium on Information Theory*, Budapest, 1973, pp. 267-281.

[6] J. Rissanen, "Modeling by Shortest Data Description," *Automatica*, Vol. 14, pp. 465-471, 1978.

[7] M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *Ieee T Acoust Speech*, Vol. 33, pp. 387-392, 1985.

[8] V. D. Calhoun, et al., "A method for making group inferences from functional MRI data using independent component analysis," *Hum Brain Mapp*, Vol. 14, pp. 140-51, 2001.

[9] Y. O. Li, T. Adali, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Hum Brain Mapp*, Vol. 28, pp. 1251-1266, 2007.

[10] J. Chen, et al., "Multifaceted genomic risk for brain function in schizophrenia," *NeuroImage,* in press.

[11] Z. Dawy, et al., "A Novel Gene Mapping Algorithm Based on Independent Component Analysis," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processings*, Philadelphia, 2005, pp. 381-384.

[12] J. Liu, et al., "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Hum Brain Mapp*, Vol. 30, pp. 241-255, 2009.

[13] S. Purcell, et al., "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, Vol. 81, pp. 559-575, 2007.