Fine-Scale Genetic Mapping using Independent Component Analysis

Zaher Dawy, *Member, IEEE,* Michel Sarkis, *Student Member, IEEE,*
Joachim Hagenauer, *Fellow, IEEE,* and Jakob C. Mueller

Abstract

The aim of genetic mapping is to locate the loci responsible for specific traits such as complex diseases. These traits are normally caused by mutations at multiple loci of unknown locations and interactions. In this work, we model the biological system that relates DNA polymorphisms with complex traits as a linear mixing process. Given this model, we propose a new fine-scale genetic mapping method based on independent component analysis. The proposed method outputs both independent associated groups of SNPs in addition to specific associated SNPs with the phenotype. It is applied to a clinical data set for the Schizophrenia disease with 368 individuals and 42 SNPs. It is also applied to a simulation study to investigate in more depth its performance. The obtained results demonstrate the novel characteristics of the proposed method compared to other genetic mapping methods. Finally, we study the robustness of the proposed method with missing genotype values and limited sample sizes.

Index Terms

Independent component analysis (ICA), principal component analysis (PCA), single nucleotide polymorphisms (SNPs), linkage disequilibrium, complex diseases, association mapping.

Z. Dawy is with the Department of Electrical and Computer Engineering at the American University of Beirut, Riad El Solh 11-0236, Beirut 1107 2020, Lebanon. Email: zaher.dawy@aub.edu.lb

M. Sarkis is with the Institute for Data Processing (LDV) at the Munich University of Technology (TUM), Arcisstr. 21, 80290 Munich, Germany. Email: michel@tum.de

J. Hagenauer is with the Institute for Communications Engineering (LNT) at the Munich University of Technology (TUM), Arcisstr. 21, 80290 Munich, Germany. Email: hagenauer@tum.de

J. C. Mueller is with the Max Planck Institute for Ornithology, Seewiesen and the Institute for Psychiatry and Psychotherapy at the Munich University of Technology, Germany. Email: mueller@orn.mpg.de

Digital Object Indentifier 10.1109/TCBB.2007.1072

I. Introduction

One of the goals of genetic mapping is to associate single nucleotide polymorphisms (SNPs) with the influence to given traits such as the susceptibility for complex diseases. Association studies of complex traits are particularly challenging due to the influence and possible interaction of multiple loci in an unknown manner, e.g. see [1], [2], [3]. With the development of high-throughput and cost-effective genotyping techniques, the focus of genetic mapping research is shifting towards population-based association studies [4], [5], [6]. In these studies, trait variation in the general population or between cases and controls is tested for correlation with genetic markers such as SNPs. The identified SNPs can be possibly causal or in linkage disequilibrium with a causal variant in the same gene region.

Currently, genome-wide genetic mapping techniques are at the forefront and the information from the genotype data of the HapMap project is used to select uncorrelated SNP sets for efficient genotyping. However, this will not abandon the need for fine-scale genetic mapping after a genomic or gene region has been localized as being associated in several replication studies [7], [8], [9]. The aim of fine-scale genetic mapping is then to test all polymorphisms within the identified region of interest.

Standard genetic mapping methods are based on statistical techniques such as contingency table tests and regression tests. These methods try to evaluate the association of multiple markers with the trait/phenotype based on either single marker or multi-marker multivariate analysis. The main drawback of independent single marker analysis of multiple markers is the multiple testing problem which results in an inflation of false positives and the need for adjustment procedures. On the other hand, multi-marker multivariate models suffer from the high number of degrees of freedom (model terms). To overcome these limitations, stepwise regression procedures with forward and backward selection can be employed [1]. The strength of this method is its ability to provide knowledge on the relationships between neighboring SNPs. A similar conditional genetic mapping method based on Shannon's mutual information has been proposed in [9].

An alternative approach for multi-marker multivariate analysis is to investigate association between the trait and several selected subsets of SNPs [7], [8], [10], [11], [12], [13]. The subsets of SNPs are selected from within haplotype blocks, as correlated groups, or arbitrarily. Another class of methods try to model the interactions between SNPs according to interrelationships of genes within a biological pathway in relation to the disease [14]. These methods introduce new parameters and require model assumptions on the possible interactions.

Recently, there is high interest in the interdisciplinary research field of genomic signal processing which aims at applying well known techniques from signal processing to problems in the field of genetics. For example, singular value decomposition has been applied for the analysis of gene expression data [15], principal component analysis has been applied for the selection of SNP sets that capture intragenic genetic variation [16], and independent component analysis has been applied for the analysis of gene expression data [17].

In this work, we propose a novel fine-scale genetic mapping method based on independent component analysis (ICA) to locate SNPs that are associated with complex traits. ICA is a powerful signal processing technique for revealing hidden factors from multivariate statistical data, e.g. see [18], [19]. The essence of the proposed method is based on finding a suitable model that involves mixing of various sources so that ICA can be applied. To demonstrate the validity and illustrate the properties of our new method, results are presented for one clinical data set in addition to a simulation study. Moreover, we investigate the influence of the following implementation issues on the robustness of the proposed method: missing values due to genotyping failures and number of available individuals or samples.

The proposed method is formed of two stages. In the first stage, it forms independent groups of SNPs according to a linear model and selects the SNP groups (called SNP expressions) that are highly associated with the phenotype. In the second stage, it performs association analysis on each of the identified SNP groups to select individual SNPs that are highly associated with the phenotype. A potential advantage of the method is that it reduces estimation noise as regression analysis is performed on selected independent groups of SNPs that represent the major genetic variation in the given data. As a result, associated SNPs that belong to different SNP groups can be assumed to have an independent effect on the phenotype. Moreover, ICA is a well established method for determining the contributions of individual variables to a set of independently observed mixtures and is available in commonly used statistical packages.

The proposed method applies best for fine-scale mapping of a delimited genomic region which could be a linkage region, a gene region of interest, or a linkage disequilibrium block. Given all polymorphisms of such a genomic region, the method will describe the complete genetic architecture of the association between these polymorphisms and the trait. Fine-scale results are necessary to compile a list of potential causal variants, which can then be tested experimentally for their function.

This paper is organized as follows. In Section II, we formulate the problem in such a way that ICA can be applied. Section III presents the different steps of the proposed method. Section IV presents the results obtained by testing the proposed method on simulated and clinical data sets. In Section V, the effects of missing genotype values and sample size on performance are investigated. Finally, conclusions are drawn in Section VI.

II. Problem Modelling

Given a study comprising a total of N individuals divided among cases and controls where for each individual a SNP sequence of length M is provided. The aim of this work is to design a new genetic mapping method based on ICA to determine the SNPs that are highly associated with a given trait/phenotype. To apply ICA, the genetic mapping problem need to be properly modelled. Therefore, we assume that SNPs interact in a linear unknown environment to form a set of independent SNP expressions or groups that might influence the given trait. SNPs in a single SNP expression could be in linkage disequilibrium or could belong to genes that are in a common pathway. The problem model can be mathematically expressed as follows:

$$\underbrace{\begin{pmatrix} e_{11} & \dots & e_{P1} \\ \vdots & \ddots & \\ e_{1N} & & e_{PN} \end{pmatrix}}_{\mathbf{E}\in\mathbb{R}^{N\times P}} = \underbrace{\begin{pmatrix} s_{11} & \dots & s_{M1} \\ \vdots & \ddots & \\ s_{1N} & & s_{MN} \end{pmatrix}}_{\mathbf{S}\in\mathbb{R}^{N\times M}} \underbrace{\begin{pmatrix} a_{11} & \dots & a_{P1} \\ \vdots & \ddots & \\ a_{1M} & & a_{PM} \end{pmatrix}}_{\mathbf{A}\in\mathbb{R}^{M\times P}}, \tag{1}$$

where $\mathbf{S}$ contains the given SNP sequences of all individuals, $\mathbf{E}$ is the matrix of independent SNP expressions, $\mathbf{A}$ is the SNP coefficients matrix that models the linear mixing environment, and P is the number of SNP expressions to be estimated. Each column of $\mathbf{E}$ contains the individuals' contribution to one expression and $\mathbf{A}$ contains the magnitude of the contribution of the different SNPs to the expressions.

However, the given input data is the matrix $\mathbf{S}$. Therefore, it is more convenient to mathematically express the problem model given in (1) as follows:

$$\mathbf{S} = \mathbf{E}\mathbf{A}^{+} = \mathbf{E}\mathbf{D}, \tag{2}$$

where + is the pseudoinverse sign and $\mathbf{D} \in \mathbb{R}^{P\times M}$ is the pseudoinverse of $\mathbf{A}$. The problem transforms to estimating the SNP expressions matrix $\mathbf{E}$ along with the mixing matrix $\mathbf{D}$ from the SNP matrix $\mathbf{S}$. The reason the model was not derived initially as in (2) is the biological relevance; normally, the SNPs determine the SNP expressions and not vice versa. The SNP expressions can be considered as independent sources or hidden factors in the SNPs' sequences. Once $\mathbf{D}$ is estimated, the SNP coefficients matrix $\mathbf{A}$ can be obtained by performing the pseudo-inverse operation.

III. Method Description

The proposed genetic mapping method is composed of two main stages. In the first stage, the SNPs are clustered into independent groups and each group is tested for association with the phenotype. In the second stage, association analysis is performed on the SNPs in each of the groups that are highly associated with the phenotype. The proposed method is composed of several steps as shown in Figure 1.

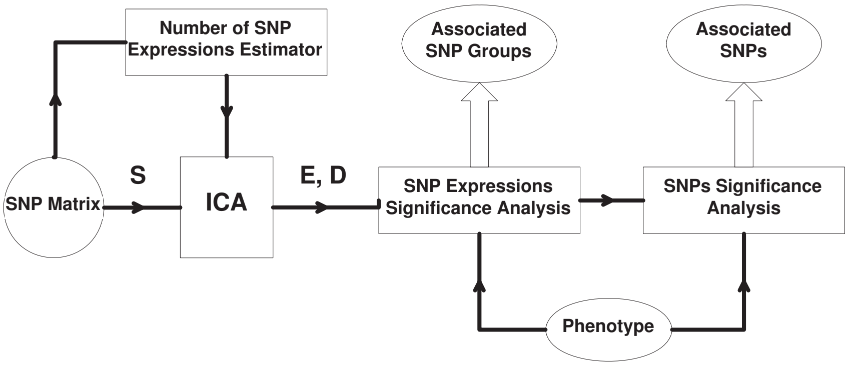

Number of SNP Expressions Estimator
Associated SNP Groups
Associated SNPs
S
E, D
SNP Matrix
ICA
SNP Expressions Significance Analysis
SNPs Significance Analysis
Phenotype

Fig. 1. Steps of the proposed genetic mapping method.

The central step in the method is performing ICA. Basically, the ICA algorithm can be applied directly on the input SNP matrix **S** to estimate the SNP expressions which would be by default equal to the number of variables M. However, in order to reduce the computational complexity and to reduce the estimation noise, it is common practice to perform data centering followed by whitening and dimension reduction before applying ICA. Centering is performed by removing the mean of the data and whitening is performed by normalizing the variance of the data. Whitening and dimension reduction are performed can be performed by applying the principal component analysis algorithm [18]. In order to perform dimension reduction, there is a need to estimate the optimal number of components P. The number of components need to be carefully selected as it determines the number of SNP expressions to be estimated in order to represent the genetic variation in the given data.

In practice, the input matrix **S** might contain missing values due to genotyping failures. A simple solution to this problem is to omit missing value samples from the analysis. However since our method requires matrix operations, all the samples of individuals with missing values need to be omitted from the data. Therefore, this solution is not favorable since the amount of available data is normally scarce in clinical data sets. Alternative solutions have been proposed in the literature to estimate the missing values [20], [21], [22], [23]. The influence of missing values on the proposed method is dealt with in depth in Section V. In the sequel, we assume that the matrix **S** is given as input to the method without any missing values. The following sections explain the different steps of the proposed method.

A. Number of Expressions Estimator

The first step is to calculate the number of components P that should be estimated by the ICA algorithm. Selecting all components of the data will increase the noise since the eigenvalues with low power comprise noise more than useful information [18]. One possibility would be to choose all components that have eigenvalues larger than one [24]. This possibility is not favorable since it does not consider the structure of the data and it results in over reduction of the data's dimension in many cases.

In the proposed method, we use a new estimator for the number of components by using residual based statistical fit [25]. The idea is based on minimizing the difference between the given data and its approximation after dimension reduction. Statistical fit has also been used in other applications which include regression tests and factor analysis [26].

Given the SNP matrix $\mathbf{S}$, the singular value decomposition (SVD) of its covariance matrix $\mathbf{C}_{\mathrm{s}} \in \mathbb{R}^{M \times M}$ can be expressed as:

$$\mathbf{C}_{\mathrm{s}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{T}, \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{M \times M}$ is an orthogonal matrix and $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ is a diagonal matrix containing the singular values. Let $\tilde{\mathbf{C}}_{\mathrm{s}} \in \mathbb{R}^{M \times M}$ be the obtained covariance matrix by selecting the m significant singular values of (3):

$$\begin{aligned} \tilde{\mathbf{C}}_{\mathrm{s}} &= \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{U}}^{T} \\ &= \left(\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}^{1/2}\right)\left(\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}^{1/2}\right)^{T} \\ &= \mathbf{L}\,\mathbf{L}^{T}, \end{aligned} \tag{4}$$

where the matrices $\tilde{\mathbf{U}}$, $\tilde{\mathbf{\Sigma}}$, and $\mathbf{L}$ are of dimensions $(M \times m)$, $(m \times m)$, and $(M \times m)$, respectively. Let r_{jk} be the correlation between SNP j and SNP k, $j \neq k$, and $\tilde{r}_{jk}$ be a correlation term after dimension reduction which can be expressed as follows:

$$\tilde{r}_{jk} = \sum_{i=1}^{m} l_{ji} l_{ki}, \tag{5}$$

where l_{ji} is an entry in $\mathbf{L}$. The obtained covariance matrix $\tilde{\mathbf{C}}_{\mathrm{S}}$ is considered a good estimate of $\mathbf{C}_{\mathrm{S}}$ if the standard deviation of the distribution of $r_{jk} - \tilde{r}_{jk}$ (the residual correlations) is less than or equal to the standard deviation of a distribution with zero correlation, i.e. $\sigma_{r=0} = N^{-1/2}$ where N is the total number of samples [25]. Consequently, the smallest singular value is set to zero at each step in $\tilde{\mathbf{\Sigma}}$ until a dimension m is found where the standard deviation of the residual correlations is greater than $N^{-1/2}$. Hence, the number of SNP expressions P that has to be estimated by ICA is equal to $m+1$.

B. Application of ICA

The number of expressions P to be estimated is given to the PCA algorithm which performs whitening and dimension reduction. PCA is a linear transformation that transforms a given data consisting of a large number of interrelated variables to a new coordinate system. The greatest variance, i.e. the first principal component, is projected on the first basis, the second principal component on the second basis, and so on. The output of the PCA algorithm $\mathbf{T} \in \mathbb{R}^{N \times P}$ after whitening and dimension reduction can be expressed as:

$$\mathbf{T} = \mathbf{S}\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}^{-1/2}, \tag{6}$$

where $\tilde{\mathbf{\Sigma}} \in \mathbb{R}^{P \times P}$ is a diagonal matrix containing the P most important eigenvalues and $\tilde{\mathbf{U}} \in \mathbb{R}^{M \times P}$ is the matrix containing the P most important eigenvectors. The reduced dimension matrix $\mathbf{T}$ is then given to the ICA algorithm to estimate the SNP expressions matrix $\mathbf{E}$ and the pseudoinverse matrix $\mathbf{D}$ of the SNP coefficients matrix $\mathbf{A}$. A fundamental assumption for ICA to function properly is that the expressions follow a non-Gaussian distribution [18]. ICA works by trying to maximize the non-Gaussianity of each of the components. The kurtosis and the negentropy are two commonly used measures of non-Gaussianity.

C. SNP Expressions Significance Analysis

The obtained SNP expressions do not relate yet to the phenotype under study. Therefore, the next step is to perform a significance analysis between the SNP expressions and the phenotype. In the proposed method, regression analysis is used to measure the distance of each SNP expression to the phenotype. Logistic regression can be applied for case-control studies and linear regression can be applied for quantitative trait loci studies.

The inputs to the regression analysis are the phenotype vector $\mathbf{t} \in \mathbb{R}^{N\times1}$ and the SNP expressions matrix $\mathbf{E} \in \mathbb{R}^{P\times N}$, and the output is the regression vector $\mathbf{k} \in \mathbb{R}^{P\times1}$. The SNP coefficients matrix $\mathbf{A}$ contains the contribution of the SNPs to each of the estimated SNP expressions. To obtain the contribution of each SNP to the SNP expressions taking into account the phenotype, each term of $\mathbf{k}$ should be multiplied by the corresponding column of the SNP coefficients matrix as follows:

$$\mathbf{W} = \left(\mathbf{a}_1 \cdot k_1 \ \ldots \ \mathbf{a}_P \cdot k_P \right), \tag{7}$$

where matrix $\mathbf{W} \in \mathbb{R}^{M\times P}$ is the weighted SNP coefficient matrix, $\{\mathbf{a}_1, \ldots, \mathbf{a}_P\}$ are the columns of matrix $\mathbf{A}$, and $\{k_1, \ldots, k_P\}$ are the calculated regression coefficients.

This shows that the more a SNP expression is related to the phenotype, the larger is the magnitude of its regression coefficient and, thus, the larger is the corresponding SNPs contribution to the phenotype. The next step is to choose the columns of the matrix $\mathbf{W}$ that are most relevant to the phenotype. This is done by choosing the columns in which the regression coefficients have p-values less than a threshold for a given significance level. This threshold p-value will be denoted as *components p-value* in the sequel. Each of the chosen columns of $\mathbf{W}$ contains the weighted SNPs contribution to the phenotype. Thus, the entries of each of the columns determine how much each SNP has an effect on the phenotype. These entries will be denoted as *SNP factors* in the results presented in Section IV.

D. SNPs Significance Analysis

The final step in the method is to find out the significant SNPs in each of the selected associated SNP expressions (selected columns of $\mathbf{W}$). This is performed using permutation test with a sufficient number of rounds [27]. In each round, the permutation test can be implemented either by permuting the given SNP genotype data and performing all the steps of the method or by permuting the phenotype vector and performing only the steps of the method after ICA. The latter option is selected since it requires less computations.

The significance level associated with every SNP is determined in a permutation test where the case/control labels in the phenotype vector $\mathbf{t}$ are permuted. In each round of the permutation test, regression analysis is performed between the permuted phenotype vector and the SNP expressions matrix $\mathbf{E}$ to compute a new weighted SNP coefficients matrix $\mathbf{W}$ using (7). Then, the maximal SNP factor among all SNP expressions is selected. After running a large number of rounds, a frequency histogram is constructed and used to calculate a threshold SNP factor based on a target p-value that corresponds to a given significance level. This target p-value will be denoted as *SNPs p-value* in the sequel. The selected maximal SNP factors are global as just one value is taken out of each round. Therefore, we have adjusted for multiple testing by the permutation test procedure.

IV. Results and Analysis

This section presents the results of the proposed method applied to a case-control simulation study in addition to a real clinical data set. For the case-control simulation, a haplotype population was generated by a coalescent approach allowing for random mutations and recombinations (recombination parameter $4N_e r = 100$ for 100 kb) [28]. Haplotypes were characterized by a sequence of SNP alleles. SNPs with a minor allele frequency of less than 0.05 were excluded and two loci with minor allele frequencies between 0.1 and 0.3 were chosen as the causal variants. A two-locus multiplicative association model with allelic relative risk of 1.5 and a phenocopy rate of 0.01 was specified [29]. A genotyping error, i.e. a misspecified allele, of rate 0.002 was allowed. After removing the causal SNPs (at positions between 11 and 12 and between 37 and 38), case-control pairs with a given number of SNPs per individual were drawn. We used the explained simulation procedure to generate two simulated data sets. The first data set is denoted as *sim2loci* and is composed of 2,000 samples divided equally among cases and controls with 50 SNPs per sample. Similarly, the second data set is composed of 10,000 samples divided equally among case and controls with 50 SNPs per sample. On the other hand, the real data set contains (unpublished) clinical data of individuals with the Schizophrenia disease and of age, sex, and geographically matched control subjects. The data set is composed of 184 cases and 184 controls with 42 SNPs per individual.

The SNPs input data is coded as follows. For a binary phenotype, cases are given the code 1 while controls are given the code 0. For a genotype with example allele combinations from the set {AA, AT, TA, TT}, the coding scheme is presented in Table I. The genotypes AA and TT are homozygous and hence are given two different codes. However, we assume that the haplotype phase is unknown and, thus, the heterozygous genotypes AT and TA are given the same code. Note that the proposed method extends automatically to data sets where the haplotype phase has been estimated. Moreover, we assume that missing values due to genotyping failures are either removed or estimated. In the simulated data sets, there are no missing values. In the real clinical data set, around 3% of the data are missing values which are estimated using the BPCA algorithm as described in Section V. Using simulated data sets for the evaluation of the proposed method has several advantages. First, the exact number and position of the causal SNPs are known. Besides, the data sets are homogeneous, so that spurious associations among badly matched case-control samples will not present a problem. Moreover, the controlled introduction of missing values in the simulated data sets can help in testing the performance of missing value estimation algorithms (see Section V).

TABLE I
GENOTYPE CODING FORMAT.

Genotype	Code
AA	1
TT	2
AT and TA	3
Genotyping failure	0

The obtained results with the proposed method are compared with two other standard genetic mapping methods: a single marker logistic regression method [30] and a stepwise multiple logistic regression method [1].

A. The sim2loci Data Set

The *sim2loci* data set is based on a full multiplicative risk model where the two loci affect the phenotype without interaction. The causal polymorphisms are located between positions 11-12 and 37-38, respectively and were removed from the data. This is usually done to test the robustness of the methods since in real life the causal SNPs are not known during genotyping; however, the causal SNPs can still be determined by way of correlation between SNPs (linkage disequilibrium).

The number of expressions estimator resulted in $P = 15$ components which contain 68% of the variance of the total data. Applying dimension reduction with $P = 15$ followed by ICA, Figure 2 presents the 15 estimated SNP expressions in addition to the p-values of the regression coefficients of the estimated expressions (components). Two SNP expressions (9 and 10) are selected as significant with component p-values less than 0.01 and, thus, are plotted in Figure 3 along with the outcome of the single marker logistic regression and stepwise multiple regression methods.

For the three methods, the SNPs significance level for 99% confidence is also plotted in order to determine the decision border upon which a SNP is selected as associated or not. It can be noticed that the three methods have successfully determined the locations of the causal regions. The highest peaks in all three methods are at locations 14 and 34. One might think that the methods are not accurate since the real locations are defined between 11-12 and 37-38. However, the causal SNPs were removed from the simulated data and the peaks found correspond to the SNP locations that are in high correlation with the skipped ones. Note that the proposed method performed better than single marker logistic regression and similar to stepwise regression. Moreover, it was able to additionally determine the independent effects of the two causal regions since each one belongs to a different component (SNP expression). This demonstrates a novel feature of the proposed method.

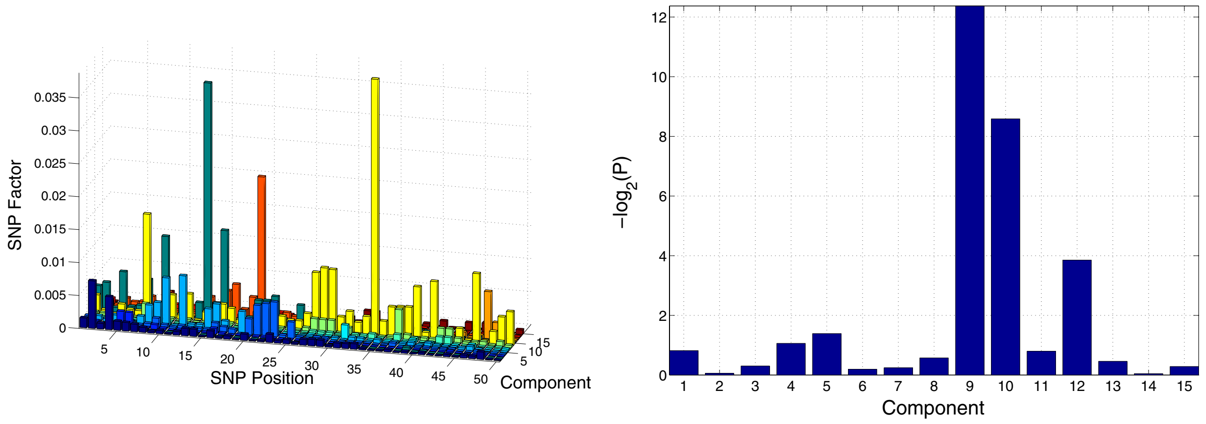


Fig. 2. Results for the *sim2loci* data set. Left: Plot of all 15 estimated components. Right: Plot of the p-values (logarithm transformed) of the regression coefficients of the components.

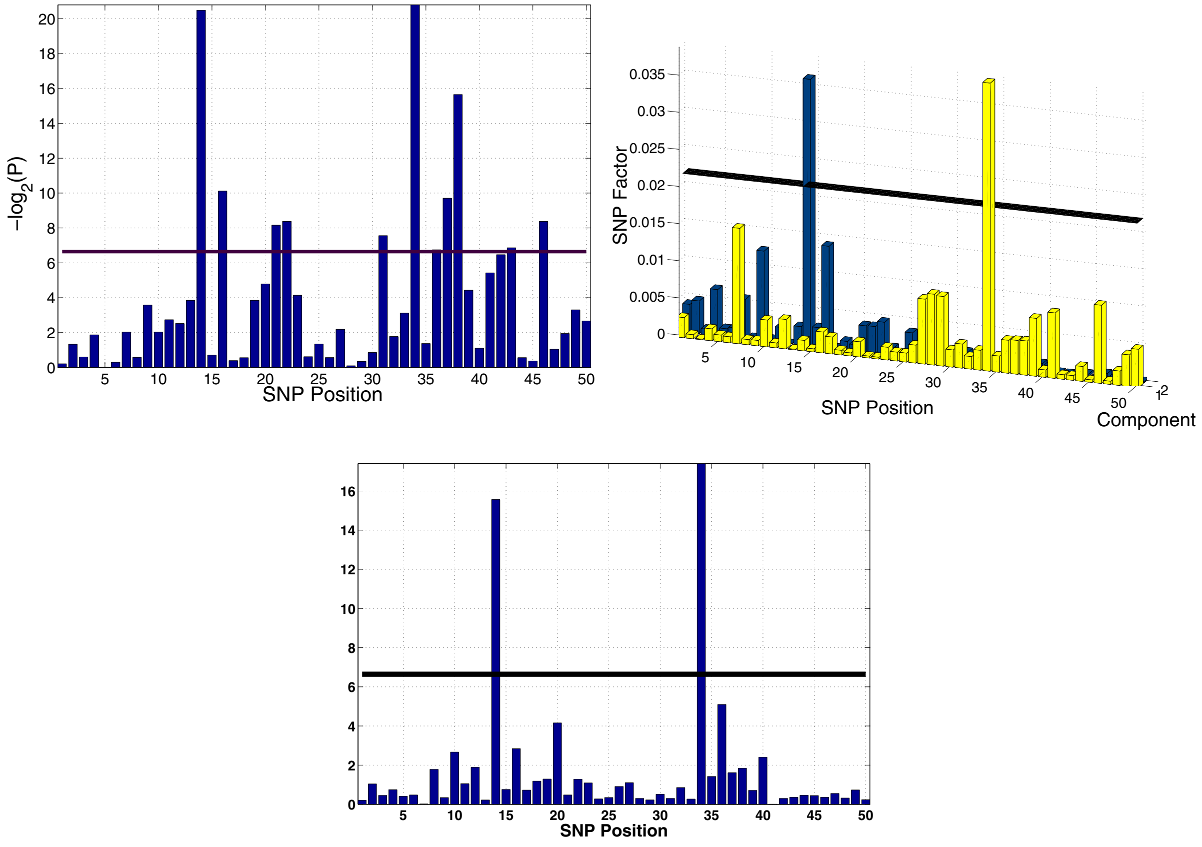


Fig. 3. Results for the *sim2loci* data set. Upper left: Plot of the p-values (logarithm transformed) of ordinary logistic regression. Upper right: Plot of the SNP factors of the highly associated components of the proposed method. Bottom: Plot of the p-values (logarithm transformed) of stepwise regression.

B. Simulation Study

In order to further test the proposed method, we used the second simulated data set composed of 10,000 samples with 50 SNPs per sample. From this large data set, 100 data sets were chosen at random where each contains 1,000 samples equally divided between cases and controls. The proposed method in addition to single marker logistic regression and stepwise multiple logistic regression were applied on the 100 generated data sets to locate the associated SNPs with the phenotype.

For fair comparison and multiple testing adjustment, permutation testing was applied on each data set for all three methods in order to determine the thresholds of global significance (p-value thresholds for regression methods and SNP factor threshold for proposed method) based on a given significance level. For each data set, 1000 permutation rounds were performed by permuting the phenotype vector in order to build frequency histograms for SNPs with maximum association. In the simulated data set, there are two regions of phenotype-associated SNPs; these are SNP region 11-15 and SNP region 34-38. SNPs outside these regions are considered as unassociated with the phenotype.

For each method, the analysis done consisted of counting the number of times where exactly the two associated SNPs were found to be significant, the two associated SNPs were found to be significant along with false positives (FPs), exactly one of the associated SNPs was found to be significant, and only one of the associated SNPs was found to be significant along with false positives. Table II presents the outcome of this study for 95% and 99% SNPs significance levels. These significance levels correspond to *SNPs p-values* of 0.05 and 0.01, respectively.

TABLE II
PERFORMANCE RESULTS OF SIMULATION STUDY.

	SNPs p-value=0.05				SNPs p-value=0.01			
Method	Exactly 2	2 + FPs	Exactly 1	1 + FPs	Exactly 2	2 + FPs	Exactly 1	1 + FPs
Proposed method	61	7	25	7	25	0	72	3
Logistic regression	15	5	69	11	8	1	83	8
Stepwise regression	55	9	28	8	42	3	50	5

A false negative SNP is a SNP that belongs to the associated regions, but is detected by the method to be unassociated. A false positive SNP is a SNP that does not belong to the associated regions, but is detected by the method to be associated. Table III presents the false positive and false negative results of the three methods. Table III can be derived from Table II as follows: the number of false positives is the sum of the second and fourth columns and the number of false negatives is the sum of the third and fourth columns.

TABLE III

FALSE POSITIVES AND FALSE NEGATIVES OF SIMULATION STUDY.

	SNPs p-value=0.05		SNPs p-value=0.01	
Method	False positives	False negatives	False positives	False negatives
Proposed method	14	32	3	75
Logistic regression	16	80	9	91
Stepwise regression	17	36	8	55

For 95% SNPs significance level, the proposed method detected exactly the two associated SNPs 61% of the time while the other two methods detected exactly two associated SNPs 55% and 15% of the time. However, for 99% SNPs significance level, the proposed method was able to detect only 25% of the time exactly the two associated SNPs which resulted in a high rate of false negatives. In general, the performance of the proposed method is relatively comparable with stepwise multiple logistic regression while the performance of single marker logistic regression is worse.

In Table IV, we investigate the effect of the choice of the *components p-value* on selecting the highly associated SNP expressions with the phenotype (see Section III-C). These results demonstrate the importance of proper selection of the *components p-value* parameter on the performance of the proposed method.

TABLE IV

EFFECT OF COMPONENTS P-VALUE ON FALSE POSITIVE RESULTS.

	SNPs p-value=0.05			
Components p-value	Exactly 2	2 + FPs	Exactly 1	1 + FPs
0.05	61	7	25	7
0.01	53	4	35	8
0.001	36	0	60	4

C. Schizophrenia Data Set

The Schizophrenia data set is composed of 42 candidate SNPs collected from different coding regions and different chromosomes. The causal SNPs are not known yet for this data set. The outcome of the proposed method is depicted in the left plot of Figure 4 for 12 components and a retaining variance of 63%, while the right plot presents the p-values of the regression coefficients of the 12 components. As can be noticed, the second component has the highest p-value equal to 0.0005 while the second highest p-value is only 0.0873. Figure 5 shows the final results of the three methods with the significance level for 99% confidence. In the proposed method, the significant SNPs have the locations 29 and 30. These two peaks do not contradict with the ones in the regression based methods. However, the output of the logistic regression method seems to have more significant peaks (e.g. at positions 10 and 16). It can also be seen that the proposed method contains less side peaks than the two other methods. This is due to the fact that in the first stage of the proposed method, ICA forms first independent groups of SNPs and then SNP expressions regression analysis eliminates those groups that are insignificant. In other words, the other peaks were probably eliminated since they belong to insignificant groups.

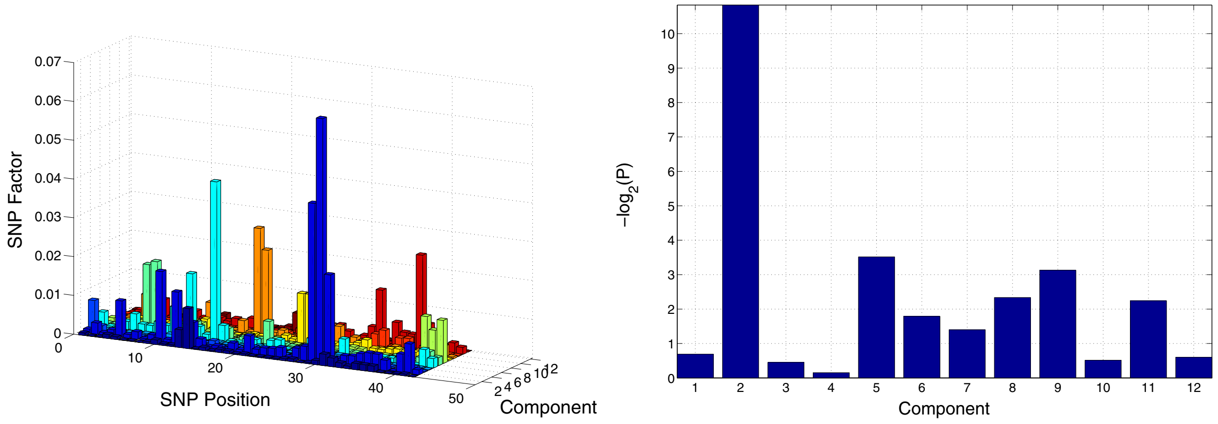


Fig. 4. Results for the Schizophrenia data set. Left: Plot of all the components. Right: Plot of the p-values (logarithm transformed) of the regression coefficients of the components.

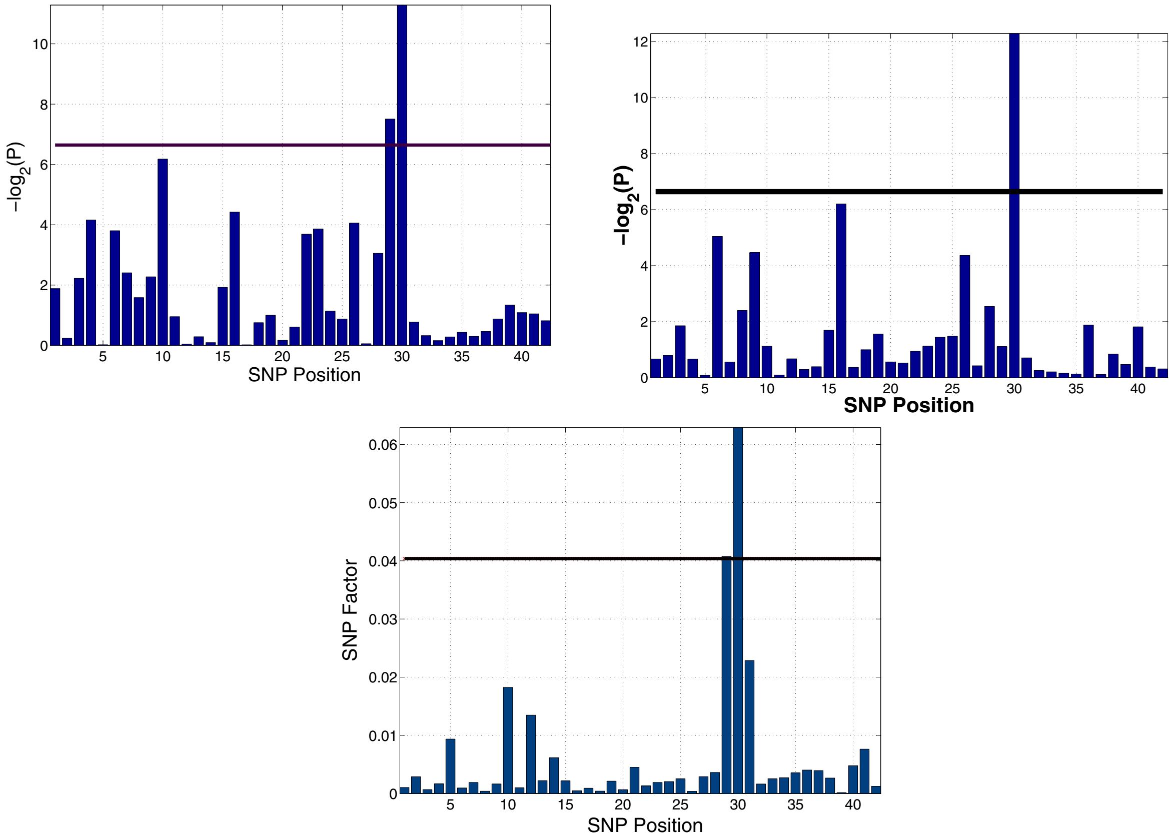


Fig. 5. Results for the Schizophrenia data set. Upper left: Plot of the p-values (logarithm transformed) of ordinary logistic regression. Upper right: Plot of the p-values (logarithm transformed) of stepwise regression. Bottom: Plot of the SNP factors of the highly associated components of the proposed method (component 2).

V. Implementation Issues

In this section, we investigate the effects of missing values and sample size on the performance of the proposed method in addition to some comments on its computational complexity.

A. *Missing Values*

In the given SNP matrix, the values of some SNPs may be missing due to genotyping failures. To overcome this problem, the missing values should be either estimated or their samples skipped. The latter option is not favorable since the amount of available data in clinical data sets is normally scarce. Several solutions have been proposed in the literature to estimate the missing values [20], [21], [22]. The method that we adopt is Bayesian PCA (BPCA) because it gives lower estimation error rates when compared to the other methods, does not require any assumption of an underlying model, and converges almost always to one solution [23].

In order to apply BPCA, the SNP matrix is divided into two matrices: one matrix contains all samples without missing values and the other matrix contains all samples with missing values. BPCA starts by performing PCA to both matrices after replacing the missing values with the SNP-wise averages to determine the principal basis. The next step is to estimate the posterior distribution of the missing values. This requires an iterative solution which is efficiently performed using the variational Bayes algorithm. Finally, the missing values are filled according to the estimated distribution, see [23] for more details on the derivation. Note that BPCA was originally designed for microarray data, so it outputs continuous values. Hence, we modify the algorithm by quantizing the outputs to the nearest integer since the SNPs belong to a finite set of values.

In order to evaluate the performance of BPCA, artificial missing entries are introduced in the *sim2loci* data set. These entries are chosen by selecting a specific percentage and removing them randomly from the SNP matrix. The percentage of the introduced missing values is fixed and the experiment is repeated a large number of times in order to obtain average results. Then, the error rate of the missing value estimation is obtained by evaluating the Mean Bit Error Rate (mBER) for the used percentage value. The mBER is calculated by taking the ratio of the mean number of error estimates to that of the total estimates. The mBER approaches its minimum value 0 when the estimation is error free. In the other extreme case, i.e. when the estimation is too poor due to large amount of noise, mBER approaches 1.

We compare BPCA to two other methods. In the first method, the missing values are replaced by the SNP-wise averages and then rounded to the nearest integer. The other method consists of filling the missing values according to an estimated SNP-wise empirical probability distribution. Inserting 5% missing values in the data set *sim2loci*, we obtained the following mBER results: 0.68 for the empirical probability method, 0.4 for the averaging method, and 0.22 for the used BPCA method. The mBER of BPCA is the lowest due to the fact that BPCA tries to find the distribution that best fits the whole data. Moreover, the outcome of the averaging method is shown to be better than that of the SNP-wise empirical probability distribution. In the latter case, the missing values are filled arbitrarily according to the estimated probabilities and this introduces in some cases more errors in the results.

In order to test the impact of missing value estimation on the performance of the proposed genetic mapping method, the following study was performed. A percentage of missing values was introduced in the SNP matrix of the *sim2loci* data set, and then the proposed method was applied for the following two cases: removing all individuals with missing values versus estimating the missing values using BPCA. Table V presents the obtained results. It is formed of the following columns: the percentage of introduced missing values, the number of samples left after removing all individuals with missing values, genetic mapping outcome if samples with missing values are removed, the missing value estimation error with BPCA (number of errors out of total number of introduced missing values), and genetic mapping outcome if missing values are estimated. Note that since the proposed method requires matrix operations, we cannot omit only missing value entries but we have to omit all the individuals (samples) with missing values, i.e. remove the whole row from the SNP matrix.

TABLE V
GENETIC MAPPING RESULTS WITH MISSING VALUES.

% missing values	remaining samples	genetic mapping results	estimation error	genetic mapping results
0.5	1558	both SNPs	41 out of 500	both SNPs
1	1207	both SNPs	79 out of 1000	both SNPs
1.5	954	only one SNP	118 out of 1500	both SNPs
2	727	only one SNP	160 out of 2000	both SNPs
2.5	563	none	211 out of 2500	both SNPs

It can be seen that missing value estimation is very beneficial and helps in improving the accuracy of the proposed genetic mapping method. Moreover, it is seen that the percentage of missing value estimation errors is nearly constant as the percentage of missing values increases. This is due to the fact that BPCA tries to find the distribution that best fits the whole data. Running more simulations, we found out that starting from around 40% missing values, the proposed method was not always able to correctly find both associated SNPs even after missing value estimation.

B. Sample Size

The second implementation issue that we considered is the effect of the the available sample size (number of individuals) on the performance of the proposed method. The given sample size affects the performance of the number of expressions estimator, the performance of the ICA algorithm, and the regression analysis to select the associated SNP expressions. The effect of the sample size will be investigated by running experiments with variable sample sizes.

The dependency of the number of components on the sample size is first investigated. The number of SNP expressions for the *sim2loci* data set is computed for different sample sizes (see Figure 6). As the number of samples increases, the number of components also increases with an envelope inversely proportional to $\sigma_{r=0}$ (see Section III-A). This is also verified in the figure by the stair-case plot of $N^{1/2}$. Note that the stair-case shape of the output is expected because the number of components is limited to the set of integer numbers. As N tends to infinity, $\sigma_{r=0}$ tends to zero and, hence, dimension reduction will not be necessary anymore. This is justified from the fact that as the number of samples increases, the statistical analysis of the data improves eliminating the need of reducing the dimension. For example, the number of components to be estimated is 15 for 2000 samples, whereas it is 2 for 200 samples and 50 when the number of samples tends to infinity. Note also that the number of components to be estimated depends on the data itself since the covariance matrix of the data is involved in the computations as shown in (4). Thus, for a specific number of samples, the number of components to be estimated will differ from one data set to another.

The variation in the number of components and in the sample size will certainly affect the output of the ICA algorithm. To test this effect, the outcome of the method is tested on the *sim2loci* data set for different sample sizes. For comparison purposes, the same test is conducted on the stepwise multiple

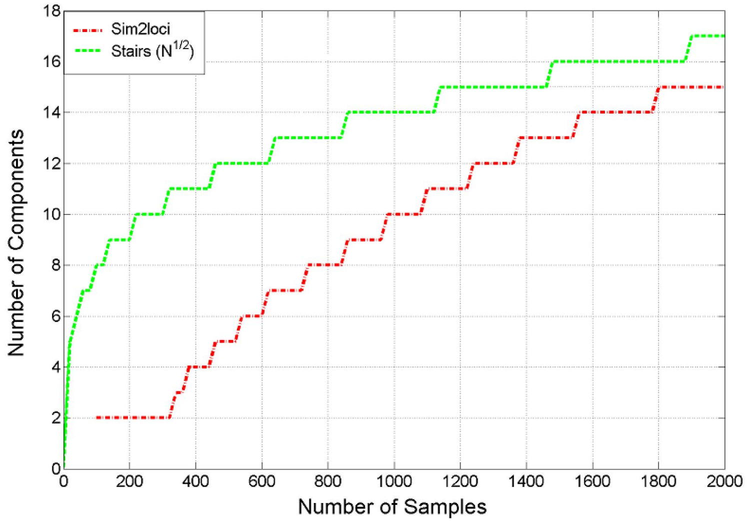


Fig. 6. Sample size effect on the number of expressions (components) estimator.

logistic regression method. Figure 7 and Figure 8 present the outcome of the two methods for 1,000 samples and 200 samples, respectively (note that original data set size is 2,000 samples). Compared to Figure 3, it can be seen that with 1,000 samples both methods are able to pinpoint the correct causal regions. However, when the number of samples is reduced to 200, the stepwise regression method performs worse due to the presence of other significant peaks. The proposed method has performed better in this case because the reduction of the number of samples is compensated by a decrease in the number of SNP expressions to be estimated. With 200 samples, only two components were estimated compared to 15 components with 2,000 samples. This study demonstrates yet another advantage of the proposed method as it proves that the proposed method is more robust to estimation noise when the number of available samples is limited.

Finally, we consider the effect of sample size on the SNP expressions regression analysis. When the number of samples decreases, the fitting will not be as representative as before because the sum of the square errors also known as Mean Square Error (MSE) is inversely proportional to the sample size N. The MSE can be calculated as:

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} \epsilon_n^2, \tag{8}$$

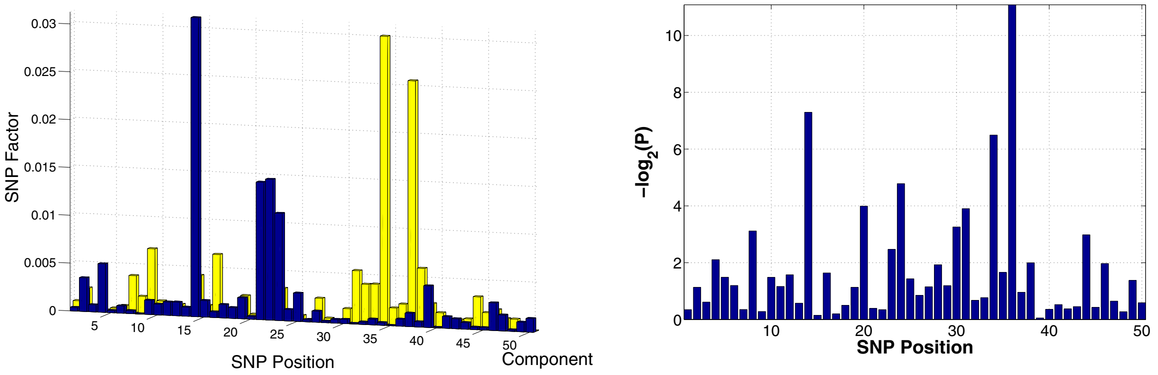


Fig. 7. Results of *sim2loci* data set with 1,000 samples. Left: Outcome of the proposed method. Right: Outcome of the stepwise regression method.

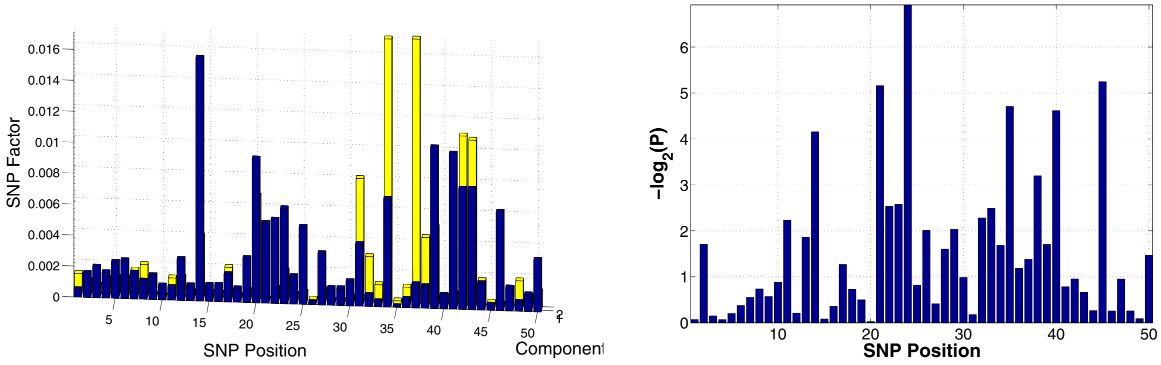


Fig. 8. Results of *sim2loci* data set with 200 samples. Left: Outcome of the proposed method. Right: Outcome of the stepwise regression method.

where ϵ is the residual error, i.e. the difference between the real and fitted values. The MSE was calculated for the Schizophrenia data set with 368 samples and with 184 samples (half sample size). The obtained MSE values were $1.07 \cdot 10^{-4}$ and $2.5 \cdot 10^{-3}$, respectively. This shows the effect of the sample size on the accuracy of the regression analysis.

C. Computational Complexity

The computational complexity of the proposed method can be divided into two main parts. The first part is the ICA block with dimension reduction and SNP expressions estimation. Several low complexity implementations have been proposed for ICA to speed its execution time, see [31] for a detailed complexity analysis of ICA with comparisons between different implementations. In our method, running the ICA block on a Pentium IV PC with 1.5 GHz requires around 10 sec. for the Schizophrenia data set. On the other hand, the second part is regression analysis (SNP expressions regression analysis and SNPs regression analysis with permutation test). The computational complexity of this part is comparable to ordinary genetic mapping methods based on regression analysis with permutation testing. Moreover, it depends on the number of rounds of the permutation test.

The proposed method is implemented in Matlab language. The ICA computations with dimension reduction are performed using the FastICA package [32]. Note that FastICA packages are freely available on the web for R and Matlab languages. The regression analysis to select associated SNP expressions is performed using the function *robustfit* in Matlab. The single marker logistic regression method is implemented with the help of the function *glmfit* in Matlab. The stepwise multiple logistic regression method is implemented with the help of the function *stepwisefit* in Matlab. Finally, missing value estimation using BPCA is performed using the Matlab toolbox provided by [23].

VI. Conclusions

In this work, we propose a novel method for fine-scale genetic mapping based on Independent Component Analysis (ICA). ICA is a well established algorithm that is available in commonly used statistical packages and is proven to be successful for a wide spectrum of applications [18] . The proposed method is composed of two stages. In the first stage, it forms independent groups of SNPs according to a linear model and selects the SNP groups that are highly associated with the phenotype. In the second stage, it performs association analysis on each of the identified SNP groups to select individual SNPs that are highly associated with the phenotype. Therefore, the method assumes that SNPs get mixed in an unknown linear environment to produce independent SNP expressions which affect a given phenotype.

The proposed method is tested on a real clinical data set for the Schizophrenia disease in addition to several simulated data sets. The obtained results are compared with a single marker logistic regression method and a stepwise multiple logistic regression method. It is shown that the proposed method performs better than logistic regression and comparably to stepwise regression. These comparisons prove the validity of the method in terms of accuracy of results and demonstrate its novel characteristics. The proposed method reduces estimation noise as regression analysis is performed on selected independent groups of SNPs that represent the major genetic variation in the given data. As a result, associated SNPs that belong to different SNP groups can be assumed to have an independent effect on the phenotype. Moreover, it accounts for correlations between markers and is robust for data sets with limited number of samples. As a summary, the proposed method is shown to be a successful alternative to other multi-locus genetic mapping methods with its own advantages and disadvantages.

The proposed method can be extended to include prior knowledge about the phenotype in the implementation of the ICA algorithm. This prior knowledge would then shape the ICA solution. Such an extension is useful for situations in which multiple genes influence the phenotype but the effect of each individual gene is weak or in situations where SNPs interact indirectly through a gene network to influence the phenotype. Another possible extension is to apply stepwise regression with forward and backward selection in order to pick the expressions that are highly associated with the phenotype.

Genotyping errors have an effect on all genetic mapping and linkage disequilibrium methods and thus will also affect our proposed method, e.g. see [33]. Spurious association signals due to cryptic population substructures are also a general problem of population-based genetic mapping methods. We do not try to tackle this problem in this work and, thus, assume homogeneous samples.

APPENDIX

In this section, the various steps of the proposed method are demonstrated by applying it on a simple example. The input data is composed of $N = 5$ individuals where for each individual a SNP sequence of size $M = 3$ is given. The SNP matrix is given by:

$$\mathbf{S} = \begin{pmatrix} 1 & 3 & 3 \\ 2 & 2 & 2 \\ 3 & 2 & 3 \\ 2 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix}$$

The covariance matrix of $\mathbf{S}$ and its singular value decomposition (SVD) can be expressed as:

$$\begin{aligned} \mathbf{C}_S &= \begin{pmatrix} 0.5 & -0.25 & 0 \\ -0.25 & 0.5 & 0.5 \\ 0 & 0.5 & 0.7 \end{pmatrix} \\ &= \begin{pmatrix} -0.25 & 0.89 & -0.38 \\ 0.65 & 0.14 & -075 \\ 0.72 & -0.43 & 0.54 \end{pmatrix} \cdot \begin{pmatrix} 1.15 & 0 & 0 \\ 0 & 0.54 & 0 \\ 0 & 0 & 0.01 \end{pmatrix} \cdot \begin{pmatrix} -0.25 & 0.89 & -0.38 \\ 0.65 & 0.14 & -075 \\ 0.72 & -0.43 & 0.54 \end{pmatrix} \end{aligned}$$

Taking out the lowest singular value, the obtained covariance matrix can be expressed as:

$$\begin{aligned} \tilde{\mathbf{C}}_S &= \begin{pmatrix} -0.25 & 0.89 \\ 0.65 & 0.14 \\ 0.72 & -0.43 \end{pmatrix} \begin{pmatrix} 1.15 & 0 \\ 0 & 0.54 \end{pmatrix} \begin{pmatrix} -0.25 & 0.65 & 0.72 \\ 0.89 & 0.14 & -0.43 \end{pmatrix} \\ &= \begin{pmatrix} 0.498 & -0.25 & 0.002 \\ -0.25 & 0.49 & 0.5 \\ 0.002 & 0.5 & 0.697 \end{pmatrix} \end{aligned}$$

The standard deviation of the distribution of the elements of $\mathbf{C}_S - \tilde{\mathbf{C}}_S$ is equal to 0.0035 which is certainly less than $\left(1/\sqrt{5}\right) = 0.448$. Hence, the data can be represented by two dimensions instead of three. This dimension number will be given to the ICA algorithm which will perform dimension reduction using PCA and then estimate the SNP expressions matrix $\mathbf{E}$ along with the SNP coefficients matrix $\mathbf{A}$ which are given by:

$$\mathbf{E} = \begin{pmatrix} -4.72 & -1.07 \\ -3.48 & -2.52 \\ -4.66 & -4.23 \\ -3.47 & -2.53 \\ -1.98 & -2.62 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} -0.25 & -1.36 \\ -0.55 & 0.44 \\ -0.94 & -0.34 \end{pmatrix}$$

Assume that the phenotype vector $\mathbf{t} = (0\ 0\ 0\ 1\ 1)^T$ where T is matrix transpose operation. The obtained regression coefficients are -0.211 and 0.06. According to Equation (7), the weighted SNP coefficient matrix is given by:

$$\mathbf{W} = \begin{pmatrix} -0.0156 & 0.286 \\ -0.035 & -0.0932 \\ -0.059 & 0.073 \end{pmatrix}$$

The elements of $\mathbf{W}$ give the contribution of each SNP to the phenotype for both components (SNP expressions). For this simple example, it can be easily noticed from the magnitude of the regression coefficients that the effect of the first component is negligible compared to the second one (0.06 versus 0.211). Therefore, the second component is selected as associated with the phenotype. Moreover, the entry of the first SNP in the second component has the highest magnitude (SNP factor equal to 0.286) among the three SNPs. Therefore, the first SNP would most probably be selected as the associated SNP with the given phenotype.

References

[1] H. Cordell and D. Clayton, "A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in Type 1 diabetes," *Am. J. Hum. Genet.*, vol. 70, no. 1, pp. 124–141, January 2002.

[2] R. Zee, J. Hoh, S. Cheng, R. Reynolds, M. Grow, A. Silbergleit, K. Walker, L. Steiner, G. Zangenberg, A. Fernandez-Ortiz, C. Macaya, E. Pintor, A. Fernandez-Cruz, J. Ott, and K. Lindpaintner, "Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease," *The Pharmacogenomics Journal*, vol. 2, pp. 197–201, 2002.

[3] J. C. Mueller, J. Fuchs, A. Hofer, A. Zimprich, P. Lichtner, T. Illig, D. Berg, U. Wuellner, T. Meitinger, and T. Gasser, "Multiple regions of α-Synuclein are associated with parkinsons disease," *Ann Neurol*, vol. 57, pp. 535–541, April 2005.

[4] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nat Rev Genet*, vol. 2, pp. 91–99, 2001.

[5] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nat Rev Genet*, vol. 4, pp. 701–709, 2003.

[6] D. J. Balding, M. Bishop, and C. Cannings, *Handbook of statistical genetics*. Chichester: John Wiley and Sons, 2001.

[7] M. S. McPeek and A. Strahs, "Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping," *Am. J. Hum. Genet.*, vol. 65, pp. 858–875, 1999.

[8] A. P. Morris, J. C. Whittaker, and D. J. Balding, "Bayesian fine-scale mapping of disease loci, by hidden Markov models," *Am. J. Hum. Genet.*, vol. 67, pp. 155–169, 2000.

[9] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. Mueller, "Gene mapping and marker clustering using Shannon‘s mutual information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, pp. 47–56, January-March 2006.

[10] D. Zaykin, P. Westfall, S. Young, M. Karnoub, M. Wagner, and M. Ehm, "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals," *Human Heredity*, vol. 53, no. 2, pp. 79–91, May 2002.

[11] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," *Genome Research*, vol. 11, pp. 2115–2119, 2001.

[12] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland, "Score tests for association between traits and haplotypes when linkage phase is ambiguous," *Am. J. Hum. Genet.*, vol. 70, pp. 425–434, 2002.

[13] M. R. Nelson, S. L. R. Kardia, R. E. Ferrell, and C. F. Sing, "A combinatorial partitioning method to identify multilocus genotype partitions that predict quantitative trait variation," *Genome Research*, vol. 11, pp. 458–470, March 2001.

[14] C. Kooperberg and I. Ruczinski, “Identifying interacting SNPs using Monte Carlo logic regression,” *Genetic Epidemiology*, vol. 28, pp. 157–170, February 2005.

[15] O. Alter, P. Brown, and D. Botstein, “Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms,” *PNAS*, vol. 100, no. 6, pp. 3351–3356, March 2003.

[16] B. D. Horne and N. J. Camp, “Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation,” *Genetic Epidemiology*, vol. 26, pp. 11–21, January 2004.

[17] W. Liebermeister, “Linear modes of gene expression determined by independent component analysis,” *Bioinformatics*, vol. 18, no. 1, pp. 51–60, February 2002.

[18] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York, USA: Wiley, 2001.

[19] P. Comon, “Independent component analysis, a new concept?” *Elsevier Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[20] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, June 2001.

[21] L. Liu, D. Hawkins, S. Gosh, and S. Young, “Robust singular value decomposition analysis of microarray data,” *PNAS*, vol. 100, no. 23, pp. 13 167–13 172, November 2003.

[22] K. Gabriel and S. Zamir, “Lower rank approximation of matrices by least squares with any choice of weights,” *Technometrics*, vol. 21, no. 4, pp. 489–498, November 1979.

[23] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, “A Bayesian missing value estimation method for gene expression profile data,” *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, November 2003.

[24] H. F. Kaiser, “The varimax criterion for analytic rotation in factor analysis,” *Psychometrika*, vol. 23, pp. 187–200, 1958.

[25] M. Sarkis, Z. Dawy, F. Obermeier, and K. Diepold, “Automatic model-order selection for PCA,” in *International Confererence on Image Processing (ICIP 2006)*, October 2006.

[26] A. Machado, J. Gee, and M. Campos, “Visual data mining for modeling prior distributions in morphometry,” *IEEE Signal Processing Mag.*, vol. 21, no. 3, pp. 20–27, May 2004.

[27] G. A. Churchill and R. W. Doerge, “Empirical threshold values for quantitative trait mapping,” *Genetics*, vol. 138, no. 3, pp. 963–971, November 1994.

[28] R. Hudson, “Generating samples under a Wright-Fisher neutral model of genetic variation,” *Bioinformatics*, vol. 18, pp. 337–338, February 2002.

[29] M. Nothnagel, “Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-cotrol data by supervised learning methods,” *American Journ. of Human Genetics*, vol. 71 (Suppl.), no. A2363, October 2002.

[30] D. Collett, *Modelling Binary Data*, 2nd ed. Chapman and Hall/CRC Press, 2002.

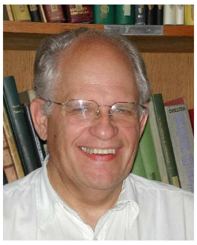
[31] S. Shwartz, M. Zibulevsky, and Y. Y. Schechner, “ICA using kernel entropy estimation with NlogN complexity,” in *Fifth International Conference on Independent Component Analysis (ICA 2004)*, Aussois, France, September 2004.

[32] A. Hyvaerinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[33] J. M. Akey, K. Zhang, M. Xiong, P. Doris, and L. Jin, “The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures,” *Am. J. Hum. Genet.*, vol. 68, pp. 1447–1456, 2001.

Zaher Dawy received the B.E. degree in Computer and Communications Engineering with high distinction from the American University of Beirut (AUB) in 1998. He received his M.Sc. and Dr.-Ing. degrees in Electrical Engineering with excellent distinction from Munich University of Technology (TUM) in 2000 and 2004, respectively. Between 1999 and 2000, he worked as a part-time Communications Engineer at Siemens AG research labs in Munich focusing on the development of enhancement techniques for UMTS. At TUM, between 2000 and 2003 he managed a research project with Siemens AG where he designed advanced multiuser receiver structures for UMTS base stations. Since September 2004, he is an Assistant Professor at the Electrical and Computer Engineering department at the American University of Beirut. His research interests are in the areas of Wireless Communications, Multiuser Information Theory, and Computational Biology.

Michel Sarkis received a Bachelor in electrical engineering at the American University of Beirut (AUB), Lebanon, in 2002 and a M.Sc. in electrical engineering at Munich University of Technology (TUM), Germany in 2004 with a Master Thesis emphasizing on Genomic Signal Processing. Currently, he is pursuing his PhD thesis in the Telepresence Teleaction project at the Institute for Data Processing (LDV) at TUM. His research interests include image processing, differential geometry, time-varying systems, computer vision, signal processing, and statistical genetics. He is a graduate student member at IEEE.



Joachim Hagenauer received his degrees from the Technical University of Darmstadt. He held a postdoctoral fellowship at the IBM T.J. Watson Research Center, Yorktown Heights, NY, working on error-correction coding for magnetic recording. Later he became a Director of the Institute for Communications Technology at the German Aerospace Research Center DLR and since 1993 he holds a chaired professorship at the TU Munich. During 1986-1987 he spent a sabbatical year as an ”Otto Lilienthal Fellow” at Bell Laboratories, Crawford Hill, NJ, working on joint source/channel coding and on trellis coded modulation for wireless systems. He served as an editor and guest editor for the IEEE and for the ”European Transactions on Telecommunications (ETT)”. Joachim Hagenauer is a Fellow and a ”Distinguished Lecturer” of the IEEE. He served as President of the IEEE Information Theory Society. Amongst other awards he received in 1996 the E.H. Armstrong-Award of IEEE COMSOC, in 2003 the IEEE ”Alexander Graham Bell Medal” and an Honorary Doctorate from the University Erlangen-Nuremberg in 2006. His research interests concentrate on the turbo principle in communications and on the application of communication principles to genetics.

Jakob C. Mueller received his PhD degree at the Department of Population Biology of the Johannes Gutenberg-University in Mainz, Germany, where he served as an assistant professor. Since 2002 he holds positions as a senior scientist at the Max Planck Institute for Ornithology (Seewiesen), at the Institute of Human Genetics (GSF, Munich), and at the Hertie-Institute for Clinical Brain Research (Tuebingen). Jakob Mueller is also affiliated to the National Center for Genetic Epidemiological Methods. His research interests include population and evolutionary genetics in animals and humans as well as gene mapping for complex traits through linkage and association studies.