Statistics and Its Interface Volume 1 (2008) 197–207

A project of applied statistical methods in China: review and outlook*

Ning-Zhong Shi, Zhi Geng†, Jianhua Guo and Jian Tao

In this paper, we review and explore some important aspects of a key project on applied statistical methods supported by the National Natural Science Foundation of China. Our discussion focuses on (1) statistical inference under order restriction, (2) structural learning of graphical models, and (3) statistical analysis on haplotypes and recombination fractions in genetics and some research results on Hardy-Weinberg equilibrium (HWE) and admixture linkage disequilibrium (ALD).

AMS 2000 subject classifications: Primary 62P10, 62F30; secondary 62-09.
Keywords and phrases: Admixture linkage disequilibrium, Graphical model, Order restriction, Statistical genetics, Structural learning.

1. INTRODUCTION

Statistics is a science for extracting information from complex and noisy data with uncertainty. Applied statistical methods help analyze data to serve specific purposes in applications. Under the support of the National Natural Science Foundation of China as a key research project in applied statistics, our research is advancing on a broad front, from the development of new statistical methods based on probability theory, to novel applications of statistical methods to new data. In this paper, we review what we have done and explore some important issues in the the following three areas: statistical inference under order restriction, structural learning of graphical models, and statistical genetics.

Statistical inference under order restriction has been studied extensively [5, 62, 70].The origins of order restricted statistical inference are usually dated back to the early 1950s. The field developed rapidly during the 1960s and early 1970s and formed the more precise theory of estimation and testing under a variety of assumptions. Today, it remains a fertile area of statistical research. In the meantime, it has been increasingly used to deal with practical problems. Clinical trials, bioassays, biomedical sciences, genetics and bioinformatics are all among the interdisciplinary fields that have a growing need for statistical inference. For example, in a dose-response study, a therapeutic response may first increase with dosage, and then decreases as the dose level increases further due to adverse effect. The fact that the utilization of ordering information increases the efficiency of statistical inference procedures is well documented. Therefore, research in this area is of important practical significance. The problems of estimation and testing are the most important part of order restricted statistical inference. In this paper, we will review some main theoretical results that we have obtained, and we will pay more attention to some efficient applications of the order restricted methods in the interdisciplinary fields, such as bioinformatics.

Graphical models (i.e., undirected independence graphs, directed acyclic graphs (DAG) and Bayesian networks) have been widely applied to many fields [16, 42, 56, 76]. Structural learning of graphical models has been discussed by many authors [9, 32, 56, 76]. We have proposed several approaches for structural learning. For an undirected independence graphical model, given a vertex, its neighbor set can be found based on the mutual information between the vertex and a potential neighbor set [13]. For a very large graphical model, [87] and [90] presented decomposition approaches in which a problem of structural learning for a large graphical model is decomposed into problems for small models. [82] combined clustering with structural learning. Clustering can be used to facilitate structural learning of a graphical model with a large number of variables. On the other hand, graphical structure of variables in a cluster can be used to explain dependencies of variables within the cluster.

With the rapid progress of the Human Genome Project, a huge volume of SNPs have been discovered in the human genome. SNPs play a very important role in modern genetic epidemiology studies, including the study of fine-mapping of Mendelian disorders and the study of mapping common complex disorders. However, the information offered by single SNP is limited, which motivates us to consider many SNPs simultaneously in the mapping of disorders. The haplotype, which is a set of closely linked SNP alleles along a region of a chromosome, contains more genetic information than a single SNP. In practice, however, what can be obtained directly is genotype data but not haplotype data. Genotypes, which are obtained through widely used large-scale genotyping technologies for SNPs, do not contain phase information. In order to use the haplotype information in

*This research was supported by the National Natural Science Foundation of China (Grant Number 10431010).
†Corresponding author.

mapping of disorders, we should first determine all possible haplotype pairs that are compatible with the observed genotypes or at least estimate haplotype frequencies.

Molecular genetics has made much progress in recent years, among which linkage analysis fulfills an important role. Statistical machinery has been used to analyze family data and to detect linkage [17, 61, 55, 79].

This article is organized as follows. We first introduce our researches of statistical inference under order restriction with its application in Section 2, and we present our works on structural learning of graphical models in Section 3, then we discuss some problems of concern in statistical genetics and show some of our works on this field in Section 4. A brief conclusion is presented in last section.

2. ORDER-RESTRICTED INFERENCE WITH APPLICATIONS

In many biostatistical studies (such as microarray experiments and dose-response studies), experimental conditions usually have some inherent orderings. The performance of statistical inference can be improved significantly if this information can be appropriately utilized in the inferential procedure [57]. Order-restricted statistical inference is an efficient tool by using ordering information. Depending on the particular practical situation, one can use different order restricted methodologies.

2.1 Tests of homogeneity of odds ratios

For a given contingency table of ordinal variables, [64] proposed a test about the homogeneity of odds ratios against a partial order restriction. The inference of these odds ratios is considered on an extended hypergeometric distribution, a conditional distribution of cell frequencies given both marginal totals. By taking a transformation, the order restriction on the odds ratios was transformed into some linear inequalities restriction. Furthermore, a test is proposed from the transformation as a one-sided likelihood ratio test in the normal case and its asymptotic null distribution is the $\bar{\chi}^2$ distribution. In practice, many odds ratios exhibit a trend. For example, there is usually a simple order on the odds ratios. In the study of dose-response relationships, a unimodal trend may be considered, which is also said to be the umbrella order and includes the simple order. The proposed test can be easily applied to test homogeneity of the odds ratios against the simple or any other partial order restricted alternatives.

Under the background of dose-response and carcinogenesis studies, we proposed a new non-model-based significance test for detecting dose-response relationship with the incorporation of historical control data. Our non-model-based test is considered simpler from a regulatory perspective because it does not require validating any modeling assumptions. Moreover, our test is especially appropriate to those studies in which the intravenous doses for the investigational chemical are labeled as, e.g., low, medium and high or the dose labels do not suggest any obvious choices of dose scores. Morever, our test can be easily adopted for detecting general dose-response shape, such as an umbrella pattern [48, 24].

2.2 Restricted estimation for normal means

Estimation problems of means and variances from normal populations under simultaneous order restrictions have a profound theoretical basis and practical significance. For k normal populations with unknown means μ_i and unknown variances σ_i^2, $i = 1, 2, \ldots, k$, assume that there are some order restrictions among the means and variances respectively, such as simple order restrictions: $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ and $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_k^2 > 0$. [65] explored some properties of maximum likelihood estimation of μ_i's and σ_i^2's, and proposed an algorithm of obtaining the maximum likelihood estimators under the order restrictions. Furthermore, for the isotonic regression problem of normal means, it is usual to assume that all variances are known or unknown but equal. [67] generalized the procedure of [65] to a general case, i.e., there are no conditions imposed on the variances.

[57] proposed a powerful method based on the theory of order-restricted inference for selecting and clustering genes according to their time-course or dose-response profiles. The proposed method necessitates the assumption of a constant variance through time or among dosages. This homoscedasticity assumption is, however, seldom satisfied in practice. As an open problem, [57] expected that the estimation procedure for mean gene expression can be modified for the situation with unknown and unequal variances along the lines of [65]. [71] carried out the above task based on a bootstrap resampling procedure. Alternatively, we proposed a new method based on the algorithm proposed by [65] and [67] to deal with the heteroscedastic situation where a bootstrap technology is also utilized to get samples [45]. Simulation results show that the proposed alternative procedure can control the false positive rate better than that developed by [71].

2.3 Restricted EM algorithm

Incomplete data and order restriction are two important fields in statistics, which have been widely applied in engineering, biology, medicine, economics or social sciences, and so on. They often occur together in applications. Therefore, research on the two problems is very important in practice. [39] discussed the estimation under linear equality restrictions with missing data. [60] discussed the estimation under nonlinear equality restrictions and solved the problem of risk assessment of quantitative responses of a new drug by that method. Moreover, the research of estimation under restrictions

$$A_0\beta \geq \mathbf{a}$$

has extensive practical backgrounds. But the existing methods, including the method of [39], are not suitable for the

general problems. Based on the above reasons, we proposed a restricted EM algorithm under linear inequality restrictions in linear models, generalized linear models and Cox models etc. Furthermore, we have discussed some convergence properties of the EM sequence and the consistency of the restricted EM estimator and a related testing problem [69, 91].

In some biological experiments, it is quite common that laboratory subjects may be different in their patterns of susceptibility to a treatment. In these situations, finite mixture model analysis becomes a useful tool. Under a finite mixture normal model, we have studied the drug risk assessment problems under the assumption that the population at risk consists of multiple sub-types of different susceptibilities. The restricted EM algorithm is utilized to obtain the maximum likelihood estimates of the model. Particularly, we have also discussed the model selection problem based on Bayes factor, which is approximated by the Schwarz criterion. The practical significance of the proposed method is illustrated with an actual dose-response data set in [78].

2.4 New estimation techniques

For generalized log-linear models, the iterative proportional fitting procedure (IPFP) or iterative proportional scaling procedure is one of the most effective methods to compute the maximum likelihood estimation of the related parameters, which was given by [73]. [12] generalized this method and proposed the generalized iterative scaling procedure, which is now named Darroch and Ratcliff's methods. It is widely applied to deal with log-linear models. In reality, one is often faced with ordinal category data and expected to construct models with ordinal information. [2] discussed the importance to model ordinal category data and summarized some recent developments. There were no effective methods to compute the maximum likelihood estimation for related parameters. We proposed an effective iterative algorithm for log-linear models with ordinal variables [19, 68, 20].

Stochastic ordering is a useful concept in order restricted inferences. Based on the idea with two consecutive supremization steps, we propose a new estimation technique for the parameters in two multinomial populations under stochastic orderings when missing data are present. In comparison with traditional maximum likelihood estimation method, our new method can guarantee the uniqueness of solution. Furthermore, it doesn't depend on the choice of initial values for the parameters, in contrast to the EM algorithm. Finally, we give the asymptotic distributions of the likelihood ratio statistics based on the new estimation method [75].

Finally, using stochastic differential equations, we have developed a new maximum likelihood technique to estimate the parameters in a randomized Logistic equation [35–37]. It is expected that this new research direction will also open novel research challenges and possibilities for the current biological approaches.

3. STRUCTURAL LEARNING OF GRAPHICAL MODELS

In this section, we introduce our works on structural learning of graphical models. We first introduce structural learning of undirected independence graphs, which can also be used to facilitate discovery of directed acyclic graphs (DAGs). Next we show a decomposition approach for learning structures of DAGs, which decomposes a problem of learning a large graph into several smaller learning subgraphs. This decomposition learning can also be used for structural learning of a DAG from multiple databases with overlapped variable sets. Then we give an approach in which variables are hierarchically clustered from small clusters into large clusters, and subgraphs of these clusters are constructed simultaneously. In the beginning, we construct a subgraph for each initial small cluster, then we piece together subgraphs whenever these clusters are combined together. We assume that the graphical models are faithful, that is, all independencies and conditional independencies among variables can be represented by graphs [76].

3.1 Learning of undirected independence graphs

A graph is a pair $G = (V, E)$ where V is a finite set of vertices and E is a subset $V \times V$ of ordered pairs of distinct vertices, called the set of edges. An edge is directed pointing from i to j if $\langle i, j\rangle \in E$. If $\langle i, j\rangle \in E$ and $\langle j, i\rangle \in E$, an edge between vertices i and j is undirected, denoted by (i, j) and depicted by a line in the graph. A graph is undirected if it contains only undirected edges. In an undirected graph, the neighbor set of a vertex i is defined as a set of vertices that have one edge connecting i in G, denoted by N_i.

Let $X = (x_1, \ldots, x_p)$ be a p-dimensional vector of random variables. Each variable x_i in X is depicted by a vertex i in G. An undirected graphical model is then a family of probability distributions P_G which has the Markov property over the undirected graph G [42], that is, variables x_i and x_j are conditionally independent given other variables (denoted by $x_i \perp\!\!\!\perp x_j \mid X_{V\setminus\{i,j\}}$) if there is no undirected edge between vertices i and j (i.e. $(i, j) \notin E$). This property is called the pairwise Markov property. Thus the existence of undirected edges can be checked with conditional independence tests. However, when the conditioning set is large, the independence test becomes inefficient, especially for the case of discrete variables. Let A, B and C be a partition of all variables in X. Suppose that at least one of variables x_i and x_j are contained in the set A. If $A \perp\!\!\!\perp B|C$, then $x_i \perp\!\!\!\perp x_j|X \setminus \{x_i, x_j\}$ if and only if $x_i \perp\!\!\!\perp x_j|(A\cup C)\setminus\{x_i, x_j\}$ [90]. Thus an undirected edge (i, j) which falls in the vertex set A or crosses A and C can be checked conditionally on a smaller set $(A\cup C)\setminus\{x_i, x_j\}$ rather than the set $X\setminus\{x_i, x_j\}$.

Alternatively we can try to find a conditioning set as small as possible such that a variable x is independent of others given the conditioning set, which is the neighbor set

of x. We use the mutual information to measure independence between variables. The mutual information for independence between X and Y is defined as

$$I(X,Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

It equals 0 if and only if X and Y are independent. It can be shown that a variable x_i has the largest mutual information with its neighbor set N_i, that is, $I(x_i, X_{N_i}) = \max_{Y \subseteq X} I(x_i, Y)$. Thus we can find the neighbor set of a variable based on the criterion of the largest mutual information, and we may use some score with penalty of the neighbor size to select a neighbor set [13].

3.2 A decomposition approach for learning DAGs

After obtaining an undirected independence graph over the full vertex set V, we can further learn a DAG. To construct DAGs from observed data, the IC algorithm searches for a separator S from all possible variable subsets such that two variables u and v are independent conditional on S, and the PC algorithm limits possible separators to vertices that are adjacent to u and v [56, 76]. [87] presented a decomposition approach for recovering structures of DAGs. The decomposition approach starts with an undirected independence graph which may not be a moral graph and may have extra edges added to the moral graph. To decompose a problem of learning a large graph into ones of learning small subgraphs, the approach first finds an undirected independence graph from data, next constructs its junction tree or its d-separation tree, then finds a local skeleton for every node of the tree, and finally combines these local skeletons together.

Below we give a simple example to illustrate the decomposition approach. Suppose that the true underlying DAG $\vec{G}_V$ is given in Figure 1 (a) and we obtain observed data from the DAG. From the observed data, we first learn an undirected independence graph as shown in Figure 1 (b). In order to construct a junction tree, we triangulate the undirected independence graph with two dashed edges as shown in Figure 1 (c) and then obtain a junction tree as shown in Figure 2, where a triangle denotes a tree node and a rectangle denotes a separator. For each node of the junction tree, we construct a local skeleton using the IC algorithm or the PC algorithm, see Figure 3 (a). Finally we combine all local skeletons together and delete those edges which are absent in a local skeleton, such as the edges $(4,6)$ and $(1,6)$, see Figure 3 (b). A v-structure $i \to j \leftarrow k$ is determined if there is a local skeleton $i - j - k$ and the edge (i,k) is deleted with a separator which does not contain j. For example, the edge $(2,3)$ is deleted with the separator $\{1\}$ which does not contain 4, and thus $2 - 4 - 3$ can be oriented as a v-structure $2 \to 4 \leftarrow 3$. To obtain a DAG, other edges in Figure 3 (b) can be oriented as long as the graph does not create a directed cycle or a new v-structure. The Markov equivalence class can be obtained by collecting all of these possible DAGs.

In many practical applications, conditional independence between variable sets can be judged with domain or prior knowledge or with incompletely observed data patterns, such as Markov chain, chain graphical models, dynamic or temporal models, file-matching for large databases and split

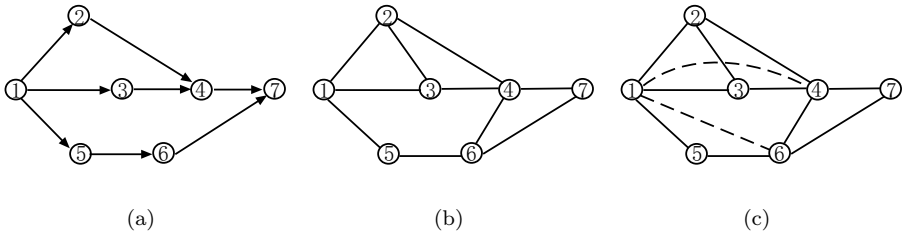

Figure 1. Illustration of the Decomposition Approach. (a) An Underlying DAG; (b) An Undirected Independence Graph; and (c) A Triangulated Graph.

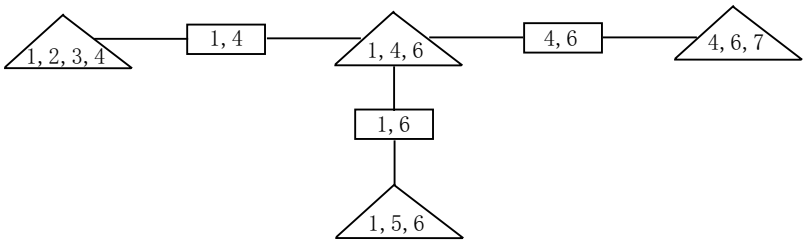

Figure 2. A Junction Tree.

questionnaire survey sampling [10, 44, 59]. This prior knowledge of conditional independence can be used for decomposition in the approach.

[86] gives a divide-and-conquer strategy in which structural learning for a large DAG is split recursively into those for subgraphs. The recursive algorithm can be depicted as a binary tree whose top node is the full set of all variables and whose other nodes are proper subsets of variables at its parent node. The algorithm consists of the top-down and the bottom-up steps. First at the top-down step, the full set of all variables at the top is decomposed into two small subsets, each of which is decomposed recursively into two smaller subsets until each node cannot be decomposed further at the bottom of the tree. At each step, the decomposition is achieved by learning an undirected graph known as independence graph for a variable subset. Next at the bottom-up step, subgraphs of leaf nodes are first constructed, and then a pair of child subgraphs are combined together into a large subgraph at their parent node until the entire graph is constructed at the top of the tree. Since the recursive algorithm reconstructs an undirected subgraph at each recursive step and the decomposition is rechecked for the subgraph, a large graph can be decomposed smaller than the decomposition algorithm proposed in [87].

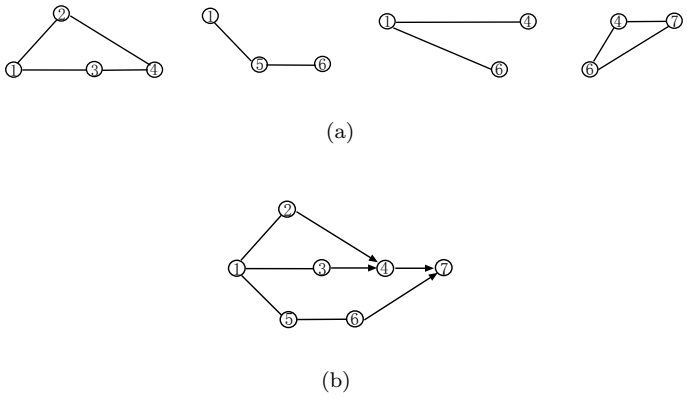

Figure 3. Skeleton and v-Structures for the DAG. (a) Local Skeletons for Every Node of the Tree; and (b) The Global Skeleton and All v-Structures.

3.3 Combination of clustering and structural learning

[82] proposed an approach which combines structural learning and cluster analysis. By the combined approach, graphical structures can be used to explain relationships among variables in each cluster. In a cluster analysis, the most correlated variables are initially grouped together, then the correlated clusters are grouped into a larger cluster, and so on. The cluster analysis, however, does not explain why and how these variables are grouped into a cluster at each step. We construct a subgraph for each cluster to explain the association relationships among variables in the cluster. On the other hand, we use hierarchical cluster approach to assist structural learning. When there are a large number of variables, searching for a separator and statistical test are not efficient. In the approach of combining structural learning and cluster analysis, a larger graph is constructed by combining several small graphs whenever the corresponding small clusters are grouped into a larger cluster. In such a way, a difficult problem of constructing a large graph is split into easy problems of constructing small graphs.

Below we illustrate the approach with a simple DAG in Figure 4 (a). We first obtain a cluster tree with two small clusters $\{x_1, x_2, x_3\}$ and $\{x_4, x_5, x_6\}$ in Figure 5. For the small clusters, we separately construct two subgraphs, as shown in Figure 4 (b). From the subgraphs, it can be seen that x_2 and x_3 are conditionally independent given x_1, and x_4 and x_6 are conditionally independent given x_5. Although x_2 and x_3 do not directly associate each other, x_2 and x_3 are grouped into a cluster since x_1 associates both x_2 and x_3. Secondly these two subclusters are grouped into a larger cluster, and then we combine two subgraphs into a larger graph in Figure 4 (c). Thirdly, we add a moral edge between X_2 and X_3 and obtain a moral network in Figure 4 (d). Finally, we find a v-structure $x_2 \to x_4 \leftarrow x_3$, and then we

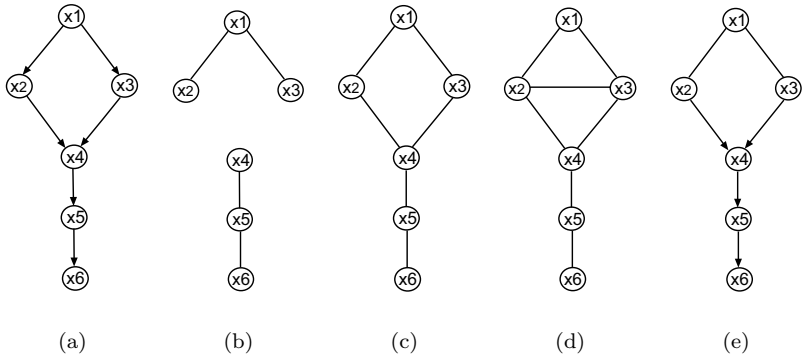

Figure 4. The Underlying DAG and Learning Process.

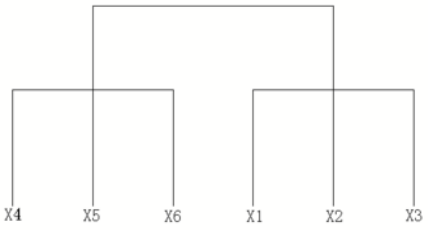

Figure 5. A Hierarchical Cluster Tree of the DAG.

can orient the edges: x_4 pointing at x_5 and x_5 pointing at x_6 to avoid new v-structures, as shown in Figure 4 (e). The directions of edges between x_1 and x_2 cannot be determined uniquely, although they must be oriented as one of $x_2 \leftarrow x_1 \rightarrow x_3$, $x_2 \leftarrow x_1 \leftarrow x_3$ or $x_2 \rightarrow x_1 \rightarrow x_3$ such that no new v-structure appears.

4. STATISTICAL GENETICS

In this section, we present our work in the field of statistical genetics. We first address some issues on haplotype inference, since the data we obtain are always genotype data but not haplotype ones. Next we show some studies of HWE which may not express homogeneity across population strata, and we also discuss the problem of admixture linkage disequilibrium across several strata. Lastly we introduce some aspects of genetic linkage analysis, with results on the inference of recombination fractions under some natural inequality restrictions.

4.1 Haplotype inference

Currently, there are primarily two categories of methods for estimating haplotype frequencies and haplotype inference: molecular or experimental methods and statistical methods. The former can not be used for large-scale studies because direct laboratory haplotyping assays are expensive and low-throughput. On the other hand, the statistical methods are less expensive, and can be widely used for large-scale studies, which contain the Clark algorithm [8], EM type algorithms [18, 31, 46, 58, 6] and Bayesian algorithms [77, 52, 89]. However, almost all of the statistical methods perform haplotype inference under the assumption that genotype data sets have errors. Nearly all large genotype data sets have errors. Furthermore, many authors pointed out that genotyping errors have enormous impact on haplotype inference [3, 22, 14, 40, 1, 74, 38]. Here, we will propose several novel strategies to reduce the impact induced by genotyping misclassification or genotyping errors in haplotype inference.

Firstly, to adjust for genotyping errors we adopt a double sampling scheme. Let $\mathbf{G} = (G_1, \ldots, G_N)$ denote the fallible genotypes of N independent individuals identified by inexpensive and large-scale genotyping technologies. We advise researchers to re-identify the genotypes of a subsample, say, $n < N$, using the high-fidelity but expensive genotyping technologies, and we consider the genotypes $\tilde{\mathbf{G}} = (\tilde{G}_1, \ldots, \tilde{G}_n)$ classified in this way to be the genotypes without errors. We can use the relations between $(G_1, \ldots, G_n)$ and $(\tilde{G}_1, \ldots, \tilde{G}_n)$ to infer the possible infallible genotypes $\tilde{\tilde{\mathbf{G}}} = (\tilde{\tilde{G}}_{n+1}, \ldots, \tilde{\tilde{G}}_N)$ of other $N-n$ individuals, and then we use an EM algorithm to infer haplotypes of individuals by considering the unobserved phase information as the missing data. The likelihood function can then be expressed as

$$\Pr(\tilde{\mathbf{G}}, \mathbf{G}) = \prod_{i=1}^{n} \left[\sum_{D \oplus = \tilde{G}_i} \Pr(D) \right] + \prod_{i=n+1}^{N} \left[\sum_{\tilde{\tilde{G}}_i} \sum_{D \oplus = \tilde{\tilde{G}}_i} \Pr(D) \Pr(G_i | \tilde{\tilde{G}}_i) \right],$$

where D denotes the possible haplotype pair that is compatible with the infallible genotype $\tilde{G}_i$ or $\tilde{\tilde{G}}_i$ of the ith individual.

Secondly, to adjust for genotyping errors we adopt a multi-genotyping scheme, that is to say, we recommend to identifying the genotype of each individual using several different inexpensive genotyping technologies. Let $G^{(1)}, G^{(2)}$ denote the two genotypes of the same individual identified by two different inexpensive genotyping technologies. $G^{(1)}, G^{(2)}$ may contain genotyping errors due to the inexpensive technologies. We consider the unobserved infallible genotype G to be a latent class variable and use a latent class model to estimate $\Pr(G^{(1)}|G)$ and $\Pr(G^{(2)}|G)$. We also use an EM algorithm to infer haplotypes of individuals by considering the unobserved phase information as the missing data. The contribution to the likelihood by one subject can then be expressed as

$$\Pr\big(G^{(1)}, G^{(2)}\big) = \sum_{G} \sum_{D \oplus = G} \Pr(D) \Pr\big(G^{(1)}|G\big) \Pr\big(G^{(2)}|G\big),$$

where D denotes the possible haplotype pair that is compatible with the infallible genotype G.

[38] advised researchers to use the raw fluorescent intensity (FI) data but not fallible genotype data in haplotype inference studies when genotype data sets have errors. However, their GS-EM algorithm is not applicable to pedigree data directly. Although some genotyping errors can be detected using Mendel's laws as a check for pedigree data, [23] showed that the error-detection rate that has been estimated using trio designs does not exceed 30% for one diallelic marker. So it is important to develop methods that incorporate genotyping errors into haplotype frequency estimation for pedigree data. To perform haplotype inference based on FI data of pedigree structure, we described a

new GenoSpectrum-EM algorithm, called GS-PEM, which estimates haplotype frequencies by taking into account of the dependence information among related individuals in pedigrees [94]. For a pedigree, let $\mathbf{X}_f = (X^1, \ldots, X^J)$ and $\mathbf{X}_n = (X_n^1, \ldots, X_n^R)$ denote the FI values of J founders and R non-founders. We assumed that all pedigrees are independent, and founders in each pedigree are treated as an independent sample from the population. The contribution to the likelihood by one pedigree can then be expressed as

$$\Pr(\mathbf{X}_f, \mathbf{X}_n) = \sum \Pr(\mathbf{D}_f, \mathbf{D}_n) \Pr(\mathbf{X}_f, \mathbf{X}_n | \mathbf{G}_f, \mathbf{G}_n),$$

where $\mathbf{G}_f, \mathbf{G}_n$ represent the possible vectors of genotypes of J founders and R non-founders, and $\mathbf{D}_f, \mathbf{D}_n$ denote the possible vectors of haplotype pairs that are compatible with $\mathbf{G}_f$ and $\mathbf{G}_n$.

After the haplotype pair of each individual is determined or haplotype frequencies are estimated, we can study the associations of haplotypes with traits. Furthermore, we can study the associations of haplotypes with traits based on unphased genotype data directly, where we need to account for haplotype ambiguity by modeling the probabilities of the possible haplotype pairs per subject [63]. Although there are many haplotype-based association methods currently, the performance of the existing methods depends on the high-fidelity genotyping technology. That is to say, almost all of the existing methods perform haplotype association analysis under the assumption that genotype data do not contain errors. As mentioned above, however, almost all large genotype data sets have errors, which can impact haplotype-based association analysis enormously. To reduce the impact induced by genotyping errors in haplotype association studies, we want to estimate the haplotype frequencies and haplotype effects based on raw FI readouts from case-control studies [93]. Firstly, we adopted a clustering algorithm based on mixtures of t distributions, presented by [38], to obtain the "GenoSpectrum" for each individual. Secondly, we proposed a likelihood-based approach incorporating the genotyping uncertainty to assess the associations of haplotypes and traits through a haplotype-based logistic regression model. Let X denote the FI value of an individual, and let Y denote a complex disorder, with values 1 or 0 corresponding to the presence or absence of the disorder. The contribution to the likelihood by one subject can then be expressed as

$$\Pr(X, Y) = \sum_G \sum_{D \oplus = G} \Pr(Y|D, \boldsymbol{\beta}) \Pr(D|\boldsymbol{\theta}) \Pr(X|G),$$

where G denotes the possible genotype of the individual, and D denotes the possible haplotype pair that is compatible with the genotype G. We can use a haplotype-based logistic regression model to model the conditional probability of the trait given haplotype pair, $\Pr(Y|D, \boldsymbol{\beta})$. $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\Pr(D|\boldsymbol{\theta})$ is the probability of haplotype pair D, where $\boldsymbol{\theta}$ is the vector of frequencies of haplotypes. Moreover, we can use the clustering algorithm based on mixtures of t distributions to estimate the conditional probability $\Pr(X|G)$.

4.2 Research on HWE in population genetics

The law of Hardy-Weinberg equilibrium (HWE) states that in a large random mating population that is not affected by the evolutionary processes of mutation, migration, or selection, both the allele frequencies and the genotype frequencies are constant from generation to generation [30, 83] Furthermore, the genotype frequencies are related to the allele frequencies by the square expansion of those allele frequencies. The original descriptions of HWE became an important landmark in the history of population genetics [11], and it is now common practice to verify whether observed genotypes conform to Hardy-Weinberg expectations [33, 53].

In a diallelic locus with alleles A_1 and A_2 across K strata, let the genotypic array of the k-th $(k = 1, \ldots, K)$ stratum be $p_{11k}A_1A_1 + p_{12k}A_1A_2 + p_{22k}A_2A_2$. Let p_k be the allelic frequency of A_1 in the k-th stratum and $q_k = 1 - p_k$ $(k = 1, \ldots, K)$. Populations with genotypic frequencies satisfying $p_{11k} = p_k^2$, $p_{12k} = 2p_kq_k$, and $p_{22k} = q_k^2$ $(k = 1, \ldots, K)$ are said to be in HWE at the locus under consideration. In studies of HWE, there are two widely used coefficients, namely the fixation and disequilibrium coefficients [84]. For stratum k $(k = 1, \ldots, K)$, the fixation and disequilibrium coefficients are defined by $f_k = 1 - p_{12k}/(2\sqrt{p_{11k}p_{22k}})$ and $D_k = p_kq_k - p_{12k}/2$, respectively. Hence, the problem of testing HWE when individuals are sampled from several strata is equivalent to testing one of the following hypotheses:

$$(1) \qquad \begin{aligned} &H_0' : \theta_k = 0 \text{ for all } k = 1, \ldots, K \quad \text{versus} \\ &H_1' : \theta_k \neq 0 \text{ for some } k, \end{aligned}$$

where $\theta_k = f_k$ or D_k. For statistical tests based on disequilibrium coefficients, one can refer to the work of [29] and [72]. For test procedures based on functions of fixation coefficients (e.g., $(1 - f_k)^2$), one can consult the work of [80, 51].

It is noteworthy that any statistical procedure for testing the null hypothesis in Eq. (1) assumes that the measure of disequilibrium (i.e., θ_k) is constant across the strata. In this regard, it is important that one should consider testing the assumption of homogeneity of the measure of disequilibrium across strata prior to any testing of the null hypothesis in Eq. (1). For this purpose, we consider the following hypotheses:

(2)
$$H_0 : \theta_1 = \cdots = \theta_K \quad \text{versus} \quad H_1 : \text{Not all } \theta_k's \text{ are equal},$$

[54] proposed a large-sample test and an exact test for verifying the null hypothesis H_0 in Eq. (2) via a function of fixation coefficients, $(1 - f_k)^2$. They also approximated the P-value of the exact test using a Markov chain Monte Carlo

approach. Although the use of fixation coefficients to describe departures from HWE has some merit, it has the disadvantage that these parameters are estimated as ratios of genotypic frequencies. It is difficult to study sampling properties of ratio statistics [33, 84]. Besides, functions of fixation coefficients such as $(1-f_k)^2$ may possess an infinite upper bound. On the other hand, there are advantages in working with a composite kind of quantity such as the disequilibrium coefficient. This is simply the difference between a frequency and its values expected when there is no association between alleles. Moreover, it is easy to show that the disequilibrium coefficient D_k satisfies $\max\{-p_k^2, -q_k^2\} \leq D_k \leq p_k q_k$. Unfortunately, a test of homogeneity of disequilibrium coefficients across several strata has not been considered in the literature. We develop an asymptotic test for this purpose [88].

Simulation results demonstrated that our homogeneity score test performs satisfactorily in the sense that its empirical size seldom exceeds the pre-chosen nominal level by more than 10 percent, even for small sample sizes. Empirical results from our simulation studies supported that our homogeneity score test is a reliable asymptotic testing procedure even for small sample sizes. However, our test may suffer from the drawback that it may be quite conservative for rare allelic probabilities (e.g., ≤ 0.1). In this case, one may require larger sample sizes to overcome the conservativeness issue.

4.3 Research on ALD in population genetics

Admixture linkage disequilibrium (ALD), a phenomenon created by gene flow between genetically distinct populations, has for some time been used as a tool in gene mapping. It is therefore important to analyze the pattern of ALD over generations. We explored two models of admixture: the gradual admixture (GA) model, in which admixture occurs at a variable rate in every generation; and the immediate admixture (IA) model, a special case of the GA model, in which admixture occurs in a single generation [25, 27]. In the case of ALD, the well-known formula of linkage disequilibrium ($\Delta^{(t)} = (1-r)^t \Delta^{(0)}$) is not applicable under these two models. We noted the effect of a random-mating population (RMP) on the gametic frequencies from the parental population to the offspring population, and provide the correct formula for ALD.

Case-control studies compare marker-allele distributions in affected and unaffected individuals, and significant results may be due to linkage but can also simply reflect population structure. To test for linkage after obtaining a significant case-control finding, within-family analysis can be performed. In a transmission/disequilibrium test (TDT), genotypes of cases are compared to those of their parents to explore whether a specific allele, or marker, at a locus of interest is transmitted to a greater degree than Mendelian inheritance would warrant. For multi-allelic markers, several authors have proposed extensions to the TDT. We proposed a TDT test that utilizes the available information of a case-control study in the grouping of alleles for multi-allelic markers, and thereby increases the statistical power of a TDT test with a small sample size [25].

4.4 Genetic linkage analysis

The degree of linkage can be measured by the recombination fraction. Many map functions under different assumptions have been derived [28, 50], from which the genetic distance and the recombination fraction can be mutually transformed. Human gene mapping is now an important field of science. A critical first step in finding gene loci that contribute to a genetic trait is to demonstrate linkage with a gene of known location (marker). So estimating the recombination fractions is important in linkage analysis.

In several respects, three-locus analysis yields more information than does two-locus analysis [85, 47]. Three-locus linkage analysis is also an important case of multi-locus problems. Methods for detecting multilocus linkage in humans and estimation of recombination have been proposed by Lathrop et al. [41]. More recently, Ott [55] considered the estimation of two-locus recombination fractions for phase-unknown triple backcross families with two offspring in each family, and presented the estimates of the two-locus recombination fractions. Wu et al. [85] considered simultaneous estimation of linkage and linkage phases in outcrossing species. However, as mentioned in [55], the estimates suggested by the author may not satisfy some natural restrictions which two-locus recombination fractions should satisfy. One may not obtain a reasonable interpretation of the recombination phenomenon based on the estimates. Furthermore, illegitimate estimates of recombination fractions may reduce the power to detect linkage. In addition, the restrictions on recombination fractions given within context are necessary in the analysis. For example, they can be applied to determine the locus order on the genome [85, 47].

Let θ_{AB}, θ_{BC} and θ_{AC}, respectively denote two-locus recombination fractions between loci A and B, between loci B and C, and between loci A and C. The θ_i's need to satisfy the following restrictions:

$$\begin{cases} \theta_{AB} \leq \theta_{AC}, \\ \theta_{BC} \leq \theta_{AC}, \\ \theta_{AC} \leq \theta_{AB} + \theta_{BC}, \\ \theta_{AC} \leq 1/2. \end{cases}$$

This estimation problem of two-locus recombination fractions in three-locus linkage analysis belongs to the constrained parameter problems which are not only important but also appear in many areas. The reader is referred to the works of [15, 62, 69]. However, the methods provided in the literatures cannot be directly applied to the above genetics problem.

For the case of phase-unknown triple backcross families with two offspring in each family, we consider the estimation of the two-locus recombination fractions under the above

restrictions, which is also a missing data problem [92]. We develop a restricted EM algorithm, which gives estimating results by taking into account of the natural inequality restrictions on the two-locus recombination fractions, and the algorithm is easy to extend to other cases.

5. CONCLUSION AND SOME FURTHER DISCUSSIONS

The undisputed facts show that the statistical inference techniques under order restrictions are powerful tools for extracting valuable information hidden in a larger amount of practical data. However, a number of practical problems are far from being really solved. Anraku [4] opens the possibility of classical model selection applied to order-restricted alternative hypotheses. He proposed an order restricted information criterion (ORIC) based on the isotonic regression theory under the restriction of simple order. In many application problems, the configuration of parameters may present an umbrella order trend instead of simple order trend. For example, in dose-response studies, the therapeutic variables may increase with dose levels at first and then decrease with the further increasing of dose levels due to adverse effect. In the context of gene selection, [57] provided a practical demand for a generalized ORIC to treat some more general order restriction cases. We hope a breakthrough can be made in this field.

In Section 3, we have introduced several approaches for structural learning of graphical models which were investigated in our project. The theoretical results which are the base of the algorithms are also important for observational and experimental designs, analysis of incomplete data and local structural discovery of graphical models. Efficient approaches for a large number of variables but a small sample size need to be investigated further. DAGs are often used to depict causal relationships among variables. Besides conditional independencies, causal relationships between causes and effects are going to be discovered from experimental data and even observational data [56, 76]. A criterion of surrogate endpoints was discussed based on DAGs in [43] and [7].

In this paper, part of our work in applied statistics has been reviewed. These theoretical results and related approaches can be applied to many fields, such as medical and epidemiological studies, data mining, artificial intelligence, and genetics. Some of our research has been applied to the study of traditional Chinese medicine. We hope that our work will be of interest to our colleagues in China and around the world.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (Grant Number 10431010), National 973 Key Project of China (2007CB311002), 863 Project of China (2007AA01Z437) and NCET-04-0310. The fourth author was partially supported by the grant of Training Fund of NENU'S Scientific Innovation Project (NENUSTC07002).

Received 19 January 2008

REFERENCES

[1] Abecasis, G. R., Cherny, S. S. and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* **9** 130–134.

[2] Agresti, A. and Coull, B. A. (2002). The analysis of contingency tables under inequality constraints. *J. Stat. Plan. Inf.* **107** 45–73. MR1927754

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

[4] Anraku K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika* **86** 141–152. MR1688078

[5] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions.* Wiley, New York.

[6] Becker, T. and Knapp, M. (2004). Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet. Epidemiol.* **27** 21–32.

[7] Chen, H., Geng, Z. and Jia, J. (2007). Criteria for surrogate end points. *J. Royal Statist. Soc. Ser. B* **69** 919–932.

[8] Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7** 111–122.

[9] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Springer, New York.

[10] Cox, D. R. and Wermuth, N. (1993). *Multivariate Dependencies: Models, Analysis, and Interpretation.* Chapman and Hall, London.

[11] Crow, J. E. (1988). Eighty years ago: The beginnings of population genetics. *Genetics* **119** 473–476.

[12] Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43** 1470–1480. MR0375574

[13] Deng, K., Liu, D., Gao, S. and Geng, Z. (2005). Structural learning of graphical models and its applications to traditional Chinese medicine. Fuzzy Systems and Knowledge Discovery, Lipo Wang and Yaochu Jin (Eds.). *Lecture Notes in Computer Science* **3614** 362–367, Springer-Verlag.

[14] Douglas, J. A., Skol, A. D. and Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am. J. Hum. Genet.* **70** 487–495.

[15] Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Am. Statist. Assoc.* **78** 837–842. MR0727568

[16] Edwards, D. (1995). *Introduction to Graphical Modelling.* Springer-Verlag, New York.

[17] Elston, R. C. and Stewart, J. (1971). A general model for the analysis of pedigree data. *Hum. Hered.* **21** 523–542.

[18] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12** 921–927.

[19] Gao, W. and Shi, N.-Z. (2003). I-projection on isotonic cones and its applications to maximum likelihood estimation for log-linear models. *Ann. Inst. Statist. Math.* **55** 251–263. MR2001863

[20] Gao, W. and Shi, N.-Z. (2005). Estimating cell probabilities under order-restricted odds ratios. *Comput. Stat. Data Anal.* **49** 77–84. MR2129165

[21] Geng, Z., Wang, C. and Zhao, Q. (2005). Decomposition of search for v-structures in DAGs. *J. Multivar. Anal.* **96** 282–294. MR2204979

[22] Goldstein, D. R., Zhao, H. and Speed, T. P. (1997). The effects of genotyping errors and interference on estimation of genetic distance. *Hum. Hered.* **47** 86–100.
[23] Gordon, D., Heath, S. C. and Ott, J. (1999). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum. Hered.* **49** 65–70.
[24] Guo, J. H., Ma, Y. P., Shi, N.-Z. and Lau, T. S. (2004). Testing for homogeneity of relative difference under inverse sampling. *Comput. Stat. Data Anal.* **44** 613–624. MR2026435
[25] Guo, W., Fung, W. K., Shi, N.-Z. and Guo, J. H. (2005). On the formula for admixture linkage disequilibrium. *Hum. Hered.* **60** 177–180.
[26] Guo, W. and Fung, W. K. (2005). Combining the case-control methodology with the small size transmission/disequilibrium test for multiallelic markers. *Eur. J. Hum. Genet.* **13** 1007–1012.
[27] Guo, W. and Fung, W. K. (2006). The admixture linkage disequilibrium and linkage inference on the gradual admixture population. *Acta Genetica Sinica* **31** 12–18.
[28] Haldane, J. B. S. (1919). The recombination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* **8** 299–309.
[29] Haldane, J. B. S. (1954). An exact test for randomness of mating. *J. Genet.* **52** 631–635.
[30] Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28** 49–50.
[31] Hawley, M. E. and Kidd, K. K. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* **86** 409–411.
[32] Heckerman, D. (1998). A tutorial on learning with Bayesian networks. *Learning in Graphical Models*, M. I. Jordan (Ed.), 301–354, Kluwer Academic Pub., Netherlands.
[33] Hernandez, J. L. and Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. *Biometrics* **45** 53–70. MR0999440
[34] Hodge, S. E., Boehnke, M. and Spence, M. A. (1999). Loss of information due to ambiguous haplotypings of SNPs. *Nat. Genet.* **21** 360–361.
[35] Jiang, D. Q. and Shi, N. Z. (2005). A note on nonautonomous logistic equation with random perturbation. *J. Math. Analysis Appl.* **303** 164–172. MR2113874
[36] Jiang, D. Q., Shi, N.-Z. and Zhao, Y. N. (2005). Existence, uniqueness and global stability of positive solutions to the food-limited population model with random perturbation. *Mathematics and Computer Model* **42** 651–658. MR2173483
[37] Jiang, D. Q., Zhang, B. X., Wang, D. H. and Shi, N.-Z. (2007). Existence, uniqueness and global attractivity of positive solutions and MLE of the parameters to the logistic equation with random perturbation. *Sci. China Ser. A.* **50** 977–986. MR2355869
[38] Kang, H., Qin, Z. S., Niu, T. and Liu, J. S. (2004). Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **74** 495–510.
[39] Kim D. K. and Taylor J. M. G. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *J. Am. Statist. Assoc.* **430** 708–716. MR1340522
[40] Kirk, K. M. and Cardon, L. R. (2002). The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur. J. Hum. Genet.* **10** 616–622.
[41] Lathrop, G. M., Lalouel, J. M., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* **81** 3443–3446.
[42] Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.
[43] Lauritzen, S. L. (2004). Discussion on causality. *Scand. J. Statist.* **31** 189–192. MR2066248
[44] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.
[45] Liu, R.-Y., Tao, J. and Biswas, A., Shi, N.-Z. (2008). Generalized order-restricted inference methods for selecting and clustering genes according to their time-course and dose-response profiles. Submitted.
[46] Long, J. C., Williams, R. C. and Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56** 799–810.
[47] Lu, Q., Cui, Y. H. and Wu, R. L. (2004). A multilocus likelihood approach to joint modelling of linkage, parent diplotype and gene order in a full-sib family. *BMC Genet.* **5** 20.
[48] Ma, Y. P., Guo, J. H., Shi, N.-Z. and Tang, M. L. (2002). On the use of historical control information for trend test in carcinogenesis. *Biometrics* **58** 917–927. MR1945024
[49] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institution* **22** 719–748.
[50] Morgan, T. H. (1928). *The Theory of Genes*. Yale University Press, New Haven.
[51] Nam, J. M. (1997). Testing a genetic equilibrium across strata. *Annals of Human Genetics* **61** 163–170.
[52] Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked singe-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70** 157–169.
[53] Olson, J. M. (1993). Testing the Hardy-Weinberg law across strata. *Annals of Human Genetics* **57** 291–295.
[54] Olson, J. M. and Foley, M. (1996). Testing for homogeneity of Hardy-Weinberg disequilibrium using data sampled from several populations. *Biometrics* **52** 971–979.
[55] Ott, J. (1999). Phase-Unkown Triple Backcross with Two Offspring. In: *Analysis of Human Genetic Linkage*, 3rd ed. The Johns Hopkins University Press: Baltimore, pp. 122–124.
[56] Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.
[57] Peddada, S., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C. and Umbach, D. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19** 834–841.
[58] Qin, Z. S., Niu, T. and Liu, J. S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with singe-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71** 1242–1247.
[59] Rassler, S. (2002). Statistical Matching. *Lecture Notes in Statistics* **168**. Springer, New York.
[60] Razzaghi, M. and Kodell, R. L. (2000). Risk assessment for quantitative responses using a mixture model. *Biometrics* **56** 519–527.
[61] Risch, N. (1990). Linkage strategies for genetically complex traits. *Am. J. Hum. Genet.* **46** 222–253.
[62] Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley, New York.
[63] Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* **27** 348–364.
[64] Shi, N.-Z. (1991). A test of homogeneity of odds ratios against order restrictions. *J. Am. Statist. Assoc.* **86** 154–158. MR1137106
[65] Shi, N.-Z. (1994). Maximum likelihood estimation of means and variances form Normal populations under simultaneous order restrictions. *J. Multivar. Analy.* **49** 282–294. MR1293047
[66] Shi, N.-Z. and Geng, Z. (1996). Multiple isotonic regression. *Journal of the Royal Statistical Society C.* **45** 266–273.
[67] Shi, N.-Z. and Jiang, H. (1998). Maximum likelihood estimation of isotonic normal means with unknown variances. *J. Multivar. Analy.* **64** 183–195. MR1621859
[68] Shi, N.-Z. and Zheng, S. R. (2004). The maximum likelihood estimates of expected frequencies under the loop order. *Statistica Sinica* **14** 283–295. MR2036773
[69] Shi, N.-Z., Zheng, S. R. and Guo, J. H. (2005). The restricted EM algorithm under inequality restrictions on the parameters. *J. Multivar. Analy.* **92** 53–76. MR2102244
[70] Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Wiley, New York.

[71] Simmons, S. and Peddada, S. (2007). Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances. *Bioinformation* **1(10)** 414–419.
[72] Smith, C. A. B. (1970). A note on testing the Hardy-Weinberg law across strata. *Annals of Human Genetics* **33** 377–383.
[73] Sinkhorn, R. (1967). Diagonal equivalence of matrices with prescribed row and column sums. *American Mathematical Monthly* **74** 402–405. MR0210730
[74] Sobel, E., Papp, J. C. and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70** 496–508.
[75] Song, H.-Y., Tao, J. and Shi, N.-Z. (2008). S-S method for stochastically ordered multinomial populations with missing data. Submitted.
[76] Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search*, 2nd ed. MIT Press, Cambridge.
[77] Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68** 978–989.
[78] Tao, J., Shi, N.-Z. and Lee, S. Y. (2004). Drug risk assessment with determining the number of sub-populations under finite mixture normal models. *Comput. Stat. Data Anal.* **46** 661–676. MR2084137
[79] Thompson, E. A. (2000). *Statistical Inference from Genetic Data on Pedigree*. Institute of Mathematical Statistics Beachwood, Ohio.
[80] Troendle, J. J. and Yu, K. F. (1994). A note on testing the Hardy-Weinberg law across strata. *Annals of Human Genetics* **58** 397–402.
[81] Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, Elsevier, Amsterdam, pp. 255–268.
[82] Wang, M., Geng, Z., Wang, M., Chen, F. and Ding, W. (2006). Combination of network construction and cluster analysis and its application to traditional Chinese medicine. *Lecture Notes in Computer Sciences* **3973** 777–785, Springer-Verlag.
[83] Weinberg, W. (translation by S. H. Boyer) (1908). On the demonstration of heredity in man. 1963. *Papers on Human Genetics*. Englewood Cliffs, Prentice-Hall, New Jersey.
[84] Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates Inc., Sunderland, Massachusetts.
[85] Wu, R. L., Ma, C. M. and Painter, I., Zeng, Z. B. (2002). Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing populations. *Theor. Pop. Biol.* **61** 349–363.
[86] Xie, X. and Geng, Z. (2008). A recursive method for structural learning of directed acyclic graphs. *J. Machine Learning Research* **9** 459–483.
[87] Xie, X., Geng, Z. and Zhao, Q. (2006). Decomposition of structural learning about directed acyclic graphs. *Artificial Intelligence* **170** 422–439.
[88] Yin, X. L., Ma, W. Q., Tang, M. L. and Guo, J. H. (2006). On test of homogeneity of Hardy-Weinberg disequilibrium across strata. *Eur. J. Hum. Genet.* **14** 1223–1230.
[89] Zhang, Y., Niu, T. and Liu, J. S. (2006). A coalescence-guided hierarchical bayesian method for haplotype inference. *Am. J. Hum. Genet.* **79** 313–322.
[90] Zhao, Q., Chen, H. and Geng, Z. (2007). Structural learning about independence graphs from multiple databases. *Advances in Knowledge Discovery and Data Mining, Lecture Notes of Aritificial Intellegengce* **4426** 1122–1130.
[91] Zheng, S. R., Shi, N.-Z. and Guo, J. H. (2005). The restricted EM algorithm under linear inequalities in a linear model with missing data. *Sci. China Ser. A.* **35** 231–240. MR2158976
[92] Zhou, Y., Shi, N.-Z., Fung, W. K., and Guo, J. H. (2008). Maximum likelihood estimates of two-locus recombination fractions under some natural inequality restrictions. *BMC Genet.* **9(1)**, to appear.
[93] Zhu, W. S. and Guo, J. H. (2006). A likelihood-based method for haplotype association studies for case-control data with genotyping uncertainty. *Sci. China Ser. A.* **49** 130–144. MR2220790
[94] Zhu, W. S., Fung, W. K. and Guo, J. H. (2007). Incorporating genotyping uncertainty in haplotype frequence estimation in pedigree studies. *Hum. Hered.* **64** 172–181.

Ning-Zhong Shi
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun, Jilin, 130024, P.R. China
E-mail address: shinz@nenu.edu.cn

Zhi Geng
School of Mathematical Sciences
Peking University
Beijing 100871, P.R. China
E-mail address: zgeng@math.pku.edu.cn

Jianhua Guo
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun, Jilin, 130024, P.R. China
E-mail address: jhguo@nenu.edu.cn

Jian Tao
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun, Jilin, 130024, P.R. China
E-mail address: taoj@nenu.edu.cn