BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 15 2010, pages 1871–1878
doi:10.1093/bioinformatics/btq290

Genetics and population analysis

Advance Access publication June 10, 2010

Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies

Xing Hua[1], Han Zhang[1,*], Hong Zhang[1,2], Yaning Yang[1] and Anthony Y.C. Kuk[3]

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China, [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA and [3]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The multifactor-dimensionality reduction (MDR) method has been widely used in multi-locus interaction analysis. It reduces dimensionality by partitioning the multi-locus genotypes into a high-risk group and a low-risk group according to whether the genotype-specific risk ratio exceeds a fixed threshold or not. Alternatively, one can maximize the χ^2 value exhaustively over all possible ways of partitioning the multi-locus genotypes into two groups, and we aim to show that this is computationally feasible.

Methods: We advocate finding the optimal MDR (OMDR) that would have resulted from an exhaustive search over all possible ways of partitioning the multi-locus genotypes into two groups. It is shown that this optimal MDR can be obtained efficiently using an ordered combinatorial partitioning (OCP) method, which differs from the existing MDR method in the use of a data-driven rather than fixed threshold. The generalized extreme value distribution (GEVD) theory is applied to find the optimal order of gene combination and assess statistical significance of interactions.

Results: The computational complexity of OCP strategy is linear in the number of multi-locus genotypes in contrast with an exponential order for the naive exhaustive search strategy. Simulation studies show that OMDR can be more powerful than MDR with substantial power gain possible when the partitioning of OMDR is different from that of MDR. The analysis results of a breast cancer dataset show that the use of GEVD accelerates the determination of interaction order and reduces the time cost for *P*-value calculation by more than 10-fold.

Availability: C++ program is available at http://home.ustc.edu.cn/~zhanghan/ocp/ocp.html

Contact: zhanghan@mail.ustc.edu.cn

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 5, 2010; revised on May 11, 2010; accepted on May 29, 2010

1 INTRODUCTION

Analysis of gene–gene interaction is of great concern in modern genetic studies (Cordell, 2009). With a large number of genes, classical statistical methods such as the logistic regression and Pearson's χ^2 test may not be appropriate due to high degree of freedom involved (Wade, 2000). Methods that can detect multi-locus interacting genes that may have been overlooked by a single-locus analysis are of central importance in genetic association studies.

Multifactor-dimensionality reduction (MDR) is a non-parametric and model-free method that was developed for detecting high-order gene–gene interaction in the absence of marginal effects (Ritchie *et al.*, 2001, 2003). Operationally, the MDR method divides the multi-loci genotype combinations into high- and low-risk groups according to whether the ratio of cases to controls exceeds a fixed threshold or not. As noted by Moore *et al.* (2006), this is a constructive induction method that changes the representation space of the data by reducing/recoding the single nucleotide polymorphism (SNP) combinations to a single binary attribute. This results in a 2 × 2 contingency table and the sparsity and the high-dimensionality problems are greatly alleviated. The Pearson's χ^2 statistic is calculated but with *P*-value determined using a permutation procedure rather than from a χ^2 distribution with one degree of freedom (df). Empirical studies show that the MDR method is a useful tool for testing high-order gene–gene interaction (Moore *et al.*, 2006; Ritchie *et al.*, 2001). Since it was proposed, the MDR method has been widely applied to multi-locus association studies (Cho *et al.*, 2004; Julia *et al.*, 2007; Ritchie *et al.*, 2001; Tsai *et al.*, 2007).

Many efforts have been made to improve the power of MDR, such as the extended MDR (EMDR) of Mei *et al.* (2005), modified MDR methods based on balanced accuracy (Velez *et al.*, 2007), likelihood ratio and normalized mutual information (Bush *et al.*, 2008) or other evaluation measures (Namkung *et al.*, 2009). Some other extensions of MDR such as the odds ratio-based MDR (OR-MDR; Chung *et al.*, 2007), model-based MDR (MB-MDR; Calle *et al.*, 2008), log-linear model-based MDR (LM-MDR; Lee *et al.*, 2007) and generalized MDR (GMDR; Lou *et al.*, 2007) were also developed. These extensions make the MDR more flexible in applications (Moore *et al.*, 2010).

In recent years, many other data mining algorithms for detecting high-order genetic interactions such as classification trees, random forests and multivariate adaptive regression splines (MARS) have been proposed (Montana, 2006). Chatterjee *et al.* (2006) developed a powerful multi-locus parametric test based on Tukey's 1-df model of interaction. Millstein *et al.* (2006) proposed the focused interaction testing framework (FITF). The performance of neural networks evolved from genetic programming is studied by Bush *et al.* (2005). Motsinger-Reif *et al.* (2008) compared the performances of six methods for testing main effects and interactions. The six methods

*To whom correspondence should be addressed.

they considered are MDR, grammatical evolution neural networks, random forests, FITF, step-wise and explicit logistic regression. Heidema *et al.* (2007) compared MDR with set association (Hoh and Ott, 2003), and random forests and suggested the use of a combination of existing multi-locus methods in genetic association studies. Bastone *et al.* (2004) demonstrated that MDR is a special case of recursive partitioning (Breiman *et al.*, 1984). Park and Hastie (2008) illustrated penalized logistic regression based on L_2 regularization used in interaction detection. The FlexTree algorithm (Huang *et al.*, 2004) is also introduced along with their method. Chen *et al.* (2008) proposed to use a support vector machine and a combinatorial optimization method in gene–gene interaction detection.

The MDR is originally designed for balanced case–control studies. It partitions the contingency table by collapsing multi-locus genotypes with risk ratios equal to or greater than the threshold 1 as high risk and others as low risk (for unbalanced case–control studies, the threshold is the ratio of the numbers of cases to controls). Such a partitioning strategy using a fixed threshold may not be optimal since it excludes many other possible ways of partitioning. Intuitively, a different threshold that yields a larger value of the χ^2 statistic could possibly lead to a more powerful test. In this article, we show that the optimal threshold for maximizing the χ^2 value can be obtained by using an ordered combinatorial partitioning (OCP) algorithm. Although our OCP algorithm maximizes over the ordered partitions only, which is what makes it computationally feasible, we prove that the algorithm has not missed anything important and will still give us the global maximal χ^2 that would have resulted from an exhaustive search over all possible partitions. We call the optimal partitioning produced by the OCP algorithm as the optimal MDR (OMDR). To reduce the computational burden of permutation in assessing significance and determining the order of interaction, we propose a procedure based on the generalized extreme value distribution (GEVD) approximation in the OMDR method.

This article is organized as follows. In Section 2.1, the limitations of the partitioning strategy used in the original MDR are illustrated by a simple example. The OCP strategy is then proposed for partitioning a contingency table. The computational complexity of the OCP and the resulting OMDR are discussed. In Section 2.2, we introduce the GEVD approximation method for interaction order determination and significance assessment in the OMDR method. Simulation studies for comparing the MDR and the OMDR are conducted in Section 3. Finally, some discussions about the OMDR and the OCP method are given in Section 4.

2 METHODS

In Sections 2.1 and 2.2, we introduce the two key ingredients of the OMDR method, namely, the efficient exhaustive partitioning method, OCP, and the GEVD method in interaction order determination and the computation of a validity measure.

2.1 OCP method

MDR partitions multi-locus genotype combinations according to their risk levels based on a fixed threshold. Here, we review this strategy in detail and point out its limitations. Suppose that n dichotomous genes such as SNP markers (each gene has three genotypes) are considered for testing interactions. The 3^n genotype combinations of cases and controls can be summarized in a 2×3^n contingency table. One can calculate the ratio of the number of cases to the number of controls for each of all 3^n genotype combinations. We further assume that the columns of the contingency table that contain no observations have been deleted, and the columns are sorted into an ascending order according to the ratios. Based on the ordered contingency table, the MDR method labels each genotype combination as ‘high risk’ if its ratio meets or exceeds some chosen threshold (e.g. the fixed threshold 1 used in the original MDR in balanced case–control studies), or ‘low risk’ otherwise. Fixing the threshold beforehand is an attractive feature of the MDR method since this can reduce the computational expense substantially. However, even in some simple situations, partitioning based on a fixed threshold as is the case in the MDR method may lead to substantial loss of power compared with an exhaustive search strategy. We illustrate this using an artificial 2×3 case–control data as shown in Table 1.

Table 1. Data for a single dichotomous gene

	aa	Aa	AA	Total
Case	1	9	110	120
Control	10	10	100	120
Total	11	19	210	240

Table 2. Collapsed tables by CP with threshold 1 (left, $P=0.088$) and threshold 0.8 (right, $P=0.006$)

	aa or Aa	AA	Total
Case	10	110	120
Control	20	100	120
Total	30	210	240

	aa	Aa or AA	Total
Case	1	119	120
Control	10	110	120
Total	11	229	240

For this dataset, the P-value of the standard 2-df Pearson’s χ^2 statistic is 0.019, which reports an insignificant result at the 0.01 level (all P-values are obtained from permutation with 10^5 replicates). Since the first two columns have ratios less than 1, while the third column has ratio greater than 1, we combine the first two columns together to get Table 2 (left) and the standard 1-df Pearson’s χ^2 statistic for this table is $240\times(10\times 100-20\times 110)^2/(120\times 120\times 30\times 210)=3.810$, and the P-value based on 10^5 replicates of Monte Carlo simulation increases to 0.088, which concludes an insignificant result even at the 0.05 level. However, if we use a different threshold, say, 0.8, and classify the last two columns as ‘high risk’ and collapse them into one column, then we have another 2×2 table (Table 2, right) with χ^2 statistic $240\times(1\times 110-10\times 119)^2/(120\times 120\times 11\times 229)=7.717$ (double the statistic of MDR, 3.810), which is the maximal 1-df χ^2 statistic of all 2×2 collapsed tables, and the P-value reduces to 0.006, which concludes a significance result at the 0.01 level.

The simple example shown above suggests that a partitioning strategy with a different threshold may improve the result of MDR substantially. In this article, we propose an exhaustive search method based on the OCP property (Theorem 1 below), which can quickly identify the threshold that corresponds to the 2×2 table with the maximal χ^2 value among all possible 2×2 tables. The computational cost of this method is much smaller than the naive exhaustive search method. Before presenting our main results, we introduce a couple of concepts.

Definition 1 (CP). *Let S be a $2\times K$ contingency table. An integer set $\mathcal{P}=\{i_1,i_2,\cdots,i_s\}\subset\{1,2,\cdots,K\}$ is called a combinatorial partition (CP) of S if the columns in $\mathcal{P}$ and columns in $\mathcal{P}$'s compliment $\mathcal{P}^c$ are, respectively, collapsed into one column, to form a new 2×2 table.*

Definition 2 (OCP). *Let S be a $2\times K$ contingency table with each column having at least one positive entry, and suppose that S is sorted into ascending*

Table 3. Upper bounds of computational complexity for different CP strategies

Gene number	Exhaustive search	OCP	CP
n	$2^{3^n-1}-1$	3^n-1	1
2	255	8	1
3	6.71×10^{7}	26	1
4	1.21×10^{24}	80	1
5	7.07×10^{72}	242	1
6	1.41×10^{219}	728	1

order by column according to the risk ratio of each column. Then the CPs $\mathcal{P}=\{1,2,\cdots,i\}, 1\le i\le K-1$, *are called OCPs.*

Obviously, for a $2\times K$ contingency table S, there are up to $2^{K-1}-1$ distinct CPs, while the number of distinct OCPs is not more than $K-1$. Without loss of generality, for n candidate genes, suppose that each gene (e.g. SNP) has three genotypes, which are labeled as $0,1$ and 2. Then the genotype combinations can be organized as a 2×3^n contingency table, with the first row containing the counts of the genotype combinations of the cases and the second row for the controls. The naive exhaustive strategy that finds the optimal table with maximal χ^2 value would involve up to $2^{3^n-1}-1$ collapsed tables and calculations of χ^2 statistics. This number is enormously large even when n is small. For example, three genes corresponds to up to $2^{3^3-1}-1=67\,108\,863$ tables. Therefore, it is not feasible to use such exhaustive strategy. The following theorem states that the exhaustive search can be accomplished efficiently by an OCP method.

THEOREM 1 (OCP optimality). *For any contingency table S, let* Ω *be the set of all CPs, and* Ω_o *be the set of all OCPs of S. Then*

$$\max_{\mathcal{P}\in\Omega}\chi^2_{\mathcal{P}}=\max_{\mathcal{P}\in\Omega_o}\chi^2_{\mathcal{P}}.$$

here $\chi^2_{\mathcal{P}}$ *is the* χ^2 *statistic for* $\mathcal{P}$.

An equivalent result has been established previously by Shih (2001), which is a corollary of a general theorem given in Breiman *et al.* (1984). During the preparation of this article, we found this result independently and an elementary proof can be found in Supplementary Material. We shall call any $\mathcal{P}_0\in\Omega_0$ maximizing the χ^2 value as the optimal OCP. According to the definition, any CP selected by MDR is also an OCP, but it is not necessarily the optimal OCP. The example in Section 2.1 sheds some lights on this.

Table 3 illustrates the upper bounds of computational complexities (number of χ^2 computed) for naive exhaustive search, OCP (for OMDR), and CP with fixed threshold (for MDR) strategies. Compared with the naive exhaustive search strategy, the OCP strategy is much more efficient. In real data analysis, the computational burden can be further reduced by using the sparsity property. Suppose the total sample size is k, then the number of columns with at least one positive count would be at most k, and the computation complexity is not more than $k-1$.

The analysis above suggests that the computational complexity of OCP procedure should be bounded by the minimum of the number of OCPs and the total sample size. In fact, this upper bound can be further relaxed. Suppose S_1 is a contingency table with positive margins and it is sorted into ascending order by column according to the ratio of the number of cases to controls. Since the columns with zero entries for cases have ratio 0, these columns can be combined into one column; likewise, the columns with zero entries for controls can be combined into another column. Let S_2 be this combined version of S_1. We have the following proposition, the proof of which can be found in Supplementary Material.

PROPOSITION 1. *Suppose* S_1 *is a contingency table and* S_2 *is the combined version of* S_1. *Then*

$$\max_{S_1}\chi^2=\max_{S_2}\chi^2.$$

We illustrate the above property by an example. We generated the genotypes of 50 cases and 50 controls based on Model 6 in Ritchie *et al.* (2003). In this example, we can find the optimal OCP of the 2×729 table with only about 11 operations according to Proposition 1, which is much smaller than the number of OCPs 728 and the total sample size 100.

2.2 Order determination and *P*-value calculation by GEVD

Adequacy of the MDR method was assessed by comparing the average cross-validation (CV) consistency (or prediction error) from the observed data to the distribution of average consistency (or prediction error) under the null hypothesis of no associations derived empirically from 10-fold CV and a large number of permutations. Thus, the computational burden is huge when the number of loci is large. In this article, in order to reduce computational burden we will use a *P*-value approximated by a GEVD (Jenkinson, 1955) as a validity measure for the OMDR procedure. A similar approach has been used in Pattin *et al.* (2008). We describe the procedure as follows.

Suppose we want to detect interactions up to order N_0, $N_0<N$. For every order n ($n=1,2,\cdots,N_0$) and every choice of $n+1$ loci out of the total of N, we use the OCP method to find the maximal 1-df χ^2 statistic. The optimal gene combination for this order of interaction is the one which maximizes the maximal χ^2 statistic and we denote this maximum by $T_n^{(0)}$. The *P*-value of $T_n^{(0)}$ is used to measure significance of the n-order interaction model. The optimal order of interaction is the one with the smallest *P*-value. Since $T_n^{(0)}$ is the maximum of many χ^2 statistics, we assume that its null distribution can be approximated by a GEVD with cumulative distribution function

$$F(x;\mu,\sigma,\xi)=\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},\xi(x-\mu)>-\sigma, \qquad (1)$$

where $-\infty<\mu,\xi<+\infty,\sigma>0$. Therefore, the *P*-value of $T_n^{(0)}$ can be approximated by $\rho_n=1-F(T_n^{(0)};\mu,\sigma,\xi)$.

To compute ρ_n, we randomly permute the case–control labels of the original dataset to get M sets of null data. Since we have assumed the distributional form of the null distribution, the number of permutation replicates, M, can be taken to be very small (say, less than 50 for most studies). The maximal 1-df χ^2 for these M permutation datasets, $\{T_n^{(m)},n=1,2,\cdots,m=1,2,\cdots,M\}$, can be obtained in the same way as that for $T_n^{(0)}$. Then, we use the permutation sample $\{T_n^{(m)},m=1,...,M\}$ to estimate the unknown parameters in ρ_n based on the maximum likelihood principle. With this permutation sample, we can get the maximum likelihood estimates (MLEs) of the three parameters μ,ξ and σ. Notice that the asymptotic property of the MLE may not hold if the shape parameter ξ falls outside the interval $[-0.5,0.5]$. Actually, the MLE may not exist if $|\xi|\ge1$ (Smith, 1985). Therefore, we solve for the MLE with a constraint $-0.5\le\xi\le0.5$. The initial values of parameters are generated by the probability-weighted moment method (Hosking *et al.*, 1985). For each $n=1,2,\cdots,N_0$, the *P*-value, ρ_n, for testing association between any n genes and disease can then be estimated by $\hat{\rho}_n=1-F(\log T_n^{(0)};\hat{\mu}_n,\hat{\sigma}_n,\hat{\xi}_n)$, where $\hat{\mu}_n,\hat{\sigma}_n,\hat{\xi}_n$ are estimated parameters. The order of interaction is estimated by $\hat{n}=\operatorname{argmin}\{\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_{N_0}\}$, and the corresponding gene combination is chosen as the final interaction model.

Finally, we need to assess the significance of the detected model by computing the *P*-value of $\hat{\rho}_{\hat{n}}$. To do this efficiently, we apply the GEVD theory again and assume $-\log\hat{\rho}_{\hat{n}}$ has a GEVD distribution as in (1) under the null hypothesis. To estimate the parameters in this distribution, we permute the original dataset again and repeat the above procedure M_1 times and get null values $-\log\hat{\rho}^{(m)}_{\hat{n}_m},m=1,2,...,M_1$. These values are used to estimate the parameters in the GEVD distribution and hence the *P*-value of $\hat{\rho}_{\hat{n}}$. Note that due to application of the GEVD theory, the total number, MM_1, of permutations is relatively small. Figure 1 illustrates the procedure of the OMDR method.

3 RESULTS

In this section, we first show some simulation results, then we apply our method to a real dataset.

3.1 Simulation

We conducted simulations to compare the proposed OMDR with MDR. We considered 10 diallelic genetic markers (e.g. SNPs), two of which were assumed to be associated with the disease and the others are not associated. We generated genotypes of the two associated markers from four different two-locus models. Under Model 1 (Table 4), the two markers have no marginal effects but have interaction effects (Culverhouse *et al.*, 2002). Models 2, 3 and 4 (Table 5) are the epistasis model, the threshold interaction model, and the multiplicative interaction model considered in Marchini *et al.* (2005) and Pickrell *et al.* (2007). In generating genotype data for the 10 genes, we assume they obey the Hardy–Weinberg equilibrium law and are in linkage equilibrium in the population. The population minor allele frequencies are 0.25 for the associated loci and 0.5 for the other 8 loci. In what follows, we consider interactions up to the order of $N_0 = 4$ and all of the simulation results are based on 400 replications. In addition, we consider balanced case–control study only, i.e. the numbers of cases and controls are equal since the results for unbalanced design are similar. As mentioned in the introduction, there are many versions and extensions of MDR in the literature. We use the version based on consistency and predictive error as the benchmark (Ritchie *et al.*, 2003) in our comparisons.

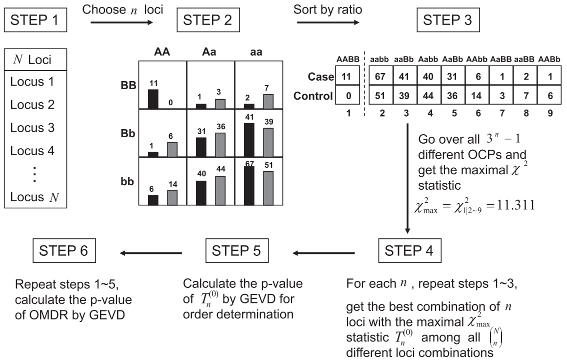
STEP 1
Choose n loci
STEP 2
Sort by ratio
STEP 3
N Loci
Locus 1
Locus 2
Locus 3
Locus 4
⋮
Locus N
AA
Aa
aa
BB
Bb
bb
11
0
1
3
2
7
1
6
31
36
41
39
6
14
40
44
67
51
AABB aabb aaBb Aabb AaBb AAbb AaBB aaBB AABb
Case 11 67 41 40 31 6 1 2 1
Control 0 51 39 44 36 14 3 7 6
1 2 3 4 5 6 7 8 9
Go over all $3^n - 1$ different OCPs and get the maximal χ^2 statistic
$\chi^2_{\max} = \chi^2_{1|2\sim9} = 11.311$
STEP 4
STEP 5
STEP 6
For each n, repeat steps 1~3, get the best combination of n loci with the maximal $\chi^2_{\max}$ statistic $T_n^{(0)}$ among all $\binom{N}{n}$ different loci combinations
Calculate the p-value of $T_n^{(0)}$ by GEVD for order determination
Repeat steps 1~5, calculate the p-value of OMDR by GEVD

Fig. 1. Procedure of OMDR method.

Table 4. Two-locus interaction models without marginal effect (entries are penetrance. p_1 and p_2 are the minor allele frequencies of two loci, $q_1 = 1-p_1$ and $q_2 = 1-p_2$; K is the population prevalence; ϕ is the tuning parameter controlling the size of interaction effect)

		L_2		
		bb	Bb	BB
L_1	aa	K	K	K
	Aa	K	$(1-\frac{p_1p_2}{4q_1q_2}(1-\phi))K$	$(1+\frac{p_1}{2q_1}(1-\phi))K$
	AA	K	$(1+\frac{p_1}{2q_1}(1-\phi))K$	ϕK

Table 6 shows the empirical Type I error rates of the OMDR method based on 400 replicates of simulation and $M = M_1 = 50$ permutation replicates in order determination and P-value calculation. We can see that all the empirical Type I error rates of OMDR are close to the nominal level 0.05. Our simulations suggest that the number of permutation replicates can even be reduced to 20 without inflating the size of test when the total sample size is between 200 and 600. Simulation results for other choices of minor allele frequencies also show that the GEVD method can well control the Type I error rate of the OMDR method (results are not shown here).

In our simulation study, we define the detection rate (DR) as the probability of successfully pinpointing the true set of interacting genes by the OCP method when the order of interaction is known *a priori*. The detected set is the one with maximal χ^2 no matter it is significant or not. Power is defined as the success probability of locating the pre-disposing genes without knowing the order of interaction while controlling the Type I error. Since determining order of interaction may have extra error, power is always less than the DR.

Figure 2 shows DRs and powers for Model 1. It is seen that, compared with the MDR, OMDR has a substantial gain in DR and

Table 6. Empirical Type I error of OMDR

Total sample size	200	400	600	800	1000	1200
Type I error rate[a]	0.054	0.051	0.055	0.047	0.065	0.059

[a] The 95% confidence interval of empirical Type I error is [0.029, 0.071]

Table 5. Two-locus interaction models with marginal effect [entries are odds. α is the baseline for genotype (aa, bb); λ_1, λ_2 are main effects of two loci; γ is the interaction effect]

		Epistasis			Threshold interaction			Multiplicative interaction		
		L_2			L_2			L_2		
		bb	Bb	BB	bb	Bb	BB	bb	Bb	BB
L_1	aa	α	α	$\alpha\lambda_2$	α	$\alpha\lambda_2$	$\alpha\lambda_2$	α	$\alpha\lambda_2^{\frac{1}{2}}$	$\alpha\lambda_2$
	Aa	α	α	$\alpha\lambda_2$	$\alpha\lambda_1$	$\alpha\lambda_1\lambda_2\gamma$	$\alpha\lambda_1\lambda_2\gamma$	$\alpha\lambda_1^{\frac{1}{2}}$	$\alpha\lambda_1^{\frac{1}{2}}\lambda_2^{\frac{1}{2}}\gamma^{\frac{1}{4}}$	$\alpha\lambda_1^{\frac{1}{2}}\lambda_2\gamma^{\frac{1}{2}}$
	AA	$\alpha\lambda_1$	$\alpha\lambda_1$	$\alpha\lambda_1\lambda_2\gamma$	$\alpha\lambda_1$	$\alpha\lambda_1\lambda_2\gamma$	$\alpha\lambda_1\lambda_2\gamma$	$\alpha\lambda_1$	$\alpha\lambda_1\lambda_2^{\frac{1}{2}}\gamma^{\frac{1}{2}}$	$\alpha\lambda_1\lambda_2\gamma$

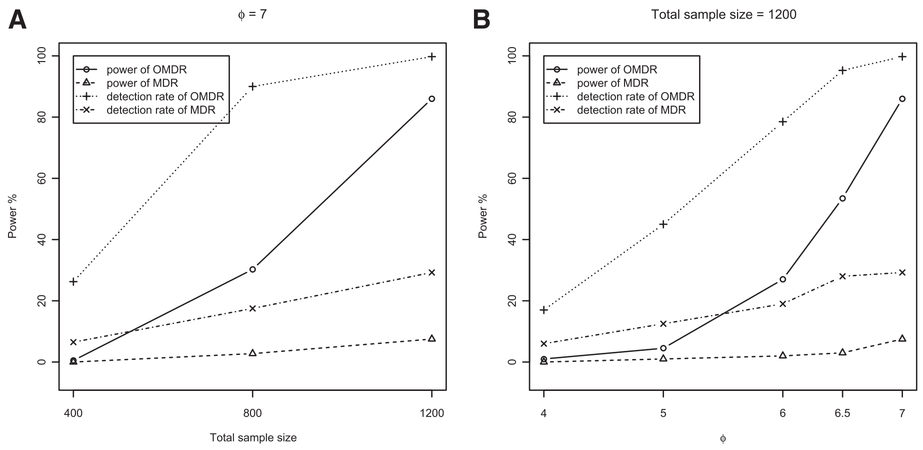

Fig. 2. Comparison of OMDR and MDR under the interaction model with no marginal effects.

power. For example, when the tuning parameter ϕ is 7 in Model 1 (Fig. 2A) the DR gains of OMDR over MDR are 0.20, 0.73 and 0.71, and the power gains are 0.05, 0.28 and 0.79 for sample sizes 400, 800 and 1200, respectively. In addition, when the total sample size is fixed to be 1200 and the size of interaction effect ϕ changes from 4 to 7 (Fig. 2B), the DR increment of OMDR ranges from 0.11 to 0.71, while the power gain ranges from 0.01 to 0.79.

We also studied the situation that the associated loci have both main and interaction effects. Table 5 lists three models of this kind that were studied in the literature, namely, the epistasis, threshold and multiplicative interaction models. Simulation results for the former two models are summarized in Figure 3. The sizes of the main effects of the two associated loci are fixed to be $\lambda=\lambda_1=\lambda_2=2.0$. Again, the DR of OMDR is uniformly higher than that of MDR. For the threshold interaction model, the DR gain of OMDR over MDR ranges from 0.22 to 0.39. Except for the case of $\gamma=2.5$ and sample size 400, the power gain of OMDR over MDR is around 0.3 for all the other five cases depicted in the lower panels of Figure 3. Figures 2 and 3 show that the power of the OMDR method can be even greater than the DR of MDR when the interaction effect is large. Analogous results but with smaller power gain are observed for the epistasis model. For example, when the total sample size is 1200, the power gains of OMDR are 0.04 and 0.07 for interaction effect $\gamma=3,4$. But when interaction effect is smaller ($\gamma=2$), no power gain is spotted. For the multiplicative interaction model and the six models used in Ritchie *et al.* (2003), the DR behavior of OMDR and MDR are similar but OMDR is slightly less powerful (see Supplementary Material), with the biggest power difference of 0.12 recorded for Model 5 (0.70 for MDR and 0.58 for OMDR). The magnitude of this loss in power by OMDR is not as big as the power gained by OMDR for the interaction model with no marginal effect (Fig. 2) and the threshold interaction model (Figure 3C and D), which can reach 0.79 and 0.31, respectively.

Upon closer scrutiny, we find out what happened was that OMDR would sometimes select a larger model containing the true model instead of the true model. MDR did not suffer from this problem because it selects order of interaction by means of an integer valued criterion called consistency and it gives preferences to lower order models in case of ties. In contrast, OMDR selects order by minimizing the estimated P-values that are continuous with no ties in general, and sometimes a higher than necessary order of interaction is chosen because its P-value is just slightly smaller. Since MDR favors small models by design and the true model used to simulate the data is of order 1 only, the fact that MDR is slightly more powerful than OMDR for the six models of Ritchie is perhaps understandable.

3.2 Real data analysis

We applied the GEVD-based OCP and the permutation distribution based OCP to find the OMDR for the real case–control dataset from the Ontario Familial Breast Cancer Registry (John *et al.*, 2004). This case-control study contains 398 breast cancer cases and 372 population controls. Onay *et al.* (2006) analyzed the data set as summarized below. Nineteen SNPs from 18 cancer-related genes passed the Hardy–Weinberg equilibrium test. Under recessive, dominant and co-dominant models, an unconditional logistic regression procedure was employed for each SNP by adjusting for age, BMI and family history. Only XPD-[Lys751Gln] showed a significant main effect at the 0.05 level according to a crude P-value. However, none of the 19 SNPs showed significant evidence of a main effect after correcting for multiple testing by applying the false discovery rate (FDR) principle (Benjamini and Hochberg, 1995). On the other hand, four two-way interactions, XPD-[Lys751Gln] and IL10-[G(-1082)A], GSTP1-[Ile105Val] and COMT-[Met108/158Val], COMT-[Met108/158Val] and CCND1-[Pro241Pro], and BARD1-[Pro24Ser] and XPD-[Lys751Gln], were detected at the 0.05 level based on FDR-adjusted P-values under a co-dominant model, although none of these four two-way interactions remained significant by Bonferroni adjustment. The joint genotype counts of the four SNP pairs are given in Table 5 in Onay *et al.* (2006).

We used $M=M_1=100$ permutation replicates to get more stable and accurate estimations of P-values. Results of two-way interaction analysis are shown in Table 7. Note that the P-values computed by GEVD are quite similar to those computed by 10^5 permutation replicates. Meanwhile, the computation time is reduced more than 10-fold by using the GEVD method in our setting. Thus, using GEVD rather than permutation with the OCP method can significantly reduce the computation time while ensuring accuracy.

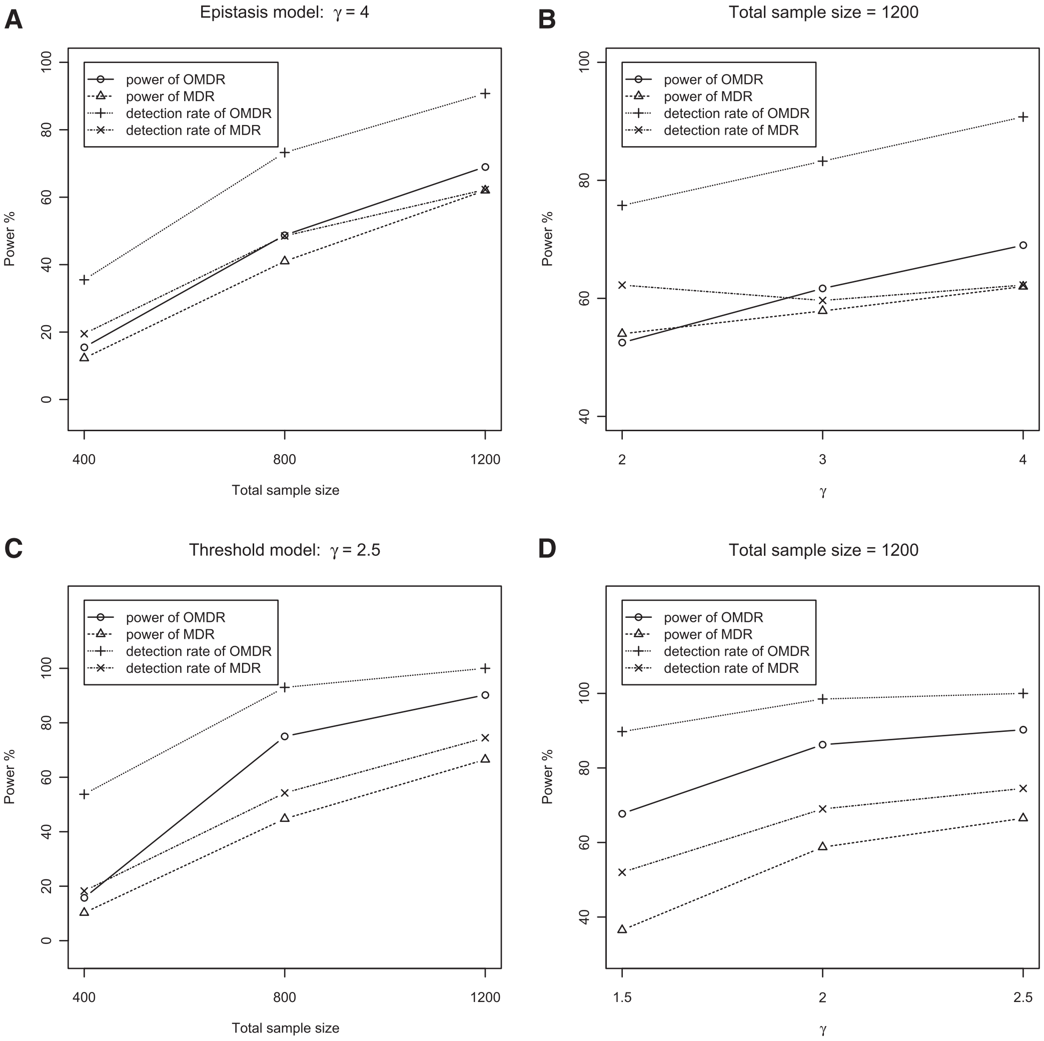

Fig. 3. Comparison of OMDR and MDR under the epistasis model and threshold model with marginal effects.

Table 7. Results for the breast cancer data

SNP–SNP interaction	OMDR_{GEV}[a]	OMDR_P[b]	MDR[c]	FDR[d]	FPRP[e]
XPD × IL10	0.005	0.010	0.008	0.007	0.092
BARD1 × XPD	0.087	0.085	0.082	0.014	0.671
COMT × CCND1	0.066	0.068	0.054	0.014	0.169
GSTP1 × COMT	0.137	0.166	0.132	0.007	0.093

[a] OCP with GEVD, $M = M_1 = 100$.
[b] OCP with permutation based on 10^5 replicates of Monte Carlo simulation.
[c] MDR with permutation based on 10^5 replicates of Monte Carlo simulation.
[d] FDR-adjusted P-values, see Table 4 in Onay *et al*. (2006).
[e] FPRP, see Table 4 in Onay *et al*. (2006).

The OCP solution, being the maximal χ^2, is always not less than the MDR solution and, in this example, they coincide with each other. Therefore, in Table 7 the P-values of OCP method using GEVD approximation or permutations are also similar to but slightly larger than the ones from the MDR method. We also provided some real data examples in the Supplementary Material to show that it is a common phenomenon in practice that the partitionings of MDR and OMDR are different in which case the maximal χ^2 statistic will be strictly larger than the MDR solution.

False positive report probabilities (FPRPs) listed in Table 7 are from Table 4 in Onay *et al*. (2006), only the analysis results adjusted for age are presented due to the highly missing proportion of BMI. FPRP is defined as the probability that the null hypothesis is true given a statistically significant result due to some test procedure, which depends on the prior probability that the null hypothesis is true, the level and statistical power of the test (Thomas and Clayton, 2004; Wacholder *et al*., 2004). According to FPRP at the 0.1 level, at least two out of the four interactions are unnoteworthy: COMT-[Met108/158Val] and CCND1-[Pro241Pro], BARD1-[Pro24Ser] and XPD-[Lys751Gln] were reported to be insignificant. Our OMDR method also reports insignificance of these two interactions at the 0.05 level (P-value = 0.066 and 0.087, respectively), while agrees with the conclusion in Onay *et al*. (2006) that the interaction of XPD-[Lys751Gln] and IL10-[G(-1082)A] plays a very important role in breast cancer ($P = 0.005$). However, OMDR suggests that the interaction of GSTP1-[Ile105Val] and COMT-[Met108/158Val] might not be significant ($P = 0.137$), which is different from the conclusion in Onay *et al*. (2006) based on FDR-adjusted *P*-value and FPRP.

4 DISCUSSION

MDR is a powerful non-parametric method in genetic interaction studies. Many empirical computer studies and real data analyses have demonstrated the impressive capability of MDR in detecting high-order interactions. The key idea of MDR is the pooling of high-dimensional predictors into two distinct groups, which reduces dimensionality and alleviates the sparsity problem. We observe that the power can be further improved by introducing the exhaustive search strategy. In this study, an OCP-based search strategy is proposed for finding the optimal contingency table with largest χ^2 statistic, which has computational complexity that is dramatically lower than that of the exhaustive search method. The GEVD theory is employed to further reduce computational cost. Both simulation studies and real data analysis show that the GEVD approximation is rather accurate and much more computationally efficient than using the permutation distribution.

Simulation studies in Ritchie *et al*. (2003) show that the MDR method performs robustly in the presence of noise due to genotype error and missing data, but may lose power substantially in the presence of phenocopy. Chen *et al*. (2008) proposed a gene–gene interaction detection approach based on the support vector machine and demonstrate several advantages of their machine learning method by simulation studies, especially the strong ability (e.g. power and model stability) to handle sample noise. Since the OMDR method always leads to higher DRs than MDR, it is conceivable that our OMDR method can deal properly with this noise while MDR does not. The performance of OMDR in the presence of these sources of noise will be explored in future studies and compared with Chen *et al*.'s (2008) method.

Another important issue is the class-imbalance problem in gene–gene interaction detection. The generalization ability of many classification methods such as logistic regression will be weakened by extremely imbalanced datasets. In balanced case–control design, the threshold used by MDR is 1 which can be naturally adjusted by the ratio of cases to controls when the design is unbalanced. Velez *et al*. (2007) proposes the balanced accuracy function defined as the arithmetic mean of sensitivity and specificity to substitute the accuracy function such as cross-validation consistency and prediction error used in previous literature (Ritchie *et al*., 2001, 2003). Simulation shows that the power of the MDR method using this balanced accuracy function with the adjusted threshold can be elevated efficiently. Since the variable threshold used in OMDR is data-driven, our OMDR method can also be applied without modification to detect gene–gene interactions in unbalanced case–control studies.

ACKNOWLEDGEMENTS

The authors would like to thank David Nott for his editorial help and three anonymous reviewers for their insightful comments and suggestions, which greatly improve the manuscript.

Funding: USTC Innovation Foundation for Graduate Students (to X.H. and H.Z.); National Natural Science Foundation of China (to H.Z. and Y.Y.).

Conflict of Interest: none declared.

REFERENCES

Bastone,L. *et al.* (2004) MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. *Hum. Hered.*, **58**, 82–92.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical approach and powerful approach for multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Breiman,L. *et al.* (1984) Classification and regression trees. Chapman & Hall, Boca Raton.

Bush,W.S. *et al.* (2005) Can neural network constraints in GP provide power to detect genes associated with human disease? *Appl. Evol. Comp. Proceed.*, **3449**, 44–53.

Bush,W.S. *et al.* (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics*, **9**, 238–254.

Calle,M.L. *et al.* (2008) Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat. Med.*, **27**, 6532–6546.

Chatterjee,N. *et al.* (2006) Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am. J. Hum. Genet.*, **79**, 1002–1016.

Chen,S.H. *et al.* (2008) A support vector machine approach for detecting gene-gene interaction. *Genet. Epidemiol.*, **32**, 152–167.

Cho,Y.M. *et al.* (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia*, **47**, 549–554.

Chung,Y. *et al.* (2007) Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics*, **23**, 71–76.

Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Culverhouse,R. *et al.* (2002) A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.*, **70**, 461–471.

Heidema,A.G. *et al.* (2007) Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. *Genet. Epidemiol.*, **31**, 910–921.

Hoh,J. and Ott,J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Genet. Rev.*, **4**, 701–709.

Hosking,J.R.M. *et al.* (1985) Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, **27**, 251–261.

Huang,J. *et al.* (2004) Tree-structured supervised learning and the genetics of hypertension. *Proc. Natl Acad. Sci. USA*, **101**, 10529–10534.

Jenkinson,A.F. (1955) The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Q. J. R. Meteorol. Soc.*, **81**, 158–171.

John,E.M. *et al.* (2004) The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res.*, **6**, R375–R389.

Julia,A. *et al.* (2007) Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics*, **90**, 6–13.

Lee,S.Y. *et al.* (2007) Log-linear model based multifactor dimensionality reduction method to detect gene–gene interactions. *Bioinformatics*, **23**, 2589–2595.

Lou,X.Y. *et al.* (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.*, **80**, 1125–1137.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Mei,H. *et al.* (2005) Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data. *BMC Genet.*, **6** (Suppl. 1), S145.

Millstein,J. *et al.* (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.*, **78**, 15–27.

Montana,G. *et al.* (2006) Statistical methods in genetics. *Brief. Bioinform.*, **7**, 297–308.

Moore,J.H. *et al.* (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.*, **241**, 252–261.

Moore,J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.

Motsinger-Reif,A.A. *et al.* (2008) A comparison of analytical methods for genetic association studies. *Genet. Epidemiol.*, **32**, 767–778.

Namkung,J. *et al.* (2009) New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. *Bioinformatics*, **25**, 338–345.

Onay,V.U. *et al.* (2006) SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer*, **6**, 114–130.

Park,M.Y. and Hastie,T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30–50.

Pattin,K.A. *et al.* (2008) A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genet. Epidemiol.*, **33**, 87–94.

Pickrell,J. *et al.* (2007) Power of genome-wide association studies in the presence of interacting loci. *Genet. Epidemiol.*, **31**, 748–762.

Ritchie,M.D. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Ritchie,M.D. *et al.* (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.*, **24**, 150–157.

Shih,Y.-S. (1999) Families of splitting criteria for classification trees. *Stat. Comput.*, **9**, 309–315.

Shih,Y.-S. (2001) Selecting the best splits for classification trees with categorical variables. *Stat. Probab. Lett.*, **54**, 341–345.

Smith,R.L. (1985) Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Thomas,D.C and Clayton,D.G. (2004) Betting odds and genetic associations. *J. Natl Cancer Inst.*, **96**, 421–423.

Tsai,C.T. *et al*. (2007) Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis*, **195**, 172–180.

Velez,D.R. *et al.* (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.*, **31**, 306–315.

Wacholder,S. *et al.* (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.*, **96**, 434–442.

Wade,M.J. (2000) Epistasis as a genetic constraint within populations and an accelerant of adaptive divergence among them. In: Wade,M.J. *et al*. (eds) *Epistasis and Evolutionary Process*. Oxford University Press, Oxford.