**A note on Linear Model implementation for Multifactor Dimensionality Reduction**

Author: Guo-Bo Chen

Address: Queensland Brain Institute, University of Queensland, St Lucia, Australia, 4072

Corresponding to: [chenguobo@gmail.com](mailto:chenguobo@gmail.com)

**Abstract**

Multifactor dimensionality reduction (MDR) and its various extension have been widely used in human genetics for detecting high-order interaction. In the last decade, under the endeavours of many researchers, MDR makes one of the most active methods in detecting gene-gene interactions. The original MDR is built on a machine-learning kernel, which is extremely flexible and enriched in topological demonstration; but it lacks of clear statistical properties such as p-values, which is often evaluated via permutation or the central-limited theory. In this work, we used linear model to reconstruct the kernel algorithm of the MDR, called LMDR. LMDR provides statistical interpretation of the results, and get p-values straightforwardly. In addition, the previously commonly used statistics, such as accuracy can directly find its statistical interpretation. LMDR is easy to implement, and compatible to most extensions of MDR. The linear model kernel could replace the previous MDR kernel as demonstrated by simulation, and can provide a much easier framework that is compatible to current MDR methods.

**Introduction**

As the marginal effect of a single locus is often small, even large-size genome-wide association studies fail to find the much heritability of complex traits. Detecting high-order gene-gene (GXG) interactions may provide a way for revealing genetic patterns associated to complex traits. Compared with the conventional epistasis methods [1, 2], which are often limited to two-locus models, multifactor dimensionality reduction (MDR) provides a novel way that can detect high-order interactions with relative less computational burden [3]. The original MDR method only handle dichotomous traits, but its *MDR kernel algorithm* is so flexible that it has been extended to many scenarios. For its extended works, the *MDR kernel algorithm* is inherited by all its offspring methods with little modification. Among its many improvements, the major one was proposed by installing MDR a generalized linear model [4], which allows MDR capable of handling both dichotomous and quantitative traits and adjusts for covariates when necessary. Family-based application [5–7], and handling even more complicated population structure [8] are made possible for MDR. Various software have been developed based on MDR kernel algorithm [4, 8–11]. In the last decade, under the endeavours of many researchers, MDR makes one of the most active fields in detecting gene-gene interactions [12].

Nevertheless, as all the MDR methods share the same kernel algorithm (will introduce in detail below), these methods show more or less similar weakness, which includes: 1) Topological interpretation. Topological interpretation itself is great for visualization, but it lacks the succinctness as parametric methods, which often summarize information in a couple of statistics. 2) Test statistics. In general, in MDR methods, accuracy is in general measured by the ration (TP: true positive; TN: true negative; FP: false positive; FN: false negative), and it does not have straightforward p-values. Although permutation and z-score approximation [4, 6] have been used to evaluate p-values for accuracy, they are computational expensive and requires many rounds of resampling to find approximate p-values.

In this study, we will introduce a linear model implementation that can replace the original MDR kernel algorithm and offers a natural way for evaluating p-values without any permutation or resampling. It should be noticed that here the linear model is to replace the MDR kernel algorithm rather than an extension as has been done via the generalized linear model, which adds flexibility in adjusting covariates but does not touch the MDR kernel algorithm. Under the linear model MDR, we will show the statistical properties of the previous used statistics, such as accuracy, and their connection with commonly genetic parameters and p-values. The LMDR offers a natural way for evaluating p-values without any permutation or resampling.

**Methods**

**The original MDR kernel algorithm**

Depending on the scheme of the analysis,samples are split into training and testing sets directly, or partitioned into even subdivisions, which can be used for the leave-one-out cross-validation. Given discrete factors, which can be either genotypes or environmental factors, choosing from discrete factors a total of distinct -dimensional grid can be generated. The purpose of MDR algorithm is to classify the -dimensional grid into high-risk or low-risk classes.

**In step one**, each such subset corresponds to a -dimensional finite grid, and each subject who is genotyped and assessed for the environmental exposures will fall into exactly one cell in this grid. The values of the subjects are averaged over each cell (a family of methods has been proposed to adjust the phenotypes [13]). Each nonempty cell is labelled either high-risk if its average statistic value is not less than some threshold , or low-risk otherwise. When the phenotype is standardized, can simply be zero by default.

**In step two**, a multilocus model is formed by pooling high- and low-risk cells into two groups (i.e., high-risk and low-risk).

**In step three**, for the model identified in step two, the classification accuracy for the training set and the testing set is assessed by the averages of the statistic values in the high-risk group and the low-risk group, respectively.

Steps one to three are iterated for all other possible combinations, and the above procedure is repeated for combinations **(Figure 1)**.

The framework above has been used for nearly all MDR methods. In spite of its flexibility for easy extension, the major problem is the evaluation of the *p*-values for the test statistic, and the proportion of variance explain by the detected interaction. For example, one of the popular test statistic is accuracy, (TP: true positive; TN: true negative; FP: false positive; FN: false negative), for which many reported interaction detected by MDR ranges between 0.56~0.62 [14]. However, it is unsure its p-value and the proportion of variance explained. In genetic research, both are important.

So far, *p*-values for test statistics in MDR methods are often evaluated by permutation. It often shuffles subjects’ phenotypes, and rerun the step one to three. In order to reach accurate *p*-value, it takes upon thousands of rounds of permutations. With enhance computational facility, exhaustive searching for -dimenstioanl interaction is made possible[8, 9]. For example, searching two-locus interaction on GWAS data with 1,000,000 loci, it generates up 1012 interactions. In order to reach nominal cut-off for 0.05 at the genome-wide level, 1014 permutations are required for each locus. It quickly drains the computational resource.

**LMDR** kernel algorithm

The LMDR has same procedure but for the step three. In step three, we introduce a coding function . If a -dimension locus genotype is classified as high-risk, , otherwise .

Given the coding scheme, a generalized linear model is implemented on the data

(Equation 1)

in which is the phenotype of the subject, is the -locus genotype. is the link function, depending on the distribution of the phenotype, is identify for continuous traits, and logit for case-control data. The is the intercept of the model, is the regression coefficient, and is the residual. The LMDR is easily integrated into the original MDR framework **(Figure 1)**.

**Scenario I: For quantitative traits**

This regression can be applied to both the training and the testing sets. However, for the training set, as the classification of the genotypes are based on the training set itself, the model will be over-fitted; whereas for the testing test, which is independent to the classification in step 2, can give the unbiased estimate of , which is interpreted as the genetic effect of -dimension interaction. As it is a linear model framework, evaluation of the p-value for the regression coefficient is straightforward.

and . As an interaction model often explains a small proportion of the total variance, , .

The z-score for is

(Equ 2)

in which is the correlation between and the genotypes after classification. After taking the square of the z-score, it becomes and follows , a chi-square test with one degree of freedom.

The proportion of the variance explained by the model can be measure by the coefficient of determinant, . The statistical significance can be evaluated by test with the degrees of freedoms and 1.

For the testing test, it may exist orphan genotypes, the classifications of which are unknown. The proportion of orphan genotypes will increase when searching interaction goes to even higher dimension. Two ways to treat unknown genotypes.

1) Those genotypes can either be discarded or randomly classified into high- or low-risk groups.

2) Discarding those genotypes, the test statistic for the regression coefficient becomes , in which , a shrunk sample size; randomly classifying them into the risk groups is equivalent to adding noise into the model, a natural penalty that increases the difficult for detecting high-order interaction under the finite sample size.

**Simulation results for quantitative traits**As a proof of principle, we set the data into training set, which has 1000 subjects, and test sets, which has 500 subjects. 200 loci, which are in linkage equilibrium, were simulated, and the two-locus interaction model was searched exhaustively. To evaluate the validation of LMDR, we simulated the data from the null distribution, under which the distribution of the *p*-value of the test statistic should follow a uniform distribution. In total it generated 19,900 two-locus interaction models. The experiment-wise p-value for the type I error rate control is .

The searching strategy is as described in the method section. For comparison, the test statistics were evaluated using both the MDR accuracy and the LMDR regression.

Figure Q1: type I error rate for LMDR

It is known that under the null distribution, the p-value follows uniform distribution. Here each p-value was from a z-score test for the regression coefficient for LMDR, which was constructed based on the classification of the high- and low-risk groups. It should distinguish the distributions of p-values from the training set and the test set. In the training set, the distribution of the p-value is skewed to low-tail, which indicates the over-fitting in the training set. It is often the case for the training set due to ascertainment, and explained why the test statistic from the training set along does not make too much sense. In contrast, the distribution of the p-value for the test set is nearly uniform. Under Bonferroni correction, only one model reached significance level of 0.05 (See red points at the right panel in Figure 3). It indicates well controlled type I error rate, which is crucial for the validation of a statistic test. This Figure demonstrated the validation of the LMDR test.

Figure Q2: Proportion of variance explained by multi-locus model

One question remained for MDR is how much variation has been explained by a multi-locus model. Here the from LMDR is used as proxy to measure the contribution of the variance explained. is a useful measure for variance explained, or heritability. It can be seen that a linear relationship between the and accuracy. Training accuracy, which is always greater than 0.5, has between 0 to 0.02, nearly linear between and accuracy. Similarly, the linear relation correlation exists between and the test accuracy. The test accuracy can be less than 0.5, and it shows symmetric distribution.

Of note, because of small sample size for the testing accuracy, variation of is greater than that in the training set.

Figure Q3: Regression coefficient and accuracy

Similarly, the regression coefficient (Eq 1) also shows its corresponding to the accuracy, but not as linear as that between and accuracy. Due to ascertainment, the training accuracy is greater than 0.5 regardless the testing accuracy is greater than 0.5 (blue) or less than 0.5 (red), and its corresponding is greater than 0; whereas the testing accuracy was less or greater than 0.5 by chance.

The simulation above indicated that LMDR captures nearly the same information as did MDR statistics, but LMDR showed clearer statistical properties and much easier for interpretation. **, rather than the regression coefficient, shows linear relationship with accuracy.**

**Statistical power for quantitative traits**

It seems that the p-value depends on sample size and the correlation between the phenotype and classification of the genotypes. , and the non-centrality parameter is . However, we only observe the classified genotypes rather than the real classification. For example, for a two-locus model, if the real high-risk groups are , however, in classification, due to power issues the classified high group may be different, say . It makes the loss of power, a phenomena similar to the linkage disequilibrium between the causal locos and the genotyped marker. So, the NCP=, in which indicates how accurate the high-risk groups has been uncovered by the MDR algorithm.

Assume for a genotypic cell , its has distribution , if , it is high-risk group, and its probability being identified as a high-risk group is , in which is the count of genotype and is the cumulative probability for given the normal distribution with and sd . Similarly, the probability for . So, a large training size will help increase accuracy, and a large frequency also increase classification accuracy.

The probability that high- and low-risk genotypes will be identified correctly is

For a two-locus interaction model, if the high-risk is , and the allele frequency is 0.5 for each locus, , and . Assuming the phenotype has been standardized, , then ; the proportion of high-risk is .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 1000 |  |  |  |  |
| 5000 |  |  |  |  |

It is clear that for the high-risk groups, it is not likely to be misclassified, whereas the low-risk class is liked to be misclassified. is determined by the discovery sample size, and it will be reduced because of low-risk groups, and eventually decrease the power of MDR.

NCP=

**Scenario II For case-control data**

The observed data can be summary into the contingency table below

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| High-risk |  |  |  |
| Low-risk |  |  |  |
|  |  |  |  |

A logistic regression should be used for LMDR.

, in which . The sampling variance depends on the sample size, prevalence of the disease.

Then, the is measured by Nagelkerke’s definition, which is scaled to the maximal Rsq could be for binary data[15].

in which is the likelihood under the alternative model and is under the null model.

Often compared with under the linear model, .

**Simulation study**

The training set has 1000 individuals, and the test set has 500 individuals. The prevalence is 0.5 for both training and testing. 200 loci in linkage equilibrium were simulated. Different from quantitative traits, here the logistic regression is used to estimate the regression coefficient, and is used to quantify the variance explained.

Figure C1: type I error rate for Logistic MDR

In the training set, the distribution of the p-value is skewed to low-tail, which indicates the over-fitting in the training set. It is often the case for the training set due to ascertainment, and explained why the test statistic from the training set along does not make too much sense. In contrast, the distribution of the p-value for the test set is nearly uniform. Under Bonferroni correction, no model reached significance level of 0.05. This Figure demonstrated the validation of the LMDR test.

Figure C2: Proportion of variance explained by multi-locus model

Here the from LMDR is used as proxy to measure the contribution of the variance explained. Training accuracy, which is always greater than 0.5, has between 0 to 0.05, nearly linear with accuracy. Similarly, the linear relation correlation exists between and the test accuracy. The testing accuracy can be less than 0.5, and it shows symmetric distribution.

Figure C3: Regression coefficient and accuracy

Due to ascertainment, the training accuracy is greater than 0.5 regardless the testing accuracy is greater than 0.5 (blue) or less than 0.5 (red), and its corresponding is greater than 0; whereas the testing accuracy was less or greater than 0.5 by chance.

**Statistical power**

Given samples, are cases, and are controls. After classification, follows binomial distribution too, and and genotypes are classified as high-risk and low-risk, respectively. The power can be assessed through fisher’s exact test for the contingent table

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| High-risk |  |  |  |
| Low-risk |  |  |  |
|  |  |  |  |

It is constrained that , and . is only possible to be 1 when .

For details, see <http://en.wikipedia.org/wiki/Fisher%27s_exact_test>

The much closer and , the higher the power will be.

**Impact of linkage disequilibrium**

If the pair of loci is in linkage equilibrium, ; if the pair of loci is in linkage disequilibrium, . LD will lift , and will make the p-value smaller (inflated type I error rate).

The upper bound of

Given the prevalence, , of the disease in the sample, the phenotypic variance is . Assume the number of cases is smaller than that of controls, and

in which is the proportion of high risk genotypes. Assume ,

, when is small.

If the training and the test set have different proportion of cases, the upper bound of will be different between for the training and the testing.

In cross-validation, as

**References**

1. Cockerham CC: **An extension of the concept of partitioning of hereditary variance for analysis of covariances among relatives when epistasis is present**. *Genetics* 1954, **39**(November):859–882.

2. Kempthorne O: **The correlation between relatives in a random mating population.** *Proc R Soc London Ser B, Biol Sci* 1954, **143**:102–13.

3. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**:138–47.

4. Lou X-Y, Chen G-B, Yan L, Ma JZ, Zhu J, Elston RC, Li MD: **A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence.** *Am J Hum Genet* 2007, **80**:1125–1137.

5. Lou X-Y, Chen G-B, Yan L, Ma JZ, Mangold JE, Zhu J, Robert C, Li MD, Elston RC: **A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies.** *Am J Hum Genet* 2008, **83**:457–467.

6. Chen G-B, Zhu J, Lou X-Y: **A faster pedigree-based generalized multifactor dimensionality reduction method for detecting gene-gene interactions.** *Stat Interface* 2011, **4**:295–304.

7. Martin ER, Bass MP, Gilbert JR, Pericak-Vance M a, Hauser ER: **Genotype-based association test for general pedigrees: the genotype-PDT.** *Genet Epidemiol* 2003, **25**:203–13.

8. Chen G-B, Liu N, Klimentidis YC, Zhu X, Zhi D, Wang X, Lou X-Y: **A unified GMDR method for detecting gene–gene interactions in family and unrelated samples with application to nicotine dependence**. *Hum Genet* 2014, **133**:139–150.

9. Zhu Z, Tong X, Zhu Z, Liang M, Cui W, Su K, Li MD, Zhu J: **Development of GMDR-GPU for Gene-Gene Interaction Analysis and Its Application to WTCCC GWAS Data for Type 2 Diabetes**. *PLoS One* 2013, **8**:e61943.

10. Hahn LW, Ritchie MD, Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions**. *Bioinformatics* 2003, **19**:376–382.

11. Lee SY, Chung Y, Elston RC, Kim Y, Park T: **Log-linear model-based multifactor dimensionality reduction method to detect gene gene interactions.** *Bioinformatics* 2007, **23**:2589–95.

12. Wei W-H, Hemani G, Haley CS: **Detecting epistasis in human complex traits**. *Nat Rev Genet* 2014, **15**(September):722–733.

13. Lou X-Y: **UGMDR: a unified conceptual framework for detection of multifactor interactions underlying complex traits.** *Heredity (Edinb)* 2014(August):1–7.

14. Chen G-B, Xu Y, Xu H-M, Li MD, Zhu J, Lou X-Y: **Practical and Theoretical Considerations in Study Design for Detecting Gene-Gene Interactions Using MDR and GMDR Approaches**. *PLoS One* 2011, **6**:9.

15. Nagelkerke NJD: **A note on a general definition of the coefficient of determination**. *Biometrika* 1991, **78**:691–692.

Figure 1 The algorithm of the MDR and LMDR.

Figure 2 P-value distributions for training accuracy and testing accuracy, respectively

Figure 3 Joint distributions for and accuracy for training and the testing sets, respectively

Figure 4 Joint distributions for LMDR regression coefficient and accuracy for training and testing sets, respectively.