Multifactor-Dimensionality Reduction in the Presence of Missing Observations

Yujin Chung[1], Seung Yeoun Lee[2], Taesung Park[1]

Abstract

An identification and characterization of susceptibility genes for common complex multifactorial diseases is a challengeable task, in which the effect of single genetic variation will be likely dependent on other genetic variations (gene-gene interaction) and environmental factors (gene-environment interaction). To address this issue, the multifactor dimensionality reduction (MDR) has been proposed and implemented by Ritchie et al. (2001), Moore et al. (2002), Hahn et al. (2003) and Ritchie et al. (2003). With MDR, multilocus genotypes effectively reduce the dimension of genotype predictors from n to one, which improves the identification of polymorphism combinations associated with disease risk. However, MDR cannot handle missing observations appropriately, in which missing observation is treated as an additional genotype category. This approach may suffer from a sparseness problem since when high-order interactions are considered, an additional missing category would make the contingency table cells more sparse. We propose a new MDR approach with minimum loss of sample sizes by considering missing data over all possible multifactor classes. We evaluate the proposed MDR by using the prediction errors and cross validation consistency.

Keyword : Multifactor-Dimensionality Reduction (MDR), Gene-Gene Interactions, Case-Control study.

1. Introduction.

In human genetics, one of the challenging problems is to identify and characterize the susceptibility genes for common, complex multifactorial human diseases, in which the effect of any single genetic variation will be likely dependent on other genetic variations (gene-gene interaction) and environmental factors (gene-environment interaction). To address this issue, gene-gene interactions in complex

[1] Department of Statistics, Seoul National University, San 56-1 Shillim-Dong, Kwanak-Gu, Seoul, Korea, 151-742
[2] Department of Applied Mathematics, Sejong University, 98 Gunja-Dong, Kwangjin-Gu, Seoul, Korea, 143-747
e-mail : Taesung Park (tspark@snu.ac.kr)

diseases have been examined by a logistic regression model, multilocus linkage disequilibrium tests and the Hardy-Weinberg equilibrium test. However, all of those methods have limitations in their general applications (Moore et al. 2002). For instance, logistic regression is less practical for dealing with high-dimensional data because there are many contingency table cells that contain the sparse or missing observations when high-order interactions are modeled (Ritchie et al. 2001). This can lead to very large coefficient estimates and standard errors (Hosmer and Lemeshow 2000). One solution to this problem is to collect very large number of samples to allow robust estimation of interaction effects; however, the magnitudes of the samples that are often required incur prohibitive expense (Ritchie et al. 2001).

One alternative method for detecting high-order gene-gene interactions is a multifactor-dimensionality reduction (MDR) method (Ritchie et al. 2001), which was inspired by the combinatorial-partitioning method (CPM) that examines multiple genes, each containing multiple variable loci, to identify partitions of multilocus genotypes that predict inter-individual variation in quantitative trait levels (Nelson et al. 2001). The MDR method is for detecting and characterizing high-order gene-gene and gene-environment interactions in case-control and discordant-sib-pair studies with relatively small samples (Ritchie et al. 2001). With MDR, multilocus genotypes are pooled into high-risk and low-risk groups, effectively reducing the genotype predictors from n dimensions to one dimension (Ritchie et al. 2001). This method is model-free, in that it does not assume any particular genetic model, and is nonparametric, in that it does not estimate any parameters (Ritchie et al. 2001).

However, when high-order interactions are considered with relatively small sample, there may be many multifactor cells with either missing data or singleton data (Ritchie et al. 2001), which yields to serious loss of data. To reduce the loss of data, we propose a new improved MDR method. The main idea of the new MDR is to make use of all of available data for any pair of SNPs. This method makes more observations be included in the analysis, while only complete observations are included for the original MDR. In other words, the new MDR method considers all available observations, while the original MDR method uses non-missing observations in entire dataset. Like the original MDR method, this new MDR method is also model-free and non-parametric. With the new MDR, a loss of information can be avoided. We compare the new MDR with the original MDR using the prediction accuracy and cross-validation accuracy.

In section 2, the original MDR is described and the new MDR is proposed in section 3. A short discussion is given in section 4.

2. The Original MDR method

As illustrated in Figure 1, the MDR method is implemented through six steps using 10-fold cross-validation. We denote N to be the number of total instances, N^o the number of non-missing instances in the entire dataset, and S_c the family of possible c

SNPs of all SNPs ($1 \le c \le$ total number of SNPs). In step 1, the data are partitioned equally for cross-validation. In this case, dataset is partitioned into 10 equal parts, allowing 9/10 of the data to be used for the training set while 1/10 of the data is used for the test set. Here, N^o_{case} is the number of cases and N^o_{ctl} is the number of controls in the complete training set. In step 2, an element s_c is selected from S_c if c-order interactions of SNPs are considered. In step 3, we tabulate c-dimensional frequency table for both of cases and controls according to each three genotypes of SNPs in the s_c. For example, when $c=2$ and s_c={SNP1, SNP2}, we denote $n^o_{ij,case}$ and $n^o_{ij,ctl}$ be the number of cases and controls, respectively, with the ith genotypes of SNP1 and jth genotypes of SNP2 in the complete dataset. In step 4, the ratio $\frac{n^o_{ij,case}}{n_{ij,ctl}}$ is calculated within each cell. Then each cell is labeled as high-risk if the ratio is equal or greater than the threshold, $\frac{N^o_{case}}{N^o_{ctl}}$, and low-risk if $\frac{n^o_{ij,case}}{n_{ij,ctl}} < \frac{N^o_{case}}{N^o_{ctl}}$. The concept of this method is that the proportion of case is larger than control in each high-risk cell and smaller in each low-risk cell. That is,

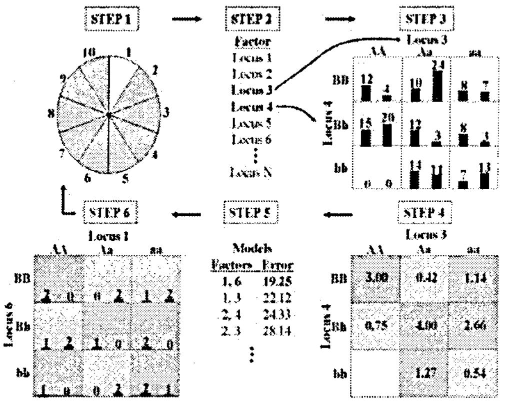

Fig.1. Summary of the general steps involved in implementing the MDR method Bars represent hypothetical distributions of cases(left) and controls(right) with each multifactor combination. Dark-shaded cells represent high-risk genotype combinations while light-shaded cells represent low-risk genotype combinations. No shading or white cells represent genotype combinations for which no data was observed.(Hahn et al. 2003)

$$\frac{n^o_{ij,case}/N^o_{case}}{n^o_{ij,ctl}/N^o_{ctl}} \ge 1 \Leftrightarrow \frac{n^o_{ij,case}}{n^o_{ij,ctl}} \ge \frac{N^o_{case}}{N^o_{ctl}} \Leftrightarrow \quad \text{high-risk}$$

$$\frac{n^o_{ij,case}/N^o_{case}}{n^o_{ij,ctl}/N^o_{ctl}} < 1 \Leftrightarrow \frac{n^o_{ij,case}}{n^o_{ij,ctl}} < \frac{N^o_{case}}{N^o_{ctl}} \Leftrightarrow \quad \text{low-risk}$$

If the dataset is balanced, the threshold is equal to 1. In step 5, ratio of correct misclassifications to the total number of instances classified within the training set for each set s_o training accuracy (1-training error), is evaluated and a set s_c with maximizing training accuracy is selected. In step 6, the prediction accuracy (1-prediction error), which is the ratio of correct classifications to the total number of instances classified within the testing set, is calculated for the single model selected in step 5. Step 1 through 6 is repeated 10 times with 10-fold cross-validation. Single best model is also selected from c -order combinations. Among this set of best multifactor models, the combination that maximize both the prediction accuracy and CV consistency is selected.

3. The New MDR method

In step 1 of Figure 1, the entire data set, not complete data, is partitioned into 10 equal parts for the 10-fold cross-validation. In step 2, an element s_c is selected from S_c and we consider complete training data only containing SNPs in the selected set s_c. We denote N^s be the number of non-missing instances in this data. Then, N^s is larger than or equal to N^o , which means more cases are used for the MDR. In step 3, when $c=2$ and s_c={SNP1, SNP2}, we also denote $n^s_{ij,case}$ be the number of cases with the ith genotypes of SNP1 and jth genotypes of SNP2 in the complete dataset and $n^s_{ij,ctl}$ the number of controls, respectively. In step 4, the new class is determined by the ratio of the proportion of cases to the proportion of controls. That is,

$$\frac{n^s_{ij,case}/N^s_{case}}{n^s_{ij,ctl}/N^s_{ctl}} \geq 1 \Leftrightarrow \frac{n^s_{ij,case}}{n^s_{ij,ctl}} \geq \frac{N^s_{case}}{N^s_{ctl}} \Leftrightarrow \quad \text{high-risk}$$

$$\frac{n^s_{ij,case}/N^s_{case}}{n^s_{ij,ctl}/N^s_{ctl}} < 1 \Leftrightarrow \frac{n^s_{ij,case}}{n^s_{ij,ctl}} < \frac{N^s_{case}}{N^s_{ctl}} \Leftrightarrow \quad \text{low-risk}$$

The steps 5 and 6 are implemented by the same way of the original MDR.

The difference between original and new MDR is how to eliminate missing cases. The complete dataset is fixed irrespective of any combination of SNPs for the cross-validation in the original MDR, while the new MDR defines the complete dataset by the selection of s_c among all possible combinations. Therefore, while the value of threshold ratio is identical for each s_c in the original MDR, the new MDR has different value of threshold ratio for each s_c depending on the selection of SNPs. In general, the different value of threshold ratio is used for each s_c, because the distribution of case is different from that of control in each s_c. The new MDR would be more efficient than the original MDR, when high-order interactions are considered, in terms of the prediction error and the cross validation accuracy.

4. Discussion

Ritchie et al. (2001) described the primary advantage of MDR; it facilitates the simultaneous detection and characterization of multiple genetic loci associated with a discrete clinical endpoint. This is accomplished by pooling genotypes from multiple loci into high-risk and low-risk groups, depending on whether they are more common in affected or in unaffected subjects. Another advantage is that it is a non-parametric approach, which avoids the problems associated with the use of parametric statistics to model high-order interactions. A third advantage is that it assumes no particular genetic model; that is, no mode of inheritance needs to be specified. This is important for diseases, such as sporadic breast cancer, in which the mode of inheritance is unknown and likely very complex. The fourth advantage is that false-positive results due to multiple testing are minimized. This is primarily due to the cross-validation strategy used to select optimal models.

On the other hand, it is also pointed out that there are some disadvantages by Ritchie et al. (2001). One important disadvantage is ability of MDR to make predictions for independent data sets when the dimensionality of the best model is relatively high and the sample is relatively small. High dimensionality and a small sample lead to many multifactor cells with either missing data or singleton data and it is a problem for estimation of the prediction error. However, a suggestion to solve these problems occurred by missing data is to use larger data set that is complete data set in each s_c rather than considering all possible missing data over a whole set of SNPs.

Alternatively, imputation of missing observations may be made by using KNN, SVD, EM algorithm or BPCA. However, these methods cannot be applied under the specific structure of the data such as haplotype block. Instead, linkage disequilibrium-based imputation and haplotype-based imputation methods may be applicable for estimating missing values of the data based on the specificity of SNP genotype data (Park et al.) Finally, investigating the relationship between the MDR method and the log-linear model is expected to be a promising approach for gene-gene interactions since the log-linear model is easy to handle and interpret.

Acknowledgement

The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation.

References

Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi 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

Hosmer DW, Lemeshow S (2000), Applied logistic regression, John Wiley & Sons, New York

Moore JH, William SM(2002) New strategies for identifying gene-gene interactions in hypertension. Ann Med 34:88-95

Nelson M, Kardia SLR, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.

Genome Res 11:458-470

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

Marylyn D. Ritchie, Lance W. Hahn, and Jason H. Moore (2003) Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity. Genetic Epidemiology 24:150-157

Yun-Ju Park, Young-Jin Kim, Jung-Sun Park, Kuchan Kim, InSong Koh, and Ho-Youl Jung. A new Method for Imputation of Missing Genotype using Linkage Disequilibrium and Haplotype Information.(submitted)