Decomposition, Gradient, and Reduction Colinearity in High-Dimension Data from a Case-Control Study

[1]Yuanzhang Li, PhD, David W. Niebuhr, MD, MPH, MS, Tianqing Liu, PhD
Division of Preventive Medicine, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, Maryland 20910

Conflict of Interest
The authors have declared no conflict of interest.

Abstract

The effect of individual predictors in a multiple regression model may be biased due to multicollinearity. Multicollinearity often occurs in longitudinal studies, especially when the objective is to study the association between disease and biomarkers. In this study, we proposed space decomposition to group potential biomarkers according their association and use the gradient-nuisance vector approach to reduce the number of biomarkers included in the high-dimension regression. The co-linearity among the biomarkers was dramatically reduced by this method. We used this approach on a US military case control data set to evaluate the association of biomarkers on risk of schizophrenia. The linear correlation among biomarkers was as high as 0.8 and after decomposition, the correlation coefficient among the vectors was less than 0.3. The predictive power of the model as a whole was not reduced. The proposed approach can help investigators to select biomarkers used to identify high risk population for diseases including schizophrenia and can be extended to other case control studies. This work was funded by the Walter Reed Army Institute of Research, Independent Laboratory In-House Research Program W8XWH-11-C-0082.

Key words: Biomarker selection, reduction; co-linearity, case control study, schizophrenia

1. Introduction

Vulnerability to mental illnesses, such as schizophrenia, bipolar disorder and anxiety disorders, has been found to be associated with genetic components. Traditional genetic studies usually search for an unknown biomarker that may cause the disease in isolated families. Development of such objective tools would help investigators to identify and stratify patients with psychiatric disorders, enable the accurate identification of schizophrenia patients early on in the disease process, improve patient outcomes and reduce healthcare costs [Wu et al., 2005]. Usually, an individual biomarker may have too small an effect size to be used as accurate classifiers. For high dimension multiple biomarkers, it is natural to find a handful of biomarkers with significant effect by chance. Detecting multiple biomarkers, each contributing only a small effect requires large sample sizes and powerful technologies that can associate genetic variations with diseases [Chakravarti, 1999]. Examining biomarkers individually could lead to a loss of valuable information.

Regression of high dimensional data is difficult at least for two reasons: sample size and collinearity. When multicollinearity exists, the coefficient estimates may change erratically in response to small changes in the model. For high dimension data, it is natural that several biomarkers are highly correlated. The high degree of multicollinearity also causes software packages to fail in performing the matrix inversion or to make the results of that inversion inaccurate.

In this study, we propose a three-step approach to identify schizophrenia cases from controls by using high dimension regression of biomarkers.

1. Decomposition of the space of X consisting of all independent variables according to their association, such that all biomarkers in any subspace are independent.
2. Find the gradient direction, which is the linear combination of the biomarkers that can best separate schizophrenia cases and controls, as well as the perpendicular vectors to the gradient in each subspace.
3. Using the biomarker general linear regression (GLR) based on the gradient direction and other significant vectors to identify cases with fewer biomarkers. For longitudinal data, the ordered GLR is used, for associations between multiple biomarkers and schizophrenia, the sensitivity and specificity in training and testing groups are studied. Our development approach was based on the application of 48 biomarkers of 294 schizophrenia cases and their associated controls.

2. Statistical method

In regression, the repressors: x_1, x_2, …, x_k are assumed to be independent. However such an assumption is usually not correct. Hence we cannot simply delete the factures, which are not significant. The multicollinerity of the regression could be solved by space decomposition. We divided the whole space into several subspaces. Among individual subspace, all biomarkers are independent. The biomarkers in different subspaces might be correlated. We checked the pair correlation among the 48 biomarkers, using a given ρ in (0,1) as threshold to decompose the space, such that: $S= S_aX\ S_bX\ S_cX$…In each sub-space, the absolute value of Pearson correlation of any pair of biomarkers is less than ρ.

Without loss of generality, let $g(y) \in R^N$ be a vector of n i.i.d. random variables which we would like to estimate g(y) to be a continuous function of y_i or categorical function of y_i. The link function should be well defined, which is noted as g(y). The independent observations are X Є R^{Nxk}, a matrix containing N independent row vectors, each of dimension k. A regression model relates g(y) to a function of **X** and **β**. For a case control study, g(y) is the logit. IfK>N, we cannot use the logistic regression. The Fisher's linear discriminate method [Johnson, 1982], can be used to find the gradient vector. Nuisance direction can find by simulation by randomly assigning case and controls. After selecting the gradient and their orthogonal vectors, we use following regression:

—— (1)

where t could be continuous or categorized into several categories. ω_i represents any combination of biomarkers, which are the gradient and the associated significant orthogonal directions.

—— (2)

We also used the proportional hazard model procedure in SAS version 9.3 (SAS Institute, Cary, NC, USA) to perform the conditional logistic model:

— (3)

The GENMOD procedure based on the binomial distribution assumption and with the logit link function for uncentered products was used to perform the sensitivity and specificity analysis. We also applied the within subject function in GENMOD, which kept all samples by their collection time order.

3. Schizophrenia biomarkers study

3.1. Data and demographic analysis

In this section, we describe the process of GNO in a multiple regression model through an application on case control data from a study of schizophrenia. Schizophrenia is a pervasive neuropsychiatric disorder of uncertain etiology. Data for US military service members who received medical discharges from the military with a diagnosis of schizophrenia from 1992 to 2005 were obtained from the Physical Disability Agencies (PDA) databases of the Army, Navy (including Marines), and Air Force [Niebuhr et al.,2011]. Those aged 18 and older who were on active duty at the time of their schizophrenia diagnosis, and who had at least one serum sample of 0.5 ml or greater in the Department of Defense Serum Repository (DoDSR) obtained prior to diagnosis were selected as potential study cases. Hospitalized cases were preferentially selected and virtually all (99%) study subjects were hospitalized with a mental disorder before their discharge from military service. The time of onset of schizophrenia was estimated as the earliest date of either the first hospitalization with mental disorder it International Classification of Disease 9th Revision (ICD-9-CM) diagnostic codes (290-319) or the date the military medical or physical evaluation boards were initiated.

Control subjects were selected from the active duty US military service population who were over the age of 18 and with no inpatient or outpatient mental disorder diagnoses. One control per case was selected for 700 males; three controls were selected for each 155 female cases. All control subjects were matched to their cases on sex, race, branch of military service, date of birth (+/-12 months), and military enlistment (+/-12 months).

Serum specimens, stored at -30°F, were obtained from the DoDSR. At least one, and up to four, matched (+/-90 days) specimens were selected for each study subject. The time of specimen collected for controls were selected by their matched cases. If the number of specimen is different for the matched case and controls, simulation and weight assignment were used to balance the data without increasing the sample power. All laboratory measurements were performed using immunological techniques. The first part of the analysis comprised Enzyme-linked immunosorbent assay (ELISA) measurement of antibodies to pre-selected infectious agents, including Toxoplasmosis *T. gondii*, Cytomegalovirus HHV-6, vaccinia and measles, and antibodies to the food borne antigens, casein and gliadin [Niebuhr et al. (2011]. Due to the cost, in the second stage of analysis, we selected subset individuals for further testing: all serum samples for 18 plates of 296 cases with their associated controls were examined in three groups, the group1 (plates 1-6), group2 (plates 7-12) and plates3 (13-18). The demonstration analysis showed that the distribution is heterogeneous. Group 1 had more female, slightly higher white and more younger subjects: (18 females, 59 younger than 25 and 49 whites), slightly higher compound to group 2 and 3.

Total 48 biomarkers were measured in all three groups. Multiple imputation approach was performed for the undetectable values, with the range of 0 to minimum detectable value for each biomarker and for the missing value with serum quantity insufficiently not enough ranged from 0 to the maximum. To avoid measurement disappearances in magnitude, all biomarkers were standardized.

3.2 Decomposition and biomarker grouping

We checked the Pearson correlation coefficients of all biomarkers, and found several pairs of them were highly correlated. For example, the correlation coefficients between Luteinizing Hormone (LH) and Apolipoprotein C-I (Apo C-I) was ρ=0.87, the the correlation coefficients between Apolipoprotein A-II (Apo A-II) and Apolipoprotein H (Apo H) was ρ=0.68. We use ρ=0.4 as the cutoff point. All 48 biomarkers were separated into three subspaces $S= S_a X\ S_b X\ S_c$. In each subspace, the Pearson correlation coefficients between any pair of biomarkers were less than 0.4.

3.3 Biomarker Selection and gradients

The biomarker selection phase of the present study was aimed at identification of molecules that were altered reproducibly in schizophrenia patients compared with their matched control. As we discussed in the method section, the gradient and their associated vectors were used to select the biomarkers backward, which treated all biomarkers equally and simultaneously. We removed the biomarkers, which had the smallest absolute coefficient on the gradient and other significant vector, if found, to choose which biomarkers should be removed from the space. Hence we did not find any other vector, except for the gradient .Figure 1 shows the sensitivity for the largest Space A.

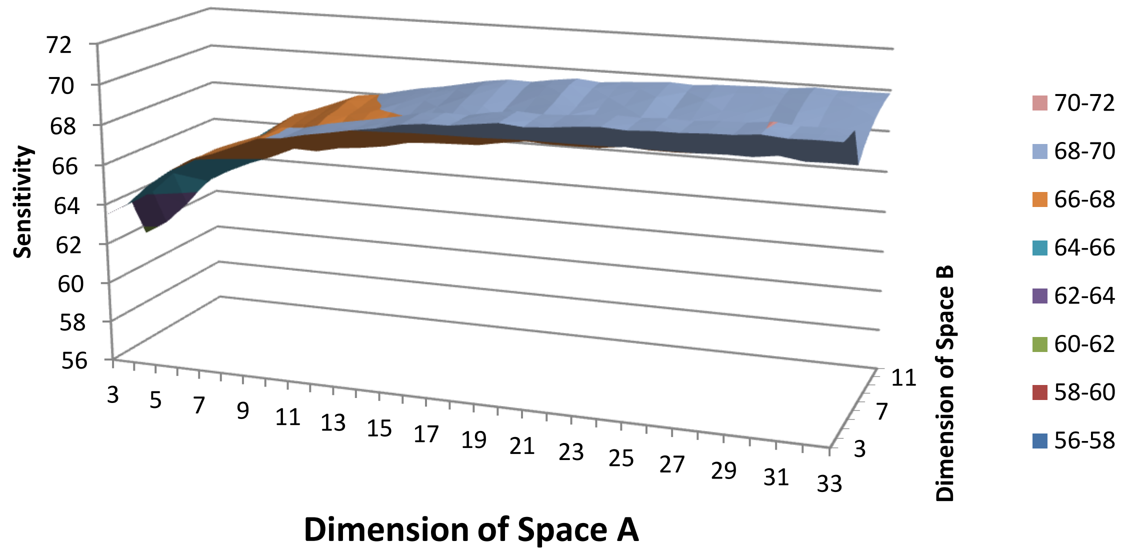

Figure 1: Average Sensitivity of the Training Group

From the above analyses, we can roughly estimate the number of the biomarkers from each space. First, Space C did not have much contribution; the associated p value of the gradient was greater than 0.4, hence all biomarkers in Space C were deleted. For both Space A and Space B, only the gradient vector showed high significance. From the graphs, we can see for Space A, when the number of biomarkers was less than 15, the sensitivity increased as the number of biomarkers increased, then the surface was almost flat for the training set, while for the testing set, 11 biomarkers sufficient. For Space B, the surface did not change much and the change was not monotonic. Selecting 4 to 6 biomarkers was good sufficient.

3.3 The joint effect and sensitivity

In this section we use the selected biomarkers to identify schizophrenia cases from controls. A perfect separation of two classes by a hyperplane is usually not possible and the association between schizophrenia status and biomarkers is complex. Using decomposition, more than one hyperplane will be used to separate the schizophrenia and controls. No matter how many hyperplanes are used, we cannot avoid mis-classification errors. The selected biomarkers may vary from study to study; hence find; a robust approach is a challenge. In order to avoid selection bias and to get more generalization conclusions, we randomly divided the cases into 10 groups. Using 9 of them and their associated controls as training and the one remaining as testing we concluded the average sensitivities and standard deviations for both training and testing groups. The sensitivity and specificity were correlated; if the specificity is higher, the sensitivity will be lower. We assumed the specificity was 0.5, and then the sensitivity in each of these 10 scenarios was calculated for both training and testing data using the same selected biomarkers and model derived from the training data. Table 1 shows the average sensitivity from the 10-fold selections.

Table 1: The sensitivity for the selected biomarkers by collection time of serum specimen to schizophrenia diagnosis and dimension of space A and B

Time	Dimension of A		Dimension of B		
		study	5	6	7
1 year before diagnosis	11	Training	72.05	72.28	72.00
			0.59	0.58	0.61
		Testing	64.17	66.27	66.28
			5.11	4.77	4.78
	12	Training	73.14	73.28	73.14
			0.58	0.58	0.54
		Testing	64.13	66.00	66.55
			4.93	4.61	4.77
	13	Training	73.29	73.57	73.13
			0.59	0.64	0.51
		Testing	65.90	65.92	65.92
			4.86	4.56	4.70
	14	Training	72.96	73.95	73.37
			0.45	0.55	0.50
		Testing	65.81	68.24	68.62
			4.13	4.70	4.93
Near diagnosis	11	Training	64.52	65.16	65.29
			0.49	0.44	0.44
		Testing	55.43	56.87	56.85
			3.53	3.79	3.72
	12	Training	64.89	65.49	65.67
			0.48	0.43	0.50
		Testing	56.50	58.10	57.39
			3.46	3.51	3.44
	13	Training	64.34	65.13	65.38
			0.48	0.47	0.51
		Testing	56.33	57.80	57.63
			3.50	3.56	3.31
	14	Training	64.14	65.06	65.15
			0.44	0.44	0.49
		Testing	56.67	58.25	58.00
			3.22	3.35	3.20

4. Conclusions

1. The selection of biomarkers is robust
2. The difference between training set and test is not large
3. The sensitivity for the training group is stable. If the sample size is large, we can expect more reliable biomarker selection
4. The variation for the testing group is not smaller, because the sample size for testing group is just about 30.
5. Since the specificity was set at 50%, and if sensitivity =70%, the odds ratio =2.33, vectors counts and 95%, CI.
6. Comparing with the results by Classification and regression trees (CART) two are comparable. Our approach is more robust on the number of biomarkers selection, and the meaning is clear and easy to explain.

The final decomposition regression model important biomarkers included the following:

From space A
Prolactin (PRL)
Interleukin-6 receptor (IL-6r)
Carcinoembryonic Antigen (CEA)
Follicle-Stimulating Hormone (FSH)
Brain-Derived Neurotrophic Factor (BDNF)
Cancer Antigen 125 (CA-125)
Prostatic Acid Phosphatase (PAP)
Immunoglobulin M (IGM)
Macrophage Migration Inhibitory Factor (MIF)
Epidermal Growth Factor Receptor (EGFR)

From Space B
Apolipoprotein H (Apo H)
Serotransferrin (Transferrin)
Sortilin
Haptoglobin
Immunoglobulin A (IgA)
Connective Tissue Growth Factor (CTGF)

5. Discussion

We have proposed a new approach to decompose the sample space X into two spaces for high dimension data regression. Our main contribution is using the gradient vector and the non-significant vector for the dependent variable to construct an orthogonal base repeatedly, and to construct a subspace with fewer biomarkers and fewer vectors. This approach can handle both categorical and continuous dependent variables by choosing different types of models. It can be applied to the multivariable response g(y) and longitudinal data. It offers a tool to study biomarker-biomarker interactions. The results from actual data and from simulation demonstrate that the proposed approach is robust and stable for the purpose of reducing the dimension of data.

Acknowledgements

The authors are grateful for Ms. Janice K. Gary, Medical Standards Analysis & Research Activity (AMSARA) staff for administrative support in preparation of the manuscripts. This work was funded by the Stanley Medical Research Institute (SMRI), Bethesda, Maryland, the ILIR (Independent Laboratory In-House Research), Department Army, W81XWH-11-C-0082 and the Department of the Army.

References

1. Wu EQ, Birnbaum HG, Shi LZ, Ball DE, Kessler RC, Moulis M, et al. The economic burden of schizophrenia in the United States in 2002. *J Clin Psychiat.* 2005;66(9):1122-9.
2. Chakravarti A. Population genetics--making sense out of sequence. *Nature genetics*. 1999;21(1 Suppl):56-60. Epub 1999/01/23.
3. Johnson, Richard A and Wichern, Dean W. Applied Multivariate Statistical Analysis, Prentice-Hall, Inc. NJ 1982.
4. Niebuhr DW, Li Y, Cowan DN, Weber NS, Fisher JA, Ford GM, et al. Association between bovine casein antibody and new onset schizophrenia among US military personnel. *Schizophr Res*. 2011;128(1-3):51-5. doi:10.1016/j.schres.2011.02.005.