ITALIAN JOURNAL OF PUBLIC HEALTH

High dimensional regression on serum analytes

Yuanzhang Li[1], Emanuel Schwarz[2], Sabine Bahn[2], Robert Yolken[3], David W Niebuhr[1]

ABSTRACT

Regression of high dimensional data is particularly difficult when the number of observations is limited. Principal Component Analysis, canonical correlation analysis and factor analysis are commonly used methods to reduce data dimensions, but usually cannot find the most significant linear combination. The goal is usually to find a particular partition of the space X consisting of all independent factors. In this paper, we propose an approach to high dimensional regression for applications where N>K or N<K, where N is the sample size, k is the dimension of space X. The approach starts by finding the most significant linear combination and one of the most insignificant directions to decompose the sample space into two subspaces and reduce the dimension. Further, we examine the contributions of individual variables to those most significant vectors by the coefficients of the combinations to reduce the total number of variables in the selected space without losing the power of the prediction. We use the proposed approach to determine the potential association of 51 serum analytes with schizophrenia using data derived from a case control study (n=208). Numerical results demonstrate that the proposed approach can significantly improve dimension reduction.

Key words: Gradient; High dimensional regression; Schizophrenia

(1) Department of Epidemiology, Walter Reed Army Institute of Research, Silver Spring, Maryland (USA)
(2) Institute of Biotechnology, University of Cambridge, Cambridge (UK)
(3) Stanley Neurovirology Laboratory, Developmental Neurobiology, Johns Hopkins University School of Medicine, Baltimore, Maryland (USA)

CORRESPONDING AUTHOR: *Yuanzhang Li; Division of Preventive Medicine, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, Maryland 20910 USA. Tel: 301-319-9728, Fax: 301-319-9104. e-mail: Yuanzhang.Li@amedd.army.mil*
DOI: ***10.2427/8672***

INTRODUCTION

The vulnerability to mental illnesses, such as schizophrenia and bipolar disorder has been found to be associated with genetic components. The traditional genetic studies usually search for an unknown gene that may cause the disease in isolated families. But it is difficult to identify common gene variants that are associated with the disease across populations. It is clear that these disorders are not caused by a single defective gene, but by the joint effects of many genes acting together with non-genetic factors (1). Research suggests that such interaction of genetic predisposition and environmental factors is common in many diseases.

Detecting multiple genes, each contributing only a small effect requires large sample sizes and powerful technologies that can associate genetic variations with diseases (2). Examining genes individually could lead to a loss of valuable information. To find the genes with a relatively larger effect, high dimensional regression studies could be used under an assumption that a specific group of genes may cause the disease. Unfortunately, it is difficult and costly to have a large sample size in studies. Regression of high dimensional data is particularly difficult when the size of the data is limited. Traditional regression methods that use the sample covariance perform poorly in this situation (3). There are no generally accepted methods for relating the number of observations versus the number of independent variables in the model. In many cases, when we study epidemiological data with biomarkers, the sample number (N) might be less than the data dimension K. Then the sample covariance is singular. Even when N>K in large-scale data mining, predictive modeling, and especially for multivariate regression exercises with a large number of possible explanatory/predictive variables, variable selection and dimension reduction is a major task.

A common method in regression analysis for dimension reduction is the stepwise regression. One of the major limitations of the algorithm is that many variables used as independent variables in a regression may have a high degree of correlation. When several of the predictive variables are highly correlated, it is difficult to distinguish their effects on the dependant variable. Therefore the estimation and the test of statistical significance are not reliable, and the assumption of independence for these tests is violated. The parameter estimates in a regression equation may change with a slight change in the data and, thus, are not stable for predicting the future estimation. In the past, regression methods that adopt regularization have been introduced, such as ridge regression (4), subset selection, and principal component analysis (PCA). Recently, there has been an increasing interest in replacing the sample covariance with some sparse estimates of the true covariance or its inverse for high dimensional regression problems (5).

PCA is a traditional statistical method commonly used to reduce the number of predictive variables and solve the multi-colinearity problem (6). PCA looks for a few linear combinations of the variables that can be used to summarize the data without losing too much information. Since PCA is a non-supervised method, it does not use information of the dependent variable for the construction of such linear combinations. Therefore, the first principal component is often not the linear combination of the input variables that is most significantly associated with the dependant variable, e.g. disease state. In this study, we propose a component decomposition of the space of X consisting of all independent variables according to their association with the dependent variable, g(y). We will now formally define and describe gradient-noise-orthogonal (GNO) and their orthogonal components and show how they can be derived.

THEORY AND APPROACH

The Gradient-noise-orthogonal base

Without loss of generality, let $g(y) \in R^N$ be a vector of n i.i.d. random variables observed, the link function g(y) could be a continuous function of y_i or categorical function of y_i. The independent observations are $X \in R^{NxK}$, a matrix containing N independent row vectors, each of dimension K. A regression model relates g(y) to a function of $\mathbf{X}$ and $\boldsymbol{\beta}$.

$$\mathbf{g(Y)} \approx \boldsymbol{f}\mathbf{(X,\beta)} \quad [1]$$

The approximation is usually formalized as

$$\mathbf{E(g(Y) \mid X) = f(X, \beta)} \quad [2]$$

To carry out regression analysis, the form of the function f must be specified. In order to perform a regression analysis the user must provide information about the dependent variable g(y): If N data points of the form $(\mathbf{g}(Y),\mathbf{X})$ are observed, where $N < K$, most classical approaches to regression analysis cannot be performed, as there is not enough data to recover $\boldsymbol{\beta}$. If exactly $N = K$ data points are observed, and the function f is linear, the equations $g(y) = f(\mathbf{X}, \boldsymbol{\beta})$ can be solved exactly.

The most common situation is $N > K$. In this case, there is enough information in the data to estimate a unique value for $\boldsymbol{\beta}$ that best fits the data in some sense, and the regression model, when applied to the data, can be viewed as an over determined system $\boldsymbol{\beta}$. However K should not be too large and the pair-wise correlations should not be too high. Unfortunately, both of these scenarios often occur in microarray data analysis. Correlations among groups of analytes

sharing the same biological "pathways", can be high (7). The ideal gene selection method should be able to achieve two objectives: eliminate the trivial genes, and automatically include whole groups of correlated predictors into the model once one gene amongst them is selected. The basic idea is to decompose the Space X into two parts: U and V. The vectors in Space U are highly associated with g(y), and the vectors in V have almost no association with g(y). Ideally, the dimension of U would be much smaller than the sample size, the dimension of Space R. All the vectors are linear or non-linear combinations of X. Most software will not tallow records with missing value; hence imputation might be needed for the missing value.

First we need to find the most significant direction on g(y), which is called the gradient. For case control studies, we use Fisher's linear discriminant analysis (LDA) (8) to find the gradient vector. In such a case, y=1 is for cases and y=0 is for controls. LDA approaches the problem by assuming that the conditional probability density functions of **X** for y=1 and y=0 are both normal probability density functions of **X** for y=1 and y=0 are both normal.

With mean and covariance parameters: (μ_1, Σ_1) and (μ_0, Σ_0) respectively. Under this assumption, it is well known that the Bayes optimal solution to predict a subject as being from the case: if the ratio of the log-likelihoods is below some threshold T is as follows:

$$(X-\mu_0)^T\Sigma_0^{-1}(X-\mu_0)+ln|\Sigma_0|-(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)-ln|\Sigma_1|<T$$

[3]

If we assume that the class covariance are identical: $\Sigma_0=\Sigma_1=\Sigma$ with full rank, we will have the solution for maximization of the square distance of the two groups by $\omega \cdot X < c$, where $\omega=\Sigma^{-1}(\mu_1-\mu_0)$.

It is obvious that ω is the most significant direction to distinguish cases and controls, denoted as ω_1, which belongs to U, denoted as $U_{K,0}$. Next we find one of the insignificant directions by simulation, which are called noise. For example, for case control study, we randomly assign Y value (0 or 1) to the subjects, and use equation 3 to get the vector ω_2, which belongs to V. In order to find the direction ω_3, satisfying $\omega_3 \perp \omega_1$, $\omega_3 \perp \omega_2$, we minimize the sum of the absolute inner products with ω_1 and ω_2.

$$\text{Min}\{|(\omega_3 \cdot \omega_1)|+|(\omega_3 \cdot \omega_2)|;\ \omega_3 \text{ in X}\}$$

[4]

If the association between the g(y) and ω_3 is significance, put it into $U_{K,0}$, otherwise put into V. Repeat the same process, we get the remain vectors, in general, we select ω_j in X, it satisfies.

$$\text{Minimize}_{\omega j}\{|(\omega_j \cdot \omega_1)|+|(\omega_j \cdot \omega_2)|+\ldots+|(\omega_{j-1} \cdot \omega_2)|;\ \omega_j \text{ in X}\}\ j=3,4,\ldots K;$$

[5]

It can be seen that ω_j is perpendicular to $U_{K,0}$ and V in the step j-1. If the association between the g(y) and ω_j is significance, put it into $U_{K,0}$, otherwise put into V. $U_{K,0}$ consists of a few vectors, which depends on the selected significance level. If we re-rank the vectors in $U_{K,0}$ by the association significance on g(y), the gradient vector would be labeled number 1.

THE GNO SEQUENCE

After $U_{K,0}$ is selected, we examine the coefficients of the linear combinations of the vectors in $U_{K,0}$, removing the variable or variables (gene or analyte) with lower value coefficients for all the vectors in $U_{K,0}$ to get the space U_{K-1} (or U_{K-m}, if m variables are removed). U_{K-1} is a subspace of X with K-1 variables of X_i's. We decompose U_{K-1} by repeating the process in the previous Section, get $U_{K-1,0}$, a subspace of U_{K-1} with a new gradient vector and a few perpendicular vectors in U_{K-1}. Continuing the same process, we get U_{K-2} and $U_{K-2,0}$; $U_{K-3,}$ and $U_{K-3,0}$ ... until $U_{j-1}=U_j$, or stop by biological judgment, which would be the smallest subspace or an acceptable subspace U_j and $U_{j,0}$ as the final U.

Adjust p value for gradient

Since gradient vector is the most significant direction in U_k among the k linear combinations for any given k, the p value of the uni-biomarker should be adjusted. However, the aim of GNO is to remove the biomarkers, which have weak effects on the outcome, it is not necessary to find the true p value of the gradient direction in each step. For the final U, only a few biomarkers remain, the classic multivariate analyses could be performed. Obviously, the gradient vector has the smallest p value among the k orthogonal vectors in U_k, hence the adjusted p value of the gradient vector could be found by using the extreme distribution of the order statistics. Among the

p values of the k vectors in U_k, the adjusted p value is $p_a=1-(1-p)^k$. Simulation might be used to find the distribution of p value of the gradient vector, if needed.

GNO sequence properties

The gradient vector identifies the most predictive combination and the noise vector identifies one of the most insignificant directions. We can choose the insignificant direction lying on the hyper-plane with the gradient as normal, and then these two vectors will be perpendicular to each other. We can construct an orthogonal decomposition or near orthogonal decomposition.

1. All vectors are perpendicular to each other, except for the directions of the gradient and noise. However, we can choose the noise vector lying on the hyper plane with the gradient as the normal, and then all vectors are perpendicular to each other.
2. If all $x_1, x_2, \ldots, x_K$ are independent, then all vectors in GNO are independent or nearly independent with one exception.
3. The gradient vector has the strongest association with g(y).
4. If g(y) is binary, then the hyper plane with gradient as normal separates g(y) better than any other hyper-planes.
5. The sequence of the subspace $U_i \subset U_{i+1} \subset U_{i+2} \subset \ldots \subset U_{K-1} \subset U_K = X$ and $U = U_{i,0} \subset U_i$
6. Y is almost independent on V. $E(g(Y)|V)$ is almost a constant or a total random on V.
7. $E(g(Y)|U) \approx E(g(Y))$, $V(g(Y)|U) \approx V(g(Y))$.
8. If ω_1 and ω_2 are not independent, the interactions among the components of X exist.

APPLICATION AND EVALUATION

Data and Method

Schizophrenia is a pervasive neuro-psychiatric 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 databases of the Army, Navy, Marines and Air Force (9). 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 in the Department of Defense Serum Repository 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. Control subjects were selected from the active duty US military service population who had no inpatient or outpatient mental health diagnoses. 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).

All laboratory measurements were performed using immunological techniques. The first part of the analysis comprised Enzyme-linked immunosorbent assay measurement of antibodies to pre-selected infectious agents. Due to the cost, in the second stage of analysis, we selected a subset individuals for further testing: the samples for 6 plates of 86 cases (18 females, 27 older than 25 and 49 whites), with 122 perfectly matched controls. Fifty one analytes shown in Table 1 were measured in every sample. All analytes were standardized. Multiple imputation was performed for the undetectable values, with the range (0, minimum value) for each analyte. Over 80% of the data were undetectable for Interleukin-10, Interleukin-11 and Interleukin-17; therefore we combine these three standardized values for multiple analytes analysis; resulting in 48 of individuals and one of the combined Interleukins were analyzed with GNO and compared to PCA.

Using GNO to examine the association of schizophrenia and serum analytes

We performed conditional logistic analyses for all analytes separately. Gender, age and the time to diagnosis were included as strata. The only significant finding was prolactin, with an adjusted hazard ratio of 1.28 (p=0.03), for increasing two standard deviations before diagnosis. Adjusting by the extreme distribution, the type I error for finding the highest significant value of 0.03 among 51 analytes was actually 83%. Table 2 shows the adjusted hazard ratio for the 10 most significant analytes. For the other analytes, the p value was greater than 0.5.

Using the GNO method, Table 3 shows the HR for the gradient and other six other vectors

TABLE 1

ANALYTES MEASURED USING MULTIPLEXED IMMUNOASSAYS			
va1	Alpha-1-Antitrypsin (AAT)	va27	Vascular Endothelial Growth Factor (VEGF)
va2	Apolipoprotein A-I (Apo A-I)	va28	Vitronectin
va3	Apolipoprotein A-II (Apo A-II)	va29	Interleukin-6 receptor (IL-6r)
va4	Apolipoprotein B (Apo B)	va30	Interleukin-7 (IL-7)
va5	Apolipoprotein C-I (Apo C-I)	va31	Kidney Injury Molecule-1 (KIM-1)
va6	Apolipoprotein H (Apo H)	va32	Luteinizing Hormone (LH)
va7	Beta-2-Microglobulin (B2M)	va33	Monocyte Chemotactic Protein 2 (M CP-2)
va8	Brain-Derived Neurotrophic Factor (BDNF)	va34	Macrophage-Derived Chemokine (MDC)
va9	Serotransferrin (Transferrin)	va35	Macrophage Migration Inhibitory F actor (MIF)
va10	Complement C3 (C3)	va36	Macrophage Inflammatory Protein-1
va11	Cancer Antigen 125 (CA-125)		alpha (MIP-1 alpha)
va12	Calbindin	va37	Matrix Metalloproteinase-2 (MMP-2)
va13	CD5 (CD5L)	va38	Prostatic Acid Phosphatase (PAP)
va14	Carcinoembryonic Antigen (CEA)	va39	Prolactin (PRL)
va15	Cortisol (Cortisol)	va40	Peptide YY (PYY)
va16	Connective Tissue Growth Factor (CTGF)	va41	Serum Amyloid P-Component (SAP)
va17	Epidermal Growth Factor Receptor (EGFR)	va42	Sortilin
va18	Endothelin-1 (ET-1)	va43	Testosterone, Total
va19	Fetuin-A	va44	Thrombopoietin
va20	Ferritin (FRTN)	va45	Tissue Inhibitor of Metalloprotei nases 1 (TIMP-1)
va21	Follicle-Stimulating Hormone (FSH)	va46	Tumor Necrosis Factor Receptor-Li ke 2 (TNFR2)
va22	Haptoglobin	va47	TNF-Related Apoptosis-Inducing
va23	Intercellular Adhesion Molecule 1 (ICAM-1)		Li gand Receptor 3 (TRAIL-R3)
va24	Immunoglobulin A (IgA)	va48	Betacellulin (BTC)
va25	Immunoglobulin M (IGM)	va49	Interleukin-10 (IL-10)
va26	Thyroid-Stimulating Hormone (TSH)	va50	Interleukin-11 (IL-11)
		va51	Interleukin-17 (IL-17)

TABLE 2

THE HAZARD RATIOS FOR BEING A SCHIZOPHRENIA CASE OF THE 10 MOST SIGNIFICANT ANALYTES				
ANALYTES	HR	95% CI		P
39	1.28	1.02	1.62	0.03
7	1.34	0.96	1.88	0.09
13	0.45	0.17	1.17	0.1
23	1.25	0.91	1.71	0.17
19	0.8	0.52	1.21	0.28
34	1.17	0.87	1.58	0.3
24	0.87	0.62	1.22	0.42
36	0.82	0.47	1.4	0.46
40	1.13	0.81	1.57	0.46
9	0.87	0.59	1.27	0.47

with relative smaller p values. For the gradient alone, Table 3 shows that the risk to being a case was more than doubled (HR=2.4) when increasing 2 standard deviations in the gradient direction. The p value was 2.33E-07 and after adjusting by the extreme distribution of the 49 vectors, the adjusted p value was 1.1E-5. The goodness of fit (-2log likelihood ratio) for the gradient vector was 1989.1 compared with the total goodness of fit of 2023.8, indicating that the gradient vector had a high contribution to the fit of the model.

Table 4 shows the coefficients of the gradient vector. Many of them are lower than others in magnitude and the smallest value was 0.002, indicating a weak association with schizophrenia status. We repeated the steps 4-12 to remove analytes with weak effects on schizophrenia status. Table 5 shows the gradient vectors of the three final spaces: U_{11},U_{10} and U_9. In U_{11}, analyte 17 had the smallest coefficient in the gradient vector in U_{11}, which was much smaller than that of other analytes. In $U_{10,}$ analyte 10 has the smallest contribution, but its coefficient was close to those of analytes 44 and 8.

Table 6 shows the two most significant effects in U_{11},U_{10} and U_9. In U_{11}, the only significant vector was the gradient, but in U_{10}, except for the gradient, vector 4 was almost significant (p=0.05), and in U_9, vector 4 was significant. This means $U_{11,0}$ could have only one vector, but $U_{10,0}$ and $U_{9,0}$ should have two vectors. Therefore we could use U_{11} as the final selected subspace of U.

For the new gradient alone, Table 7 shows that the risk to be a schizophrenia case was doubled (HR=2.1) with a p value of 8.06E-06. Considering the extreme distribution among the 11 vectors, the adjusted p value was 8.9E-5, which is a little larger than the gradient using 51 analytes. Except for the new gradient,

TABLE 3

THE HR OF BEING A SCHIZOPHRENIA CASE AND 95% CI FOR THE GRADIENT AND OTHER SIX MOST SIGNIFICANT VECTORS IN X

VECTORS	HR	95% CI		P VALUE
GRADIENT	2.40	1.72	3.34	2.33E-07
32	1.32	0.93	1.89	0.12
46	1.28	0.92	1.76	0.14
25	1.20	0.88	1.63	0.25
22	1.18	0.88	1.58	0.28
49	0.83	0.60	1.16	0.29
29	1.17	0.87	1.55	0.29

TABLE 4

THE ANALYTE COEFFICIENTS OF THE GRADIENT VECTOR IN SPACE OF X

Va1	Va2	Va3	Va4	Va5	Va6	Va7	Va8	Va9	Va10
-0.015	-0.033	-0.032	-0.018	-0.022	0.140	0.319	-0.153	-0.298	0.166
Va11	Va12	Va13	Va14	Va15	Va16	Va17	Va18	Va19	Va20
-0.014	-0.012	-0.073	0.092	0.064	0.072	-0.103	0.019	-0.053	0.048
Va21	Va22	Va23	Va24	Va25	Va26	Va27	Va28	Va29	Va30
0.078	-0.021	0.080	-0.095	0.087	0.035	0.022	-0.019	0.038	0.040
Va31	Va32	Va33	Va34	Va35	Va36	Va37	Va38	Va39	Va40
-0.022	0.062	-0.079	0.043	0.051	-0.034	-0.018	-0.017	0.158	0.008
Va41	Va42	Va43	Va44	Va45	Va46	Va47	Va48	Va49	
-0.054	0.057	0.069	-0.129	0.463	-0.592	0.052	-0.156	-0.002	

no other vectors showed significant effect on schizophrenia risk. Table 8 shows the results of interactions in U_{11}. There are no significant interactions between gradients and other vectors. Adding any other vectors had only minor effect or the gradient vector except for Vector 4.

The selected subspace U would be U_{11} or U_7 with two vectors: gradient and Vector 5 (data not shown). The individual effect of Vector 4 Table 6 was positively associated with schizophrenia. If both gradient and Vector 4 are included in the model Table 8, the adjusted

TABLE 5

GRADIENT VECTOR IN U_{11}, U_{10} AND U_9

Space	va6	va7	va8	va9	va10	va17	va39	va44	va45	va46	va48
U_{11}	0.177	0.288	-0.119	-0.403	0.124	-0.066	0.193	0.151	0.549	-0.551	-0.156
U_{10}	0.183	0.291	-0.128	-0.414	0.106		0.197	0.127	0.543	-0.555	-0.159
U_9	0.186	0.293	-0.121	-0.346			0.208	0.150	0.578	-0.562	-0.169

TABLE 6

THE TWO MOST SIGNIFICANT EFFECTS OF VECTORS IN U_{11}, U_{10} AND U_9

SPACE	VECTOR	PARAMETER	STANDARD ERROR	P VALUE
U_{11}	Gradient	0.74	0.17	8.06E-06
	V9	-0.23	0.17	0.17
U_{10}	Gradient	0.75	0.17	8.06E-06
	V4	0.25	0.13	0.05
U_9	Gradient	0.74	0.16	5.89-06
	V4	0.28	0.13	0.04

TABLE 7

THE HR AND 95% CI FOR GRADIENT AND 10 OTHER VECTORS IN U_{11}

VECTORS	HR	95% CI		P VALUE
GRADIENT	2.09	1.51	2.89	8.06E-06
9	0.79	0.57	1.11	0.17
4	1.20	0.92	1.55	0.18
6	0.82	0.58	1.15	0.25
11	0.88	0.62	1.24	0.47
3	1.10	0.79	1.54	0.56
8	0.95	0.69	1.32	0.77
2	1.06	0.68	1.67	0.79
10	1.02	0.75	1.39	0.91
5	0.99	0.71	1.39	0.96
7	1.00	0.74	1.37	0.98

effect of Vector 4 is negatively associated with schizophrenia and the adjusted gradient effect is increased. This phenomenon suggests that a non-linear combination of the individual analytes may be better.

Results by using PCA

The PCA is obtained by Eigen-value decomposition of the covariance or correlation matrix of the predictive variables under

TABLE 8

THE RISK EFFECT BY GRADIENT AND OTHER ORTHOGONAL VECTORS IN U_{11} WITH INTERACTION TERA

MODEL	VECTOR	PARAMETER	STD ERROR	P VALUE	HR
1	gradient	0.73	0.17	<.0001	2.08
	V9	-0.05	0.16	0.76	0.95
	gradient*grad9	0.08	0.30	0.79	.
2	gradient	0.73	0.17	<.0001	2.07
	V6	-0.16	0.20	0.43	0.86
	gradient*V6	0.01	0.13	0.93	.
3	gradient	0.87	0.22	<.0001	2.38
	V4	-0.02	0.19	0.93	0.98
	gradient*V4	-0.08	0.13	0.54	.
4	gradient	0.81	0.22	0.00	2.39
	V6	-0.15	0.14	0.29	0.88
	V4	-0.10	0.15	0.49	0.87
	V9	-0.01	0.17	0.95	0.97

TABLE 9

THE HR OF BEING SCHIZOPHRENIA CASES AND 95% CI FOR PRINCIPAL 1 AND TWELVE MOST SIGNIFICANT PRINCIPALS

PRINCIPALS	HR	95% CI		P VALUE
1	1.06	0.76	1.47	0.75
48	0.61	0.44	0.85	0.003
13	1.34	0.94	1.91	0.10
24	1.27	0.91	1.76	0.15
36	0.80	0.58	1.10	0.17
14	1.28	0.89	1.82	0.18
5	0.78	0.54	1.12	0.18
41	0.84	0.63	1.13	0.25
10	0.83	0.59	1.17	0.29
38	0.84	0.61	1.16	0.29
27	1.18	0.86	1.64	0.31
15	1.16	0.84	1.61	0.36
47	0.86	0.63	1.19	0.38

consideration. Most statistical software can compute the principal components. Using PCA, Table 9 shows the HR for principal 1 and 12 other most significant principals ($|\ln(HR)|=0.18$ or above). Ranking the principal HRs by their p values resulted in a rank of 32 for the 1st principal. The most significant principal was the 48th principal (p=0.003), which is more significant than that of the individual factor Prolactin, but is also lower in significance compared to the gradient. Adjusting by the extreme distribution of 49 principals, the type 1 error was approximately 14%.

To assess the model fit with the log likelihood ratio, we need to include at least the 24 most significant principals into the model to reach the fitting level of modeling gradient alone (1989.1) in the initial space U_{49}. The associated –2log likelihood was 1989.5 for the 24 most significant principals and 1988.3 for the 25 most significant principals. The total likelihood ratio was 2023.8. The GNO method can reduce the dimension significantly more than PCA. There is no iteration process when using PCA, as that in GNO

The sensitivity analysis of the selection of gradient vector

The selection of the gradient vector is essential for the GNO. In order to study the sensitivity of the selection, we randomly select 76 cases among the 86 cases and the matched controls, to construct a sub-sample. Then we find the gradients for this sub-sample in U49 and U48. We repeat this process 500 times to get 500 gradients in U49 and U48. For each gradient, we first rank the coefficients of the individual analytes by their coefficients from

TABLE 10

SENSITIVITY BY 500 RANDOM SELECTED SAMPLES

RANK TYPE		RANK BY ACTUAL VALUE WITH LARGEST ABSOLUTE COEFFICIENTS						RANK BY ABSOLUTE VALUE			
		THE SMALLEST (NEGATIVE)				THE LARGEST (POSITIVE)		THE SMALLEST (NEAR ZERO)			
ANALYTES		VA46	VA9	VA48	VA24	VA7	VA45	VA49	VA38	VA11	VA18
U49	Original[a]	49	48	47	46	1	2	49	44	46	41
	mean[b]	49(100%)	47.9(91%)	46.8(77%)	46.2(93%)	1.2(81%)	1.8(99.6%)	42.5	44.0	42.1	40.9
	std	0	0.4	0.6	0.6	0.4	0.4	4.2	4.1	3.3	3.7
	min	49	46	45	44	1	1	31	29	33	32
	Q1	49	48	47	46	1	2	40	41	40	38
	median	49	48	47	46	1	2	43	45	42	40
	Q3	49	48	47	47	1	2	45	47.5	45	43
	max	49	48	48	48	2	5	49	49	49	49
ANALYTES		VA46	VA9	VA48	VA8	VA45	VA7	VA11	VA18	VA38	VA40
U48	mean	48.0(99%)	46.8(88%)	44.6(57% ≥ rank45)	44.3(75% ≥ rank44)	1.2(87%)	1.9(99%)	40.6	39.2	37.0	36.8
	std	0.1	0.7	1.3	1.8	0.6	0.4	5.7	6.6	8.3	8.3
	min	47	41	40	36	1	1	17	20	12	13
	Q1	48	47	44	43	1	2	37	35	32	30
	median	48	47	45	45	1	2	42	41	39	39
	Q3	48	47	46	46	1	2	45	44	44	44
	max	48	48	47	47	8	3	48	48	48	48

[a]The original rank of the analytes in the original sample.

[b]Mean of the rank and the percentage of the analyte taking extreme ranking.

largest to smallest, both largest (positive) and smallest (negative) have the larger contribution in the gradient vector. The six analytes with largest effects in the gradients from the original sample are shown in Table 10. We list the mean, standard deviation, and the 5 number summaries of the 500 simulations. Following the mean of the rank, the number in the parenthesis is the percentage of the rank or ranks with the extreme ranking: largest or smallest.

For example in U49, the analyte Va46 has the original rank 49, the most negative contribution in the original sample gradient, the mean rank of the 500 simulations is 49 and the standard deviation is 0. This implies that Va49 has the most negative value in all 500 simulations. The percentage of Va46 with the 49th rank is 100%. The analyte Va48 has the original rank 47, the 3rd most negative one in the original sample. Among the 500 simulations, Va48 has mean rank of 46.8 and the standard deviation of 0.6; 77% of Va48 are ranked 47 and above (47and 48). The analyte Va7 has the largest positive coefficient in the original sample. Among the 500 simulations, the mean of the rank of Va7 is 1.2; 81% of them are ranked 1, and 19% are ranked 2nd. We also ranked the coefficients by their absolute value from largest to smallest, those analytes, which have minimal contribution to the gradient will have the highest ranks, 4 of them were listed in Table 10. Our GNO analysis results in the following observations:

1. For the most effective analytes, their contribution to the gradient is very stable and consistent. They are not sensitive to the sampling selection.
2. For those 4 analytes with minimum effect on the gradient, their ranks among the 500 simulations are relatively stable, which are mostly distributed in the lowest third rank ranges: 30th to 49th. It is to be expected that those analytes with smallest contributions have variations in their ranks.
3. Similar observations are found in U48, which eliminates the Va49 with the smallest contribution to the gradient in the original sample.
4. Those with highest contributions to the gradient in U49 still have the highest contributions in U48, except one analyte, Va24 was replaced by Va8. Va24 is the 6th highest in U49, but it is 7th highest in U48.
5. Those with the smallest absolute coefficients in U49 still have the smallest coefficients in U48, with minor rank changes.

We also used simulation to select the samples of different sizes (60 to 75) to examine the robustness of GNO, which gave similar results. For example, using simulation with selecting 70 cases 500 times, the mean HR of the gradient was 2.42 and the range of the 500 simulations was (2.23, 2.64), comparing with 2.40 for the original data.). Summarizing above, the GNO is robust and stable in the process of selection of gradients.

DISCUSSION

In this study, we proposed the GNO decomposition method for high dimensional regression. For high dimensional data, we chose the vectors of decomposition one by one according to their effect on the dependent variable. Since all vectors are perpendicular to each other, the correlation among them is usually weak. The interaction among them on the dependent variable is usually low. Therefore GNO reduces multi-colinearity, which often occurs in the regression analysis.

We examined the coefficients of the gradient vector and other vectors in U_j sequence to remove individual analytes with weak effect on outcomes. The proposed method, can find the effective analytes as well as the analyte-analyte interactions. It reduces both the dimensions of the regression and the total number of the individual analytes needed to be tested.

It is very important that the simulation results show that the selection of gradient is stable.

Compared to the popularly used PCA as a dimension reduction technique, the gradient in GNO method is the most significant vector, while the 1st principal is not and the most significant component usually is not same as anyone of the Eigen vectors. Therefore, the dimension of space U would be much less with GNO than PCA.

Many of the analytes identified as associated with schizophrenia care states a have been implicated previously in acute or chronic inflammatory conditions, endothelial cell dysfunction, cardiovascular disease, type II diabetes mellitus and metabolic disorder (10-12). Interestingly inflammatory disorders and schizophrenia share an increased prevalence of insulin resistance, metabolic syndrome and type II diabetes (13-16). Using the proposed

method, could find analytes that combined together are associated with schizophrenia care status in the general population.

The associations we found between analytes were complex, requiring multiple samplings for identification. This highlights the importance of performing longitudinal investigations of biologically relevant markers in complex brain disorders such as schizophrenia. This finding is consistent with the growing body of literature indicating that schizophrenia is likely to involve multiple etiologies and biological pathways. For example, recent genome wide association studies have identified a number of genes which confer an increased risk of schizophrenia; however, no single gene can explain more than a small number of cases (17).

Theoretically GNO is an algorithm to approximate the central dimension reduction subspace with the sufficient dimension reduction (18). There are many existing methods for dimension reduction, which are better than PCA, such as sliced inverse regression and sliced average variance estimation (19), likelihood-based sufficient dimension reduction (20), estimating the central subspace based on the inverse third moment (21), estimating the central solution space (22). We will use those approaches to analyze the same data in the future and compare to GNO.

ACKNOWLEDGEMENTS: *the authors are grateful for Dr. Michael Spain, Austin, Texas, for performing the analyses assay Ms. Janice K. Gary Walter Reed Army Institute of Research, Division of Preventive Medicine staff for administrative support in preparation of the manuscripts. This work was funded by the Stanley Medical Research Institute (SMRI), Bethesda, Maryland, and the Department of the Army.*

FUNDING: *this work was funded by the Stanley Medical Research Institute (SMRI), Bethesda, Maryland, and the Department of the Army.*

DISCLAIMER: *the views expressed are those of the authors and should not be construed to represent the positions of the Department of the Army or Department of Defense. None of the authors have any associations, financial or otherwise, that may present a conflict of interest.*

References

(1) Risch N, Spiker D, Lotspeich L, et al. A genomic screen of autism: evidence for a multilocus etiology. Am J Hum Genet 1999; 65(2): 493-507

(2) Chakravarti A. Population genetics--making sense out of sequence. Nat Genet 1999; 21(1 Suppl): 56-60

(3) Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 1996; 58(1): 267-88

(4) Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2001

(5) Witten DM, Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. J Roy Stat Soc B 2009; 71: 615-36

(6) Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Stat Assoc 2006; 101(473): 119-37

(7) Segal MR, Dahlquist KD, Conklin BR. Regression approaches for microarray data analysis. J Comput Biol 2003; 10(6): 961-80

(8) Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, New Jersey: Prentice Hall Inc, 1982

(9) Niebuhr DW, Li Y, Cowan DN, et al. Association between bovine casein antibody and new onset schizophrenia among US military personnel. Schizophr Res 2011; 128(1-3): 51-5

(10) Fessel WJ, Solomon GF. Psychosis and systemic lupus erythematosus: a review of the literature and case reports. Calif Med 1960; 92: 266-70

(11) Goldberg RB. Cytokine and cytokine-like inflammation markers, endothelial dysfunction, and imbalanced coagulation in development of diabetes and its complications. J Clin Endocrinol Metab 2009; 94(9): 3171-82

(12) Volp AC, Alfenas Rde C, Costa NM, et al. Inflammation biomarkers capacity in predicting the metabolic syndrome. Arq Bras Endocrinol Metabol 2008; 52(3): 537-49

(13) Chung CP, Avalos I, Oeser A, et al. High prevalence of the metabolic syndrome in patients with systemic lupus erythematosus: association with disease characteristics and cardiovascular risk factors. Ann Rheum Dis 2007; 66(2): 208-14

(14) De Hert M, van Winkel R, Van Eyck D, et al. Prevalence of diabetes, metabolic syndrome and metabolic abnormalities in schizophrenia over the course of the illness: a cross-sectional study. Clin

Pract Epidemiol Ment Health 2006; 2: 14

(15) Wajed J, Ahmad Y, Durrington PN, Bruce IN. Prevention of cardiovascular disease in systemic lupus erythematosus - proposed guidelines for risk factor management. Rheumatology 2004; 43(1): 7-12

(16) Shoelson SE, Lee J, Goldfine AB. Inflammation and insulin resistance. J Clin Invest 2006; 116 (7): 1793-1801

(17) Mittal VA, Ellman LM, Cannon TD. Gene-environment interaction and covariation in schizophrenia: the role of obstetric complications. Schizophr Bull. 2008; 34(6):1083-94

(18) Stefansson H, Ophoff RA, Steinberg S, et al. Common variants conferring risk of schizophrenia. Nature 2009; 460(7256): 744-7

(19) Li KC. Sliced Inverse Regression for Dimension Reduction. J Am Stat Assoc.1991; 86(414): 316-27

(20) Cook RD, Forzani L. Likelihood-Based Sufficient Dimension Reduction. J Am Stat Assoc 2009; 104(485): 197-208

(21) Yin XR, Cook RD. Estimating central subspaces via inverse third moments. Biometrika 2003; 90(1): 113-25

(22) Cook RD, Forzani L. Likelihood-Based Sufficient Dimension Reduction. J Am Stat Assoc 2009; 104(485): 197-208

*