Multifactorial Dimensionality Reduction for Disordered Trait

Alexander Rakitko
Steklov Mathematical Institute, Russian Academy of Sciences, Moscow, Russia

Keywords: GWAS, Multifactorial Dimensionality Reduction, Associated Factors, Disordered Response, Cross-validation Procedure.

Abstract: We develop our recent works concerning the identification of the factors associated with a certain complex disease. The case of disordered discrete trait is studied. We build two models (3D and 2D) for the range of response variable indicating the state of the health of a patient. In this work we consider the problem of optimal forecast for response variable depending on a finite collection of factors with values in arbitrary finite set. The quality of prediction is described by the error function involving a penalty function. The estimation of the error requires some cross-validation procedure. The developed approach provides the basis to identify the set of significant factors. Such problem arises naturally, e.g., in the genome-wide association study. Using simulated data we illustrate the efficiency of our method.

1 INTRODUCTION

The problem of the high dimension arises naturally in various stochastic models. For instance, in genetics studies the number of explanatory factors (e.g. SNP - single nucleotide polymorphism) $X_1,\ldots,X_p$ is much more than the possible sample size. But nowadays the main part of specialists share the paradigm that not all of them are significant for certain complex disease provoking. So the challenging problem is to find among huge number of factors the collection $X_{k_1},\ldots,X_{k_r}$ of factors associated with the disease.

In the previous works (Bulinski and Rakitko, 2014) we considered the case when response variable Y (Y depends on a number of factors $X_1,\ldots,X_p$ and indicates disease status) takes values in the set $\{-m,\ldots,0,\ldots,m\}$ for some natural $m \in \mathbb{N}$. However, in many applications it is impossible to introduce a linear order for traits under consideration. In this paper we study the model which assumes that Y can take values in an arbitrary finite set with no assumption about its serializability.

In medical and biological studies there exists a special research domain called the *genome-wide association studies* (GWAS). This branch of bioinformatics has already included a number of different approaches for identification of significant factors. Among powerful statistical methods applied in GWAS one can find the principal component analysis (Lee et al., 2012), logic and logistic regression (Ruczinski et al., 2003), (Sikorska et al., 2013), LASSO (Tibshirani and Taylor, 2012) and various methods of machine learning (Hastie et al., 2001). In our work we concentrate on the development of *multifactorial dimensionality reduction* (MDR) method. For the first time this method was implemented in the paper by M. Ritchie (Ritchie et al., 2001). Our variant of this method is based on the estimation of the error functions. Details are given in the next section.

The paper is organized as follows. In the second section we describe our method and build two models for response variable. Moreover, we introduce the procedure of simulation of genetics data. Some statistical results for proposed estimations are given as well. In the third section we discuss the results of application of our method to the analysis of generated data. Here we compare two different models and give some recommendations about the choice of the size of significant collection. Summary of the work is given in conclusion.

2 MATERIALS AND METHODS

Let $X = (X_1,\ldots,X_p)$ be a random vector with components $X_k : \Omega \to \{0,1,\ldots,s\}$ where $k = 1,\ldots,p$ and $s,p \in \mathbb{N}$. All random variables are defined on a probability space $(\Omega,\mathcal{F},\mathsf{P})$. In general for different k one could consider different s as it is of no importance. For instance, X_k can characterizes single nu-

Rakitko A.
Multifactorial Dimensionality Reduction for Disordered Trait.
DOI: 10.5220/0005285302320236
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOSTEC 2015)*, pages 232-236
ISBN: 978-989-758-070-3

cleotide polymorphism (SNP) and takes values in the set $\{0,1,2\}$ (corresponding to the number of minor alleles) whereas X_l, $l \neq k$, can be binary and indicates smoking addiction.

We assume that $X : \Omega \to \mathbb{X}$ (e.g. $\mathbb{X} = \{0,1,\ldots,s\}^p$) and $Y : \Omega \to \mathbb{Y}$. In our model a response variable Y depends on factors $X_1,\ldots,X_p$ and describes the state of the health of a patient. In some recent papers (Bulinski and Rakitko, 2014) the case of linearly ordered set Y was studied. For example, if $\mathbb{Y} = \{-1,0,1\}$ then $Y = 1$ or $Y = -1$ mean that person is sick or healthy, respectively. The value 0 one can interpret as ”intermediate”. In other words, in this ”grey zone” corresponding to 0 one cannot make conclusion whether disease appear or not. However, in some applications it is difficult (or even impossible) to introduce a linear order of the set $\mathbb{Y}$. For instance, Y can indicates the subtype of acute ischemic stroke according TOAST classification (Adams et al., 1993): large-artery atherosclerosis, cardioembolism or small-artery occlusion. Besides, we should take into account the other two groups corresponding to stroke of other determined etiology and stroke of undetermined etiology. So we add an extra value of Y which is responsible for such uncertainty.

2.1 2D and 3D Models

Here we consider two models for linearly unordered set $\mathbb{Y}$. The following theory with little effort could be extend onto the case of any finite capacity of $\mathbb{Y}$. But for simplicity we assume that $\mathbb{Y} = \{y_0,y_1,y_2,y_3\}$ where y_1, y_2, y_3 correspond, as example, to one of the three subtypes of ischemic acute stroke and y_0 indicates unclassified patient.

3D-Model. In this model all elements of $\mathbb{Y}$ are equidistant from each other. It means that $\{y_i\}_{i=0}^4$ are located at the vertices of tetrahedron (Figure 1). Without loss of generality let edge of tetrahedron equals 1.

2D-Model. Let us put elements y_1, y_2, y_3 at the vertices of the regular triangle with unit edge. The element y_0 is located in the middle of triangle (Figure 2).

2.2 Methods

In this subsection we describe new modification of MDR-EFE method (Bulinski and Rakitko, 2014) concerning introduced 2D- and 3D-Models.

MDR-EFE Method. To predict Y we use deterministic function $f : \mathbb{X} \to \mathbb{Y}$ of factors $X_1,\ldots,X_p$. The quality of such f is determined by means of error function

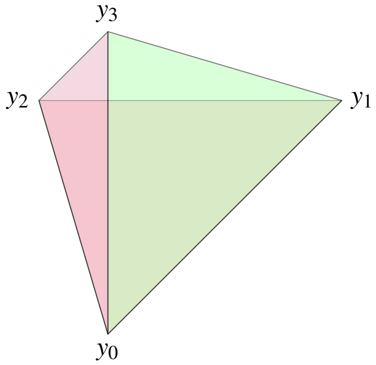

Figure 1: 3D-case.

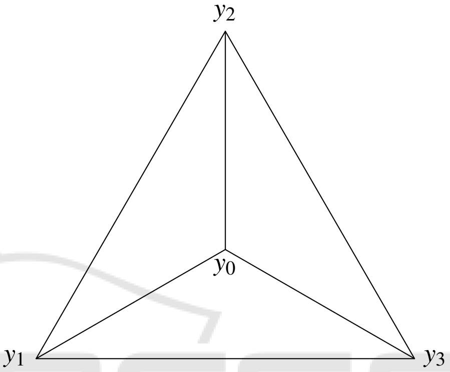

Figure 2: 2D-case.

$Err(f)$ involving a penalty function $\psi : \mathbb{Y} \to \mathbb{R}_+$:

$$Err(f) := \mathsf{E}|\mathsf{Y}-\mathsf{f}(\mathsf{x})|\psi(\mathsf{Y}). \tag{1}$$

The trivial case $\psi \equiv 0$ is excluded. In fact, the choice of penalty functions ψ gives an additional degree of freedom. But for further analysis we take function

$$\psi(y) = \frac{c}{\mathsf{P}(\mathsf{Y}=\mathsf{y})}, \quad y \in \mathbb{Y},\ c = const > 0, \tag{2}$$

proposed by Velez (Velez et al., 2007). Assuming here that $\mathsf{P}(\mathsf{Y}=\mathsf{y}) > 0$ for $y \in \mathbb{Y}$ one can take $c = 1$ without loss of generality. In (Bulinski, 2014) it was explained that this choice is natural.

For $r = 1,\ldots,p$ set $\mathbb{X}_r = \{0,1\ldots,s\}^r$. Then $\mathbb{X} = \mathbb{X}_p$. We write $\alpha = (k_1,\ldots,k_r)$, $X_\alpha = (X_{k_1},\ldots,X_{k_r})$ and $x_\alpha = (x_{k_1},\ldots,x_{k_r})$ where $x_i \in \{0,\ldots,s\}$, $i = 1,\ldots,p$. In many models it is natural to assume that response variable Y depends significantly only on a certain collection of factors $X_{k_1},\ldots,X_{k_r}$ where $1 \leq k_1 < \ldots < k_r \leq p$. In other words, for $x \in \mathbb{X}$, $\mathsf{P}(\mathsf{X}=\mathsf{x}) > 0$ and $y \in \mathbb{Y}$ the following relation holds true

$$\mathsf{P}(Y = y|X = x) = \mathsf{P}(Y = y|X_\alpha = x_\alpha). \tag{3}$$

Here $\mathsf{P}(X = x_\alpha) \geq \mathsf{P}(X = x) > 0$.

In medical and biological studies the factors $X_{k_1},\ldots,X_{k_r}$ can be viewed as essential for provoking complex disease whereas the impact of other can be neglected. Any collection of such indexes $\{k_1,\dots,k_r\}$ from (3) is called *significant*.

Fortunately, it is possible to describe all functions $f:\mathbb{X}_r\to\mathbb{Y}$ which are the solution of the problem $Err(f)\to\inf$. It is natural to approximate Y by means of one of this *optimal* functions. So we take a certain function f^{β}_{opt} defined below from the whole class of optimal functions.

Optimal function f^{β}_{opt} in 3D-Model. Simple arithmetic conversions conclude that in 3D-Model the following f^{β}_{opt} is optimal. Let us define the system of sets B^{β}_{y}, $y\in\mathbb{Y}$ as shown

$$x\in B^{\beta}_{y}\Longleftrightarrow \mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y})>\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{z})$$

for all $z\neq y$, $z\in\mathbb{Y}$. If for some $z\neq y$, $z,y\in\mathbb{Y}$ and $x\in\mathbb{X}$ one has $\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y})=\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{z})$ then we add this x to one of B^{β}_{y} or B^{β}_{z} at one's discretion. Augmented sets B^{β}_{y} we denote as A^{β}_{y} for any $y\in\mathbb{Y}$. Obviously, $\{A^{\beta}_{y}\}_{y\in\mathbb{Y}}$ is a partition of the set $\mathbb{X}$. Then the following function is optimal

$$f^{\beta}_{opt}(x)=\sum_{y\in\mathbb{Y}}y\mathbb{I}\{x\in A^{\beta}_{y}\}. \tag{4}$$

Optimal function f^{β}_{opt} in 2D-Model. In this case let's define the sets $\{B^{\beta}_{y}\}_{y\in\mathbb{Y}}$ by the following way

$$x\in B^{\beta}_{y_0}\Longleftrightarrow\begin{cases}\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_0)>\\ \quad(1-\sqrt{3})\big(\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_2)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_3)\big)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_1),\\ \mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_0)>\\ \quad(1-\sqrt{3})\big(\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_1)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_2)\big)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_3),\\ \mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_0)>\\ \quad(1-\sqrt{3})\big(\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_1)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_3)\big)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_2).\end{cases} \tag{5}$$

And for $i=1,2,3$ and $k,l\in\{1,2,3\}\setminus i$, $k\neq l$

$$x\in B^{\beta}_{y_i}\Longleftrightarrow\begin{cases}\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_i)>\\ \quad(\sqrt{3}-1)\big(\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_k)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_l)\big)+\\ \qquad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_0),\\ \mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_i)>\\ \quad\mathsf{P}(\mathsf{X}_{\beta}=\mathsf{x}_{\beta}|\mathsf{Y}=\mathsf{y}_j),\\ \qquad j\in\{1,2,3\}\setminus i.\end{cases} \tag{6}$$

Then after augmentation of the sets $\{B^{\beta}_{y}\}_{y\in\mathbb{Y}}$ we come to the sets $\{A_y\}_{y\in\mathbb{Y}}$ and the optimal function

$$f^{\beta}_{opt}(x)=\sum_{y\in\mathbb{Y}}y\mathbb{I}\{x\in A^{\beta}_{y}\}. \tag{7}$$

In fact, function f^{β}_{opt} depends only on x_{β}. Moreover, if the collection of indexes α is significant then the property of optimality of f^{α}_{opt} implys for any collection $\beta=(m_1,\dots,m_r)$, where $1\leq m_1\leq\dots\leq m_r\leq p$ the following relation

$$Err(f^{\alpha})\leq Err(f^{\beta}). \tag{8}$$

Let $\xi^1,\xi^2,\dots$ be a sequence of independent identically distributed (i.i.d.) random vectors having the same law as (X,Y). For $N\in\mathbb{N}$, set $\xi_N=(\xi^1,\dots,\xi^N)$. We will use approximation of $Err(f)$ by means of ξ_N (as $N\to\infty$) and a *prediction algorithm* (PA). This PA employs a function $f_{PA}=f_{PA}(x,\xi_N)$ defined for $x\in\mathbb{X}$ and ξ_N and taking values in $\mathbb{Y}$. More exactly, we operate with a *family of functions* $f_{PA}(x,v_p)$ (with values in $\mathbb{Y}$) defined for $x\in\mathbb{X}$ and $v_t\in(\mathbb{X}\times\mathbb{Y})^t$ where $t\in\mathbb{N}$, $t\leq N$. To simplify the notation we write $f_{PA}(x,v_t)$ instead of $f_{PA}(x,v_t)$.

Following (Bulinski and Rakitko, 2014) we can construct an estimate of $Err(f)$ involving ξ_N, prediction algorithm defined by f_{PA} and K-cross-validation (on cross-validation we refer, e.g., to (Arlot and Celisse, 2010)).

Theorem 1. *Let $\alpha=(k_1,\dots,k_r)$ where a significant collection $\{k_1,\dots,k_r\}\subset\{1,\dots,n\}$. Then, for any $\varepsilon>0$ and each $\beta=(m_1,\dots,m_r)$ with $\{m_1,\dots,m_r\}\subset\{1,\dots,n\}$, the following inequality holds*

$$\widehat{Err}_K(\widehat{f}^{\alpha}_{PA})\leq\widehat{Err}_K(\widehat{f}^{\beta}_{PA})+\varepsilon\ \ a.s. \tag{9}$$

for all N large enough.

Theorem 1 shows that it is quite natural to take for further analysis as significant such collection of indexes $\{k_1,\dots,k_r\}\subset\{1,\dots,n\}$ that $\widehat{Err}_K(\widehat{f}^{\alpha}_{PA},\xi_N)$ with $\alpha=(k_1,\dots,k_r)$ has the minimal value (or near the minimal value) among all $\widehat{Err}_K(\widehat{f}^{\beta}_{PA},\xi_N)$ where $\beta=(m_1,\dots,m_r)$ and $\{m_1,\dots,m_r\}\subset\{1,\dots,n\}$.

2.3 Simulation

Our following aim is to test MDR-EFE-algorithm on generated data. So in this subsection we introduce the way for simulation of genetic markers (X) and corresponding phenotypes (Y).

First of all we generate an array of genotypes. We define the marginal distributions of each factor by selecting alleles' frequencies in a certain way and add prescribed correlation structure responsible for Linkage Disequilibrium (LD). Let us assume that there are

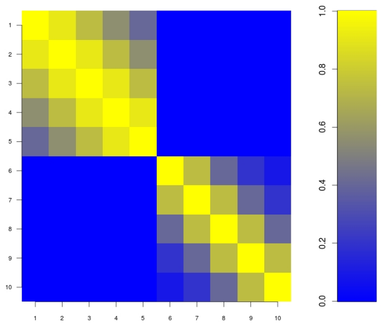

Figure 3: Correlation matrix.

N patients and for each of them we observe p factors. Then set $X_i = (X_{i,1}, \dots, X_{i,p})$ to be the genetic information about the i-th person. And the disease status for this patient we will defile below. Every factor takes values in the set $\{0,1,2\}$. For all $j \in \{1,\dots,p\}$ we generate numbers $q_j^{(1)}$ and $q_j^{(2)}$ such that $q_j^{(1)} \sim \mathcal{U}[0,1]$ and $q_j^{(2)} \sim \mathcal{U}[q_j^{(1)},1]$. Let's consider the collections $p_j = (p_{j,AA}, p_{j,Aa}, p_{j,aa}) = [q_j^{(1)}, q_j^{(2)} - q_j^{(1)}, 1 - q_j^{(2)}]$ (here square brackets $[\cdot]$ mean an ordering by ascending) as the marginal distribution of the j-th factor. For any j we can write

$$p_{j,A} = p_{j,AA} + \frac{p_{j,Aa}}{2}.$$

Here "A" and "a" denote major and minor allele correspondingly.

We sample matrix X of dimention $N \times p$ with independent elements such that $A_{i,j} \sim \mathcal{Ber}(p_{j,A})$. To obtain desired correlations between factors we apply Iman&Conovers' (Iman and Conover, 1982) algorithm to matrix A. We take block-diagonal correlation matrix where different blocks could be corresponded to different genes. For each block the correlations are constructed by means of some positive defined function

$$LD(i,j) = \exp\{-|i-j|/c\} * (1 + |i-j|/c).$$

In such a way we get alleles on the first chromosome. Using the same procedure we generate matrix B with the only difference that $B_{i,j} \sim \mathcal{Ber}(p_{j,A|A_{i,j}})$ where $p_{j,A|A_{i,j}}$ is the conditional probability to find allele A on the second chromosome given allele $A_{i,j}$ on the first one. One can see that in this case

$$X_{i,j} = A_{i,j} + B_{i,j} = \begin{cases} 0, \text{with probability} & p_{j,AA}, \\ 1, \text{with probability} & p_{j,Aa}, \\ 2, \text{with probability} & p_{j,aa}. \end{cases}$$

Parameters of Simulation:

1. $N = 6000$
2. $p = 10$
3. Correlation matrix consists of two blocks of the size 5×5, $c_1 = 2$, $c_2 = 1$
4. $Y \in \{y_1, y_2, y_3\}$ with $\mathsf{P}(\mathsf{Y} = \mathsf{y_k}) = \theta_\mathsf{k}$, $k = 1,2,3$. To define $\theta = (\theta_1, \theta_2, \theta_3)$ we introduce

$$\begin{cases} \alpha_{i,1} = \exp\{1 + 2X_{i,1} - 1.5X_{i,2} + X_{i,3}\}, \\ \alpha_{i,2} = \exp\{1 + 0.5X_{i,1} + 2X_{i,2} - 1.5X_{i,3}\}, \\ \alpha_{i,3} = \exp\{1 + 0.5X_{i,1} + 0.6X_{i,2} + 0.7X_{i,3}\}. \end{cases}$$

and assume that $(\theta_1, \theta_2, \theta_3) \sim \mathcal{Dir}(\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3})$.

To add more noise and uncertainty in our data we change the disease status of each patient onto y_0 with probability 0.1. After simulations the following distribution of response function Y were acquired: 611, 1856, 1169, 2364.

3 RESULTS

Here we demonstrate the results obtained by applying our algorithm. In the Table 1 one can see the top 10 collections (with the minimal estimation of the error function Err). The estimations are counted assuming 3D-model.

Table 1: 3D-case.

n_1	n_2	n_3	$Error$
1	**2**	**3**	**2.3794**
2	3	5	2.4929
2	3	9	2.4967
2	3	4	2.5034
2	3	6	2.5043
2	3	8	2.5088
2	3	10	2.5111
2	3	7	2.5285
1	2	10	2.6117
1	2	5	2.6178

In the Table 2 the results for 2D-model are listed.

In both models the significant collection is defined correctly. But one can see that the gap between the first and the second collection in the 2D-case almost two times bigger than the gap in the 3D-case. It may implicate that the model with triangle and its center works better with the data involving uncertainty. It should be noted that, in fact, we are interested not in the minimization of the absolute value of error, but in the growth of the gap between significant and other collections.

Table 2: 2D-case.

n_1	n_2	n_3	$Error$
1	**2**	**3**	**2.5394**
2	3	9	2.7503
2	3	8	2.7539
2	3	5	2.7595
2	3	10	2.7647
2	3	6	2.7662
2	3	7	2.7682
2	3	4	2.7697
1	2	5	2.7754
1	2	6	2.8143

Table 3: r-variation.

n_1	n_2	n_3	n_4	n_5	$Error$
3					2.95372
2	3				2.74907
1	2	3			2.56917
1	2	3	4		2.57371
1	2	3	5	9	2.56517

There is no formal rule for the choice of optimal r (the size of significant collection). But it seems quite natural to stop increase r if it doesn't decrease the estimation of the error. In the Table 3 one can find the results for another one application of our method to simulated data. In this table estimations of Err for different r are listed. It is not difficult to conclude that the error does not decrease when r exceeds 3 (the number of significant factors). Besides, one more reason to choose the $r = 3$ is that for $r = 4$ the gap between the first and the second collections is just 0.0003 (the error of collection (X_1,X_2,X_3,X_4) equals 2.57371 and the error of collection (X_1,X_2,X_3,X_6) is equal to 2.57396). It means that factors X_6 and X_4 are not in strong association with the trait in contrast to the factors X_1, X_2, X_3.

4 CONCLUSIONS

In this paper we studied the problem of identification of the collection of significant factors determining some disordered complex trait. We introduced two models for the set of possible values of response variable and developed multifactorial dimensionality reduction approach based on estimation of error function. Using simulated data we demostrated the efficiency of our method. Further research remains a comparison our algorithm with other methods of dimensionality reduction (e.g., Discriminant Principal Component Analysis).

ACKNOWLEDGEMENTS

This work is partially supported by RSF grant 14-21-00162.

REFERENCES

Adams, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., and Marsh (1993). Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke*, 24(1):35–41.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bulinski, A. (2014). On foundation of the dimensionality reduction method for explanatory variables. *Journal of Mathematical Sciences*, 199(2):113–122.

Bulinski, A. and Rakitko, A. (2014). Estimation of nonbinary random response. *Doklady Mathematics*, 89(2):225–229.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Iman, R. L. and Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11(3):311–334.

Lee, S., Epstein, M. P., Duncan, R., and Lin, X. (2012). Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. *Genetic Epidemiology*, 36(4):293–302.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138 – 147.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511.

Sikorska, K., Lesaffre, E., Groenen, P. F. J., and Eilers, P. H. C. (2013). Gwas on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, pages 166–166.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., and Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology*, 31(4):306–315.