Variable Selection Methods for Population-based Genetic Association Studies: SPLS and HSIC

Variable Selection Methods for Population-based Genetic Association Studies: SPLS and HSIC

By

Maochang Qin, MSc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

Master of Science	McMaster University
(Statistics)	Hamilton, Ontario

TITLE:	Variable Selection Methods for Population-based Genetic Association Studies: SPLS and HSIC
AUTHOR:	Maochang Qin
SUPERVISOR:	Dr. Narayanaswamy Balakrishnan Dr. Changchun Xie
NUMBER OF PAGES:	xi, 92

Acknowledgements

My sincere gratitude and appreciation will given to my supervisor Dr. Narayan aswamy Balakrishnan and Dr. Changchun Xie, for their support, patience, inspiring guidance, encouragement and wise recommendation in the preparation of this thesis and studies.

I would like to thank Dr. Joseph Beyene for his insightful comments and valuable suggestions on this thesis.

I would like to thank Yuqing Bai, Debanjan Mitra, Huh Ick, Weiping Tang and Hoihuen Wang for their helpful discussion and kind works throughout my graduate studies. My special thanks will give to Hadi Zarkoob, Sündüz Keleş and Population Health Research Institute for their sincere help in the process of running programs. I would also like to thank all my colleagues and friends who have been standing by me during my graduate studies.

Last, but not least, I would like to thank my family, especially my wife and my son, for their unconditional support and love.

Abstract

This project aims to identify the single nucleotide polymorphisms(SNPs), which are associated with the muscle size and strength in Caucasian. Two methods sparse partial least squares (SPLS) and sparse Hilbert-Schmidt independence criterion (HSIC) were applied for dimension reduction and variables selection in the Functional SNPs Associated with Muscle Size and Strength(FAMuss) Study. The selection ability of two methods was compared by simulations. The genetic determinants of skeletal muscle size and strength before and after exercise training in Caucasian were selected by using these two methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The real dataset we need to handle in this project is made up of 1397 observations and 345 factors. An attempt is made to select a few important predictor variables from a larger number of predictor variables, which make major contribution to the muscle size and strength of observations. This project belongs to the genetic association studies of genetic epidemiology. Genetic epidemiology was defined by Morton (1982) as "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations". Genetic epidemiology is closely related with both molecular epidemiology and statistical genetics.

Traditionally genetic epidemiology focuses on familial aggregation studies, segregation studies, linkage studies and association studies. Each tries to find solutions for a special question. As extensive information about Deoxyribonucleic acid (DNA) became accessible, the range of genetic epidemiology now has been extended to include common diseases to which many genes each only make a minor

contribution. As mentioned earlier, our project belongs to association studies. In fact, association studies can be divided into family-based association studies and population-based association studies. One remarkable difference between these two studies is the relationships among individuals. In family-based association studies, data collected on multi individuals with the same family unit, while in population-based association studies, data collected on unrelated individuals. Another second remarkable difference between these two studies is about allelic phase. Allelic phase is determined in family-based studies but can often be unobservable in population-based studies. Recent research indicates that the second difference might disappear in the future association studies. Strictly speaking, our project belongs to population-based association studies.

There are a lot of references about genetic epidemiology. To make our paper readable, some key concepts in genetic epidemiology will be provided here. It is well known, the development and functioning of all known living organisms with the exception of some viruses depend on the information stored in DNA. DNA molecules consist of two long polymers coiled in the shape of a double helix. The simpler units of polymers are called **nucleotides**. Each nucleotide consists of a backbone and one of four types of molecules called **bases**. The four bases founded in DNA are adenine (abbreviated A), Thymine (T), guanine (G) and cytosine (C). A **gene** is a unit of heredity located in certain regions of DNA, which contains the transcribed codes for a typical protein. In the genetics, a **locus** is the specific location of a gene or DNA sequence on a chromosome. A variant of the DNA sequence at a given locus is called an **allele**. DNA sequence can be classified in different ways. Microsatellites and single nucleotide polymorphisms (**SNPs**) are two important structural classes.

The **genotype** is defined as the pair of DNA bases observed at a location on the organism's genome. For a single nucleotide polymorphisms, its minor allele frequency (**MAF**) is the frequency of the SNP's less frequent allele in a given population. With these concepts in mind, it will be easy to understand this paper.

In this project, we aim to identifying association among SNPs and the skeletal muscle size and strength of Caucasian. It means that we need to identify the determinant from 225 SNPs of the FMS_data, which make major contribution to the muscle size and strength of observations. The data dimension reduction and variables selection are two challenges for this project. Dimension reduction is different from variables selection. Some methods can reduce dimension, but does not have variable selection. There are various typical techniques can be used for data dimension reduction. These methods can be classified two major: linear and non-linear.

Principal component analysis (PCA) first proposed by Pearson (1901) and refer to the book of Jolliffe (1986) for more information, and factor analysis (FA) are two most widely used linear dimension reduction techniques. These two methods are based upon the second-order statistics. About FA may refer to Mardia et al.(1995) for more information. Projection pursuit (PP) and independent component analysis (ICA) are two linear higher-order dimension reduction methods appropriate for non-Gaussian dataset [see Hyvaïinen (1999) for detail]. Line dimension reduction methods may not suitable if there exists nonlinear relationships among the variables of data. The results obtained by using line dimension reduction methods are usually stable than those obtained by using non-line dimension reduction methods.

The nonlinear dimensionality reduction techniques are more appropriate if the original dataset includes nonlinear relationship. The typical non-linear dimension re-

duction techniques include: principal curves (PC), self organizing maps , topographic maps, Neural networks (NN), Vector quantization and Genetic and evolutionary algorithms (GEAs). In addition, some regression methods such as projection pursuit regression, generalized linear and additive models, neural network models, sliced inverse regression (SIR) and principal hessian directions (PHD) can also be used for dimension reduction. Some non-linear methods, such as Principal curves, self organizing maps and topographic maps can be incorporated into ICA. Hastie and Stuetzle (1989) discussed the applications of principal curves. About the self organizing maps may refer to the review of Malthouse (1996). Spierenburg (1997) compared principle component analysis, vector quantization, and Neural networks. Kambhaltla and Leen (1994) introduced the Vector quantization technique and its application in dimension reduction. There are a lot of references, for example Goldberg (1989) and Raymer et al. (2000) described how the GEAs can be used for dimension reduction. About dimension reduction methods related to regression may refer to Huber (1985) for projection pursuit, McCullagh and Nelder (1989) and Hastie and Tibshirani (1990) for generalized linear models and Li (2000) for SIR/PHD. The non-linear dimension reduction techniques are sensitive to noise. Therefore, the results obtained from same data by the non-linear dimension reduction techniques may be different as the affect of noise.

The above mention methods can only be used for data dimension reduction. In this project, we need to consider both dimension reduction and variables selection. We focus on two linear methods sparse partial least squares and feature selection based on sparse Hilbert-Schmidt independence criterion. These two methods can be adopted to accomplish dimension reduction and variables selection simultaneously.

In what follows, these two methods will be introduced concisely respectively.

1.1 Sparse Partial Least Squares

In this section, we introduce sparse partial least squares. The sparse partial least squares based upon the partial least square (SPL). SPL is a new technique which combines and generalizes the strength of principal component analysis and multiple regression. Now, it has been widely used for dataset dimension reduction. The basic idea of SPL is suppose there exits a latent (hidden) component $T_{n\times k}$ such that

$$\begin{aligned} X &= TP^{\mathrm{T}} + E, \\ Y &= TQ^{\mathrm{T}} + F \end{aligned} \tag{1.1.1}$$

where $X_{n\times p}$ is predictor variables, $Y_{n\times q}$ is response variables, $P_{p\times k}$ and $Q_{q\times k}$ are coefficient matrix, $E_{n\times p}$ and $F_{n\times q}$ are errors, and superscript T is a matrix transpose operator.

Form the first equation (1.1.1), we suppose there exists a director matrix W such that $T = XW$, the usual way for finding the latent components T is transferred to find the direction columns of director matrix $W = (w_1, w_2, \cdots, w_k)$ by solving a series of optimization problems. If response variable Y is univariate, the k-th direction vector w_k can be obtained by solving the following constrained optimization problem

$$w_k = \arg\max_w \{\rho^2_{Y,Xw}\mathrm{var}(Xw)\} \text{ with } w^{\mathrm{T}}w = 1,\ w^{\mathrm{T}}\Sigma_{XX}w_j = 0, \tag{1.1.2}$$

for $j = 1, \cdots, k-1$, where Σ_{XX} represents the covariance of X. From (1.1.2), it is easy for us to see the director matrix W obtained by SPL relates to both predictor

variables X and response variables Y. Frank and Friedman (1993) regarded the director vectors W has also captured the most variable directions in the X-space.

When response Y is multivariate, there exists two formulas can be used to find these direction vectors. One was known as SIMPLES proposed by de Jong (1993), which directly uses the univariate PLS formula. The SIMPLES formula is given by

$$w_k = \text{argmax}_w\{w^{\mathrm{T}}\sigma_{XY}\sigma_{XY}^{\mathrm{T}}w\} \text{ with } w^{\mathrm{T}}w = 1,\ w^{\mathrm{T}}\Sigma_{XX}w_j = 0, \tag{1.1.3}$$

for $j = 1, \cdots, k-1$, where σ_{XY} is covariance of X and Y. The other is nonlinear iterative partial least squares (NIPALS) algorithm proposed by Wold (1966), but he did not give a specific formula. Ter Braak and de Jong (1998) gave the following 'SPL2' formula

$$w_k = \arg\max_w\{w^{\mathrm{T}}\sigma_{XY}\sigma_{XY}^{\mathrm{T}}w\} \text{ with } w^{\mathrm{T}}(I_p - W_{k-1}W_{k-1}^{-1})w = 1,\ w^{\mathrm{T}}\Sigma_{XX}w_j = 0, \tag{1.1.4}$$

for $j = 1, \cdots, k-1$, where I_p is a $p \times p$ identity matrix and W_{k-1}^{-1} is a unique Moore-Penrose inverse of $W_{k-1} = (w_1, w_2, \cdots, w_{k-1})$. And they proved the direction vector obtained by using formula (1.1.4) are exactly what solved by using NIPALS algorithm. The direction vectors obtained by using formula (1.1.3) and (1.1.4) may be different due to the variation of constraints. Their prediction performance was mainly determined by the nature of real data. De Jong (1993) pointed out both algorithms gave same direction vectors for univariate response Y.

For different response Y, the corresponding latent components T can be obtained from (1.1.2), (1.1.3) or (1.1.4), respectively. Once the latent components T obtained, the coefficients matrix Q can be estimated by solving extreme-value problem $\min_Q||Y - TQ^{\mathrm{T}}||_2$. Once we obtained latent components (direction vectors) and coefficient estimators $\hat{Q}^{\mathrm{T}}$, the parameters of the final model can be estimated via $\hat{\beta} = \hat{W}_K\hat{Q}^{\mathrm{T}}$ and the final model is $Y = \hat{\beta}X$.

A threshold for $\hat{\beta}$ was made by Huang et al.(2004) via adding sparse constraint to the procedure of finding $\hat{Q}$. Later, Chun and Keleş proposed sparse partial least square by imposing sparsity constraint in the process of dimension reduction instead of only for finding $\hat{Q}$. Dimension reduction and variables selection are accomplished at the same time in SPLS. SPLS is equivalent to solve the following constrained extreme-value problem

$$\min_{w,c}\{\kappa w^{\mathrm{T}}Mw + (1-\kappa)(c-w)^{\mathrm{T}}M(c-w) + \lambda_1|c|_1 + \lambda_2|c|_2^2\}, \text{ with } w^{\mathrm{T}}w = 1, \quad (1.1.5)$$

where $M = X^{\mathrm{T}}YY^{\mathrm{T}}X$. In (1.1.5), c is a surrogate of the original direction vector w. The (1.1.5) can be solved by alternatively iterating. In the first step solve w for fixed c, in the subsequent step solve c for fixed w, and so on.

For fixed c, if $0 < \kappa < \frac{1}{2}$, $w = \frac{1-\kappa}{1-2\kappa}(M + \lambda^* I)^{-1}Mc$ is the solution of (1.1.5) obtained by using the method of Lagrange multiplier, where λ^* solved from $c^T M(M+\lambda I)^{-2}Mc = (\frac{1-\kappa}{1-2\kappa})^2$. If $\kappa = \frac{1}{2}$, $w = UV^{\mathrm{T}}$ where U and V are singular value decomposition of Mc. (See Zou et al. (2006) for detail).

For fixed w, the solution of (1.1.5) can obtained by using the least angle regression spline algorithm (LARS) refer to Efron et al.(2004). Obtained the latent components, repeat the same procedure as that in SPL, the model parameters can be estimated by using the direction vectors solved from (1.1.5). For convenience, Chun and Keleş also provided a free R package SPLS which can be download at

1.2 Feature Selection Based on Sparse Hilbert-Schmidt Independence Criterion

In previous section, we introduced SPLS. In this section, we will introduce the feature selection method based on sparse Hilbert-Schmidt independence criterion. Before that, we first concisely address the nonnegative matrix factorization (NMF). NMF aims to express a nonnegative matrix $A \in \mathbf{R}^{m \times n}$ as a product of two nonnegative factors WH^T, where $W \in \mathbf{R}^{m \times k}$, $H \in \mathbf{R}^{n \times k}$, and $k \leq \min(m, n)$. Nonnegative matrix factorization can be traced back to the works of Gregory and Pullman (1983), Cohen and Rothblum (1993) and Paatero and Tapper (1994). Hofmann (1999) showed the NMF can be used for text retrieval. NMF was widely used as a data mining tool after Lee and Seung investigated the properties of the heuristic NMF algorithms and published several simple and useful algorithms for two types of factorizations in (1999) and (2001) respectively. Only a local minimum rather than a global minimum of the cost function $||A - WH^T||$ was guaranteed to be found by their algorithms, but a local minimum may be good enough for many practical data mining applications.

There are a lot of ways in which the factors W and H may be found, but no general method was known for NMF up to now. So it is not surprise that there no optimal algorithms was known for NMF. Partly based upon the Jordan's algorithm for the singular vectors decompensation (SVD), Biggs, Ghodsi and Vavasis (2008) gave a NMF algorithm called rank-one downdate (R1D). Compared to the local search methods which are sensitive to initialization and sometimes difficult to control convergence, R1D does not require the initial guess.

Hadi (2010) proposed a fast multivariate feature selection technique for gene express information based upon the Hilbert-Schmidt independence criterion (HSIC), R1D and SVD. According to the Hilbert-Schmidt independence criterion refer to Gretton et al. (2005) for more information, we know if two random variables X and Y with joint probability distribution ρ_{XY} are independent, then any bounded continuous functions defined on them is uncorrelated. The following squared Hilbert-Schmidt norm of cross-covariance operator

$$\mathrm{HSIC}(\rho_{XY}, F, G) \triangleq ||C_{XY}||^2_{\mathrm{HS}} \tag{1.2.6}$$

was used by Gretton et al. to measure the dependence of two variables. In (1.2.6), F is any separable reproducing kernel Hilbert space (RKHS) function from the support set of X to $\mathbf{R}$ with universal kernel $k(\cdot, \cdot)$. Similarly, G is any separable RKHS function from the support set of Y to $\mathbf{R}$ with universal kernel $b(\cdot, \cdot)$. Substituting kernel functions into (1.2.6) to compute HSIC. Hadi used the following estimator which was given by Gretton et al.

$$\mathrm{HSIC}(\rho_{xy}, F, G) = (n-1)^{-2}\mathrm{Tr}(KHBH), \tag{1.2.7}$$

where $H, K, B \in \mathbf{R}^{n\times n}$, $K_{i,j} \triangleq k(x_i, x_j)$, $B_{i,j} \triangleq b(x_i, x_j)$, $H_{i,j} \triangleq \mathrm{I} - n^{-1}\mathbf{1}\mathbf{1}^T$ (Here $\mathbf{1}$ is vector ones) called the centering matrix. In practical applications, in order to maximize the independence between two random variables X and Y, one needs to enlarge the value of $\mathrm{Tr}(KHBH)$ as likely as possible.

To obtain an estimator of HSIC, one needs to determine kernel functions $k(\cdot, \cdot)$ and $b(\cdot, \cdot)$. Hadi applied linear kernel in his article, that is $K(X, X) = X^{\mathrm{T}}X$ and $B(Y, Y) = Y^{\mathrm{T}}Y$. Suppose X represent a genomic microarray data with m gene and n sample and Y is a discrete or continuous response variable. Response variable Y

depends mainly on a projection $S = u^{\mathrm{T}}X$ of predictor variable X. It is easy to see, if u is a sparse vector, then response variable Y depends mainly on a sub set of X.

Using (1.2.7), the dependence between the projection of S and response variable Y can be measured via

$$\begin{aligned} \mathrm{Tr}(KHBH) &= \mathrm{Tr}(HX^{\mathrm{T}}uu^{\mathrm{T}}XHY^{\mathrm{T}}Y) \\ &= \mathrm{Tr}(u^{\mathrm{T}}XHY^{\mathrm{T}}YHX^{\mathrm{T}}u). \end{aligned} \tag{1.2.8}$$

(1.2.8) can be made as large as possible by increasing the magnitude. Without loss generality, one constraint $u^{\mathrm{T}}u = 1$ is added. Then the problem was transformed into the following constrained extreme-value problem

$$\max_u \mathrm{Tr}(u^{\mathrm{T}}XHBHX^{\mathrm{T}}u) \quad \text{with} \quad u^{\mathrm{T}}u = 1, \quad u \text{ is sparse.} \tag{1.2.9}$$

Use the results of Lütkepohl (1997), if the symmetric and real matrix $Q = XHBHX^T$ has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and the corresponding eigenvectors $v_1, \cdots, v_n$, then the maximum value of (1.2.9) is λ_n and the optimal solution is $u = v_n$. What we need to do is solve v_n. Since the singular vectors of XHY^{T} are equivalent to the eigenvector of $XHY^{\mathrm{T}}YHX^{\mathrm{T}}$, therefore, u can be expressed as the first singular vector of XHY^{T}.

As motioned earlier, in order to realize feature selection u should be sparse. A typical technique to add sparsity to u is to use the approach similar to the Lasso refer to Tibshirani (1994) via adding the L_1 penalty $\sum_{i=1}^{n} |u_i|$ to the cost function. Another technique is to give a threshold and retain only those elements of u that larger than the threshold.

The power method, which proposed by Stewart (1993), is a classical algorithm that can be used for computing the first singular vector of matrix. Hadi proposed

sparse power method for computing u by combining HSIC and power method. The final model obtained by using method HSIC is $Y = uX$.

The thesis was made up of four chapters. The background and methods are introduced in Chapter one. In Chapter two, we concentrate on the data generalization and simulations. We make an attempt to compare the selection ability of SPLS and HSIC via generalized data. In the Chapter three, SPLS and HSIC were used to the real FMS_data for dimension reduction and variables selection. The SNPs that make major contribution to the skeletal muscle and strength of Caucasian before and after exercise were selected. In the Chapter four, we summarize the main results and give the further work.

Chapter 2

Simulations

2.1 Introduction

For any real data, unless being given more information, it is difficult for us to know whether the predictor variables selected by using SPLS and HSIC are its true latent components or not. On the contrary, it is much easier for us to test the selection ability of SPLS and HSIC by using the data with known latent components. In this chapter, what we are going to do is compare the selection ability of SPLS and HSIC by using the data with known latent components.

First, we need to generate data. To do that, we chose 20 SNPs and 1000 observations from a real dataset. 8 SNPs were used to give signal to compute response AUC via the following identity

$$\text{AUC} = \beta_1\text{SNP1} + \beta_2\text{SNP2} + \cdots + \beta_8\text{SNP8} + \beta_9\text{BMI} + \beta_{10}\text{rannor}(n), \qquad (2.1.1)$$

where β_i are coefficients for $i = 1, \cdots, 10$, BMI is an abbreviation of body mass index, which is a heuristic measure of body weight based on a person's weight and height.

Table 2.1: The Form of Generated Dataset

BMI	rs7903146_1	···	rs934187_1	AUC1	···	AUC1000	Id
31.39212482	2	···	2	1.6736792369	···	1.5533388891	1
32.473710627	2	···	0	1.5174010482	···	1.6468266478	2
31.85961992	1	···	0	1.4491208198	···	1.4803386391	3
⋮	⋮	···	⋮	⋮	···	⋮	⋮
32.909448997	1	···	1	1.4446833844	···	1.588805091	999
47.069686474	0	···	2	1.2842626388	···	1.3332840049	1000

The term rannor(n) is noise which satisfies normal distribution, n is an integer which is called seeds in software SAS. The process of computing AUC via formula (2.1.1) is called data generation and this will be done by using software SAS. Obliviously, the signal SNPs can be treated as the latent component of simulation datasets.

In the procedure of dimension reduction and variables selection. Each model includes 1000 datasets. The predictor variables are a matrix of 1000×21 and keep same for all datasets. The response variables is 1000×1 vector. That is, the predictor variables X in our simulation is fixed for each dataset, only the response variables Y is different. The form of generated dataset is shown in Table 2.1.

The original predictor variables formed a 1000×21 matrix, but it can not be handled directly by SPLS because some of them have zero variance. Therefore, we deleted some predictor variables whose variance less than 0.01 in the process of data pretreatment. For convenience, the variances of predictor variables are shown in Table 2.2. Importing the pretreat data, we interested in counting the total number of SNPs selected from 18 SNPs and that selected from 8 signal SNPs. In order to compare the selection ability of SPLS and HSIC, we only need to compare the latter under the condition of the former is same.

Table 2.2: The Variance of Predictor Variables Generated Dataset

BMI	rs7903146_1	rs4132670_1	rs816627_1	rs10466907_1	rs3794284_1	rs2274410_1
35.13130434	0.43587487	0.45245245	0.26341441	0.10077978	0.09607107	0.44064464
rs899494_1	rs873706_1	rs8181793_1	rs238223_1	rs9551418_1	rs11568820_1	rs6489358_1
0.26223724	0.14249349	0.12890390		0.22760260	0.32212112	
rs1029629_1	rs4980845_1	rs124440_1	rs1111875_1	rs7091020_1	rs4390270_1	rs934187_1
0.43543944		0.25063463	0.49087487	0.07429830	0.51770170	0.47940340

Note: In this table, 3 SNPs (red color) are deleted because their variance less than 0.01.

2.2 Comparing the Selection Ability of SPLS and HSIC

There are lots of methods can be applied for data generation. Here, only several simple cases were considered. Some models will be used to display which method has strong ability to select the known latent components (signal SNPs).

Model 9031 includes 1000 datasets generated from following identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.12 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i) \end{aligned} \tag{2.2.2}$$

by changing the seeds. Form (2.2.2), it is easy to see that we only considered the main terms of signal SNPs in model 9031.

Input the pretreat datasets of model 9031, we count the total number of SNPs selected from 18 SNPs and the number of SNPs selected from 8 signal SNPs for each dataset.

From each dataset showed in Table 2.3 and 2.4, it is easy for us to know which method is better if the total number of SNPs selected by two methods is same. But

Table 2.3: The Number of SNPs Selected by Using SPLS from Model 9031

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	11	18	15	9	9	8	8	3	3	3
	NSSNPs	8	8	8	8	8	7	7	2	2	2
2	TSSNPs	8	9	8	18	7	14	7	3	3	3
	NSSNPs	7	8	7	8	6	8	6	2	2	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	8	10	12	18	7	7	3	3	3
	NSSNPs	7	7	8	8	8	6	6	2	2	2
1000	TSSNPs	9	9	18	15	11	8	7	3	3	3
	NSSNPs	8	8	8	8	8	7	6	2	2	2

In the above table and what follows, SN is the abbreviation for data set number, TSSNPs is used to present the total number of selected SNPs, NSSNPs present the number of selected signal SNPs. $\eta \in [0, 1)$ is called the sparsity turning parameter in SPLS. η_1 presents the optimal values chosen from the sequence $[0, 0.1, 0.2, \cdots, 0.9]$, η_2 presents the optimal values chosen from the sequence $[0, 0.1, 0.2, \cdots, 0.9] - \eta_1$ and so on. The η_i maybe get different value for different sample, take η_1 for instance, $\eta_1 = 0.5$ for sample 1 while it equal to 0.9 for sample 2. In addition, SPLS has another turning parameter K which present the number of latent components. Both η and K can be chosen by (v-fold) cross-validation using the function 'cv.spls'. K can be any integers between 1 and $\min\{p, (v-1)n/v)\}$, where p is the number of predictors and n is the sample size. For our generalized data, K is either 1 or 2.

Table 2.4: The Number of SNPs Selected by Using HSIC from Model 9031

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
S1	TSSNPs	8	5	5	6	6	6	8	8	11	14
	NSSNPs	7	4	4	5	5	5	7	7	8	8
2	TSSNPs	9	7	8	8	8	9	9	9	11	12
	SSSNPs	7	6	7	7	7	7	7	7	8	8
⋮	NSNPs	...	...	...	...	...	...	...	...	...	...
	NSSNPs	...	...	...	...	...	...	...	...	...	...
999	TSNPs	8	8	8	8	8	8	8	10	10	12
	NSSNPs	7	7	7	7	7	7	7	7	7	7
1000	TSSNPs	8	8	8	8	8	8	8	9	11	12
	NSSNPs	7	7	7	7	7	7	7	7	8	8

it is difficult for us to compare the selection ability of SPLS and HSIC only with one fixed total selected number of SNPs for the whole data. Therefore, we choose the total selected number between 7 and 9 to do compare. We obtained the Table 2.5 see below.

We use the Result of SPLS subtract the corresponding Result of HSIC, the difference shown in the Difference. If the difference in the odd row is zero, add up the corresponding number in the even row and the sum displayed in the Sum column. If the number in the even rows of Sum is larger than zero, then it means SPLS selected more signal SNPs than HSIC. In this case, we call SPLS is better that HSIC or SPLS has a stronger ability to select signal SNPs than HSIC. On the contrary, if the number in the even rows of Sum is less than zero, then it means SPLS selected less signal SNPs than HSIC. In this case, we regard HSIC is better that SPLS or HSIC has a stronger ability to select signal SNPs than SPLS. The third case, the number

Table 2.5: The Total Number of Selected SNPs Between 7 and 9 of Model 9031

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSSNPs	9	8	0	0	8	0	9	0	0	0
	NSSNPs	8	7	0	0	7	0	8	0	0	0
2	TSSNPs	9	8	7	9	8	7	0	0	0	0
	NSSNPs	8	7	6	7	7	6	1	0	0	1
3	TSSNPs	9	8	7	9	0	0	0	8	7	0
	NSSNPs	8	7	6	8	0	0	0	7	6	0
4	TSSNPs	9	8	0	0	0	0	9	8	0	0
	NSSNPs	8	7	0	0	0	0	8	7	0	0
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...

in the even rows of Sum is zero, this means SPLS and HSIC select the same number of signal SNPs. In this case, we regard there no distinguish between the selection ability of SPLS and HSIC or SPLS has same ability to select signal SNPs as SPLS. Take four datsets shown in Table 2.5 for instance, only the sum of the second dataset is 1 larger than zero. The sums of the other cases equal to zero. That is only for second dataset we know SPLS has a stronger selection ability than HSIC, we can not see any distinguish in the other datasets.

Counting the results displayed in the Sum column of Table 2.5, we obtained the Table 2.6 corresponding to the different number of datasets.

Table 2.6 is used to reflect the variation of selection ability of SPLS and HSIC corresponding to the different number of datasets and the different total number of selected SNPs. It is easy to see, under the condition of the total number of datasets is 250, if we choose total number of selected SNPs between 7 and 9 to do compare, there are 139 datasets we know SPLS is better than HSIC, only 1 dataset HSIC is

Table 2.6: The Selection Ability of SPLS and HSIC Base Upon Model 9031

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	139	152	152	286	303	303	414	447	447	522	574	574
SPLS = HSIC	110	95	95	211	191	191	331	295	295	470	415	415
SPLS < HSIC	1	3	3	3	6	6	5	8	8	8	11	11
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Note: In this table and what follows. SPLS > HSIC means SPLS shows a stronger selection ability than HSIC. SPLS = HSIC means SPLS shows the same selection ability as HSIC. And SPLS < HSIC means HSIC shows a stronger selection ability than SPLS. ND is an abbreviation of the number of datasets.

better than SPLS, and 110 datasets two methods have same selection ability. If we choose total number of selected SNPs between 7 and 10 to do compare, then there are 152 datasets we know SPLS is better than HSIC, only 3 datasets HSIC is better than SPLS, and 95 datasets two method without distinction. Therefore, We regard SPLS is better than HSIC at selecting signal SNPs. Enlarge the number of data sets, we obtain the same conclusion. Let the number of datasets equal to 1000, if we still choose total number of selected SNPs between 7 and 9 to do compare, there are 522 datasets we know SPLS is better than HSIC, only 8 datasets HSIC is better than SPLS, and 470 datasets without distinction. If we choose the total number of selected SNPs between 7 and 10 to do compare, there are 574 datasets we know SPLS is better than HSIC, only 11 datasets HSIC is better than SPLS, and 415 datasets without distinction. We also obtain SPLS is better than HSIC at selecting signal SNPs.

Based upon our simulation of model 9031, we regard SPLS has a stronger

Table 2.7: The Number of SNPs Selected by Using SPLS from Model 9032

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	11	15	18	9	8	7	7	3	3	3
	NSSNPs	8	8	8	8	7	6	6	2	2	2
2	TSSNPs	8	7	7	7	8	18	14	7	3	3
	NSSNPs	7	6	6	6	7	8	8	6	2	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	11	10	8	18	3	3	3	3	3
	NSSNPs	7	8	8	7	8	2	2	2	2	2
1000	TSSNPs	9	18	15	9	9	3	3	3	3	3
	NSSNPs	8	8	8	8	8	2	2	2	2	2

selection ability than HSIC.

In model 9031, we use two main terms 0.12·rs7903146_1 and 0.016·rs4132670_1. In the model 9032, we delete these two main terms and add the cross-product term $0.12 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1}$. The AUC in the model 9032 is computed via the identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.12 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.3}$$

Repeat the same procedure as we handle model 9031, from model 9032, we obtained the Tables 2.7 and 2.8.

Combining the Tables 2.7 and 2.8 gave the results Table 2.9.

Compare Table 2.3 and 2.7, replaced main terms by their cross term did not result in an obvious affect for SPLS. Take the first dataset for instance, for model 9031 the smallest total number required to select total signal SNPs is 9, for model 9032

Table 2.8: The Number of SNPs Selected by Using HSIC from Model 9032

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
1	TSSNPs	7	5	5	5	5	5	7	7	10	14
	NSSNPs	6	4	4	4	4	4	6	6	7	8
2	TSSNPs	8	6	7	7	7	7	8	9	11	12
	NSSNPs	6	5	6	6	6	6	6	7	8	8
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	6	6	8	8	8	8	8	9	10
	NSSNPs	7	6	7	6	7	7	7	7	7	7
1000	TSSNPs	9	8	8	8	8	9	11	11	11	13
	NSSNPs	7	7	7	7	7	7	8	8	8	8

Table 2.9: The Total Number of SNPs Selected Between 7 and 9 of Model 9032

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			sum
1	TSNPs	9	8	7	0	0	7	9	8	0	0
	NSSNPS	8	7	6	0	0	6	8	7	0	0
2	TSNPs	0	8	7	9	8	7	-9	0	0	0
	NSSNPS	0	7	6	7	6	6	-7	0	0	0
3	TSNPs	9	8	7	9	8	0	0	0	7	0
	NSSNPS	8	7	6	8	7	0	0	0	6	0
4	TSNPs	9	8	0	9	0	0	0	8	0	0
	NSSNPS	8	7	0	7	0	0	1	7	0	1
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.10: The Selection Ability of SPLS and HSIC Base Upon Model 9032

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	133	133	133	286	282	282	436	442	442	545	574	574
SPLS = HSIC	113	113	113	209	213	213	309	303	303	450	420	420
SPLS < HSIC	4	4	4	5	5	5	5	5	5	5	6	6
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

it is also 9. Consider the second dataset, for model 9031 the smallest total number required to select 7 signal SNPs is 9, while for model 9032 it is 8. Analyzes Table 2.4 and 2.8, we can obtain same conclusion. We think the possible reason is there exists a strong correlation between SNP rs7903146_1 and rs4132670_1. The signal was strengthened by replaced main terms with their cross-product term because the parameter of SNP rs4132670_1 was increased form 0.0164 to 0.12. In fact, the correlation parameter between SNP rs7903146_1 and rs4132670_1 is 0.4161.

Counting the last column of Table 2.9, we obtain Table 2.10 corresponding to the different number of datasets.

It is easy for us to see from Table 2.10, if we choose the total number of selected SNPs between 7 and 9 to do compare, there are 133 datasets we know SPLS is better than HSIC, only 4 datasets HSIC is better than SPLS, and 113 datasets without distinction. We obtain same conclusion if we choose the total number of selected SNPs between 7 and 10 or between 6 and 10 to do compare. Based upon Table 2.10, we regard SPLS is better than HSIC at selecting signal SNPs because for most datasets we obtained clear conclusion. Our replacement did not result in a fundamental affect on the selection ability of SPLS and HSIC. One of possible reason is there exists a strong correlation between SNP rs7903146_1 and rs4132670_1.

Table 2.11: The Number of SNPs Selected by Using SPLS from Model 9033

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	10	18	8	15	6	7	6	3	3	3
	NSSNPs	7	8	7	8	5	6	5	2	2	2
2	TSSNPs	7	7	8	18	13	6	5	3	3	3
	NSSNPs	6	6	7	8	7	5	4	2	2	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	10	6	6	7	18	13	3	3	3
	NSSNPs	6	6	5	5	6	8	7	2	2	2
1000	TSSNPs	8	8	9	18	14	5	7	3	3	3
	NSSNPs	7	7	7	8	7	4	6	2	2	2

In the procedure of generate datasets of model 9031, we use two signal SNPs rs899494_1 and rs873706_1 with same coefficient 0.033. Here, we want to test whether there exits any variation if these two main term $0.033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1}$ replaced by their cross-product term $0.033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1}$. Model 9033 is generated via the following identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.12 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.4}$$

Using the same procedure as we handle model 9031, from model 9033, we obtained Tables 2.11 and 2.12.

Compare Table 2.3 and 2.11, replaced main terms $0.033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1}$ with their cross term $0.033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1}$ did result in an obvious affect on SPLS. Take the first dataset for instance, for Model 9031 the smallest total number required to select total signal SNPs is 9, while for Model 9033

Table 2.12: The Number of SNPs Selected by Using HSIC from Model 9033

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
1	TSSNPs	7	4	4	4	5	6	8	9	10	13
	NSSNPs	6	3	3	3	4	5	7	7	7	7
2	TSSNPs	8	7	8	8	8	8	8	10	11	11
	NSSNPs	7	6	7	7	7	7	7	7	7	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	8	8	8	8	8	8	8	10	11
	NSSNPs	7	7	7	7	7	7	7	7	7	7
1000	TSSNPs	8	8	8	8	8	8	8	8	9	10
	NSSNPs	7	7	7	7	7	7	7	7	7	7

it is 15. Consider the second dataset, for Model 9031 the smallest total number required to select 8 signal SNPs is 9, while for model 9033 it is 18. Analyzes Table 2.4 and 2.8, we can obtain same conclusion. I think the possible reason is there exist a negative correlation between SNP rs899494_1 and rs873706_1. In fact, the correlation parameter is -0.0169. The signal was weaken by replaced main terms by their cross-prodct term even if parameter keep unchangeable.

Counting the last column of Table 2.13, we obtain the Table 2.14 respect to the different number of datasets.

Compare Table 2.6 and 2.14, replaced main terms $0.033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1}$ with their cross-product term $0.033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1}$ did result an obvious affect on both SPLS and HSIC. The number of SPLS better than HSIC is reduced dramatically while the number of HSIC better than SPLS was increased dramatically. However, the selection ability of SPLS still better than that of HSIC.

Table 2.13: The Total Number of SNPs Selected Between 7 and 9 of Model 9033

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	8	7	9	8	7	-9	0	0	0
	NSSNPS	0	7	6	7	7	6	-7	0	0	0
2	TSNPs	0	8	7	0	8	7	0	0	0	0
	NSSNPS	0	7	6	0	7	6	0	0	0	0
3	TSNPs	0	8	0	0	8	0	0	0	0	0
	NSSNPS	0	7	0	0	7	0	0	0	0	0
4	TSNPs	0	8	7	9	8	0	-9	0	7	0
	NSSNPS	0	7	6	7	6	0	-7	1	6	1
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.14: The Selection Ability of SPLS and HSIC Base Upon Model 9033

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	69	54	54	138	103	103	194	149	149	243	201	201
SPLS = HSIC	159	175	175	313	345	345	472	513	513	639	678	678
SPLS < HSIC	22	21	21	49	52	52	84	88	88	118	121	121
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table 2.15: The Number of SNPs Selected by Using SPLS from Model 90311

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	16	18	15	14	7	6	7	6	9	4
	NSSNPs	8	8	8	8	6	5	6	5	6	3
2	TSSNPs	8	8	7	10	9	13	7	18	7	5
	NSSNPs	7	7	6	7	7	8	6	8	6	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	8	9	9	8	13	18	6	5	4
	NSSNPs	7	7	8	8	7	8	8	5	4	3
1000	TSSNPs	8	11	11	18	13	5	5	4	5	4
	NSSNPs	7	8	8	8	8	4	4	3	4	3

Based upon our simulation of Model 9031, 9032 and 9033. SPLS showed a stronger ability than HSIC to select signal SNPs.

In model 90311, we want to see whether there are any variations if we weaken the signal of some SNPs. First, we still consider the linear combination of signal SNPs but weaken signal by reducing coefficients β_1 and β_8 compare to the data model 9031. AUC in the model 90311 was computed via identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.012 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.033 \cdot \text{rs899494\_1} + 0.0033 \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.5}$$

Based upon model 90311, we obtained Tables 2.15 and 2.16.

Compare Table 2.3 and 2.15, it is easy to see the signal was weaken obviously for SPLS by reducing coefficients β_1 and β_8. Take the first dataset for instance, for model 9031 the smallest total number required to select total signal SNPs is 9, while for model 90311 it is 14. Consider the second dataset, for model 9031 the smallest

Table 2.16: The Number of SNPs Selected by Using HSIC from Model 90311

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
1	TSSNPs	11	7	7	8	9	9	12	12	12	14
	NSSNPs	7	6	6	6	7	7	7	7	7	7
2	TSSNPs	9	9	9	9	9	9	9	9	11	11
	NSSNPs	7	7	7	7	7	7	7	7	7	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	8	7	8	7	8	8	8	9	9	10
	NSSNPs	7	6	7	6	7	7	7	7	7	7
1000	TSSNPs	9	6	9	8	9	9	9	9	10	10
	NSSNPs	7	5	7	7	7	7	7	7	7	7

total number required to select total signal SNPs is 9, while for model 90311 it is 13.

Compare the Table 2.4 and 2.16, it seems the signal is minor weaken for HSIC by reducing coefficients β_1 and β_8. Take the first dataset for instance, for model 9031 the smallest total number required to select 7 signal SNPs is 8, while for model 90311 it is 9. Consider the third dataset, for model 9031 the smallest total number required to select 7 signal SNPs is 8 same as that for model 90311.

Summarize the results shown in Table 2.15 and 2.16, we have the Table 2.17 corresponding to the total number selected SNPs between 7 and 9.

Adding up the results displayed in the Sum column of Table 2.17, we obtain the Table 2.18 due to the different number of datasets.

It is easy for us to see from the Table 2.18, under the condition of the number of datasets is 250, if we use the total number of selected SNPs between 7 and 9 to do compare, there are 27 datasets we have SPLS is better than HSIC, 23 datasets

Table 2.17: The Total Number of SNPs Selected Between 7 and 9 of Model 90311

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			sum
1	TSNPs	9	0	7	9	8	7	0	-8	0	0
	NSSNPS	6	0	6	7	6	6	-1	-6	0	-1
2	TSNPs	9	8	7	9	0	0	0	8	7	0
	NSSNPS	7	7	6	7	0	0	0	7	6	0
3	TSNPs	9	8	0	0	8	0	9	0	0	0
	NSSNPS	7	7	0	0	7	0	7	0	0	0
4	TSNPs	9	8	7	9	0	0	0	8	7	0
	NSSNPS	7	7	6	7	0	0	0	7	6	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.18: The Selection Ability of SPLS and HSIC Base Upon Model 90311

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	27	25	25	63	57	57	102	94	94	131	124	124
SPLS = HSIC	200	196	196	395	388	388	590	580	580	799	779	779
SPLS < HSIC	23	29	29	42	55	55	58	76	76	70	97	97
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

HSIC is better than SPLS, and 200 datasets without distinction. If we choose the total number of selected SNPs between 7 and 10 to do compare, there are 25 datasets we have SPLS is better than HSIC, 29 datasets HSIC is better than SPLS, and 196 cases without distinction. Base upon our simulation, we only can say two methods are much close. Enlarge the number of datasets, we obtain same conclusion. Enlarge the number of datasets to 1000, if we choose the total number of selected SNPs between 7 and 9 to do compare, there are 131 datasets we have SPLS is better than HSIC, 70 datasets HSIC is better than SPLS, and 799 datasets without distinction. If we choose the total number of selected SNPs between 7 and 10 to do compare, there are 124 datasets we have SPLS is better than HSIC, 97 datasets HSIC is better than SPLS, and 779 datasets without distinction.

Based on simulation in Table 2.18, we only can say SPLS has a slightly stronger ability than HSIC to select signal SNPs. Only minor distinction exists between two methods. Compared with model 9031, reducing coefficients β_1 and β_8 results in a fundamental affect on the selection ability of SPLS and HSIC.

The model 90321 is based on model 90311, we replace two main term $0.012 \cdot$ rs7903146_1 and $0.016 \cdot$rs4132670_1 by their cross-product term $0.012 \cdot$rs7903146_1$\cdot$ rs4132670_1. The AUC of model 90321 is computed by the following identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.012 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.033 \cdot \text{rs899494\_1} + 0.0033 \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.6}$$

We obtained the following results from model 90321

From the results in Table 2.7 and 2.19, we see the signal is obviously weakened for SPLS by replacing term $0.12 \cdot$rs7903146_1$\cdot$rs4132670_1 with $0.012 \cdot$rs7903146_1$\cdot$

Table 2.19: The Number of SNPs Selected by Using SPLS from Model 90321

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	16	15	18	11	11	7	7	4	4	4
	NSSNPs	8	8	8	6	6	6	6	3	3	3
2	TSSNPs	8	8	10	12	13	18	7	7	5	3
	NSSNPs	7	7	7	7	8	8	6	6	4	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	6	9	6	8	10	14	6	18	5	4
	NSSNPs	5	8	5	7	8	8	5	8	4	3
1000	TSSNPs	10	8	11	18	13	5	5	5	4	4
	NSSNPs	8	7	8	8	8	4	4	4	3	3

rs4132670_1 and $0.033 \cdot$ rs873706_1 with $0.0033 \cdot$ rs873706_1. Take the first dataset for instance, for model 9032 the smallest total number required to select total signal SNPs is 9, while for model 90321 it is 15. Compare the results in Table 2.15 and 2.19, we see the signal is minor weakened for SPLS by replacing term $0.012 \cdot$rs7903146_1+ $0.016 \cdot$rs4132670_1 with their cross-product term $0.012 \cdot$rs7903146_1$\cdot$rs4132670_1.

Compare Table 2.8 and 2.20, it seems the signal is weakened obviously for HSIC by replaced $0.12 \cdot$rs7903146_1$\cdot$rs4132670_1 by $0.012 \cdot$rs7903146_1$\cdot$rs4132670_1 and $0.033 \cdot$rs873706_1 by $0.0033 \cdot$rs873706_1. Take the first dataset for instance, for model 9032 the smallest total number required to select 8 signal SNPs is 14, while for model 90321 it should bigger than 14 because only 7 signal SNPs was selected out when the total number SNPs is 14. Compare the results in Table 2.16 and 2.20, we see the signal is minor weakened for SPLS by replacing term $0.012 \cdot$ rs7903146_1 + $0.016 \cdot$rs4132670_1 with their cross-product term $0.012 \cdot$rs7903146_1$\cdot$rs4132670_1.

Table 2.20: The Number of SNPs Selected by Using HSIC from Model 90321

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
1	TSSNPs	9	7	7	7	7	9	9	12	12	14
	NSSNPs	7	6	6	6	6	7	7	7	7	7
2	TSSNPs	9	9	9	9	9	9	9	9	10	11
	NSSNPs	7	7	7	7	7	7	7	7	7	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	6	4	4	4	4	6	7	8	8	8
	NSSNPs	5	3	3	3	3	5	6	7	7	7
1000	TSSNPs	8	6	6	6	7	8	8	8	8	11
	NSSNPs	7	5	5	5	6	7	7	7	7	7

Table 2.21: The Total Number of SNPs Selected Between 7 and 9 of Model 90321

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			sum
1	TSNPs	0	0	7	9	0	7	-9	0	0	0
	NSSNPS	0	0	6	7	0	6	-7	0	0	0
2	TSNPs	0	8	7	9	0	0	-9	8	7	0
	NSSNPS	0	7	6	7	0	0	-7	7	6	0
3	TSNPs	0	8	0	0	8	0	0	0	0	0
	NSSNPS	0	7	0	0	7	0	0	0	0	0
4	TSNPs	9	0	0	9	0	0	0	0	0	0
	NSSNPS	7	0	0	7	0	0	0	0	0	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Counting the last column of the Table 2.21, we obtain the Table 2.22 respect to different number of datasets.

Table 2.22: The Selection Ability of SPLS and HSIC Base Upon Model 90321

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	18	19	19	38	41	41	57	58	58	80	79	79
SPLS = HSIC	211	209	209	420	408	408	636	621	621	832	819	819
SPLS < HSIC	21	22	22	42	51	51	57	71	71	88	102	102
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

From the Table 2.10, we can obtain clear conclusion that SPLS is better than HSIC, while we can not obtain a clear conclusion based upon the Table 2.22. We only can see two methods are very close.

Compare the results in the Table 2.18 and 2.22, we see two methods are very close. The difference is in Table 2.18, the number of datasets of SPLS better than HSIC is a bit larger than that of HSIC better than SPLS. While from the Table 2.22, the number of SPLS better than HSIC is a bit less than that of HSIC better than SPLS. This was resulted by replacing main terms $0.012 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1}$ by their cross-product term $0.012 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1}$. This means the replacement result in a obvious affect on the selection ability of two methods.

The model 90331 is generated via the following identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.012 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.0033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1} + 0.01 \cdot \text{BMI} + 0.1 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.7}$$

Using the same procedure as we handled model 9031, we obtain the results of Table

Table 2.23: The Number of SNPs Selected by Using SPLS from Model 90331

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	10	15	18	6	8	7	6	3	3	3
	NSSNPs	7	8	8	5	7	6	5	2	2	2
2	TSSNPs	7	7	8	18	13	6	5	3	3	3
	NSSNPs	6	6	7	8	7	5	4	2	2	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	7	9	7	10	6	13	18	4	4	3
	NSSNPs	6	8	6	8	5	8	8	3	3	2
1000	TSSNPs	9	18	9	13	12	3	3	3	3	3
	NSSNPs	7	8	7	8	8	2	2	2	2	2

2.23 and 2.24 from genedata90331

Compare Table 2.14 and 2.27, replaced main terms 0.033·rs899494_1+0.0033·rs873706_1 by their cross-product term $0.0033 \cdot$ rs899494_1 $\cdot$ rs873706_1 did result in an obvious affect on selection ability of both SPLS and HSIC. The number of SPLS better than HSIC was reduced dramatically while the number of HSIC better than SPLS was increased dramatically.

Summarized the above simulation obtained from Model 9031, 9032 and 9033, we can draw a conclusion SPLS is better than HSIC at selecting signal SNPs. However, using the results obtained from model 90311, 90321 and 90331, we only can say the selection ability of SPLS and HSIC is very close.

From the Table 2.3 and 2.15, it is easy for us to see the smallest total number required to select 8 signal SNPs is only 9, which means that we give so strong signal in the procedure of data generation. We want to know if there exists any affect on

Table 2.24: The Number of SNPs Selected by Using HSIC from Model 90331

SN	η	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
1	TSSNPs	8	5	5	5	5	7	8	11	12	14
	NSSNPs	5	4	4	4	4	5	5	7	7	7
2	TSSNPs	8	7	7	7	7	8	8	8	9	11
	NSSNPs	6	6	6	6	6	6	6	6	6	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	7	5	5	5	5	7	7	8	8	8
	NSSNPs	6	4	4	4	4	6	6	7	7	7
1000	TSSNPs	7	7	7	7	7	7	7	7	8	10
	NSSNPs	6	6	6	6	6	6	6	6	7	7

Table 2.25: The Total Number of SNPs Selected Between 7 and 9 of Model 90331

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	8	7	9	8	7	-9	0	0	0
	NSSNPS	0	7	6	7	7	6	-7	0	0	0
2	TSNPs	0	8	7	0	8	7	0	0	0	0
	NSSNPS	0	7	6	0	7	6	0	0	0	0
3	TSNPs	0	8	0	0	8	0	0	0	0	0
	NSSNPS	0	7	0	0	7	0	0	0	0	0
4	TSNPs	0	8	7	9	8	0	-9	0	7	0
	NSSNPS	0	7	6	7	6	0	-7	1	6	1
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.26: The Total Number of SNPs Selected Between 7 and 9 of Model 90331

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	8	7	0	8	7	0	0	0	0
	NSSNPS	0	7	6	0	5	5	0	2	1	3
2	TSNPs	0	8	7	9	8	7	-9	0	0	0
	NSSNPS	0	7	6	6	6	6	-6	1	0	1
3	TSNPs	0	8	0	0	8	0	0	0	0	0
	NSSNPS	0	7	0	0	7	0	0	0	0	0
4	TSNPs	0	8	7	9	8	0	-9	0	7	0
	NSSNPS	0	7	6	6	6	0	-6	1	6	1
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.27: The Selection Ability of SPLS and HSIC Base Upon Model 90331

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	68	65	65	145	128	128	213	188	188	252	228	228
SPLS = HSIC	161	163	163	304	321	321	465	492	492	654	675	675
SPLS < HSIC	21	22	22	51	51	51	72	70	70	94	97	97
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table 2.28: The Ratio of Signal and Noise of Some Models

Data \ i	1	2	3	4	5	6	7	8
Model 9031	0.62766	0.01159	0.01647	4.03119e05	0.05261	0.01274	0.02856	0.01552
Model 90311	0.00628	0.01159	0.01647	4.03119e05	0.05261	0.01274	0.02856	0.00016
Model 9025	0.00100	0.00185	0.00263	6.44991e06	8.41736e05	0.00204	4.56922e05	2.48281e05
Model 9050	0.00025	0.00046	0.00066	1.61248e06	2.10434e05	0.00051	1.14231e05	6.20702e06

Note: The ratio of signal and noise is calculate by $\frac{\mathrm{Var}(\beta_i \mathrm{SNP}_i)}{\beta_{10}^2}$, which can be used as a indicator of signal intensity.

selection ability of SPLS and HSIC if we weaken the signals. We can weaken the signal by adjusting the ratio of signal and noise (RSN).

The model 9025 was generated by the following identity

$$\begin{aligned} \mathrm{AUC}_i = {} & 0.952 + 0.012 \cdot \mathrm{rs7903146\_1} + 0.016 \cdot \mathrm{rs4132670\_1} + 0.025 \cdot \mathrm{rs816627\_1} \\ & + 0.002 \cdot \mathrm{rs10466907\_1} + 0.0074 \cdot \mathrm{rs3794284\_1} + 0.017 \times \mathrm{rs2274410\_1} \\ & + 0.0033 \cdot \mathrm{rs899494\_1} + 0.0033 \cdot \mathrm{rs873706\_1} + 0.001 \cdot \mathrm{BMI} + 0.25 \cdot \mathrm{rannor}(5i). \end{aligned} \tag{2.2.8}$$

Compare (2.2.8) with (2.2.2) and (2.2.5), $\beta_1, \cdots, \beta_9$ are obviously reduced and β_{10} enlarged.

Table 2.29: The Number of SNPs Selected by Using SPLS from Model 9025

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	15	7	7	8	18	14	7	7	9	6
	NSSNPs	6	3	3	4	8	6	3	3	4	3
2	TSSNPs	9	10	9	9	2	3	3	13	18	7
	NSSNPs	4	5	4	4	2	2	2	6	8	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	1	1	2	2	3	4	18	17	10
	NSSNPs	1	1	1	2	2	3	3	8	8	6
1000	TSSNPs	7	18	17	7	3	16	12	6	10	6
	NSSNPs	5	8	8	5	2	8	6	4	6	4

Table 2.30: The Number of SNPs Selected by Using HSIC from Model 9025

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	6	8	7	8	8	9	10	10	10	10
	NSSNPs	3	3	3	3	3	4	5	5	5	5
2	TSSNPs	5	6	6	7	8	8	8	8	8	9
	NSSNPs	3	3	3	4	4	4	4	4	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	2	2	2	3	3	4	4	4	5
	NSSNPs	0	0	0	0	1	1	2	2	2	2
1000	TSSNPs	1	1	1	1	3	5	6	6	9	9
	NSSNPs	0	0	0	0	1	2	3	3	5	5

Table 2.31: The Total Number of SNPs Selected Between 7 and 9 of Model 9025

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	9	8	7	9	8	7	0	0	0	0
	NSSNPS	4	4	3	4	3	3	0	1	0	1
2	TSNPs	9	0	7	9	8	7	0	-8	0	0
	NSSNPS	4	0	4	4	4	4	0	-4	0	0
3	TSNPs	0	8	0	0	0	0	0	8	0	0
	NSSNPS	0	7	0	0	0	0	0	7	0	0
4	TSNPs	0	0	0	0	0	7	0	0	-7	0
	NSSNPS	0	0	0	0	0	2	0	0	-2	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.32: The Selection Ability of SPLS and HSIC Base Upon Model 9025

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	55	47	47	115	107	107	158	146	146	209	198	198
SPLS = HSIC	184	192	192	366	374	374	562	573	573	755	767	767
SPLS < HSIC	11	11	11	19	19	19	30	31	31	36	35	35
Sample size	250	250	250	500	500	500	750	750	750	1000	1000	1000

The model 9225 was generated by the following identity

$$\begin{aligned} \mathrm{AUC}_i = {} & 0.952 + 0.012 \cdot \mathrm{rs7903146\_1} \cdot \mathrm{rs4132670\_1} + 0.025 \cdot \mathrm{rs816627\_1} \\ & + 0.002 \cdot \mathrm{rs10466907\_1} + 0.0074 \cdot \mathrm{rs3794284\_1} + 0.017 \times \mathrm{rs2274410\_1} \\ & + 0.0033 \cdot \mathrm{rs899494\_1} + 0.0033 \cdot \mathrm{rs873706\_1} + 0.001 \cdot \mathrm{BMI} + 0.25 \cdot \mathrm{rannor}(5i). \end{aligned} \tag{2.2.9}$$

Table 2.33: The Number of SNPs Selected by Using SPLS from Model 9225

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	15	7	15	11	18	7	5	10	14	5
	NSSNPs	7	3	6	4	8	3	2	4	6	2
2	TSSNPs	9	9	9	6	10	5	7	18	14	3
	NSSNPs	4	4	4	3	5	3	4	8	6	2
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	1	1	2	2	3	4	18	15	16
	NSSNPs	1	1	1	2	2	3	3	8	6	6
1000	TSSNPs	12	14	14	6	10	7	7	18	3	7
	NSSNPs	6	6	6	4	6	5	5	8	2	5

Table 2.34: The Number of SNPs Selected by Using HSIC from Model 9225

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	3	3	3	3	9	5	5	5	5	5
	NSSNPs	1	1	1	1	4	3	3	3	3	3
2	TSSNPs	5	6	7	7	8	8	8	8	8	8
	NSSNPs	3	3	4	4	4	4	4	4	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	3	3	5	5	5	6
	NSSNPs	0	0	0	0	1	1	3	3	3	3
1000	TSSNPs	1	1	1	1	3	4	6	7	9	9
	NSSNPs	0	0	0	0	1	1	3	3	5	5

Table 2.35: The Total Number of SNPs Selected Between 7 and 9 of Model 9225

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	0	7	9	0	0	-9	0	7	0
	NSSNPS	0	0	3	4	0	0	-4	0	3	0
2	TSNPs	9	0	7	0	8	7	9	-8	0	0
	NSSNPS	4	0	4	0	4	4	4	-4	0	0
3	TSNPs	0	0	7	9	0	0	-9	0	7	0
	NSSNPS	0	0	6	5	0	0	-5	0	6	0
4	TSNPs	9	0	7	9	0	7	0	0	0	0
	NSSNPS	4	0	3	3	0	2	1	0	1	2
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.36: The Selection Ability of SPLS and HSIC Base Upon Model 9225

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	56	49	49	116	105	105	161	151	151	211	201	201
SPLS = HSIC	188	195	195	367	380	380	558	569	569	738	754	754
SPLS < HSIC	6	6	6	17	15	15	31	30	30	51	45	45
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

The model 9325 was generated by the following identity

$$\begin{aligned} \mathrm{AUC}_i = {} & 0.952 + 0.012 \cdot \mathrm{rs7903146\_1} + 0.016 \cdot \mathrm{rs4132670\_1} + 0.025 \cdot \mathrm{rs816627\_1} \\ & + 0.002 \cdot \mathrm{rs10466907\_1} + 0.0074 \cdot \mathrm{rs3794284\_1} + 0.017 \times \mathrm{rs2274410\_1} \\ & + 0.0033 \cdot \mathrm{rs899494\_1} \cdot \mathrm{rs873706\_1} + 0.001 \cdot \mathrm{BMI} + 0.25 \cdot \mathrm{rannor}(5i). \end{aligned} \tag{2.2.10}$$

It is easy to see from the Tables 2.31, 2.35 and 2.39, The signal is clearly weakened by reducing the parameters. Because the total number of selected signal SNPs is 4 less than the half the total number of selected SNPs. From the results displayed in Tables 2.32, 2.36 and 2.40. SPLS still shows a slightly stronger ability than HSIC to select signal SNPs. Replaced main terms $0.012 \cdot \mathrm{rs7903146\_1} +$

Table 2.37: The Number of SNPs Selected by Using SPLS from Model 9325

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	7	7	8	15	6	9	18	7	14	11
	NSSNPs	3	3	4	6	3	4	8	3	6	4
2	TSSNPs	9	3	9	2	10	2	13	18	7	9
	NSSNPs	4	2	4	2	5	2	6	8	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	1	1	2	2	3	15	16	17	18
	NSSNPs	1	1	1	2	2	3	7	7	8	8
1000	TSSNPs	10	7	18	17	7	16	6	3	12	4
	NSSNPs	6	5	8	8	5	8	4	2	6	2

Table 2.38: The Number of SNPs Selected by Using HSIC from Model 9325

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	6	6	8	8	8	9	10	10	10	11
	NSSNPs	3	3	3	3	3	4	5	5	5	6
2	TSSNPs	5	5	6	7	8	8	8	8	8	9
	NSSNPs	3	3	3	4	4	4	4	4	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	2	2	2	3	3	4	4	4	5
	NSSNPs	0	0	0	0	1	1	2	2	2	2
1000	TSSNPs	1	1	1	1	3	4	6	6	9	9
	NSSNPs	0	0	0	0	1	1	3	3	5	5

Table 2.39: The Total Number of SNPs Selected Between 7 and 9 of Model 9325

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	9	8	7	9	8	0	0	0	7	0
	NSSNPS	4	4	3	4	3	0	0	1	3	1
2	TSNPs	9	0	7	9	8	7	0	-8	0	0
	NSSNPS	4	0	4	4	4	4	0	-4	0	0
3	TSNPs	0	8	0	9	0	0	-9	8	0	0
	NSSNPS	0	7	0	5	0	0	-5	7	0	0
4	TSNPs	0	0	0	9	0	7	-9	0	-7	0
	NSSNPS	0	0	0	3	0	2	-3	0	-2	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.40: The Selection Ability of SPLS and HSIC Base Upon Model 9325

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	59	46	46	118	99	99	158	138	138	204	185	185
SPLS = HSIC	182	195	195	365	387	387	561	583	583	756	779	779
SPLS < HSIC	9	9	9	17	14	14	31	29	29	40	36	36
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table 2.41: The Number of SNPs Selected by Using SPLS from Model 9050

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	6	7	16	16	18	6	4	8	16	4
	NSSNPs	2	3	8	8	8	2	1	3	8	1
2	TSSNPs	8	9	5	8	8	7	5	13	18	15
	NSSNPs	3	4	2	3	3	3	2	6	8	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	2	3	18	6	18	15
	NSSNPs	2	2	2	2	2	2	8	3	8	6
1000	TSSNPs	3	3	3	5	18	16	16	6	10	12
	NSSNPs	2	2	2	3	8	8	8	4	5	6

$0.016 \cdot \text{rs4132670\_1}$ by their cross-product term $0.012 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1}$ and $0.0033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1}$ by $0.0033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1}$, respectively, did not result in an obvious variation on the selection ability of SPLS and HSIC.

In the following three models, we continue to weaken the signal by keeping $\beta_1, \cdots, \beta_9$ same as these in model 9025 while enlarging β_{10} from 0.25 to 0.5. The model 9050 was generated by the following identity

$$\begin{aligned} \text{AUC}_i = {} & 0.952 + 0.012 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1} + 0.025 \cdot \text{rs816627\_1} \\ & + 0.002 \cdot \text{rs10466907\_1} + 0.0074 \cdot \text{rs3794284\_1} + 0.017 \times \text{rs2274410\_1} \\ & + 0.0033 \cdot \text{rs899494\_1} + 0.0033 \cdot \text{rs873706\_1} + 0.001 \cdot \text{BMI} + 0.5 \cdot \text{rannor}(5i). \end{aligned} \tag{2.2.11}$$

Table 2.42: The Number of SNPs Selected by Using HSIC from Model 9050

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	5	5	6	8	8	8	8	10	12	12
	NSSNPs	3	3	4	6	6	6	6	6	6	6
2	TSSNPs	8	8	8	8	8	8	8	9	10	13
	NSSNPs	4	4	4	4	4	4	4	4	4	5
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	3	4	5	6	6	6	7	9	12	13
	NSSNPs	1	2	2	3	3	3	4	5	6	6
1000	TSSNPs	4	5	7	8	8	8	8	8	12	14
	NSSNPs	1	2	3	4	4	4	4	4	7	7

Table 2.43: The Total Number of SNPs Selected Between 7 and 9 of Model 9050

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	8	7	0	8	0	0	0	7	0
	NSSNPS	0	3	3	0	6	0	0	-3	3	-3
2	TSNPs	9	8	7	9	8	0	0	0	7	0
	NSSNPS	4	3	3	4	4	0	0	-1	3	-1
3	TSNPs	9	0	0	0	0	7	9	0	-7	0
	NSSNPS	5	0	0	0	0	1	5	0	-1	0
4	TSNPs	0	0	0	9	8	7	-9	-8	-7	0
	NSSNPS	0	0	0	3	2	1	-3	-2	-1	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.44: The Selection Ability of SPLS and HSIC Base Upon Model 9050

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	52	53	53	97	93	93	153	144	144	200	196	196
SPLS = HSIC	185	183	183	372	375	375	553	561	561	735	742	742
SPLS < HSIC	13	14	14	31	32	32	44	45	45	65	62	62
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table 2.45: The Number of SNPs Selected by Using SPLS from Model 9250

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	16	5	4	5	5	4	16	16	16	18
	NSSNPs	8	1	1	1	1	1	8	8	8	8
2	TSSNPs	8	8	8	4	9	4	5	13	18	15
	NSSNPs	3	3	3	2	4	2	2	6	8	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	2	2	18	17	18	16
	NSSNPs	2	2	2	2	2	2	8	8	8	7
1000	TSSNPs	3	3	2	4	18	16	16	5	14	15
	NSSNPs	2	2	1	2	8	8	8	3	8	8

The model 9250 was generated by the following identity

$$\begin{aligned} \mathrm{AUC}_i = {} & 0.952 + 0.012 \cdot \mathrm{rs7903146\_1} \cdot \mathrm{rs4132670\_1} + 0.025 \cdot \mathrm{rs816627\_1} \\ & + 0.002 \cdot \mathrm{rs10466907\_1} + 0.0074 \cdot \mathrm{rs3794284\_1} + 0.017 \times \mathrm{rs2274410\_1} \\ & + 0.0033 \cdot \mathrm{rs899494\_1} + 0.0033 \cdot \mathrm{rs873706\_1} + 0.001 \cdot \mathrm{BMI} + 0.5 \cdot \mathrm{rannor}(5i). \end{aligned} \tag{2.2.12}$$

Table 2.46: The Number of SNPs Selected by Using HSIC from Model 9250

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	4	4	4	4	4	5	5	6	8	9
	NSSNPs	2	2	2	2	2	3	3	4	6	6
2	TSSNPs	6	6	6	6	8	8	8	9	8	8
	NSSNPs	3	3	3	3	4	4	4	4	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	3	3	4	6	6	9
	NSSNPs	0	0	0	0	1	1	2	3	3	5
1000	TSSNPs	1	1	1	1	3	4	6	8	10	10
	NSSNPs	0	0	0	0	0	1	2	4	6	6

Table 2.47: The Total Number of SNPs Selected Between 7 and 9 of Model 9250

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	0	0	9	8	0	-9	-8	0	0
	NSSNPS	0	0	0	6	6	0	-6	-6	0	0
2	TSNPs	9	8	0	0	8	0	9	0	0	0
	NSSNPS	4	3	0	0	4	0	4	-1	0	-1
3	TSNPs	9	0	0	0	0	0	9	0	0	0
	NSSNPS	5	0	0	0	0	0	5	0	0	0
4	TSNPs	9	0	0	0	8	7	9	-8	-7	0
	NSSNPS	3	0	0	0	2	1	3	-2	-1	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.48: The Selection Ability of SPLS and HSIC Base Upon Model 9250

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	51	41	41	87	77	77	141	120	120	192	166	166
SPLS = HSIC	190	199	199	391	399	399	578	600	600	752	783	783
SPLS < HSIC	9	10	10	22	24	24	31	30	30	56	51	51
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table 2.49: The Number of SNPs Selected by Using SPLS from Model 9350

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	16	6	4	5	16	7	18	16	4	16
	NSSNPs	8	2	1	1	8	3	8	8	1	8
2	TSSNPs	8	8	9	8	13	15	5	7	5	18
	NSSNPs	3	3	4	3	6	7	2	3	2	8
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	2	3	17	6	18	16
	NSSNPs	2	2	2	2	2	2	7	3	8	7
1000	TSSNPs	3	3	3	4	18	16	16	5	12	10
	NSSNPs	2	2	2	2	8	8	8	3	6	5

The model 9350 is generated by the following identity

$$\begin{aligned} \mathrm{AUC}_i = {} & 0.952 + 0.012 \cdot \mathrm{rs7903146\_1} + 0.016 \cdot \mathrm{rs4132670\_1} + 0.025 \cdot \mathrm{rs816627\_1} \\ & + 0.002 \cdot \mathrm{rs10466907\_1} + 0.0074 \cdot \mathrm{rs3794284\_1} + 0.017 \times \mathrm{rs2274410\_1} \\ & + 0.0033 \cdot \mathrm{rs899494\_1} + 0.0033 \cdot \mathrm{rs873706\_1} + 0.001 \cdot \mathrm{BMI} + 0.5 \cdot \mathrm{rannor}(5i). \end{aligned} \tag{2.2.13}$$

Similarly, it is easy to see from the Tables 2.43, 2.47 and 2.51, the signal is clearly weakened by increasing coefficient β_{10} from 0.25 to 0.50. Because the total number of selected signal SNPs is 3 or 4 less than the half the total number of selected SNPs. From the results displayed in Tables 2.44, 2.48 and 2.52. SPLS

Table 2.50: The Number of SNPs Selected by Using HSIC from Model 9350

SN	η	1.45	1.4	1.35	1.3	0.6	0.5	0.4	0.35	0.3	0.2
1	TSSNPs	4	4	4	4	4	5	5	5	8	10
	NSSNPs	2	2	2	2	2	2	3	3	6	6
2	TSSNPs	6	6	6	6	8	8	8	8	8	9
	NSSNPs	3	3	3	3	4	4	4	4	4	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	2	2	2	2	3	3	4	5	6	9
	NSSNPs	0	0	0	0	1	1	2	2	3	5
1000	TSSNPs	1	1	1	1	3	3	5	6	8	8
	NSSNPs	0	0	0	0	0	0	2	2	4	4

Table 2.51: The Total Number of SNPs Selected Between 7 and 9 of Model 9350

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	0	7	0	8	0	0	-8	7	0
	NSSNPS	0	0	3	0	6	0	0	-6	3	0
2	TSNPs	9	8	7	9	8	0	0	0	7	0
	NSSNPS	4	3	3	4	4	0	0	-1	3	-1
3	TSNPs	9	0	0	0	0	7	9	0	-7	0
	NSSNPS	5	0	0	0	0	1	5	0	-1	0
4	TSNPs	0	0	7	9	8	7	-9	-8	0	0
	NSSNPS	0	0	2	3	2	1	-3	-2	1	1
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.52: The Selection Ability of SPLS and HSIC Base Upon Model 9350

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	48	40	40	96	75	75	143	111	111	196	151	151
SPLS = HSIC	190	200	200	379	402	402	566	605	605	745	799	799
SPLS < HSIC	12	10	10	29	23	23	41	34	34	59	50	50
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

still shows a slightly stronger ability than HSIC to select signal SNPs. Replaced main terms $0.012 \cdot \text{rs7903146\_1} + 0.016 \cdot \text{rs4132670\_1}$ by their cross-product term $0.012 \cdot \text{rs7903146\_1} \cdot \text{rs4132670\_1}$ and $0.0033 \cdot \text{rs899494\_1} + 0.033 \cdot \text{rs873706\_1}$ by $0.0033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1}$, respectively, did not result in an obvious variation on the selection ability of SPLS and HSIC.

2.3 Discussion and Conclusions

In previous section, 12 models were used to compare the selection ability of SPLS and HSIC. Based upon our simulation, a short summary is given here. We considered the main terms of the signal SNPs in the models 9031, 90311, 9025 and 9050. We gave strong signals in first models and weak signal in the last two models. From the simulations, we know SPLS has a stronger ability than HSIC to trap signal. This was shown clearly in Tables 2.6, 2.32 and 2.44. And from Table 2.18, we see the selection ability of SPLS is slightly stronger than HSIC.

We considered the cross-product term of SNPs rs7903146_1 and rs4132670_1 and main terms of the other signal SNPs in the models 9032, 90321, 9225 and 9250. We also gave strong signals in first models and weak signal in the last two models.

From the simulations, except these from the model 90321, we know SPLS has a stronger ability to trap signal than HSIC. This was shown clearly in Tables 2.10, 2.36 and 2.48. From the simulations based on the model 90321, we see the selection ability of HSIC is slightly stronger than SPLS. And the corresponding results are shown in Table 2.22.

In the models 9033, 90331, 9325 and 9350, we considered the cross-product term of SNPs rs899494_1 and rs873706_1 and main terms of the other signal SNPs. We also gave strong signals in first models and weak signal in the last two models. From the simulations, we know SPLS has a stronger ability to trap signal than HSIC. This was shown clearly in Tables 2.14, 2.27, 2.40 and 2.52.

From one model, we know the selection ability of HSIC is slightly stronger than that of SPLS. While from the other eleven models, SPLS showed stronger ability to select signal SNPs than HSIC. Its reasonable for us to regard that SPLS has a stronger selection ability than HSIC.

In the process of data pretreatment, we delete three SNPs because their variance less than 0.01. If we included these three SNPs in and import data into HSIC, we obtain the same results as those obtained by using the pretreat data. This will lead to the following two question. On the hand, this means that we can delete some predictor variables whose variance less than certain value when SPLS and HSIC be used for data dimension reduction and variables selection. This is very important when the predictor variables X of data are so large that we can set a threshold and delete those variables in X that variance less than threshold. The process of dimension reduction and variables selection will be shorten without doubt. On the other hand, this means that we face a problem how to select these predictor variables which

make major contribution to response variables Y but with small variance. Since all of them being ignored by both SPLS and HSIC. The case like this is rarely, but really exists. This is a trouble we can not dodge when SPLS and HSIC were used for dimension reduction and variables selection.

In order to compare the selection ability of SPLS and HSIC, to make the total amount of selected SNPs by HSIC is same as that obtained by SPLS, we need to find proper parameters of HSIC. We found proper parameters for the models 9031 and 9032, because the number of SPLS equal to HSIC in these two models is less than half of the total number of datasets. However, we did not find proper parameters for the other models, since the number of SPLS equal to HSIC in other models is larger than half of the total number of datasets. Finding proper parameters that make the number of SPLS equal to HSIC less than half of the total number of datasets is a big trouble in our simulations.

Chapter 3

Analysis of Real Data

3.1 Introduction

In previous Chapter, we use some models to compare the selection ability of SPLS and HSIC. In this Chapter, SPLS and HSIC will be used for dimension reduction and determinant SNPs selection in a real dataset from FAMuss (The Functional SNPs Associated with Muscle Size and Strength). FAMuSS was funded by the National Institute of Neurological Disorder and Stroke, with jointly sponsored by National Institute of Aging, and National Institute of Arthritis and Musculoskeletal Disease in April 2001. This research started in the Fall of 2001 and completed data information collection at the beginning of Summer 2003. FAMuSS is a large-scale study aimed to test the affect of genetic factors on physiologic response to resistance exercise training. This research aimed at identifying genetic determinants of skeletal muscle size and strength of human. A total observation of n=1397 college students, and data on 225 SNPs across multiple genes and a lot of clinical and de-

Table 3.1: The Sample of FMS_data

acdc_rs1501299	···	visfatin_10953502	Gender	Age	Race	NDRM.CH	DRM.CH
CA	···	NA	Female	27	Caucasian	40	40
CA	···	NA	Male	36	Caucasian	25	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
NA	···	NA	NA	NA	NA	NA	NA

Note: In real data FMS_data and follows, NA is used to present missing value. The percentage changes in muscle strength before and after exercise training are given by NDRM.CH for non-dominant arm and DRM for dominant arm. Race include African American, American Indian, Asian, Caucasian, Hispanic and others. In our project, we only consider Caucasian.

mographic factors were collected. The data are saved in a tab-delimited text file entitled FSMS_data.txt. Interested reader may refer to Foulkers (2009) for more information of FMS_data and Thompson et al. (2004) for FAMuSS study. The data set FSMS_data.txt can be input R directly from the textbook website by using following commands:

1. fmeURL<-"http://people.umass.edu/foulkes/asg/data/FMS_data.txt".

2. read.delim(file=fmsURL, header=T,sep="\t").

Furthermore, Foulkers introduced some very useful R commands in her book. It is easy for us to handle genetic information with the help of these commands.

3.2 Data Pretreat and Results

In the process of dimension reduction and variables selection of the FMS_data,

225 SNPs, Gender, Age and Race are treated as predictor variables X and NDRM.CH and DRM.CH are treated as response variables. From Table 3.1, it is easy to see the predictor variables X belong to character. That can not be handled directly by SPLS and HSIC. In order to digitalize the predictor variables X we need to identifying the minor allele for each SNP in the FMS_data. This can be done by the following three steps:

a. First we count the number of observations with each genotype for object SNP.

b. Calculate the frequencies of genotype.

c. Calculate the frequencies of allele.

In what follows, the SNP acdc_rs1501299 is chosen as an example to illustrate the process of identifying the minor allele. From the Table 3.2 see below, we see $n_{AA} = 79$ observations have the AA genotype, $n_{CA} = 506$ observations have the CA genotype and $n_{CC} = 623$ observations have the CC genotype. An additional $n_{NA} = 189$ observations are missing this genotype that will be omitted for simplicity in the process of calculating the frequencies of genotype. The genotype frequencies for AA, CA and CC then will be given respectively by $f_{AA} = \frac{n_{AA}}{n_{AA}+n_{CA}+n_{CC}} = \frac{79}{79+506+623}$, $f_{CA} = \frac{n_{CA}}{n_{AA}+n_{CA}+n_{CC}} = \frac{506}{79+506+623}$ and $f_{CC} = \frac{n_{CC}}{n_{AA}+n_{CA}+n_{CC}} = \frac{623}{79+506+623}$. The frequencies of the A and C allele are computed as $f_A = \frac{2f_{AA}+f_{CA}}{2} = 0.2748$ and $f_C = \frac{2f_{CC}+f_{CA}}{2} = 0.7252$. Therefore, A is the minor allele for SNP acdc_rs1501299 and the frequency is 0.2748. Once the minor allele was identified, we only need to replace genotype AA, CA and CC by number 2, 1 and 0 respectively to build a additive genetic model for this SNP. Similarly, T is the minor allele for SNP actn_r577x with a frequency of 0.4898. And we only need to replace genotype CC, CT and TT

Table 3.2: The Number of Observation of Each Genotype of SNPs of FMS_data

SNP& the Number of Observations	Genotypes			
acdc_rs1501299	AA	CA	CC	NA
the Number of Observations	79	506	623	189
actn_r577x	CC	CT	TT	NA
the Number of Observations	216	318	201	662
actn_rs540874	AA	GA	GG	NA
the Number of Observations	226	595	395	181
⋮	...	...	...	...
the Number of Observations	...	...	...	...

by number 0, 1 and 2 respectively to build its genetic model. Repeat the above procedure we can replaced all genotypes by numbers.

About non-SNP variables, such as Gender, Male is replaced by 1 and Female is replaced by 0. Then predictor variables X of FMS_data are now either numbers or missing values.

In what follows, we handle missing value. For simplicity, we delete missing values since they are being treated as non-informative. There are so many missing values included in the FMS_data. We can obtain different data due to the variation of techniques used to handle missing values in the process of data pretreatment. We plan to apply three different techniques to handle missing values.

95% method: This method can be accomplished by the following three steps. In the first step, we deleted SNPs whose missing value was more than 5% of its total observation. This will reduce the number of SNPs. In the second step, we deleted the observations which still included missing value. This will reduce the number of observation. There no missing values in the original data after the above

two steps. In the last step, we delete predictors variables whose variance less than certain value. The reason is if the predictor variable with variance equal to zero, it can not be handled properly by SPLS.

Method one: In this method, we first deleted any observation if it includes missing value. All missing values include in original data were deleted in this procedure. In the subsequent step, we deleted predictors variables whose column variance less than certain value.

Method two. In this method, we deleted any SNPs if there is any missing values or its variance is less than certain value. In the subsequent step, we deleted any observation if it still includes missing value.

These three techniques have been successful used for dimension reduction and variables selection of another real dataset with minor missing values. There exists a minor difference from the results obtained by using different methods. But for FMS_data, only 95% method is valid because it includes so many missing values. Using Method one or Method two, all SNPs will be deleted in the process of data pretreatment.

The process of data pretreatment is finished now, all the variables in the FMS_data is number. In FMS_data, the response variables include NDRM.CH and DRM.CH. In the process of dimension reduction and variables selection, we only choose one as our response variables.

First, we choose NDRM.CH as the response variable. Besides Gender and Age, SPLS select the following 7 SNPs: acdc_rs1501299, akt1_g22187a, c8orf68_rs69 83944, esr1_rs2228480, myod1_rs2249104, p2ry2_rs1783596 and ppara_1800206. And HSIC select the following 7 SNPs: acdc_rs1501299, akt1_g22187a, fbox32_rs487

Table 3.3: The Numerical sample of FMS_data

acdc_rs1501299	actn3_rs540874	···	vdr_rs731236	Gender	Age	NDRM.CH	DRM.CH
0	0	···	0	0	20	40	10
0	1	···	0	0	22	0	0
2	2	···	1	0	19	57.1	12.5
0	2	···	0	0	20	77.8	0.0
0	0	···	0	1	19	27.3	0.0
⋮	⋮	···	⋮	⋮	⋮	⋮	⋮
0	0	···	1	1	20	62.5	0.0

Note: In the process of data pretreatment, Only 82 SNPs are left and the smallest variance is 0.012.

Figure 3.1: The Mean Squared Prediction Error (MSPE) Plot when $\eta = 0.2$ and $K = 1$.

1385, igf2_rs680, myod1_rs2249104, p2ry2_rs1783596 and rs302964. It is easy to see 4 SNPs acdc_rs1501299, akt1_g22187a, myod1_rs2249104 and p2ry2_rs1783596 are selected out by both SPLS and HSIC. See Tables 3.4 and 3.5 for detail.

In SPLS, The function 'cv.spls' can be used to obtain a heatcap-type plot of mean squared prediction error (MSPS) and the optimal vales for turning parameter η and K. Figure 3.1 is obtained by use the command cv<-cv.spls(X,Y, eta=seq(0,0.2,0.1),K=c(1,2)). The bootstrapped confidence intervals for the coefficients of the selected predictors can be obtained by use the function 'ci.spls'. Figure 3.2 is obtain by use the command ci.f<-ci.spls(f, plot.it=T, plot.fix="y", plot.var=1). The corresponding bootstrapped confidence intervals are given in Table 3.6

Figure 3.2: Plot of the Confidence Intervals of Coefficients.

Table 3.4: SNPs Selected by SPLS from FMS_data when Response is NDRM.CH

η	0.6	0.9	0.8	0.5	0.4	0.7	0.3	0.2	0.0	0.1
N	0	0	0	0	0	0	1	7	80	60
acdc_rs1501299								×	×	×
akt1_g22187a								×	×	×
c8orf68_rs6983944								×	×	×
esr1_rs2228480								×	×	×
myod1_rs2249104							×	×	×	×
p2ry2_rs1783596								×	×	×
ppara_1800206								×	×	×

Table 3.5: SNPs Selected by HSIC from FMS_data when Response is NDRM.CH

η	0.08	0.075	0.0073	0.07	0.068	0.065	0.0407	0.0406	0.0405
N	3	4	5	6	7	8	9	10	11
acdc_rs1501299	×	×	×	×	×	×	×	×	×
akt1_g2375a_g233a									×
akt1_g22187a	×	×	×	×	×	×	×	×	×
c8orf68_rs6983944						×	×	×	×
fbox32_rs4871385	×	×	×	×	×	×	×	×	×
igf2_rs680			×	×	×	×	×	×	×
il5ra_3136618								×	×
myod1_rs2249104		×	×	×	×	×	×	×	×
p2ry2_rs1783596					×	×	×	×	×
rs302964				×	×	×	×	×	×
tcf172_12255372							×	×	×

Note: Four SNPs (red color) are selected by both SPLS and HSIC.

Table 3.6: The Bootstrapped Confidence Interval of Selected Predictor Variables when Response is NDRM.CH

SNPs	Bootstrapped Confidence Interval	
acdc_rs1501299	-5.9239943	-0.4121616
akt1_g22187a	-5.0873395	0.3193530
c8orf68_rs6983944	-5.0053354	0.4602410
esr1_rs2228059	-0.5533171	5.4781823
myod1_rs2249104	0.6444703	6.9747711
p2ry2_rs1783596	-6.2438812	-0.2160198
ppara_1800206	-0.2703238	6.4010607
Gender	-13.8220613	-8.4689273
Age	-10.5998066	-5.0949449

In what follows, we choose DRM.CH as the response variable. In addition to Gender and Age, SPLS select the following 8 SNPs: acdc_rs1501299, akt1_g22187a, c8orf68_rs6983944, esr1_rs2228480, il15ra_2228480, il15_rs2296135, myod1_rs2249 104 and ppara_1800206. And HSIC select the following 8 SNPs: acdc_rs1501299, bcl6_3774298, bmp2_rs15705, il15_rs2296135, pik3_rs3173908, tcfl72_7903146, tcfl 72_rs12255372 and tcfl72_rs7903146. It is easy to see 2 SNPs acdc_rs1501299 and il15_rs2296135 are selected out by both SPLS and HSIC. See Tables 3.7 and 3.8 for detail.

Similarly, a heatcap-type plot of mean squared prediction error (MSPS) and the optimal vales for turning parameter η and K obtained by the function 'cv.spls'. And Figure 3.3 is obtained by use the command cv<-cv.spls(X,Y, eta=seq(0,0.3,0.1), K=c(1,2)). The bootstrapped confidence intervals for the coefficients of the selected predictors obtained by use the function 'ci.spls'. And Figure 3.4 is obtain by use the command ci.f<-ci.spls(f, plot.it=TRUE,plot.fix="y",plot.var=1). The correspond-

Table 3.7: SNPs Selected by SPLS from FMS_data when Response is DRM.CH

η	0.7	0.6	0.9	0.8	0.5	0.4	0.3	0.0	0.1	0.2
N	0	0	0	0	0	2	8	80	75	28
acdc_rs1501299							×	×	×	×
akt1_g22187a							×	×	×	×
c8orf68_rs6983944							×	×	×	×
esr1_rs2228059						×	×	×	×	×
il15ra_2228480							×	×	×	×
il15_rs2296135						×	×	×	×	×
myod1_rs2249104						×	×	×	×	×
ppara_1800206							×	×	×	×

Table 3.8: SNPs Selected by HSIC from FMS_data when Response is DRM.CH

η	0.25	0.183	0.1825	0.182	0.18	0.15	0.13	0.1	0.0915
N	3	4	5	6	7	8	9	10	11
acdc_rs1501299	×	×	×	×	×	×	×	×	×
bcl6_3774298					×	×	×	×	×
bmp2_rs15705		×	×	×	×	×	×	×	×
cast_rs754615									×
il15ra_2228059								×	×
il15_rs2296135	×	×	×	×	×	×	×	×	×
pik3_rs3173908						×	×	×	×
tcfl72_12255372				×	×	×	×	×	×
tcfl72_7903146							×	×	×
tcfl72_rs12255372			×	×	×	×	×	×	×
tcfl72_rs7903146	×	×	×	×	×	×	×	×	×

Note: Two SNPs (red color) are selected by both SPLS and HSIC.

Table 3.9: The Bootstrapped Confidence Interval of Selected Predictor Variables when Response is NDRM.CH

SNPs	Bootstrapped Confidence Interval	
acdc_rs1501299	-2.7359732	0.5319817
akt1_g22187a	-2.8414674	1.0465631
c8orf68_rs6983944	-3.1417751	0.3365952
esr1_rs2228059	-0.3432567	3.8613517
il15ra_2228480	-2.5476042	0.5935540
il15_rs2296135	-0.2583301	3.0087760
myod1_rs2249104	-0.4052514	2.5893809
ppara_1800206	-0.2929844	2.9639282
Gender	-4.6185849	-1.3466458
Age	-4.8782090	-1.7586059

Figure 3.3: The Mean Squared Prediction Error (MSPE) Plot when $\eta = 0.3$ and $K = 1$.

ing bootstrapped confidence intervals are given in Table 3.9

Figure 3.4: Plot of the Confidence Intervals of Coefficients.

3.3 Results and Discussion

Based upon the results obtained in previous section. We obtained following results. If we choose NDRM.CH as the response variable, acdc_rs1501299, akt1_g22187a, c8orf68_rs6983944, esr1_rs2228480, myod1_rs2249104, p2ry2_rs1783596 and ppara_1800206 are selected by SPLS as the genetic determinants of skeletal muscle and strength of Caucasian. The mean squared prediction error plot and Bootstrapped Confidence Interval of model coefficients are also obtained. While HSIC selected acdc_rs1501299, akt1_g22187a, fbox32_rs4871385, igf2_rs680, myod1_rs2249104, p2ry2_rs1783596 and rs302964 as the genetic determinants. Four SNPs acdc_rs1501299, akt1_g22187a, myod1_rs2249104 and p2ry2_rs1783596 are selected by both SPLS and HSIC.

If we choose DRM.CH as the response variable, acdc_rs1501299, akt1_g22187a, c8orf68_rs6983944, esr1_rs2228480, il15ra_2228480, il15_rs2296135, myod1_rs2249104 and ppara_1800206 are selected by SPLS as the genetic determinants of skeletal muscle and strength of Caucasian. The mean squared prediction error plot and Bootstrapped Confidence Interval of model coefficients are also obtained. And HSIC selected acdc_rs1501299, bcl6_3774298, bmp2_rs15705, il15_rs2296135, pik3_rs3173908, tcfl72_7903146, tcfl72_rs12255372 and tcfl72_rs7903146 as the genetic determinants. Only 2 SNPs acdc_rs1501299 and il15_rs2296135 are selected by both SPLS and HSIC.

It is easy to see, SNPs selected by use SPLS with response variables NDRM.CH are almost same as these selected with response variables DRM.CH. However, SNPs selected by use HSIC with response variables NDRM.CH are different from these

selected with response variables DRM.CH. About the associations between these selected SNPs and skeletal muscle and strength of Caucasian, further research work is needed.

The genetic determinants affecting muscle size and character in farm animals have been well studied for the economic importance of meat. For example, the myostatin gene was identified as the genetic determinants of muscle size and quality in cattle [See Grobet, et al. (1997), Kambadua, et al (1997), and Mcpherron and Lee (1997) for more detail]. The ryanodine receptor gene and the IGF-2 gene were proved to affect muscle size and quality in pigs. But no evidence shows the myostatin gene and IGF-2 have any affect on muscle size and character in human refer to Ferrell et al (1999) for detail.

The results about the genetic determinants affecting muscle size and character in human is focus on some special SNPs. For example, Walsh et al. (2009) considered the associations between the ciliary neurotrophic factor (CNTF)_1357_G and the muscle strength of 754 Caucasian men (40%) and women (60%). And no significant associations were founded. Kostek et al (2009) considered the relationship between SNPs myostatin (MSTN)_2379 and follistatin (FST)_5003 the muscle size and the strength response to resistance training (RT). They regarded these two SNPs associated with baseline muscle strength and size among African Americans. These special SNPs mentioned here were not selected by SPLS and HSIC either.

Chapter 4

Conclusion and Future Work

In order to identify SNPs determinants of skeletal muscle and strength in Caucasian before and after exercise training. Two techniques SPLS and HSIC are used for dimension reduction and important SNPs selection in the FSM_data. 8 signal SNPs (from 20 SNPs) are used to generalize datasets. 15 models are generalized to identify the selection ability of SPLS and HSIC. Of them 14 models showed that SPLS have a stronger ability than HSIC to trap signal and only 1 model we obtained the selection ability of HSIC is slightly stronger than that of SPLS. The corresponding information can be founded in Chapter 2 and Appendix A. Based on our simulation, it is reasonable for us to regard SPLS has stronger selection ability than HSIC.

In the process of FMS_data dimension reduction and SNPs selection. Choosing NDRM.CH as the response variable, we have the following conclusions. Based on the results of SPLS, we regard acdc_rs1501299, akt1_g22187a, c8orf68_rs6983944, esr1_rs2228480, myod1_rs2249104, p2ry2_rs1783596 and ppara_1800206 have sig-

nificant affecting on the skeletal muscle and strength in Caucasian. While based on the results of HSIC, we thought acdc_rs1501299, akt1_g22187a, fbox32_rs4871385, igf2_rs680, myod1_rs2249104, p2ry2_rs1783596 and rs302964 make significant contribution to the skeletal muscle and strength in Caucasian. The following 4 SNPs acdc_rs1501299, akt1_g22187a, myod1_rs2249104 and p2ry2_rs1783596 are identified as genetic determinants of skeletal muscle and strength of Caucasian by both SPLS and HSIC.

Similarly, choosing DRM.CH as the response variable, we have the following conclusions. Based on the results of SPLS, we regard acdc_rs1501299, akt1_g22187a, c8orf68_rs6983944, esr1_rs2228480, il15ra_2228480, il15_rs2296135, myod1_rs2249104 and ppara_1800206 make significant contribution to the skeletal muscle and strength in Caucasian. While based on the results of HSIC, we thought acdc_rs1501299, bcl6_3774298, bmp2_rs15705, il15_rs2296135, pik3_rs3173908, tcfl72_7903146, tcfl72_rs12255372 and tcfl72_rs7903146 have significant affecting on the skeletal muscle and strength in Caucasian. SNPs acdc_rs1501299 and il15_rs2296135 are identified as genetic determinants of skeletal muscle and strength of Caucasian by both SPLS and HSIC.

Further works are needed to identify the true genetic determinants of skeletal muscle and strength from these selected. Other studies are required to confirm our finding in the paper.

Appendix A

Supplement Models

In the following three models 1019, 1020 and 1021, we continue to weaken the signal by reduce coefficients $\beta_1, \cdots, \beta_9$ and let $\beta_{10} = 0.5$. Based upon our simulation, SPLS still showed a stronger ability than HSIC to select signal SNPs. The corresponding simulation were shown in the following Tables.

The model 1019 is generated by the following identity

$$\begin{aligned} \text{AUC}_i = 0.952 &+ 0.0012 \cdot \text{rs7903146\_1} + 0.0016 \cdot \text{rs4132670\_1} + 0.0025 \cdot \text{rs816627\_1} \\ &+ 0.0002 \cdot \text{rs10466907\_1} + 0.00074 \cdot \text{rs3794284\_1} + 0.0017 \times \text{rs2274410\_1} \\ &+ 0.00033 \cdot \text{rs899494\_1} + 0.00033 \cdot \text{rs873706\_1} + 0.0001 \cdot \text{BMI} + 0.5 \cdot \text{rannor}(5i). \end{aligned}$$

Table A.1: The Number of SNPs Selected by Using SPLS from Model 1019

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	14	8	4	15	7	5	16	4	18	6
	NSSNPs	6	3	1	7	3	1	8	1	8	2
2	TSSNPs	5	2	5	3	5	7	9	13	18	15
	NSSNPs	0	0	0	0	0	2	4	6	8	7
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	1	3	4	4	5	18	11	9	16
	NSSNPs	1	1	3	4	4	4	8	6	6	8
1000	TSSNPs	3	15	15	5	6	3	18	12	11	13
	NSSNPs	2	7	7	3	4	2	8	6	6	7

Table A.2: The Number of SNPs Selected by Using HSIC from Model 1019

SN	η	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.04	0.01	0.1
1	TSSNPs	6	6	7	8	10	11	11	11	12	11
	NSSNPs	4	4	4	5	5	5	5	5	5	5
2	TSSNPs	7	7	7	9	9	9	10	13	13	9
	NSSNPs	3	3	3	4	4	4	4	5	5	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	5	6	8	9	9	10	12	12	12	10
	NSSNPs	2	3	5	5	5	6	6	6	6	6
1000	TSSNPs	6	7	8	8	9	11	11	11	12	11
	NSSNPs	3	3	4	4	5	6	6	6	7	6

Table A.3: The Total Number of SNPs Selected Between 7 and 9 of Model 1019

SN	NSNPs	Result of SPLS			Result of HSIC			Difference			Sum
1	TSNPs	0	8	7	0	8	7	0	0	0	0
	NSSNPS	0	3	3	0	5	4	0	-2	-1	-3
2	TSNPs	9	0	7	9	0	7	0	0	0	0
	NSSNPS	4	0	2	4	0	3	0	0	-1	-1
3	TSNPs	0	8	0	0	8	0	0	0	0	0
	NSSNPS	0	5	0	0	2	0	0	3	0	3
4	TSNPs	9	0	0	0	0	0	9	0	0	0
	NSSNPS	4	0	0	0	0	0	4	0	0	0
⋮	TSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	NSSNPS	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.4: The Selection Ability of SPLS and HSIC Base Upon Model 1019

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	45	44	44	87	85	85	135	125	125	176	168	168
SPLS = HSIC	182	189	189	364	372	372	545	555	555	735	742	742
SPLS < HSIC	23	17	17	49	43	43	70	70	70	89	90	90
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table A.5: The Number of SNPs Selected by Using SPLS from Model 1020

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	8	7	15	5	4	14	18	18	6	4
	NSSNPs	3	3	7	1	1	6	8	8	2	1
2	TSSNPs	4	3	5	5	7	5	9	15	13	18
	NSSNPs	0	0	0	0	2	0	4	7	6	8
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	1	3	2	4	4	5	9	16	11	18
	NSSNPs	1	3	2	4	4	4	6	8	6	8
1000	TSSNPs	2	15	3	3	18	15	12	6	11	7
	NSSNPs	1	7	2	2	8	7	6	4	6	5

The model 1020 was generated by the following identity

$$\begin{aligned}\mathrm{AUC}_i = {} & 0.952 + 0.0012 \cdot \mathrm{rs7903146\_1} \cdot \mathrm{rs4132670\_1} + 0.0025 \cdot \mathrm{rs816627\_1} \\ & + 0.0002 \cdot \mathrm{rs10466907\_1} + 0.00074 \cdot \mathrm{rs3794284\_1} + 0.0017 \times \mathrm{rs2274410\_1} \\ & + 0.00033 \cdot \mathrm{rs899494\_1} + 0.00033 \cdot \mathrm{rs873706\_1} + 0.0001 \cdot \mathrm{BMI} + 0.5 \cdot \mathrm{rannor}(5i).\end{aligned}$$

Table A.6: The Number of SNPs Selected by Using HSIC from Model 1020

SN	η	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.04	0.01	0.1
1	TSSNPs	6	6	7	8	10	11	11	11	12	11
	NSSNPs	4	4	4	5	5	5	5	5	5	5
2	TSSNPs	7	7	7	8	9	9	10	13	13	9
	NSSNPs	3	3	3	4	4	4	4	5	5	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	5	6	8	9	9	10	12	12	12	10
	NSSNPs	2	3	5	5	5	6	6	6	6	6
1000	TSSNPs	6	7	8	9	9	11	11	11	11	11
	NSSNPs	3	3	4	4	5	6	6	6	6	6

Table A.7: The Total Number of SNPs Selected Between 7 and 9 of Model 1020

SN	NSNPs	Result of SPLS	Result of HSIC	Difference	Sum
1	TSNPs	0 8 7	0 8 7	0 0 0	0
	NSSNPS	0 3 3	0 5 4	0 -2 -1	-3
2	TSNPs	9 0 7	9 8 7	0 -8 0	0
	NSSNPS	4 0 2	4 4 3	0 -4 -1	-1
3	TSNPs	0 8 0	0 8 0	0 0 0	0
	NSSNPS	0 5 0	0 2 0	0 3 0	3
4	TSNPs	9 0 0	0 0 0	9 0 0	0
	NSSNPS	4 0 0	0 0 0	4 0 0	0
⋮	TSNPS	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮
	NSSNPS	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮

Table A.8: The Selection Ability of SPLS and HSIC Base Upon Model 1020

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	51	52	52	89	91	91	143	134	134	190	179	179
SPLS = HSIC	179	183	183	370	373	373	540	549	549	725	737	737
SPLS < HSIC	20	15	15	41	36	36	67	67	67	85	84	84
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Table A.9: The Number of SNPs Selected by Using SPLS from Model 1021

SN	η	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}
1	TSSNPs	7	8	7	9	6	9	14	18	18	4
	NSSNPs	3	3	3	3	2	3	6	8	8	8
2	TSSNPs	5	4	3	5	5	9	12	15	15	18
	NSSNPs	0	0	0	0	0	4	6	7	7	8
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	4	1	3	1	4	5	9	18	11	16
	NSSNPs	4	1	3	1	4	4	6	8	6	8
1000	TSSNPs	3	18	15	5	15	3	6	7	12	11
	NSSNPs	2	8	7	3	7	2	4	5	6	8

The model 1021 is generated by the following identity

$$\begin{aligned}\text{AUC}_i = {} & 0.952 + 0.0012 \cdot \text{rs7903146\_1} + 0.0016 \cdot \text{rs4132670\_1} + 0.0025 \cdot \text{rs816627\_1} \\ & + 0.0002 \cdot \text{rs10466907\_1} + 0.00074 \cdot \text{rs3794284\_1} + 0.0017 \times \text{rs2274410\_1} \\ & + 0.00033 \cdot \text{rs899494\_1} \cdot \text{rs873706\_1} + 0.0001 \cdot \text{BMI} + 0.5 \cdot \text{rannor}(5i).\end{aligned}$$

Table A.10: The Number of SNPs Selected by Using HSIC from Model 1021

SN	η	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0.04	0.01	0.1
1	TSSNPs	6	6	7	8	10	11	11	11	12	11
	NSSNPs	4	4	4	5	5	5	5	5	5	5
2	TSSNPs	7	7	8	9	9	9	10	13	13	9
	NSSNPs	3	3	4	4	4	4	4	5	5	4
⋮	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...
999	TSSNPs	5	6	8	9	9	10	12	12	12	10
	NSSNPs	2	3	5	5	5	6	6	6	6	6
1000	TSSNPs	6	7	8	8	9	11	11	11	11	11
	NSSNPs	3	3	4	4	5	6	6	6	6	6

Table A.11: The Total Number of SNPs Selected Between 7 and 9 of Model 1021

SN	NSNPs	Result of SPLS	Result of HSIC	Difference	Sum
1	TSNPs	9 8 7	0 8 7	9 0 0	0
	NSSNPS	3 3 3	0 5 4	0 -2 -1	-3
2	TSNPs	9 0 0	9 8 7	0 -8 -7	0
	NSSNPS	4 0 0	4 4 3	0 -4 -3	0
3	TSNPs	0 8 0	0 8 0	0 0 0	0
	NSSNPS	0 5 0	0 2 0	0 3 0	3
4	TSNPs	9 0 0	0 0 0	9 0 0	0
	NSSNPS	4 0 0	0 0 0	4 0 0	0
⋮	TSNPS	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮
	NSSNPS	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮

Table A.12: The Selection Ability of SPLS and HSIC Base Upon Model 1021

TSSNPs	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10	7-9	7-10	6-10
SPLS > HSIC	46	47	47	88	86	86	140	131	131	182	171	171
SPLS = HSIC	181	181	181	366	368	368	542	544	544	731	735	735
SPLS < HSIC	23	22	22	46	46	46	68	75	75	87	94	94
ND	250	250	250	500	500	500	750	750	750	1000	1000	1000

Bibliography

[1] Biggs, M, Ghodsi, A. and Vavasis, S. (2008). Nonnegagetive matrix factorization via rank-ong downdate. In ICML'08: Preceedings of the 25th international conference on Machine Learning, 64-71. New York.

[2] Carreira-Perpinan, M. A. (1997). A review of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield.

[3] Chun, H. and Keleş, S. (2010). Sparese partial least squares for simultaneous dimension reduction and variables selection. *Journal of Royal Statistical Society: Series B* **72**, 3-25.

[4] Cohen, J. and Rothblum, U. (1993). Nonnegative ranks, decompositions and factorizations of nonnegative matrix. *Linear Algebra and its Application* **190**, 149-168.

[5] Cox, T. F. and Cox, M. A. A. (2001). Multidimensional Scaling. Chapman and Hall/CRC.

[6] de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251-263.

[7] Dobson, A. J. (1990). An introduction to generalized linear models. Chapman and Hall/CRC.

[8] Eastie, T. Johnstone, I. and Tibshirani, R. (2004). Least square regression. *Ann. Statist.* **32**, 407-499.

[9] Ferrell, R. E., Cconte, V., Lawrencee, C., Roth, S. M., Hagberg, J. M. and Hurley, B. F. (1999). Frequent sequence variation in the human myostatin (GDF8) gene as a marker for analysis of musclerelated phenotypes. *Genomics* **62**, 203-207.

[10] Foulkers, A. S. (2009). Applied Statistical Genetics with R: for Population-based Association Studies. Springer. New York.

[11] Frank, I. E. and Friedman, J. H. (1993). a statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

[12] Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

[13] Gregory, D. A. and Pullman, N, J. (1983). Semiring rank: Boolean rank and nonnegative matrix rank. *J. Combin. Inform. System Sci.* **3**, 223-233.

[14] Gretton, A. Bousquet, O. Smola, A Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *In algorithm learning Theory, 16th*

International Conference, ALLT 2005, Singapore, october 2005, Proceedings, volum 3734 of Lecture Notes in Artificial Intelligence, 63-77. Spriger.

[15] Grobet, L., Martin, L. J. and Ponceletd, et al. (1997). A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat. Genet.* **17**, 71-74.

[16] Hadi, Z. (2010). Feature selection for gene expression data based on Hilbert-Schmidt independence criterion. http://hdl.handle.net/10012/5247.

[17] Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Am. Stat. Assoc.* **84**, 502-516.

[18] Hastie, T. J. and Tibshirani, R. J. (1990). Generalized Additive Models, volume 43 of Monographs on Statistics and Applied Probability. Chapman and Hall/CRC.

[19] Huang, X. Pan, W. Parks, S. Han, X. Miler, L.W. and Hall, J.(2004). Modelling the relationship betweent Ivad support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatrics* **20**, 888-894.

[20] Huber, P. J. (1985). Projection pursuit. *Ann. Stat.* **13**, 435-475.

[21] Hyvaïinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys* **2**, 94-128.

[22] Jackson, J. E. (1991). A User's Guide to Principal Components. John Wiley and Sons. New York.

[23] Jolliffe, I. T. (1986). Principal Component Analysis. Springer-Verlag. New York.

[24] Kambadua, R., Sharma, T., Smith, P. and Bass, J. J. (1997). Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. *Genome Res.* **7**, 910-916.

[25] Kambhaltla, N. and Leen, T. K. (1994). Fast non-linear dimension reduction. In Advances in Neural Information Processing Systems, pages 152-159. Morgan Kaufmann.

[26] Kostek, M. A. Angelopoulos, T. J. Clarkson, P. M. Gordon, P. M. Moyna, N. M. Visich, P. S. Zoeller, R. F. Price, T. B. Seip, R. L. Thompson, P. D. Devaney, J. M. Gordish-Dressman, H. Hoffman, E. P. Pescatello, L. S. (2009) Myostatin and follistatin polymorphisms interact with muscle phenotypes and ethnicity. *Med. Sci. Sports Exerc.* **41**, 1063-1071.

[27] Lee, D. and Seung, H. (1999). Learning the parts of ob?ects by non-negative matrix factorization. *Nature* **401**, 788-791.

[28] Lee, D. and Seung, H. (2001). Algorithms for Non-negative Matrix Factorization. Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. 556-562. MIT Press.

[29] Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. http://www.stat.ucla.edu/ kcli/, April 2000. Lecture notes in progress.

[30] Malthouse, E. (1996). Some thcoretical results on nonlinear principal component analysis. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.1317.

[31] Mardia, K. V. Kent, J. T. and Bibby, J.M. (1995). Multivariate Analysis. Probability and Mathematical Statistics. Academic Press.

[32] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models. Chapman and Hall/CRC.

[33] Mcpherron, A. C. and Lee, S. J. (1997). Double muscling in cattle due to mutations in the myostatin gene. *Proc. Natl. Acad. Sci. USA* **94**, 12457-12461.

[34] Michael B. Ali, G. Stephen, V. (2010). Nonnegative matrix fectorization via Rank-one downdate. Proceeding of the 25-th International Conference on Machine Learning, Helsinki, Finland, 2008.

[35] Morton, N. E.(1982). Outline of Genetic Epidemiology. Karger AG, Basel

[36] Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111-126.

[37] Pearson, K. (1901). On lines and planes of closet fit to system of points in space. *Phil. Mag. Ser. B* **2**, 559-572.

[38] Raymer, M. L. et al. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* **4**, 164-171.

[39] Spierenburg, J. A.(1997). Dimension reduction of images using neural networks. Master's thesis, Leiden University.

[40] ter Braak, C. J. F. and de Jong, S. (1998). The objective function of partial least-squares regression. *Journal of Chemometrics* **12**, 41-54.

[41] Thompson, P. D. Moyna, N. Seip, R. et al. (2004). Functional polymorphisms associated with Human muscle size and strength. *Medicine and Scinec in sports and Exercise* **36** 1132-1139.

[42] Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical society, Series B* **58**, 267-288.

[43] Walsh, S. Kelsey, B. K. Angelopoulos, T. J. Clarkson, P. M. Gordon, P. M. Moyna, N. M. Visich, P. S. Zoeller, R. F. Seip, R. L. Bilbie, S. Thompson, P. D. Hoffman, E. P. Price, T. B. Devaney, J. M. Pescatello, L. S. (2009). CNTF 1357 G: A polymorphism and the muscle strength response to resistance training . *J. Appl. Physiol.* **107**, 1235-1240.

[44] Wold, H. (1966). Estimation pf principal components and related models by iterative least square. New York: Academic Press.

[45] Zou, H. Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Computnl Graph. Statist.* **15**, 265-286.