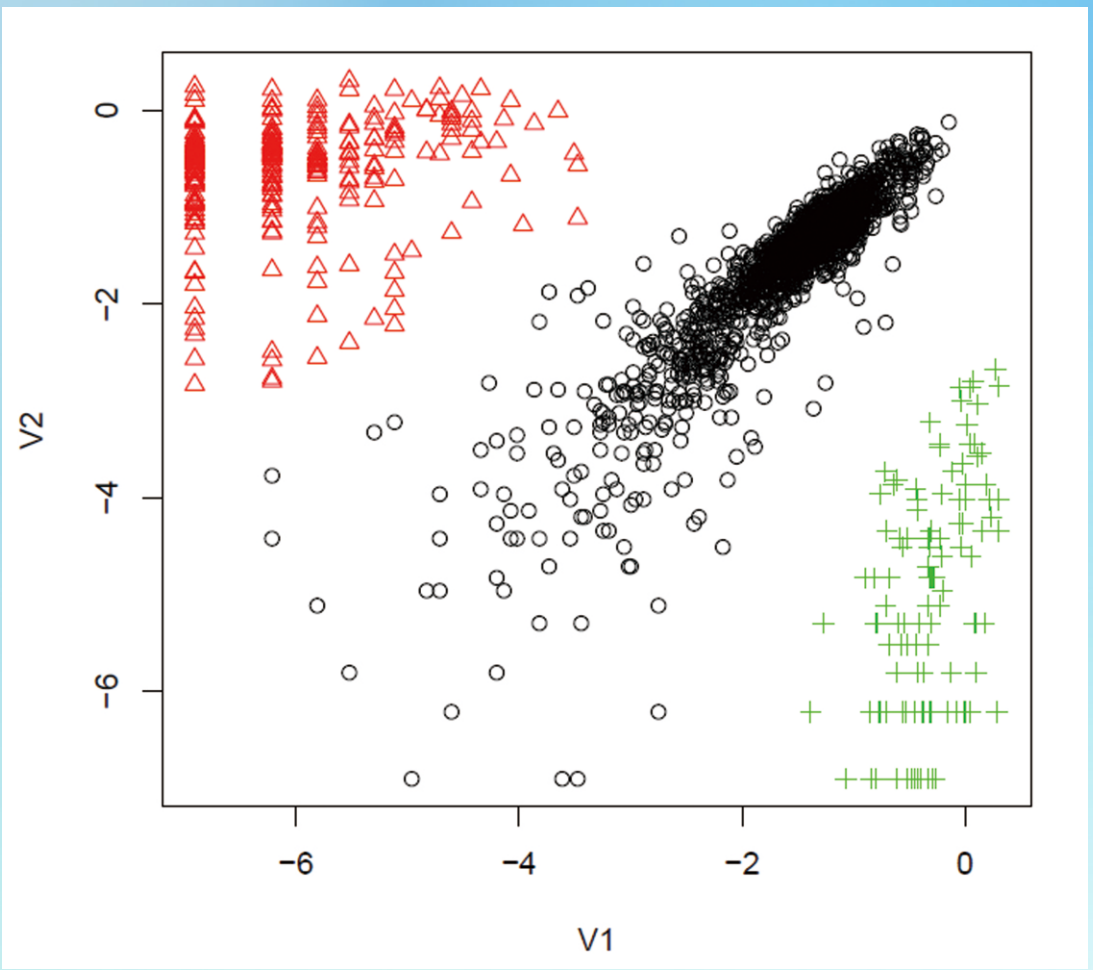
ISSN: 2161-718X Vol. 4, No. 10, December 2014

Scientific Research

Open Journal of Statistics

Special Issue on Statistical Modeling and Computation



ISSN: 2161-718X
9 772161 718001 10

www.scirp.org/journal/ojs

Journal Editorial Board

ISSN 2161-718X (Print) ISSN 2161-7198 (Online)
http://www.scirp.org/journal/ojs

Editor-in-Chief

Editorial Board

Guest Reviewers

Open Journal of Statistics, 2014, 4, 803-878
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs

Scientific Research

Table of Contents

Volume 4 Number 10 December 2014

The figure on the front cover is from the article published in Open Journal of Statistics, 2014, Vol. 4, No. 10, pp. 826-836 by Yansong Cheng and Surajit Ray.

Open Journal of Statistics (OJS)

Journal Information

SUBSCRIPTIONS

The *Open Journal of Statistics* (Online at Scientific Research Publishing, www.SciRP.org) is published bimonthly by Scientific Research Publishing, Inc., USA.

Subscription rates:

Print: $69 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ojs@scirp.org

Open Journal of Statistics, 2014, 4, 803-813
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410076

Scientific Research

Cusp Catastrophe Polynomial Model: Power and Sample Size Estimation

Ding-Geng Chen[1,2,3], Xinguang Chen[4,5], Feng Lin[1,6], Wan Tang[2], Yuhlong Lio[7], Yuanyuan Guo[8]

[1]Center of Research, School of Nursing, University of Rochester Medical Center, Rochester, NY, USA
[2]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA
[3]Institute of Data Sciences, University of Rochester, Rochester, NY, USA
[4]Department of Epidemiology, University of Florida, Gainesville, FL, USA
[5]School of Public Health, Wuhan University, Wuhan, China
[6]AD-CARE, Department of Psychiatry, University of Rochester Medical Center, Rochester, NY, USA
[7]Department of Mathematical Sciences, University of South Dakota, Vermillion, SD, USA
[8]Department of Statistics, Central South University, Changsha, China
Email: din_chen@urmc.rochester.edu

Received 6 October 2014; revised 26 October 2014; accepted 8 November 2014

Abstract

Guastello's polynomial regression method for solving cusp catastrophe model has been widely applied to analyze nonlinear behavior outcomes. However, no statistical power analysis for this modeling approach has been reported probably due to the complex nature of the cusp catastrophe model. Since statistical power analysis is essential for research design, we propose a novel method in this paper to fill in the gap. The method is simulation-based and can be used to calculate statistical power and sample size when Guastello's polynomial regression method is used to do cusp catastrophe modeling analysis. With this novel approach, a power curve is produced first to depict the relationship between statistical power and samples size under different model specifications. This power curve is then used to determine sample size required for specified statistical power. We verify the method first through four scenarios generated through Monte Carlo simulations, and followed by an application of the method with real published data in modeling early sexual initiation among young adolescents. Findings of our study suggest that this simulation-based power analysis method can be used to estimate sample size and statistical power for Guastello's polynomial regression method in cusp catastrophe modeling.

Keywords

Cusp Catastrophe Model, Polynomial Regression Method, Statistical Power Analysis, Sample Size

How to cite this paper: Chen, D.-G., Chen, X.G., Lin, F., Tang, W., Lio, Y. and Guo, Y.Y. (2014) Cusp Catastrophe Polynomial Model: Power and Sample Size Estimation. *Open Journal of Statistics*, **4**, 803-813.
http://dx.doi.org/10.4236/ojs.2014.410076

Determination

1. Introduction

Popularized in the 1970's by Thom [1], Thom and Fowler [2], Cobb and Ragade [3], Cobb and Watson [4], and Cobb and Zack [5], catastrophe theory was proposed to understand a complicated set of behaviors including both gradual and continuous changes and sudden and discrete or catastrophical changes. Computationally, there are two directions to implement this theoretical catastrophe theory. One direction is operationalized by Guastello [6] [7] with the implementation into a polynomial regression approach and another direction by a stochastic cusp catastrophe model from Cobb and his colleagues [5] with implementation in an R package in [8]. And this paper is to discuss the first direction on polynomial cusp catastrophe regression model due to its relative simplicity and ease for implementation as simple regression approach. This model has been used extensively in research. Typical examples include modeling of accident process [7], adolescent alcohol use [9], changes in adolescent substance use [10], binge drinking among college students [11], sexual initiation among young adolescents [12], nursing turnover [13], and effect of HIV prevention among adolescents [12] [14].

Even though this polynomial regression method has been widely applied in behavioral studies to investigate the existence of cusp catastrophe, to the best of our knowledge, no reported research has addressed the determination of sample size and statistical power for this analytical approach. Statistical power analysis is an essential part for researchers to efficiently plan and design a research project as pointed out in [15]-[17]. To assist and enhance application of the polynomial regression method in behavioral research, this paper is aimed to fill this method gap by reporting the Monte-Carlo simulation-based method we develope to conduct power analysis and to determine sample size.

The structure of the paper is as follows. We start with a brief review of the cusp catastrophe model (Section 2), followed by reporting our development of the novel simulation-based approach to calculate the statistical power (Section 3). This approach is then verified through Monte Carlo simulations and is further illustrated with data derived from published study (Section 4). Conclusions and discussions are given at the end of the paper (Section 5).

2. Cusp Catastrophe Model

2.1. Overview

The cusp catastrophe model is proposed to model system outcomes which can incorporate the linear model with extension to nonlinear model along with discontinuous transitions in equilibrium states as control variables vary. According to the catastrophe systems theory [1] [18]-[20], the dynamics for a cusp system outcome is expressed by the time derivative of its state variable (often called behavioral variable within the context of catastrophe theory) to the potential function: $V(z;x,y)=1/4z^4-1/2z^2y-zx$ The first derivative of V will consist of the equilibrium plane of the cusp catastrophe:

$$\partial V(z,x,y)/\partial z = z^3 - yz - x = 0 \tag{1}$$

where x is called asymmetry or normal control variable and y is called bifurcation or splitting control variable. In the model, the two control variables x and y co-vary to determine the behavior outcome variable z. **Figure 1** depicts the equilibrium plane which reflects the response surface of the outcome measure (z) at various combinations of x and y. It can be seen from the figure that the dynamic changes in a behavior measure (z) has two stable regions (attractors), the lower area in the front left and the upper areas in the front right. Beyond these two regions, behavior z becomes unstable. This characteristic can be further revealed by projecting the unstable region to the x and y control plane as a cusp region. The cusp region is characterized by two lines, line O-Q (the ascending threshold) and line O-R (the descending threshold) of the equilibrium surface. In this region, the outcome measure becomes highly unstable, and sudden change or jumping in behavior status will occur, because a very small change in x or y or both will lead z to cross either the threshold line O-Q or O-R.

Furthermore, the paths A, B, and C in **Figure 1** depict three typical but different pathways of change in the

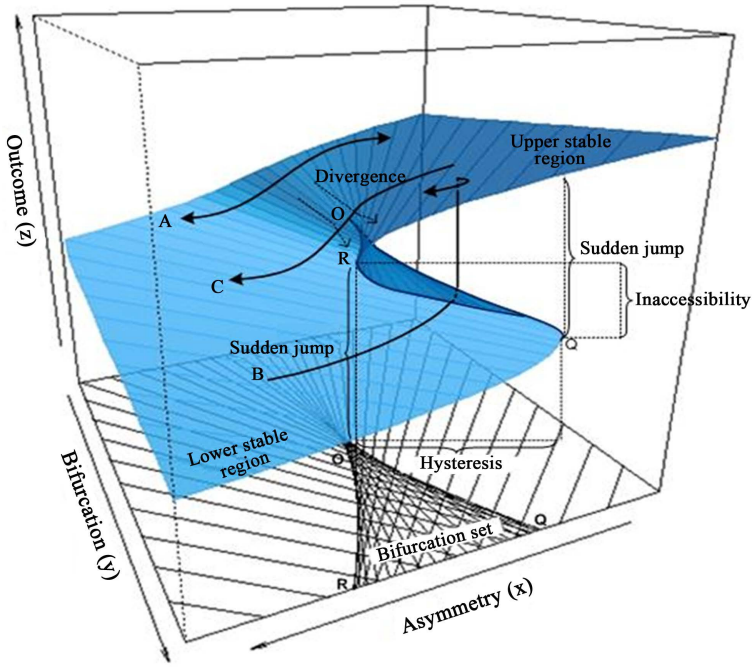


Figure 1. Cusp catastrophe model for outcome measures (Z) in the equilibrium plane with asymmetry control variable (X) and bifurcation control variable (Y). (Annotated by the authors with the original graph produced by Grasman's R package "cusp").

outcome measure (z). Path A shows that in any situations where $y < \mathrm{O}$, there is a smooth relation between outcome measure (z) and the asymmetry variable (x); path B shows that in any situations where $y > \mathrm{O}$, if the asymmetry variable x increases to reach and pass the ascending threshold link O-Q, outcome measure (z) will increase suddenly from the low stable region to the upper stable region of the equilibrium plane; and Path C shows a sudden drop in outcome measure (z) as x declines to reach and pass the descending threshold line O-R .

From the affirmative description, it is clearly that a cusp model differs from a linear model in that: 1) A cusp model allows the forward and backward progression follows different paths in the outcome measure and both processes can be modeled simultaneously (see Paths B and C in **Figure 1**) while a linear model only permits one type of relationship; 2) A cusp model covers both a discrete component and a continuous component of a behavior change while a linear model covers on continuous process (Path A). In this case a linear model can be considered as a special case of the cusp model; 3) A cusp model consists of two stable regions and two thresholds for sudden and discrete changes. Therefore, the application of the cusp modeling will advance the linear approach and better assist researchers to describe the behavior data while evidence obtained from such analysis, in turn, can be used to advance theories and models to better explain a behavior.

2.2. Guastello's Cusp Catastrophe Polynomial Regression Model

To operationalize the cusp catastrophe model for behavior research, Guastello [6] [7] developed the polynomial regression approach to implement the concept of cusp model. Since the first publication of this method, it has been widely used in analyzing real data as we described in the Introduction. In this study, we referred the method as Gastello's polynomial cusp regression. According to Gustello, this approach is derived by inserting regression β coefficients into the Equation (1), with change scores $\Delta z = z_2 - z_1$ (the differences in the measurement scores of a behavior assessed at time 1 and time 2) as a numerical approximation of $\mathrm{d}z$:

$$\Delta z = \beta_0 + \beta_1 \times z_1^3 + \beta_2 \times z_1^2 + \beta_3 \times y \times z_1 + \beta_4 \times x + \beta_5 \times y + \varepsilon \tag{2}$$

where β_0 is the intercept and ε is the normally distributed error term. Two additional term $\beta_2 \times z_1^2$ and $\beta_5 \times y$ are added to capture potential deviations of the data from the equilibrium plane. When conducting modeling analysis, a cusp is indicated ONLY if the estimated β_1 for the cubic term, plus β_3 (for the interaction term) or β_4 (for control variable *x*) in Equation (2) are statistically significant.

To demonstrate the efficiency of the polynomial regression approach in describing behavioral changes that are cusp, Guastelly [7] recommended a comparative approach. In this approach, two types, four alternative linear models can be constructed and used in modeling the same variables:

1) Change scores linear models

$$\Delta z = \beta_0 + \beta_1 z_1 + \beta_4 x + \beta_5 y \tag{3}$$

$$\Delta z = \beta_0 + \beta_1 z_1 + \beta_3 y z_1 + \beta_4 x + \beta_5 y \tag{4}$$

2) Pre-and post-linear models

$$z_2 = \beta_0 + \beta_1 z_1 + \beta_4 x + \beta_5 y \tag{5}$$

$$z_2 = \beta_0 + \beta_1 z_1 + \beta_3 y z_1 + \beta_4 x + \beta_5 y \tag{6}$$

These alternative linear models add another analytical strategy to strength the polynomial regression method. A better data-model fitting (or a larger R^2) of the cusp model (2) than the alternative linear models (3) through (6) is often used as additional evidence supporting the hypothesis that the dynamics of a study behavior follows the cusp catastrophe. Fitting Guastello's cusp regression model and the four alternative models can all be conducted with commonly available statistical software, including SAS, SPSS, STATA and *R*. More recent discussions and applications of the cusp catastrophe modeling methods can be found in [21].

3. Simulation-Based Power Analysis Approach for Guastello's Cusp Regression

3.1. A Brief Introduction to Statistical Power

In statistics, power is defined as the probability of correctly rejecting the null hypothesis. Stated in common language, power is the fraction of the times that the specified null-hypothesis value will be rejected from statistical tests. Operationally based on this definition, if we specify an alternative hypothesis H_1, a desired type-I error rate α, and a desired power $(1-\beta)$, then we can calculate the required sample size n. Alternatively, we can calculate the statistical power $(1-\beta)$ as a function of sample size n under a specified alternative hypothesis H_1 and a desired type-I error rate α. There are extensive literatures on sample size calculation as well as statistical power analysis, see the seminal books from [15]-[17] for power analysis for behavioral sciences.

As detailed in Chapter 7 in [17], five factors related to research design interplay with each other to determine the statistical power and sample size for a simple *t*-test: 1) the rate of type-I error α; 2) the desired statistical power $1-\beta$, 3) the expected treatment effect size of δ, 4) the standard error s^2 for the expected effect size, and 5) the sample size n. The mathematical formula can then be derived as $n \geq 2\left(s^2/\delta^2\right)\left[z_{1-\alpha} + z_{1-\beta}\right]^2$. Therefore, to determine the required sample size n, we would need to provide data for four of the five design characteristics. Typically, the type-I error α is set at 0.05 and the desired power $(1-\beta)$ is chosen to be 0.85 (or 0.80). The other two will be treatment effect size δ and its standard error s^2. Depending on actual research questions, different values are often selected for these two characteristics.

Extending the same concept described above for Guastello's polynomial cusp regression, we would need to specify the corresponding parameter effect size for all βs in Equation (2), the standard deviation of the error term ε. In addition, we need to specify the distribution of the two control variables, the asymmetry x and the bifurcation y; and the distribution of the outcome variable z at time 1 (*i.e.* z_1). With these parameters and variables being specified, the required sample size for a significant Cuastello's cusp regression model can be determined and statistical power can be analyzed.

3.2. Simulation-Based Approach for Power Analysis and Sample Size Determination

Power analysis and sample size determination can be developed for specific purpose. Typically, it is developed to detect treatment effect as in clinical trials or to detect the effect of specific risk factor as in regression. Similar development can be done to Guastello's cusp regression model for specific repressor in asymmetry variable

(x) or the bifurcation variable (y) if they are linked to multiple regressors or even to the overall goodness-of-fit index of R^2. However, we aim to tackle a more complicated problem to determine whether we can detect a significant overall cusp model. The complexity of cusp catastrophe model makes it rather challenging, if not impossible to derive an analytical formula to determine the statistical power for Guastello's cusp regression. To deal with this difficult, we propose a Monte-Carlo simulation-based approach. In this method the statistical power is calculated as the fraction of the times that the specified null-hypothesis of "no cusp" is rejected at the given level of type I error. Stated in another way, if there is a cusp, the statistical power will be, among 100 simulations, how many times can we detect the cusp given the sample size and type I error? The detailed steps of the simulation-based approach are outlined as follows:

1) Simulate data with sample size (n) (*i.e.* the number of observations for Guastello's cusp regression modeling) for the asymmetry variable x, bifurcation variable y and outcome variable at time 1 (*i.e.* z_1). Data are generated under required specifications for desired study, such as normal distribution with specific means and standard deviations. Guastello's cusp regression requires that all variables be standardized before data analysis and modeling. In this case, the standard normal distribution can be used to generate data for x, y and z_1;
2) Specify model parameter effect size $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and the standard deviation σ of the error term of ε (Equation (2)) obtained from prior knowledge;
3) Calculate $z_2 = z_1 + \beta_0 + \beta_1 z_1^3 + \beta_2 z_1^2 + \beta_3 y z_1 + \beta_4 x + \beta_5 y + \varepsilon$ using the data obtained in the previous two Steps. Also generate $\Delta z = z_2 - z_1$;
4) Fit the Guastello's cusp regression model (Equation (2)) with least squares method using the data generated for Δz, x, y, and z_1. After model fitting, a significant test is conducted to determine whether the data fit Guastello's cusp regression model satisfactorily according to the decision rules proposed by Guastello (1982): 1) the estimated β_1 for the cubic term and 2) β_3 (for the y and z_1 interaction term) or β_4 (for control variable x) must be are statistically significant;
5) Repeat Steps 1 to 4 a large number of times (typically 1000) and calculate the proportion of simulations which satisfy the Guastello's decision rules. This proportion then provides an estimate of the statistical power for the pre-specified sample size and the study specifications given in Steps 1 and 2;
6) With the above established five steps for power assessment, sample size is then determined to reach a pre-specified level of statistical power. This is carried out by running Steps 1 to 5 with a range of sample sizes (n) first to obtain the corresponding values of statistical power. Then a statistical power curve is constructed for these ranges of sample sizes. With this power curve, the sample size is determined through back-calculation for a pre-specified power, such as power = 0.85.

The simulation-based approach described above is implemented in free R package and the computer program is available up request from the authors.

4. Simulation Study and Real Example

4.1. Monte-Carlo Simulation Analysis

4.1.1. Rationale

To verify the novel approach proposed in Section 3, we simulated four scenarios with $n = 100$ observations for each using Guastello's cusp polynomial regression model (2). The four scenarios represent four cases of σ with different measurement errors (*i.e.* $\sigma = 1$, 2, 3, and 4). We hypothesized that data with smaller measurement errors will fit the cusp model better than the data with larger errors if the Guastello's cusp polynomial regression method is used to detect cusp catastrophic changes. Consequently, a larger sample size would be needed to detect a cusp for data with greater measurement errors.

4.1.2. Data Generation

Data are generated with the asymmetry variable x, bifurcation variable y and outcome variable at time 1 (*i.e.* z_1) being set as standard normal distribution. The parameter effect size vector is set as $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$. To illustrate the impact of measurement errors on sample sizes, we generate the error term ε following the normal distribution as $\varepsilon \sim N(0, \sigma^2)$ with increasing measurement error standard deviation of $\sigma = 1$, 2, 3, and 4 for each of the four scenarios.

With the generated x, y and z_1 along with the input values of β and σ, Δz is generated using the Guastello's polynomial regression model. This is achieved by plugging in all values of x, y, z_1, β, σ and ε into the following equation:

$$\Delta z = \beta_0 + \beta_1 z_1^3 + \beta_2 z_1^2 + \beta_3 y z_1 + \beta_4 x + \beta_5 y + \varepsilon$$

Figure 2 illustrates one realization of the data generation with $\sigma = 1$ in a pair plot. It can be seen from the figure that the distributions for x, y and z^1 are random (the upper left 3 by 3 plots). Furthermore, Δz is linearly related to x as seen from the upper right plot. The second plot on the right-side illustrates the linear relationship between Δz and y under fixed z^1 and the third plot on the right-side illustrates the cubic relationship between Δz and z^1. For $\sigma = 2$, 3, and 4 (data not shown in figure), the corresponding pair plots would have larger variations.

4.1.3. Simulation Analysis

Four data sets for the four scenarios (e.g., $\sigma = 1$, 2, 3, and 4) are simulated first. The simulated data are then fitted with Guastello's cusp regression model using least squares method. The summary statistics of the analyses are given in Table 1. It can be seen from the table that for the Scenario where $\sigma = 1$, all the parameters of the polynomial regression model are statistically highly significant $(p < 0.001)$ with $R^2 = 0.763$, indicating adequate data-cusp model fitting and F-statistic = 60.71 indicating highly significance of the polynomial regression model. The estimated $\hat{\sigma} = 1.053$, slightly greater than the true $\sigma = 1$. Since β_1, β_3 and β_4 are all highly significant, we conclude that the Guastello's polynomial regression method is sufficient to detect the specified cusp.

Results of other three scenarios in **Table 1** indicate that as σ increases, the goodness of data-model fitting declines. In the scenario where $\sigma = 2$, the R^2 drops to 0.454, F-statistic drops to 15.61 (still significant), and the estimated $\sigma = 2.107$, close to the true $\sigma = 2$. In this case, both β_1 and β_3 remain significant, indicating

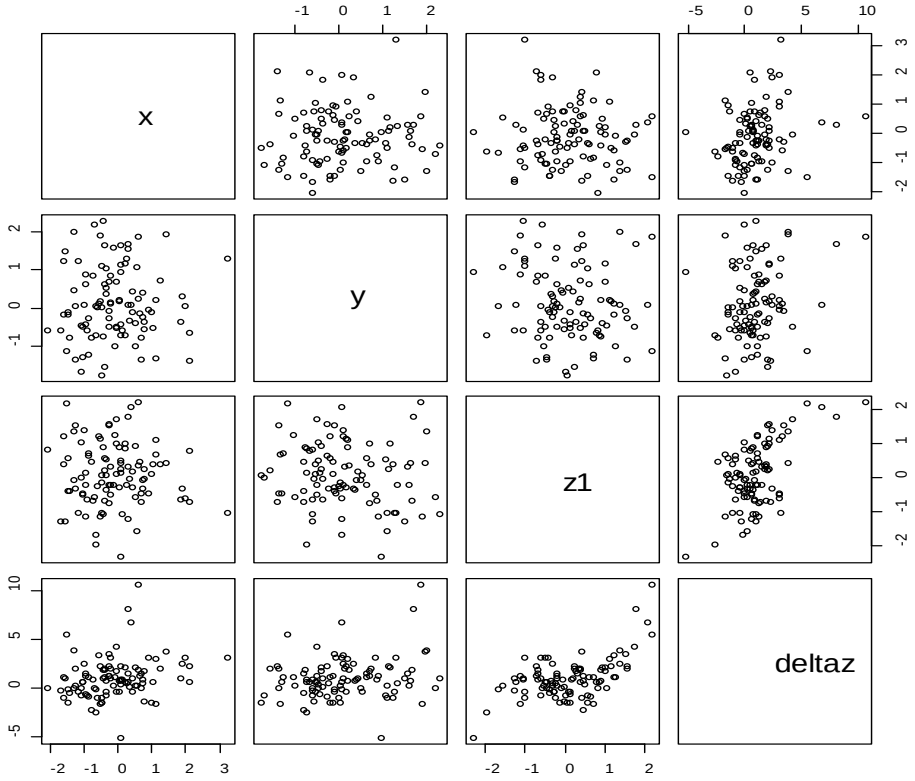


Figure 2. Example of simulated data when $\sigma = 1$ where the distributions of x y, z_1 are standard normal (the upper left 3 by 3 plots) and the relationships between Δz to x (as linear), to y (as linear) and to z_1 (as cubic).

the existence of a cusp. With regard to Scenario 3 where $\sigma = 3$, the R^2 further drops to 0.278 and F-statistic to 7.227. The estimated $\sigma = 3.160$, again close to its true $\sigma = 3$. In this case, only β_1 is highly significant and β_3 marginally significant, indicating that a cusp is likely. In Scenario 4 where $\sigma = 4$, none of the estimated parameters required to support the cusp is statistically significant. Therefore, we could not be able to determine if the data contain a cusp. A power analysis is needed to assess if the sample size $(n = 100)$ is adequate.

4.1.4. Sample Size Estimation

To demonstrate the proposed novel simulation method, we estimate sample sizes needed for each of the four scenarios to achieve 85% statistical power employing this method and the estimated parameter $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and the estimated σ from **Table 1** in the previous step. **Figure 3** summarizes the results. Data in **Figure 3** indicate that with 85% statistical power to detect the underlying cusp, the required sample sizes for Scenarios 1 through 4 are 36, 101, 195 and 293, respectively. The required sample size varies proportionately with measurement errors. This result adds more evidence supporting the validity of the simulation-based approach we proposed for power analysis.

4.1.5. Reverse-Verification

If the novel simulation-based approach is valid, the sample size estimates for each of the four scenarios described in previous section will allow approximately 85% chance to detect the underlying cusp. Therefore, we took a reverse approach to compute statistical power by applying the calculated sample size as input for each of the four scenarios. Results in **Figure 3** indicated that for Scenario 1, a sample size of 36 observations will be adequate to detect the cusp with 85% statistical power.

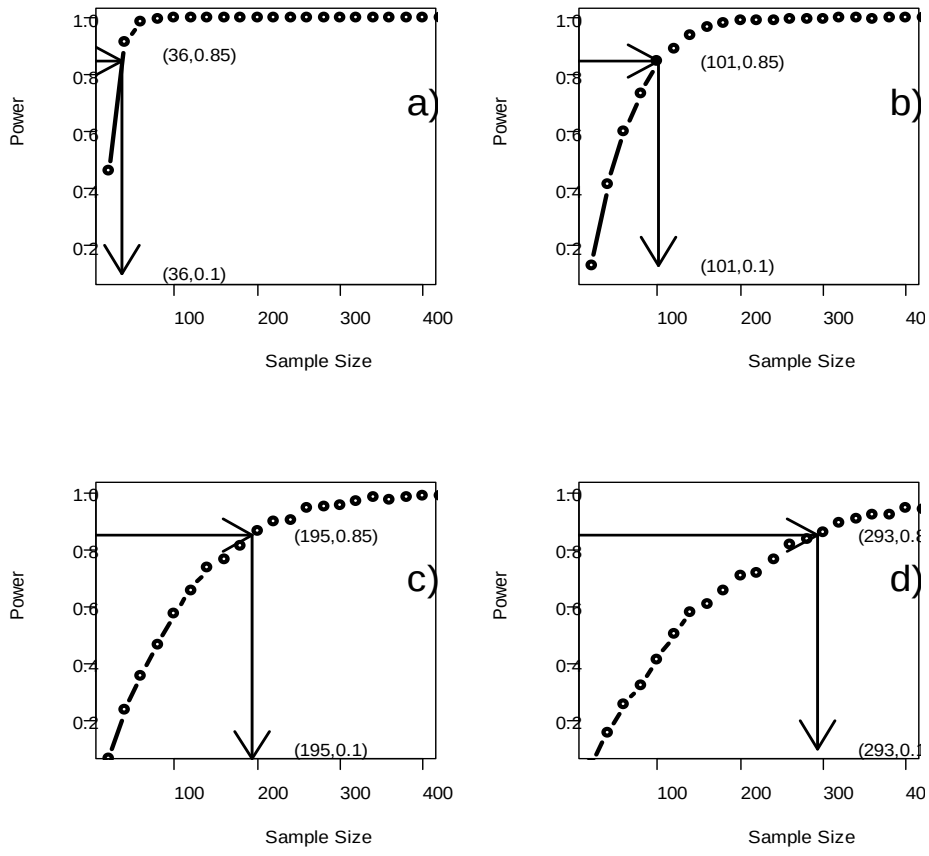


Figure 3. Statistical power curves corresponding to $\sigma = 1$ in plot a), $\sigma = 2$ in plot b), $\sigma = 3$ in plot c) and $\sigma = 4$ in plot d). The arrows illustrate the sample size determination from power of 0.85 to calculate the sample size required.

Table 1. Parameter estimates, R^2, Estimated σ^2 and F-Statistic from four simulations with $\sigma = 1$, 2, 3 and 4. The rows bolded are corresponding to the cusp determination.

	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$
β_0 (Intercept)	0.487***	0.473.	0.459	0.446
$\beta_1(z_1^3)$	**0.540*** **	**0.581*** **	**0.621*** **	**0.661*** **
$\beta_2(z_1^2)$	0.456***	0.411*	0.367	0.323
$\beta_3(y^* z_1)$	**0.360****	**0.221**	**0.081**	**−0.058**
$\beta_4(x)$	**0.563*** **	**0.626****	**0.689***	**0.753**
$\beta_5(y)$	0.468***	0.435.	0.403	0.371
R^2	0.763	0.454	0.278	0.1856
Estimated σ^2	1.053	2.107	3.160	4.214
F-Statistic with df = (5, 94)	60.71***	15.61***	7.227***	4.286*

Significant codes: *** p-value < 0.00001, ** p-value < 0.001, * p-value < 0.01, "." (p-value < 0.05).

To demonstrate this result, we make use Monte-Carlo procedure and randomly sample 36 observations from the simulate data $(n = 100)$ used for Scenario 1 $(\sigma = 1)$. We then fit the data to the Guastello's cusp regression model. We use the same criteria (significant β_1, plus either β_3 or β_4) to assess the detection of a cusp. Among 1000 repeats of the Monte-Carlo simulations with sample size $n = 36$, we found 833 times (83.3%) significant. This result indicates that the power analysis of the simulation method we proposed is close to 85%. In another word, the method we proposed is slightly conservative, which is good for research design. The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do NOT post your job titles, positions, academic degrees, zip codes, names of building/street/district/province/state, etc.). This template was designed for two affiliations.

4.2. Verification with Published Data

The best approach to demonstrate the validity of the simulation approach would be to test it with observed data. To use our approach, we need two sets of data from any reported study: parameter estimates as effect size $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and estimated mean error of model fitting $\hat{\sigma}$. However, we experienced difficulties in finding such data from all the studies we accessed in the published literature database. For example, all β coefficients were reported by all studies but β_0 was not; furthermore, data-model fitting error fitting $\hat{\sigma}$ was never reported in any of the published studies using Guastelle's cusp polynomial regression method. Fortunately, one author of this paper [12] published a study that modeled early sexual initiation among young adolescents using this polynomial regression approach.

Briefly, in Chen's study participants were 469 virgins in the control group for a randomized controlled trial to assess the effect of an HIV behavioral prevention intervention program [22] [23]. The participants in grade 6 in the Bahamian public schools were randomly assigned to receive either intervention or control conditions. They were followed every 6 months up to 24 months at the time when the analysis was conducted. A participant was categorized as having initiated sex if he or she had the first penile-vagina sexual intercourse during the follow-up period. In addition to sexual initiation, the likelihood to initiate sex was also assessed using a 5-point rating scale with 1 = very unlikely to have sex in the next 6 months and 5 = very likely to have sex. A sexual progression index (SPI) was thus created as the dependent variable for modeling analysis was defined as the first time. SPI = 1 for participants who never had sex and reported very unlikely to have sex; SPI = 2 for participants who never had sex but unsure if they are going to have sex in the next 6 months; SPI = 3 for participants who never had sex but reported very like to have sex in the next 6 months; and SPI = 4 for participants who initiated sex. In addition to SPI, age was used as the asymmetry variable x, and self-efficacy not to have sex (scale score based on 5 items) was used as the bifurcation y.

To verify the simulation-based method, the parameter effect size estimates were obtained from the paper with $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = c(-0.0309, 0.0726, -0.4819, -0.1236, 0.0613, -0.2693)$, and the data-model fitting error $\hat{\sigma} = 0.5033$ was obtained by accessing to the original computing records. With these estimates, the simulation-based approach in Section 3.2 is applied. **Figure 4** presents the sample size-power curve. From the figure it can be seen that the estimated sample size is 153 to achieve 85% power. This sample size is much smaller than the sample $(n = 469)$ in the original study.

5. Discussions

In the case where analytical solution to power analysis and sample size determination is difficult, simulation represents an ideal alternative as recommended in [16] [17] [24]. In this paper, we reported a novel simulation-based approach we developed to estimate the statistical power and to compute sample size for Gustello's polynomial cusp catastrophe model. The method was developed based on statistical power theory and our understanding of Guastello's cusp polynomial regression modeling approach. The computing method is programmed using the R software. Results from 1000 repeats of Monte Carlo simulation and empirical data analysis suggest that the method we proposed is valid and can be used in practice to conduct power analysis and to estimate sample size for Guastello polynomial cusp modeling method.

With this approach, researchers can compute statistical power and estimate sample size if they plan to conduct cusp modeling analysis using Gustallo's polynomial regression method. A detailed introduction to the method can be found in [6] [7] [21]. Data needed for our methods included parameter effect size estimates for the intercept and five model parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and a data-model fitting error σ or its estimate. With the specification of these data, power can be computed for any given sample sizes. In addition to computer power, the commonly used sample size-power curve can be generated to provide a visual presentation between sample size and statistical power. With such power curve, sample size can be estimated for specified power in design and analysis data from cusp catastrophe model.

To make the presentation easier, we confined this novel simulation approach to the situation of one regressor for each control variable in the cusp model. This approach can be easily adopted and extended to multiple regressors for each of the asymmetric (x) and bifurcation (y) variables where the Guastello's cusp polynomial regression model would need to be extended.

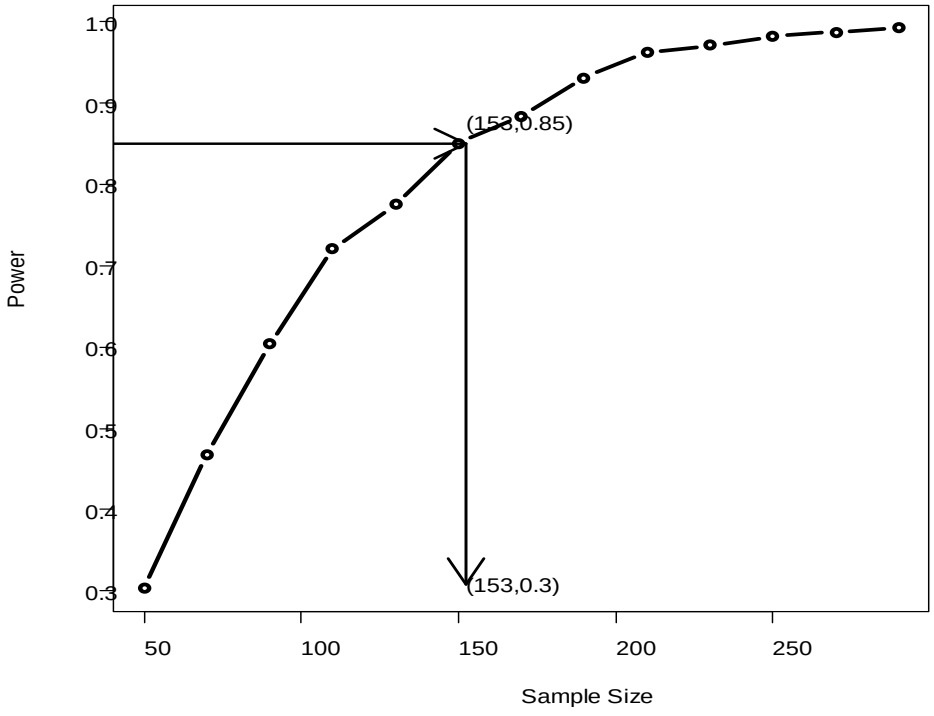


Figure 4. Power curve for Chen *et al*. (2010). The estimated sample size for power of 0.85 is 153.

More and more data suggest the utility of cusp modeling approach in characterizing a number of human behaviors, particularly health risk behaviors, such as tobacco smoking, alcohol consumption, hardcore drug use, dating violence, and unprotected sex [10] [11] [14] [21] [25] [26]. The methods we reported in this paper provide a useful tool for researchers to more effectively design their research to investigate these risk behaviors and to assess intervention programs for risk reduction.

By conducting this study, we also note that previous studies published in the literature do not report adequate information for power analysis. We highly recommend that journal editors ask authors to report all parameter estimates, including β_0, and data-model fitting error (mean square of error). In addition to power analysis and sample size estimation, such data are also useful for readers to statistically assess appropriateness of the reported results.

There are a number of strengths with the method we present in this study. The principle and the computing process are not difficult to follow; the data used for the computing can be obtained; the computing software is written with R, available from the authors by request for collaboration; and the computing does not require much time (several seconds to half minutes). We are encouraged on the results from this research and work on extending the results into stochastic catastrophe model in [4] [19]. Despite many advantages, further application of the method in practice is needed.

Acknowledgements

This research was support in part by two NIH grants, one from the National Institute On Drug Abuse (NIDA, R01 DA022730, PI: Chen X) and another from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD, R01HD075635, PIs: Chen X and Chen D).

References

[1] Thom, R. (1975) Structural Stability and Morphogenesis. Benjamin-Addison-Wesley, New York.

[2] Thom, R. and Fowler, D.H. (1975) Structural Stability and Morphogenesis: An Outline of a General Theory of Models. W. A. Benjamin, Michigan.

[3] Cobb, L. and Ragade, R.K. (1978) Applications of Catastrophe Theory in the Behavioral and Life Sciences. *Behavioral Science*, **23**, 291-419. http://dx.doi.org/10.1002/bs.3830230511

[4] Cobb, L. and Watson, B. (1980) Statistical Catastrophe Theory: An Overview. *Mathematical Modelling*, **1**, 311-317. http://dx.doi.org/10.1016/0270-0255(80)90041-X

[5] Cobb, L. and Zacks, S. (1985) Applications of Catastrophe Theory for Statistical Modeling in the Biosciences. *Journal of the American Statistical Association*, **80**, 793-802. http://dx.doi.org/10.1080/01621459.1985.10478184

[6] Guastello, S.J. (1982). Moderator Regression and the Cusp Catastrophe: Application of Two-Stage Personnel Selection, Training, Therapy and Program Evaluation. *Behavioral Science*, **27**, 259-272. http://dx.doi.org/10.1002/bs.3830270305

[7] Guastello, S.J. (1989) Catastrophe Modeling of the Accident Processes: Evaluation of an Accident Reduction Program Using the Occupational Hazards Survey. *Accident Analysis and Prevention*, **21**, 61-77. http://dx.doi.org/10.1016/0001-4575(89)90049-3

[8] Grasman, R.P., van der Mass, H.L. and Wagenmakers, E. (2009) Fitting the Cusp Catastrophe in R: A Cusp Package Primer. *Journal of Statistical Software*, **32**, 1-27.

[9] Clair, S. (1998) A Cusp Catastrophe Model for Adolescent Alcohol Use: An Empirical Test. *Nonlinear Dynamics, Psychology, and Life Sciences*, **2**, 217-241. http://dx.doi.org/10.1023/A:1022376002167

[10] Mazanov, J. and Byrne, D.G. (2006) A Cusp Catastrophe Model Analysis of Changes in Adolescent Substance Use: Assessment of Behavioural Intention as a Bifurcation Variable. *Nonlinear Dynamics, Psychology, and Life Sciences*, **10**, 445-470.

[11] Guastello, S.J., Aruka, Y., Doyle, M. and Smerz, K.E. (2008) Cross-Cultural Generalizability of a Cusp Catastrophe Model for Binge Drinking among College Students. *Nonlinear Dynamics, Psychology and Life Sciences*, **12**, 397-407.

[12] Chen, X., Lunn, S., Harris, C., Li, X., Deveaux, L., Marshall, S., *et al.* (2010) Modeling Early Sexual Initiation among Young Adolescents Using Quantum and Continuous Behavior Change Methods: Implications for HIV Prevention. *Nonlinear Dynamics, Psychology and Life Sciences*, **14**, 491-509.

[13] Wagner, C.M. (2010) Predicting Nursing Turnover with Catastrophe Theory. *Journal of Advanced Nursing*, **66**, 2071-2084.

[14] Chen, X., Lunn, S., Deveaus, L., Li, X., Brathwaite, N., Cottrell, L. and Stanton, B. (2008) A Cluster Randomized

Controlled Trial of an Adolescent HIV Prevention Program among Bahamian Youth: Effect at 12 Months Post-Intervention. *AIDS and Behavior*, **13**, 495-508.

[15] Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences. 2nd Edition, Lawrence Berbaum Associates, Hillsdale.

[16] Chow, S., Shao, J. and Wang, H. (2008) Sample Size Calculations in Clinical Research. 2nd Edition, Chapman and Hall/CRC, Boca Raton.

[17] Chen, D.G. and Peace, K.E. (2011) Clinical Trial Data Analysis Using R. Chapman and Hall/CRC, Boca Raton.

[18] Saunders, P.T. (1980) An Introduction to Catastrophe Theory. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9781139171533

[19] Hartelman, A.I. (1997) Stochastic Catastrophe Theory. University of Amsterdam, Amsterdam.

[20] Iacus, S.M. (2008) Simulation and Inference for Stochastic Differential Equations with R Examples. Springer, Berlin. http://dx.doi.org/10.1007/978-0-387-75839-8

[21] Guastello, S.J. and Gregson, A.M. (2011) Nonlinear Dynamic Systems Analysis for the Behavioral Sciences Using Real Data. CPC Press, Boca Raton.

[22] Gong, J., Stanton, B., Lunn, S., Devearus, L., Li, X., Marshall, S., Brathwaite, N.V., Cottrell, L., Harris, C. and Chen, X. (2009) Effects through 24 Months of an HIV/AIDS Prevention Intervention Program Based on Protection Motivation Theory among Preadolescents in the Bahamas. *Pediatrics*, **123**, 917-928. http://dx.doi.org/10.1542/peds.2008-2363

[23] Chen, X., Stanton, S., Chen, D.G. and Li, X. (2013) Is Intention to Use Condom a Linear Process? Cusp Modeling and Evaluation of an HIV Prevention Intervention Trial. *Nonlinear Dynamics, Psychology and Life Sciences*, **17**, 385-403.

[24] Bolker, B. (2008) Ecological Models and Data in R. Princeton University Press, Princeton.

[25] Mazanov, J. and Byrne, D.G. (2008) Modeling Change in Adolescent Smoking Behavior: Stability of Predictors across Analytic Models. *British Journal of Health Psychology*, **13**, 361-379. http://dx.doi.org/10.1348/135910707X202490

[26] West, R. and Sohal, T. (2006) "Catastrophic" Pathways to Smoking Cessation: Findings from National Survey. *British Medical Journal*, **332**, 458-460. http://dx.doi.org/10.1136/bmj.38723.573866.AE

Open Journal of Statistics, 2014, 4, 814-825
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410077

Regularization and Estimation in Regression with Cluster Variables

Qingzhao Yu[1], Bin Li[2]

[1]School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA, USA
[2]Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA, USA
Email: qyu@lsuhsc.edu, bli@lsu.edu

Received 14 October 2014; revised 5 November 2014; accepted 15 November 2014

Abstract

Clustering Lasso, a new regularization method for linear regressions is proposed in the paper. The Clustering Lasso can select variable while keeping the correlation structures among variables. In addition, Clustering Lasso encourages selection of clusters of variables, so that variables having the same mechanism of predicting the response variable will be selected together in the regression model. A real microarray data example and simulation studies show that Clustering Lasso outperforms Lasso in terms of prediction performance, particularly when there is collinearity among variables and/or when the number of predictors is larger than the number of observations. The Clustering Lasso paths can be obtained using any established algorithm for Lasso solution. An algorithm is proposed to construct variable correlation structures and to compute Clustering Lasso paths efficiently.

Keywords

Clustered Variables, Lasso, Principal Component Analysis

1. Introduction

We are often interested in finding important variables that are significantly related to the response variable and can be used to predict quantities of interest in regressions and classification problems. Important variables are often shown in clusters where variables in the same cluster are highly correlated and have similar pattern relating to the response variable. For example, a major application of microarray technology is to discover important genes and pathways that are related to clinical outcomes such as the diagnosis of a certain cancer. Typically, only a small proportion of genes from a huge bank have significant influence on the clinical outcome of interest. In addition, expression data frequently have cluster structures: the genes within a cluster often share

the same pathway and are therefore similarly related to the outcome. When regression is adapted in this setting, we often face the challenge from multi-collinearity of covariates. An ideal variable selection procedure should be able to find all genes of important clusters rather than just some representative genes from the clusters. Typically, two characteristics, pointed out by [1], evaluate the quality of a fitted model: accuracy of prediction on new data and interpretation of the model. For the latter, the sparse model with fewer selected covariates is preferred for interpretation due to its simplicity. However, when multiple variables share the same mechanism for explaining the response, all the involved variables should have an equal chance of being selected, and should exhibit the same relationship to the outcome in the fitted model, for scientific reasoning.

It is well known that the ordinary least square estimate (OLS) in linear regression often performs poorly when some of the predictors are highly correlated. OLS would generate unstable results where the estimates have inflated variances. Regularizations have been proposed to improve OLS. For example, ridge regression [2] penalizes the model complexity by the l_2 penalty of the coefficients. This method was proposed to solve the collinearity problem by adding a constant to the diagonal terms of $X'X$, where X is the observation or design matrix. Ridge regression stabilizes the estimates through the bias-variance trade-off. It can often improve the predictions but cannot select variables. [3] proposed the Lasso method by imposing an l_1-penalty on the regression coefficients. Lasso is a promising method, as it can improve prediction and produce sparse models simultaneously. However, when high correlations among predictors are present, the predictive performance of Lasso is dominated by ridge regression [3]. Moreover, when there is a cluster of variables, in which each variable associates with the response variable similarly, Lasso tends to arbitrarily select one variable from the cluster instead of identifying the cluster [1]; see also Section 2 for more discussion. Elastic Net, proposed by [1], combines both l_1 and l_2 penalties of the coefficients as the regularization criterion. The method is promising in that it encourages cluster effects and shows improved predictive performance over Lasso. Elastic Net can automatically choose cluster variables and estimate parameters at the same time. Many other methods can be used to choose clustered variables, such as principal component analysis (PCA). [4] defined "eigen-arrays" and "eigen-genes" in this way. But PCA can not choose sparse models. [5] proposed sparse principal component analysis (SPCR), which formulated PCA as a regression-type optimization problem, and then obtained sparse loadings by imposing the Elastic Net constraint. SPCR can successfully yield exact zero loadings in principal components. However, for each principal component, a regularization parameter has to be selected, which results in an overwhelming computational burden when the number of parameters is large. Other penalized regression methods have been proposed for group effect [6]-[13]. However, these methods either pre-suppose a grouping structure or assume that each predictor in a group shares an identical regression coefficient.

In practice, we often have some prior knowledge about the structure of variables and would like to make use of a priori information in analysis. For example, in gene analysis, we know the pathways and genes involved in these pathways. Therefore, we would like to group the involved variables in the same pathway together. Another example is in spatial analysis, we would like to keep a certain correlation structure among the spatial error terms. For example, sometimes we would like to fit a different coefficient for a certain variable at different regions (e.g., if the variable has different effect at different regions) but keep a correlation structure among the coefficients at neighborhood regions. The conditional autoregressive model (CAR, [14]) is one of the methods that can be used to keep such correlation structure.

In this paper, we propose a method that encourages cluster variables to be selected together and can incorporate available prior information on coefficient structures in variable selection. When there is no prior information on coefficient structure, we propose a data augmentation algorithm to find the structure. Moreover, the method uses the Lasso regularization to choose sparse models. The proposed method can be solved by any efficient Lasso algorithm such as least angle regression (LARS, [15]) and the coordinate-wise descent algorithm (CDA, [16]). We call our method the Clustering Lasso (CL).

The rest of the paper is organized as follows. In Section 2, we review the Lasso method and discuss its limitation in identifying clustered variables. Then we propose the Clustering Lasso in a Bayesian setting. Its counterparts in the Frequentist setting and computational strategies are discussed in Section 3. Sections 4 and 5 demonstrate the predictive and explanatory performance of CL through real examples and simulations. Finally, conclusions and future work are discussed in Section 6.

2. Clustering Lasso in Bayesian Setting

Consider linear regression settings with the response vector $\boldsymbol{y} = (y_1, \cdots, y_n)'$ and $n \times p$ dimensional input

matrix $\boldsymbol{X}$. The $\boldsymbol{y}$ and columns of $\boldsymbol{X}$ are centered and standardized to have the same l_2 norm. The Lasso estimates $\tilde{\boldsymbol{\beta}}_{\text{lasso}}$ are calculated by minimizing

$$\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j| \tag{1}$$

The solution of Lasso can be obtained through LARS or CDA. Compared with ordinary linear regressions, Lasso shows superior predictive performance and more stable estimates. Moreover, Lasso can often select variables and estimate coefficients simultaneously.

Group effect has been defined by [1] in the linear regression setting. Let x_i be the i th predictor. The estimates of coefficients have the group effect if $x_i = x_j$ would result in the estimated coefficients $\hat{\beta}_i = \hat{\beta}_j$. [1] further proved that if the solution for estimation is to minimize the objective function of the form:

$$\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda J(\boldsymbol{\beta}) \tag{2}$$

and the penalty term, $J(\cdot)$, is strictly convex, then the estimates from Equation (2) enjoy the group effect property. In Lasso, $J(\cdot)$ is l_1 norm of $\boldsymbol{\beta}$, which is not strictly convex. Zou and Hastie proved that in this case Lasso estimates do not have the group effect. This is also understandable through the Lasso solution path from LARS. In LARS, suppose a variable x_i is selected in the model. Its coefficient solution path will move in a direction to reduce the correlation between x_i and the current residual, $\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}$, until another variable, say x_k, has the same correlation to the current residual as does x_i. At this point, variable x_k is added into the model. If x_j is highly correlated with x_i, when the correlation between x_i and the residual decreases, so does that between x_j and the residual. Therefore, if x_i has been included in the model, Lasso is less likely to select the highly correlated variable x_j in the model. Consequently, Lasso cannot select clustered variables.

In a Bayesian setting, if $\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}$, [3] showed that the Lasso solution is identical to the posterior mode of the coefficients when the prior distributions of the coefficients are set as independent double exponential distributions, where

$$y_i|\boldsymbol{x}_i,\boldsymbol{\beta},\sigma \sim N\left(\boldsymbol{x}_i'\boldsymbol{\beta},\sigma^2\right), \quad \text{for } i=1,\cdots,n,$$

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right)=\prod_{j=1}^{p}\frac{\lambda}{2\sigma^2}\mathrm{e}^{\lambda|\beta_j|/\sigma^2}.$$

In Lasso, the penalty term of model complexity is $\sum_{j=1}^{p}|\beta_j|$. Because each coefficient is penalized equally, each one can be shrunk to zero independently. When the variables are clustered, an ideal solution path should be that the clustered variables are selected together. Therefore, we would like to penalize the coefficients with a restriction that keeps the correlation structure among the variables. With the penalization, if the coefficient of one variable is nonzero, those variables in the same cluster are less likely to be zero. For this purpose, we assume a correlation structure, specified as the structural correlation matrix $\boldsymbol{R}$, of the coefficients $\boldsymbol{\beta}$.

For simplicity, assume that the variance of the random error, σ^2 $(\sigma>0)$, and the structural correlation matrix $\boldsymbol{R}$ are known. Then the likelihood and prior distributions can be set as:

$$y_i|\boldsymbol{x}_i,\beta,\sigma \sim N\left(\boldsymbol{x}_i'\beta,\sigma^2\right), \quad \text{for } i=1,\cdots,n,$$

$$\boldsymbol{\beta}=\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\beta}^*, \text{ and}$$

$$\pi\left(\boldsymbol{\beta}^*\middle|\sigma^2\right)=\prod_{j=1}^{p}\frac{\lambda}{2\sigma^2}\mathrm{e}^{\lambda|\beta_j^*|/\sigma^2}.$$

Therefore, the posterior distribution of $\boldsymbol{\beta}$ has the form

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{R}) &\propto l(\boldsymbol{y}|\boldsymbol{\beta})\cdot\pi\left(\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{\beta}\right)\left|\boldsymbol{R}^{-\frac{1}{2}}\right| \\ &\propto \exp\left(-\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2\Big/2\sigma^2\right)\exp\left(-\lambda\left\|\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{\beta}\right\|_1\Big/\sigma^2\right),\end{aligned} \tag{3}$$

with a vector $V' = (V_1, \cdots, V_p)$, $\|V\|_1 = \sum_{i=1}^{p} |V_i|$. The posterior mode of β in the distribution (3) is the solution to

$$\operatorname{argmin}_{\beta} \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \left\| \boldsymbol{R}^{-\frac{1}{2}} \boldsymbol{\beta} \right\|_1 \tag{4}$$

Relating Equation (4) to the Bayesian Lasso solution to (1), we naturally infer the Clustering Lasso in Frequentist setting.

3. Clustering Lasso

3.1. Clustering Lasso and Its Grouping Effect

In Frequentist setting, we modify the penalization function in Lasso to retain a presumed correlation structure among coefficients. Let $\boldsymbol{\beta}^* = \boldsymbol{R}^{-\frac{1}{2}} \boldsymbol{\beta}$ and β_j^* be the jth element of $\boldsymbol{\beta}^*$. The Clustering Lasso estimate is defined as the solution to

$$\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j^*| \tag{5}$$

where $\lambda \geq 0$ is the regularization parameter. Note that instead of restricting $\sum |\beta_i|$, we restrict $\sum |\beta_i^*|$. Therefore, $\boldsymbol{\beta}$'s are not penalized independently and clustered variables could be chosen. Let $\boldsymbol{r}_j^{-\frac{1}{2}}$, with dimension $1 \times p$, be the jth row of $\boldsymbol{R}^{-\frac{1}{2}}$ and let $\boldsymbol{R}_j^{-1} = \left(\boldsymbol{r}_j^{-\frac{1}{2}}\right)' \boldsymbol{r}_j^{-\frac{1}{2}}$, a $p \times p$ matrix. The penalty term used in expression (5) can also be written as $\lambda \sum_{j=1}^{p} \left(\boldsymbol{\beta}' \boldsymbol{R}_j^{-1} \boldsymbol{\beta}\right)^{1/2}$, which is intermediate between the l_1 penalty and the l_2 penalty. When $\boldsymbol{R}$ is an identity matrix, the Clustering Lasso is identical to the ordinary Lasso method. Otherwise, the penalty function is strictly convex. Using Lemma 2 developed by [1], the solution to Expression (5) has the group effect. Therefore, Clustering Lasso can select variables by clusters.

Figure 1 illustrates the Clustering Lasso penalty contours with two predictors. The right figure shows the penalty contour when the two predictors are correlated and the left one shows the contour when the two predictors are independent, which is identical to the Lasso method. The sums of the squared errors have elliptical contours, centered and minimized at the full least squares estimate. The constraint region of Lasso is the diamond region $|\beta_1| + |\beta_2| \leq c$, while that for the Clustering Lasso is the parallelogram region defined by $\left|\boldsymbol{r}_1^{-\frac{1}{2}} \boldsymbol{\beta}\right| + \left|\boldsymbol{r}_2^{-\frac{1}{2}} \boldsymbol{\beta}\right| \leq c$. The optimal estimates are realized at the place where the elliptical contours first hit the constraint regions. The sides of the parallelogram are decided by the structural correlation matrix $\boldsymbol{R}$.

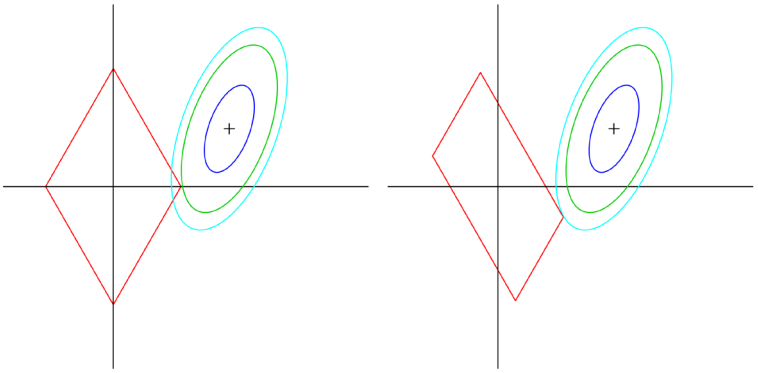


Figure 1. Estimation picture for the Clustering Lasso when two predictors are independent (left, as lasso) and when two predictors are clustered (right).

3.2. Computation

The Clustering Lasso is an extension of the Lasso method. Let $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{R}^{\frac{1}{2}}$. So the solution to Expression (5) is $\tilde{\boldsymbol{\beta}} = \boldsymbol{R}^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}^*$, where $\tilde{\boldsymbol{\beta}}^*$ is

$$\operatorname{argmin}_{\beta^*} \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right)'\left(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right) + \lambda\sum_{j=1}^{p}\left|\beta_j^*\right| \tag{6}$$

Therefore, all the established algorithms for Lasso solution, such as the least angle regression (LARS, [15]), could be used for Clustering Lasso.

3.3. The Clustering Lasso Algorithm

We can incorporate prior knowledge of clustering into a structural correlation matrix. For example, Kyoto Encyclopedia and Genes and Genomes (KEGG) and many other biological databases can be referred to in gene analysis to construct the structural correlation matrix. It is required that the structural correlation matrix be symmetric. When no prior information is readily adaptable, a natural method is to use the modified correlation matrix of the observed data, meaning that the coefficients should have a correlation structure that is similar to how the covariates are correlated. There are several well-established potential choices such as partial correlation matrix [17]. In this paper, we propose to use a modified correlation matrix so that if two variables x_i and x_j are not significantly correlated, R_{ij}, the ith row and jth column element of R, is set to be zero. As the solution for β is $\tilde{\boldsymbol{\beta}} = \boldsymbol{R}^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}^*$, zero elements in $\boldsymbol{R}^{\frac{1}{2}}$ are desired so that when β_i^* s are shrunk to zero, which is possible by the Lasso property, some β_j s could also be shrunk to exact zero.

In detail, we develop Algorithm 1—the Clustering Lasso algorithm. Let p_{val}, p_2, and m, in $[0,1]$ be three prespecified numbers, and $\boldsymbol{CORR}$ be a $p \times p$ matrix.

Algorithm 1 *Clustering Lasso*

1. For $i = 1, \cdots, p-1$

for $j = 2, \cdots, p$

do correlation test between x_i and x_j, let

$$\boldsymbol{CORR}[i,j] = \begin{cases} \operatorname{cor}\left(x_i, x_j\right) & \text{if } p \text{ value} < p_{\text{val}} \text{ and } \left|\operatorname{cor}\left(x_i, x_j\right)\right| \geq m \\ 0 & o.w. \end{cases}$$

2. Do eigen decomposition on $\boldsymbol{CORR}$ so that $\boldsymbol{CORR} = \boldsymbol{UQU}'$ and let $Q_{ii} = 0$ if $\dfrac{Q_{ii}}{\sum_{j=1}^{p} Q_{jj}} < p_2$ for $i = 1, \cdots, p$.

3. Let $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{U}'$ and $\boldsymbol{X}^* = \boldsymbol{XT}$.

4. Do Lasso on $\left(\boldsymbol{y}, X^*\right)$ and get the coefficient solution $\tilde{\boldsymbol{\beta}}^*$.

5. $\tilde{\boldsymbol{\beta}} = T\tilde{\boldsymbol{\beta}}^*$ is the solution to Clustering Lasso.

Note that only when some elements of $\boldsymbol{T}$ are set to be zero, could β s be shrunk to exact zero when β^*'s are shrunk to zero by Lasso. A special case is when $\boldsymbol{R}$ is a block diagonal matrix. To choose sparse models, we need to identify clusters of covariates, where variables in the same cluster are assumed to be correlated while those from different clusters are independent. For this purpose, there are two shrinkage steps in Algorithm 1. Step (1) shrinks the correlation coefficients to zero if there is no significant correlation between the pair of covariates at the significance level p_{val} or if the magnitude of correlation is smaller than a pre-set value m. When two covariates are not correlated, there is little chance that the two variables relate to the response variable with the same underlying pathway. Therefore, the coefficients of the two variables can be estimated independently. Step (2) shrinks some eigen-values of $\boldsymbol{CORR}$ to zero if the corresponding eigenvector explains less than p_2 times the total variance of $\boldsymbol{CORR}$. The two shrinkage steps cannot guarantee that some elements of T be zero. Subjective intervention can help for this purpose. One resolution is to cluster the covariates first

and then calculate the correlation matrices for each cluster, which in turn used to build the diagonal blocks of $\boldsymbol{R}$. In addition to building a diagonal block matrix, another resolution is to adapt shrinkage methods in the eigen decomposition process of $\boldsymbol{R}$, so that some loadings of the eigenvectors might degenerate to 0. ScotLASS [18] and sparse principal component analysis (Zou *et al*., 2006) can serve this purpose. However, these methods require extra computations for each principal component, which brings in high computational costs. The nonzero elements of the jth row of $\boldsymbol{T}$ imply that the corresponding covariates belong to the jth cluster. Ideally, their values should be proportional to the contributions of each covariate to the cluster in explaining the outcome. As pointed out by a referee of the paper, clusters in the proposed method are identified by rows of $\boldsymbol{R}^{1/2}$, where $\boldsymbol{R}^{1/2}$ is defined as $\boldsymbol{UQ}^{1/2}\boldsymbol{U}'$ with $\boldsymbol{Q}$ being the diagonal matrix of eigenvalues and $\mathbf{U}$ columns of eigenvectors of $\boldsymbol{R}$. As in principal component analysis, the nonzero elements of $\boldsymbol{R}^{1/2}$ are difficult to interpret in practice. The referee recommends setting the elements of $\boldsymbol{R}^{1/2}$ to be 0 or 1 based on the absence or presence of non-zero elements, respectively. In the paper, we set the estimate of β to be zero, if its estimated value is very close to zero, *i.e.* if $|\beta_i| < 0.005$.

3.4. Choice of Tuning Parameters

Four parameters, $(p_{\text{val}}, m, \lambda, p_2)$, are to be chosen for Algorithm 1. p_{val} is the significance level used to decide whether the correlations between a pair of covariates should be considered to restrict the estimation of their coefficients. We usually select the significance level at 0.05, the traditional significance level. When the data set is large, we can reduce the significance level. Since the correlation would be always significant when little correlation exists and the number of observations is large, we set another restriction on the magnitude of correlation- m, above which we would like to use the correlation as a restriction to the coefficient parameters. m is chosen subjectively by researchers. Algorithm 1 Step (2) is similar to the principal component analysis except that the eigen decomposition is based on the correlation matrix modified by Step (1). p_2 specifies the minimum proportion of variance explained by the eigen vector, below that, the eigen vector will not be used for further analysis. p_2 is set at a small value, typically $0.01/p$, where p is the total number of covariates.

The last parameter to be tuned is λ. In Lasso, the conventional tuning parameter is the fraction (s) of the l_1-norm. There are well-established methods for choosing s. Tenfold cross-validation (CV) on training data is the method we used in this paper. The training dataset is divided into ten folds randomly. One fold of the data is used as validation data, on which the prediction error is calculated based on the model fitted from the other nine folds of data. s is tested on a fine grid on $[0,1]$. It takes the value that minimizes the averaged prediction error from CV. We can also use ten-fold CV to tune m and p_2. We found that only a few representative values for $m \times p_2$ need to be cross validated to obtain good results, which are $\{0, 0.5\} \times \left\{0, 0.05, \frac{0.01}{p}\right\}$.

4. Microarray Data Example

We used the proposed method on an Affymetrix gene expression dataset. The data were collected by Singh *et al*. [19] and consists of 12,600 genes, from 52 prostate cancer tumor samples and 50 normal prostate tissue samples. The goal is to construct a diagnostic rule based on the 12,600 gene expressions to predict the occurrence of prostate cancer. Support vector machine (SVM, [20]), Ridge, Lasso, Elastic Net, Weighted Fusion (w.fusion) and Clustering Lasso were all applied to this dataset. We tried four types of Clustering Lasso methods:

1. CL1: $p_{\text{val}} = 0.05$, $m = 0$, and $p_2 = 0$;
2. CL2: $p_{\text{val}} = 0.05$, $m = 0$, and $p_2 = 0.05$;
3. CL3: $p_{\text{val}} = 0.05$, $m = 0$, and $p_2 = \dfrac{0.01}{\#\text{ of covariates}}$;
4. CL4: $p_{\text{val}} = 0.05$, $m = 0.5$, and $p_2 = 0.05$.

To apply these methods, we first coded the presence of prostate cancer as a 0-1 (no and yes) response y. The classification function is I (fitted value > 0.5), where $I(\cdot)$ is the indicator function. For comparison, we randomly select 52 samples as training data, based on which the diagnostic rules are constructed, and the rules are in turn tested on the remaining 50 samples.

The dataset was split 20 times. For each repetition, a 1000-gene set was preselected based on the training data

to make the computation manageable. The genes are those "most significantly" related to the response, tested by individual *t*-statistics. **Figure 2** shows the boxplots of the misclassification rates on the test data sets from different classifiers. The misclassification rates are summarized in **Table 1**. Overall, the misclassification rate from Clustering Lasso is competitive with Elastic Net and Ridge, and is better than Lasso, Weighted Fusion, and SVM. For the computational time, Clustering Lasso is comparable to the Lasso method and is much more efficient than Elastic Net and Weighted Fusion. Within the four Clustering Lasso methods, the ones with more restrictions on eigenvalues and the magnitudes of correlations perform a little bit worse.

Table 2 shows the average number of genes selected from the 20 repetitions based on different methods. The analyses were based on 1000 genes and 52 observations. We see that Lasso selected fewer than 52 genes. Elastic Net eliminated few genes—the average number of selected genes was close to 1000. Cluster Lasso identified about 25% genes as important. However, we do not know whether the chosen genes are, in fact, important or not. The efficiency of variable selection is further assessed by simulation studies.

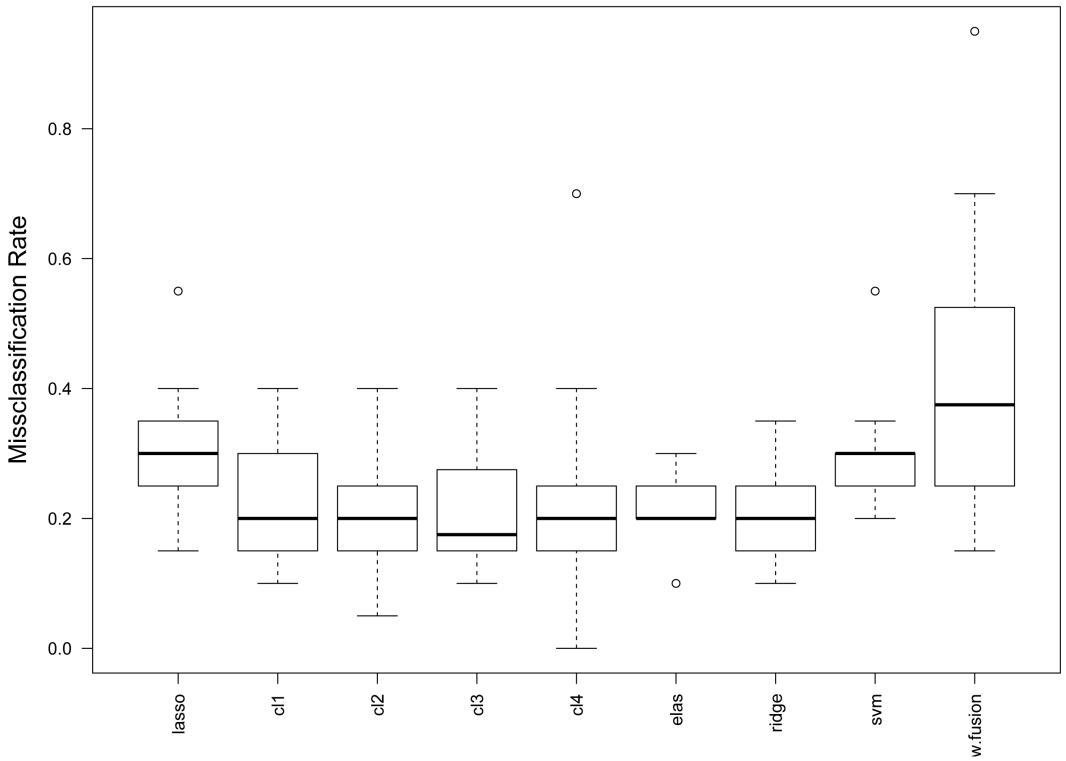


Figure 2. Misclassification rates on singh data. ELAS stands for Elastic Net.

Table 1. Summary of Misclassification Rates on Singh data.

Methods	SVM	Ridge	Elastic Net	Lasso	CL1	CL2	CL3	CL4	W. fusion
Mean	5.75	4.25	4.3	6.05	4.45	4.2	4.2	4.55	8.2
Median	6	4	4	6	4	4	3.5	4	7.5
SD	1.48	1.33	0.98	1.79	1.64	1.74	1.64	2.86	4.03

Table 2. Average number of genes selected by each method.

Methods	Elastic Net	Lasso	CL1	CL2	CL3	CL4	W. fusion
# of genes	999.25	42.25	278.35	221.50	287.05	160.40	856.65

5. Simulation Studies

We applied Clustering Lasso on some simulations to test its prediction accuracy in regressions when compared with Ridge, Lasso, Elastic Net, and Weighted Fusion. The first three simulations are adapted from the Elastic Net paper [1]. To begin, datasets are simulated from the true model:

$$y = X\beta + \sigma\epsilon, \qquad \epsilon \sim N(0,1)$$

For each scenario, we simulated 100 data sets, each consisting of a training data set and an independent test data set. Here are the details of the four scenarios.

1. In example one, we simulated 40 observations as training data and 200 observations as test data. We let $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$ and $\sigma = 3$. The pairwise correlation between x_i and x_j was set to be $\text{corr}(i, j) = 0.5^{|i-j|}$.

2. In Example two, we simulated 200 training data and 400 testing data. There are 40 predictors such that

$$\beta = \left(\underbrace{0,\cdots,0}_{10}, \underbrace{2,\cdots,2}_{10}, \underbrace{0,\cdots,0}_{10}, \underbrace{2,\cdots,2}_{10} \right), \quad \sigma = 15, \text{and } \text{corr}(i, j) = 0.5$$

3. Example 3 has the group setting that $\beta = \left(\underbrace{3,\cdots,3}_{15}, \underbrace{0,\cdots,0}_{25} \right)$ and $\sigma = 15$, where the predictors are generated as

$$\begin{aligned}
&x_i = z_1 + \epsilon_i^x, \ z_1 \sim N(0,1), \quad i = 1,\cdots,5;\\
&x_i = z_2 + \epsilon_i^x, \ z_2 \sim N(0,1), \quad i = 6,\cdots,10;\\
&x_i = z_3 + \epsilon_i^x, \ z_3 \sim N(0,1), \quad i = 11,\cdots,15;\\
&x_i \overset{\text{iid}}{\sim} N(0,1), \quad i = 16,\cdots,40;\\
&\epsilon_i^x \overset{\text{iid}}{\sim} N(0,0.01), \quad i = 1,\cdots,15.
\end{aligned}$$

As explained by [1], three groups are equally important groups, and each group contains five covariates. We created 100 observations as training data and 400 as testing data.

The fourth simulation is a modification of the third example to emphasize the group effects. The true model has the form $y = z_1 + 0.5z_2 + \epsilon$ where $\epsilon_i \sim N(0,1)$. The predictors we observed are

$$\begin{aligned}
&x_i = z_1 + \epsilon_i^x, \ z_1 \sim N(0,1), \quad i = 1,\cdots,5;\\
&x_i = z_2 + \epsilon_i^x, \ z_2 \sim N(0,1), \quad i = 6,\cdots,10;\\
&x_i = 0.6z_2 + \epsilon_i^x, \quad i = 11,\cdots,15;\\
&x_i = z_3 + \epsilon_i^x, \ z_3 \sim N(0,1), \quad i = 16,\cdots,20;\\
&\epsilon_i^x \overset{\text{iid}}{\sim} N(0,0.5), \quad i = 1,\cdots,20.
\end{aligned}$$

The latent variables, z_1 and z_2, directly relate to the response variable, where z_1 is more important than z_2. A nuisance variable z_3, does not related to y. $x_i s$ relate to zs at different levels. In terms of gene analysis, we can think of z_1, z_2 and z_3 as underlying pathways, some of which are related to the disease measured by y. We observed the gene expression levels, x_i, and would like to identify the related pathways.

We used all four Clustering Lasso methods. In all examples, the results from the four Clustering Lasso methods are close to each other. The prediction results from Lasso, CL2, CL4, Elastic Net, Ridge, and Weighted Fusion are summarized in **Table 3** and **Figure 3**. In **Figure 3**, relative MSE was defined as the MSE of the corresponding method divided by the minimum MSEs from all the methods. We see that Clustering Lasso always performs better than the Lasso method, and it is close to or better than Ridge, Weighted Fusion and Elastic Nets, even under collinearity and group effect situations.

Table 4 shows the results of variable selection. The two numbers in each cell are the proportion of times an important factor is chosen and the proportion of times a false factor is chosen, respectively. We see that compared with Elastic Net, Weighted Fusion and Lasso, Clustering Lasso is superior at selecting important factors. However, like Weighted Fusion, it is more likely to over select variables than Elastic Net. In Example 2,

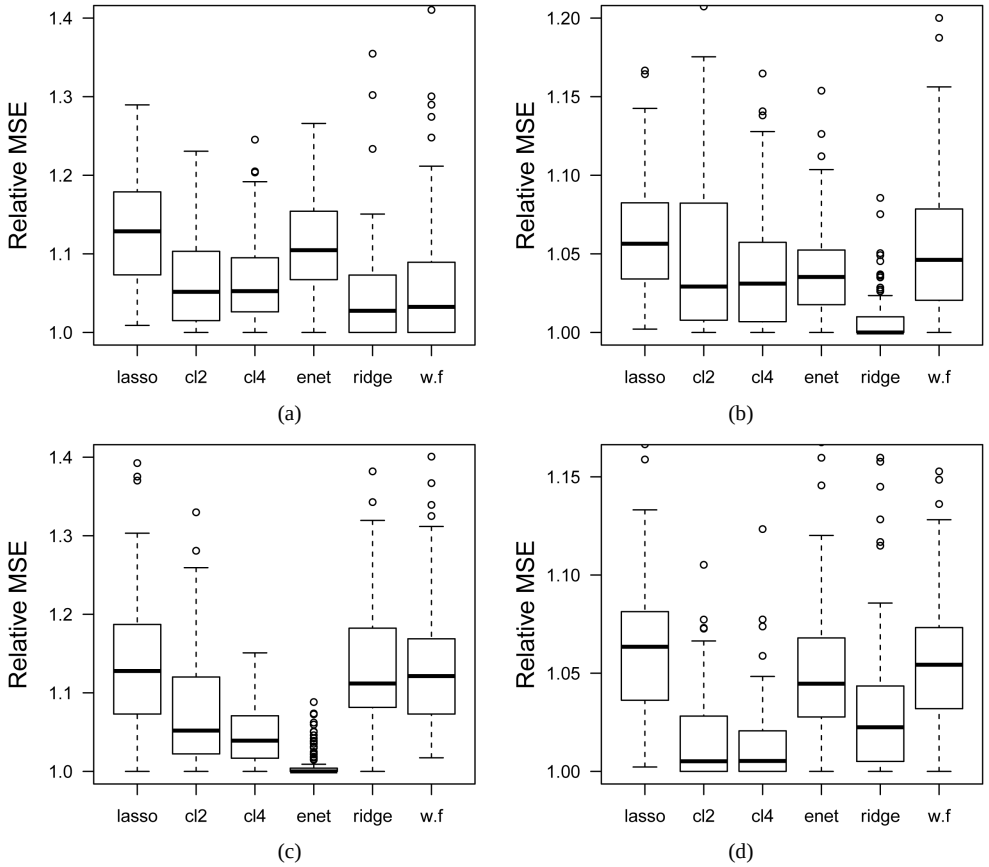


Figure 3. Comparing the simulation results from the four examples. (a)-(d): Example 1-4.

Table 3. Mean (standard deviation) of MSE for the simulated examples based on the 100 iterations.

Methods	Example 1	Example 2	Example 3	Example 4
Lasso	11.50 (1.81)	256.40 (19.05)	279.75 (30.08)	1.151 (0.09)
Elastic Net	11.23 (1.60)	251.01 (18.85)	248.42 (24.45)	1.138 (0.10)
Ridge regression	10.55 (1.46)	243.47 (15.91)	278.09 (25.81)	1.125 (0.10)
Clustering Lasso 2	10.68 (1.46)	253.75 (19.11)	265.33 (28.82)	1.097 (0.08)
Clustering Lasso 4	10.70 (1.35)	250.82 (17.78)	257.33 (23.46)	1.094 (0.08)
Weighted Fusion	10.68 (1.69)	257.07 (24.22)	287.85 (61.14)	1.141 (0.09)

Table 4. Variable selection results for the simulated examples based on the 100 iterations. In each cell, the first number is the proportion of times a true factor is chosen and the second number is the proportion of times a false factor is chosen.

Methods	Example 1	Example 2	Example 3	Example 4
Lasso	0.840, -	0.811, 0.389	0.235, 0.736	0.544, 0.186
Elastic Net	0.870, -	0.838, 0.488	0.958, 0.134	0.585, 0.124
Clustering Lasso 2	0.995, -	1.00, 0.998	1.00, 0.873	0.991, 0.630
Clustering Lasso 4	0.985, -	1.00, 0.997	1.00, 0.493	0.995, 0.460
Weighted Fusion	0.990, -	0.992, 0.975	1.00, 0.997	0.792, 0.574

since all variables are highly correlated, Clustering Lasso cannot identify the most important variables. In comparison, Clustering Lasso performs very well in Examples 3 and 4, when clusters of variables play an important role in real model.

Finally, to show how Clustering Lasso chooses covariates in groups and the behavior of the coefficients for the selected variables, we illustrate the differences between Lasso and Clustering Lasso by a modified example from [1]. Let z_1, z_2 and z_3 be three independent variables with the uniform $(0,20)$ distribution. The response variable is generated as $y \sim N(z_1 + 0.2z_2, 1)$. With the random error terms $\epsilon_i \overset{\text{ind}}{\sim} N(0, 1/16)$, the nine observed predictors are

$$x_1 = z_1 + \epsilon_1, \quad x_2 = -z_1 + \epsilon_2, \quad x_3 = z_1 + \epsilon_3,$$
$$x_4 = z_2 + \epsilon_4, \quad x_5 = -z_2 + \epsilon_5, \quad x_6 = z_2 + \epsilon_6,$$
$$x_7 = z_3 + \epsilon_7, \quad x_8 = -z_3 + \epsilon_8, \quad x_9 = z_3 + \epsilon_9.$$

The variables x_1, x_2 and x_3 are from group 1, with the direct effect z_1. x_4, x_5 and x_6 are from group 2, with the direct effect z_2. The effect from z_2 on y is much smaller than from z_1—the coefficient for z_1 is 1 compared with 0.2 for z_2. Variables x_7, x_8 and x_9 are from z_3, which does not relate to the response variable. The within-group correlations are almost 1, while the between group correlations are almost 0. **Figure 4** shows the solution paths for Lasso, Elastic Net and CL2.

We also use this simulation to compare the sensitivity and specificity of the listed methods in finding significant covariates. The simulation is repeated 100 times. **Table 5** summarizes the number of times that the coefficients of x_i are not zero. We find that the proposed Clustering Lasso of all versions can uniformly identify the important covariates while is less likely to select non-significant covariates than Lasso, Elastic Net and Weighted Fusion.

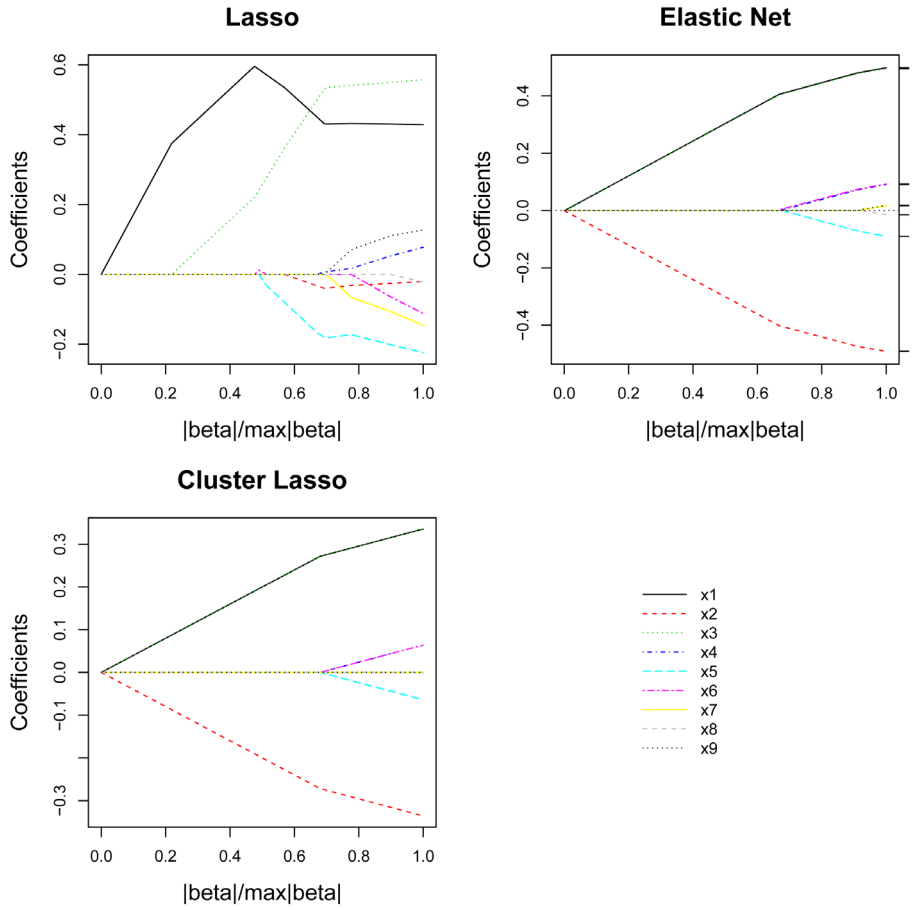


Figure 4. Comparing the solution paths from Lasso, Elastic Net and Clustering Lasso.

Table 5. Number of times the coefficients are not zero based on the 100 repetitions.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Lasso	86	84	89	69	75	64	73	61	66
Elastic Net	93	93	94	88	91	85	40	42	37
Clustering Lasso 1	100	100	100	100	100	100	58	59	61
Clustering Lasso 2	100	100	100	100	100	100	32	31	30
Clustering Lasso 3	100	100	100	100	100	100	33	35	32
Clustering Lasso 4	100	100	100	100	100	100	33	33	33
Weighted Fusion	100	99	100	100	95	100	85	83	85

6. Conclusions and Future Works

We find that the Clustering Lasso, is a novel predictive model that produces sparse model with good predictive performance, while encouraging group effects. The empirical results from a two-class microarray data classification problem and several simulation studies on regression problems show that Clustering Lasso has very good predictive performance and is superior to the Lasso method.

The method was proposed to encourage group effects so that clustered variables are selected together in a model. Clustering Lasso can automatically select groups of variables. If the structural correlation matrix used for regularization is block diagonal matrix, Clustering Lasso is equivalent to the group Lasso proposed by [7]. However, if the relationships among the variables are complicated, we have to simplify the structural correlation matrix to obtain sparse models. We proposed some shrinkage steps to build the desired structural correlation matrix. Rotating the eigen vectors or adapting techniques such as sparse component analysis can also help for this purpose. As a next step, we will use the Clustering Lasso method in the spatial analysis, so that we can maintain the important spatial correlations while selecting sparse models.

Acknowledgements

We thank Mrs. Patricia Andrews for editing the paper.

References

[1] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Application to Nonorthogonal Problems. *Technometrics*, **12**, 69-82. http://dx.doi.org/10.1080/00401706.1970.10488635

[2] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B*, **58**, 267-288.

[3] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B*, **67**, 301-320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x

[4] Alter, O., Brown, P. and Botstein, D. (2000) Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 10101-10106.

[5] Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286. http://dx.doi.org/10.1198/106186006X113430

[6] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society*: *Series B*, **67**, 91-108. http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x

[7] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*: *Series B*, **68**, 49-67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

[8] Bondell, H.D. and Reich, B.J. (2008) Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, **64**, 115-123. http://dx.doi.org/10.1111/j.1541-0420.2007.00843.x

[9] Daye, Z.J. and Jeng, X.J. (2009) Shrinkage and Model Selection with Correlated Variables via Weighted Fusion. *Computational Statistics & Data Analysis*, **53**, 1284-1298. http://dx.doi.org/10.1016/j.csda.2008.11.007

[10] Jenatton, R., Obozinski, G. and Bach, F. (2010) Structured Sparse Principal Component Analysis. *International Con-*

ference on Artificial Intelligence and Statistics (*AISTATS*).

[11] Jenatton, R., Audibert, J.Y. and Bach, F. (2011) Structured Variable Selection with Sparsity-Inducing Norms. *Journal of Machine Learning Research*, **12**, 2777-2824.

[12] Huang, J., Ma, S. and Zhang, C.H. (2011) The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression. *Annals of Statistics*, **39**, 2021-2046. http://dx.doi.org/10.1214/11-AOS897

[13] Buhlmann, P., Rutimann, P., van de Geer, S. and Zhang, C.H. (2013) Correlated Variables in Regression: Clustering and Sparse Estimation. *Journal of Statistical Planning and Inference*, **143**, 1835-1858. http://dx.doi.org/10.1016/j.jspi.2013.05.019

[14] Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion). *Journal of the Royal Statistical Society*, *Series B*, **36**, 192-236.

[15] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004) Least Angel Regression. *Annals of Statistics*, **32**, 407-499. http://dx.doi.org/10.1214/009053604000000067

[16] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007) Pathwise Coordinate Optimization. *Annals of Applied Statistics*, **1**, 302-332. http://dx.doi.org/10.1214/07-AOAS131

[17] Schafer, J. and Strimmer, K. (2005) A Shrinkage Approach to Large-Scale Covariance Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 32.

[18] Jolliffe, I.T., Trendafilov, N.T. and Uddin, M. (2003) A Modified Principal Component Technique Based on the Lasso. *Journal of Computational and Graphical Statistics*, **12**, 531-547. http://dx.doi.org/10.1198/1061860032148

[19] Singh, D., Febbo, P., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamaryo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) Gene Expression Correlates of Clinical Prostate Cancer Behavior. *Cancer Cell*, **1**, 203-209. http://dx.doi.org/10.1016/S1535-6108(02)00030-2

[20] Guyon, I., Weston, J., Barnhill, S. and Vaapnik, V. (2002) Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, **46**, 389-422. http://dx.doi.org/10.1023/A:1012487302797

Open Journal of Statistics, 2014, 4, 826-836
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410078

Scientific Research

Parallel and Hierarchical Mode Association Clustering with an *R* Package *Modalclust*

Yansong Cheng[1], Surajit Ray[2]

[1]Quantitative Science, Glaxo Smith Kline, King of Prussia, King of Prussia, Pennsylvania, USA
[2]School of Mathematics and Statistics, Glasgow University, Glasgow, UK
Email: yansong.x.cheng@gsk.com, surajit.ray@glasgow.ac.uk

Received 24 September 2014; revised 20 October 2014; accepted 5 November 2014

Abstract

***Modalclust* is an *R* package which performs Hierarchical Mode Association Clustering (HMAC) along with its parallel implementation over several processors. Modal clustering techniques are especially designed to efficiently extract clusters in high dimensions with arbitrary density shapes. Further, clustering is performed over several resolutions and the results are summarized as a hierarchical tree, thus providing a model based multi resolution cluster analysis. Finally we implement a novel parallel implementation of HMAC which performs the clustering job over several processors thereby dramatically increasing the speed of clustering procedure especially for large data sets. This package also provides a number of functions for visualizing clusters in high dimensions, which can also be used with other clustering softwares.**

Keywords

Modality, Kernel Density Estimate, Mode, Clustering

1. Introduction

Cluster analysis is a ubiquitous technique in statistical analysis that has been widely used in multiple disciplines for many years. Historically cluster analysis techniques have been approached from either a fully parametric view, e.g. mixture model based clustering, or a distribution free approach, e.g. linkage based hierarchical clustering. While the parametric paradigm provides the inferential framework and accounts for the sampling variability, it often lacks the flexibility to accommodate complex clusters and are often not scalable to high dimensional data. On the other hand, the distribution free approaches are usually fast and capable of uncovering complex clusters by making use of different distance measures. However, the inferential framework is distinctly missing in the distribution free clustering techniques. Accordingly most clustering packages in *R* also fall under the two above mentioned groups of clustering techniques.

How to cite this paper: Cheng, Y.S. and Ray, S. (2014) Parallel and Hierarchical Mode Association Clustering with an *R* Package *Modalclust*. *Open Journal of Statistics*, **4**, 826-836. http://dx.doi.org/10.4236/ojs.2014.410078

This paper describes a software program for cluster analysis that can knead the strengths of these two seemingly different approaches and develop a framework of parallel implementation for clustering techniques. For most model based approaches to clustering, the following limitations are well recognized in the literature: 1) the number of clusters has to be specified; 2) the mixing densities have to be specified, and as estimating the parameters of the mixture models is often computationally very expensive, we are often forced to limit our choices to simple distributions such as Gaussian; 3) computational speed is inadequate especially in high dimensions and this coupled with the complexity of the proposed model often limits the use of model-based techniques either theoretically or computationally; 4) it is not straightforward to extend model-based clustering to uncover heterogeneity at multiple resolutions, similar to the one offered by to the model free linkage based hierarchical clustering.

Influential work towards resolving the first three issues has been carried out in [1]-[7]. Many previous approaches have focused on model selection of mixtures by choosing the number of components, merging existing components or by determining the covariance structure of the mixture density under consideration, see [8]-[10]. They work efficiently if the underlying distribution is chosen correctly, but none of these model based approaches is designed to handle a completely arbitrary underlying distribution (see **Figure 5** for one such example). That is, we think that limitations due to issues (3) and (4), above often necessitate the use of model-free techniques.

This paper describes a software program for cluster analysis that can knead the strengths of these two seemingly different approaches and develop a framework of parallel implementation for clustering techniques. The hierarchical mode association clustering—HMAC [11], which is constructed by first determining modes of the high-dimensional density and then associating sample points to those modes, is the first multivariate model based clustering approach resolving many of the drawbacks of standard model-based clustering. Specifically, it can accommodate flexible subpopulation structures at multiple resolutions while retaining the desired natural inferential framework of parametric mixtures. [12] developed the inference procedure to the number of clusters of this approach. *Modalclust* is the package implemented in *R* (R Development Core Team, 2010) for carrying out HMAC along with its parallel implementation (PHMAC) over several processors. Though mode-counting or mode hunting has been extensively used as a clustering technique, most implementation are limited to univariate data. Generalization to higher dimensions was limited both due to the computational complexity of finding modes in higher dimension and the lack of any natural framework to study the inferential properties of modes in higher dimensions. The HMAC provides a computationally fast iterative algorithm for calculating the modes and thereby providing a clustering approach which is scalable to high dimensions. This article provides the description of the *R* package that implements HMAC and additionally provides an wide array visualization tools for representing clusters in high dimensions. Further, we propose a novel parallel implementation of the approach which dramatically reduces the computational time especially for large data sets, both in data dimensions and the number of observations.

This paper is organized as follows: Section 2 briefly introduces the algorithm of Modal Expectation Maximization (MEM) and builds the notion of mode association clustering technique. Section 3 describes a parallel computing framework of HMAC along with computing time comparisons. Section 4 illustrates the implementation of clustering functions in the *R* package *Modalclust* along with examples of the plotting functions especially designed for objects of class *hmac*. Section 5 provides the conclusion and discussion. Comparison of Modal clustering with other popular model based and model free techniques are provided in the supplementary document.

2. Modal EM and HMAC

The main challenge for using mode-based clustering in high dimensions is the cost of computing modes, which are mathematically evaluated as local maximas of the density function with support on $\mathbb{R}^D$, D being the data dimension. Traditional techniques of finding local maxima, such as "hill climbing" works well for univariate data. But multivariate hill climbing is computationally expensive thereby limiting its use in high dimensions. [9] proposed an algorithm that solves a local maximum of a kernel density by ascending iterations starting from the data points. Since the algorithm is very similar to Expectation Maximization (EM) algorithm, it is named as Modal Expectation Maximization (MEM). Define the mixture density as $f(x)=\sum_{i=1}^{K}\pi_i f_i(x)$. Now, given any initial value $x^{(0)}$, the MEM solves a local maximum of the mixture density by alternating the following two

steps until it meets some user defined stopping criterion.

1. Let $p_i = \dfrac{\pi_i f_i\left(x^{(r)}\right)}{f\left(x^{(r)}\right)}, \quad i = 1, \cdots, n$

2. Update $x^{(r+1)} = \text{argmax}_x \sum_{i=1}^{n} p_i \log f_i(x)$

Details of convergence of the MEM approach can be found in [11]. The above iterative steps provide a computationally simpler approach than grid search method for "hillclimbing" from any starting point $x \in \mathbb{R}^D$ by exploiting the properties of density functions.Given a multivariate kernel K, let the density of the data be given by

$$f\left(x|\Sigma\right) = \sum_{i=1}^{n} \frac{1}{n} K\left(x - x_i|\Sigma\right)$$

where Σ is the matrix of smoothing parameters. Further, in the special case of Gaussian kernels, *i.e.*,

$$K\left(x - x_i|\Sigma\right) = \phi\left(x|x_i, \Sigma\right)$$

where $\phi(\cdot)$ is the pdf of a Gaussian distribution, the update of $x^{(r+1)}$ is simply

$$x^{(r+1)} = \sum_{i=1}^{n} p_i x_i$$

allowing us to avoid the numerical optimization of Step 2.

Now we present the HMAC algorithm. First we scale the data and use a kernel density estimator, with a normal kernel to estimate the density of the data. The variance of the kernel, Σ is a diagonal matrix with all entries σ^2 denoted by $D\left(\sigma^2\right)$, thus σ^2 is the single smoothing parameter for all the dimensions. The choice of the smoothing parameter is an area of research in itself. In the present version of the program we incorporate the strategy of using pseudo degrees of freedom, proposed in [13]. Their strategy provides us with a range of smoothing parameters and exploring them from finest to coarsest resolution provides the user with the desired hierarchical clustering. First we describe the steps of Mode Association Clustering (MAC) for a single bandwidth σ^2.

1. Given a set of data $S = \{x_1, x_2, \cdots, x_n\}$, $x_i \in \mathbb{R}^d$ form kernel density

$$f\left(x|S, \sigma^2\right) = \sum_{i=1}^{n} \frac{1}{n} \phi\left(x|x_i, D\left(\sigma^2\right)\right) \tag{1.1}$$

2. Use $f\left(x|S, \sigma^2\right)$ as the density function. Use each x_i, $i = 1, 2, \cdots, n$, as the initial value in the MEM algorithm and find one mode of $f\left(x|S, \sigma^2\right)$ for each x_i. Let the mode identified by starting from x_i be $\mathcal{M}_\sigma(x_i)$.

3. Extract distinctive values from the set $\{\mathcal{M}_\sigma(x_i), i = 1, 2, \cdots, n\}$ to form a set G. Label the elements in G from 1 to $|G|$. In practice, due to finite precision, two modes are regarded equal if their distance is below a threshold κ. In our package, we use $\kappa = 10^{-4}$.

4. If $\mathcal{M}_\sigma(x_i)$ equals the k^{th} element in G, x_i is put in the k^{th} cluster.

We note that when the bandwidth σ increases, the kernel density estimator $f\left(x|S, \sigma^2\right)$ in (1.1) becomes smoother, and thus more points tend to climb to the same mode. This suggests a natural approach for hierarchical organization (or "nesting") of our MAC clusters. Thus, given a range of bandwidths $\sigma_1 < \sigma_2 < \cdots < \sigma_L$, The clustering can be performed in the following bottom-up manner. Define the G_l as the collection of all the distinct modes obtained by MAC using the σ_l. First we perform MAC at the smallest bandwidth σ_1. At any bandwidth σ_l, the elements in G_{l-1} obtained from the preceding bandwidth are fed into MAC using the density $f\left(x|S, \sigma_l^2\right)$. The modes identified at this level form a new set of cluster is G_l. This procedure is repeated across all σ_l's. This preserves the hierarchy of clusters and thus the name Hierarchical Mode Association Clustering (HMAC). To summarize we present the HMAC procedure in the following box.

1. Start with the data $G_0 = \{x_1, \cdots, x_n\}$ and set level $l = 0$ and initialize the mode association of the i^{th} data point as $\mathcal{P}_0(x_i) = i$.

2. $l \leftarrow l + 1$.

3. Form kernel density as in (1.1) using σ_l^2.

4. Cluster the elements in G_{l-1} by using density $f\left(x|S,\sigma_l^2\right)$. Let the set of distinct modes obtained be G_l.

5. If $\mathcal{P}_{l-1}\left(x_i\right)=k$ and the k^{th} element in G_{l-1} is clustered to the k'^{th} mode in G_l, then $\mathcal{P}_l\left(x_i\right)=k'$. In another word, the cluster of x_i at level l is determined by its cluster representative in G_{l-1}.

6. Stop if $l=L$, otherwise go to Step 2.

3. Parallel HMAC

In this section we develop the method of parallel computing of HMAC (PHMAC) and its application together with some comparisons of performance of the parallel and non-parallel approach. The MAC approach is computationally expensive when the number of objects n becomes large. It requires that we use the MEM for each data point to find its local mamximum of the density. Note that for the HMAC, the steps for the level $l=2$ onwards only need to start the MEM from the modes of the previous level G_{l-1}, and hence the computational cost does not increase at the rate of n. Fortunately the MAC approach provides a natural framework for a "divide and conquer" clustering algorithm. One can simply divide the data into m partitions, perform modal clustering on each of those partitions, and pool the modes obtained from each of these partitions to form a collection G and apply the HMAC onward. If the user has access to several computing cores of the same machine or several processors of a shared memory computing cluster, the "divide and conquer" algorithm can be seamlessly parallelized. The PHMAC procedure is summarized as follows:

Step 1. Sphering transform the data $\mathbf{X}$ to form a new data set $\mathbf{Y}$.

Step 2. Let $G_0=\{y_1,\cdots,y_n\}$. Divide the data (n objects) into m partitions G_o^j randomly, $j=1,2,\cdots,m$.

Step 3. Perform HMAC on each of these subsets at the lowest resolution, *i.e.*, using h_1 and get the modes G_1^j, $j=1,2,\cdots,m$.

Step 4. Pool the modes from each subset of data to form $G_1=\bigcup_{j=1}^{m}G_1^j$.

Step 5. Perform HMAC starting from Step 2 and obtain the final hierarchical clustering.

Step 6. Transform $\mathbf{Y}$ back to $\mathbf{X}$.

Figure 1 shows one PHMAC example on the graph. In this figure, (a) shows the simulated data with four clusters along with the contour plot, where the color indicates the final clustering using PHMAC; (b) shows the four random partitions of the unlabeled data along with the modes (red asterisks) at each partition; (c) shows the mode obtained from the four partitions; (d) shows the final modes (green triangles) starting from the modes of the partitioned data. A demonstration of different steps of parallel clustering with four random partitions is given in **Figure 1**. The original data set is partitioned into 4 random subsets, and initial modal clustering is performed within the partitions. In the next step, the modes of each of these partitions are merged to form the overall modal clusters in **Figure 1(c)**.

Modes have a natural hierarchy and it is computationally easy to merge modes from different partitions. In practice, we need to decide the best choice of the partition and how many partitions to use. In this section, we provide some guidelines regarding the choices, without exploring their quality in details. In the absence of any other knowledge, one should randomly partition the data. Other choices include partitioning data based on certain coordinates which form a natural clustering, and then taking products of a few of those coordinates to build the overall partition. This strategy might increase the computational speed by restricting the modes within a relatively homogeneous set of observations. Another choice might be to sample the data and build partitions based on the modes of the sampled data.

The PHMAC we proposed uses parallel computing at the first level of HMAC and then use non-parallel computing from the second level onwards. Therefore, the number of partitions to minimize the computational time is a complex function of the number of available processors, the number of observations and the bandwidth parameter of the KDE. If one uses too many partitions, one might speed up the first step, but would have the risk of ending up with too many modes for the next level, where the hill climbing is done from the collection of modes from each partition with respect to the overall density. In contrast, for a large n, if one chooses too few partitions or no partitions, this would lead to a huge computational cost at the first step. Moverover, the choice of the smoothing parameter will also determine how many modes one needs to start from at the merged level.

We compare the computing speed of parallel versus serial clustering using 1, 2, 4, 8 and 12 multi-core processors. Tests were performed on a 64 bits 4 Quad Core AMD 8384 (2.7 Ghz each core), with 16 GB RAM

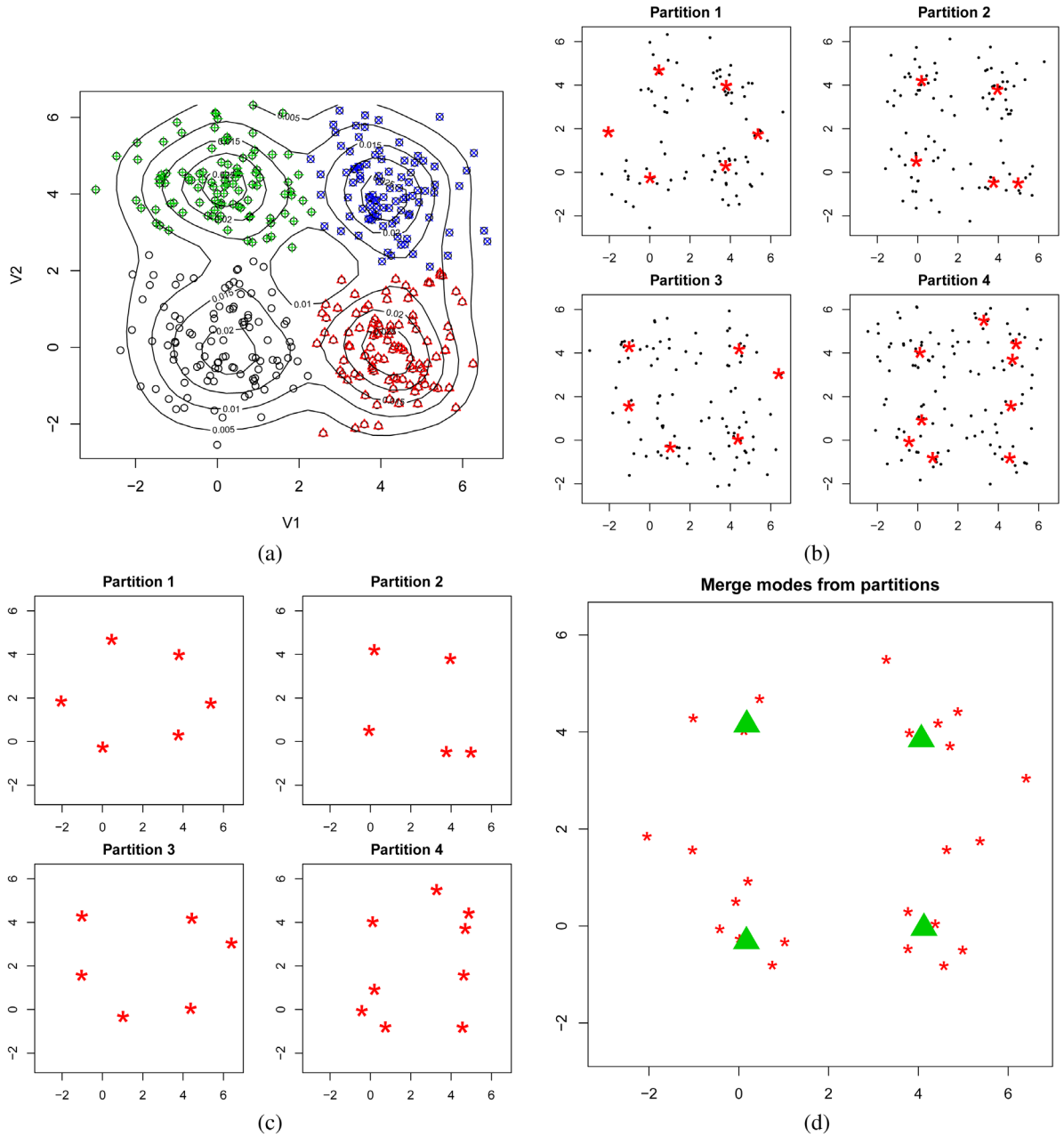


Figure 1. Steps in parallel HMAC procedure for a simulated data set.

running Linux Centos 5 and R version 2.11.0 From **Table 1**, it is clear to observe that parallel computing significantly increases the computing speed. Because the KDE is a sum of kernels centered at every data point, the amount of computation needed to identify the mode associated with a single point grows linearly with n. The computational complexity of clustering all the data by MAC is thus quadratic in n. Suppose we have p processors, then the computing complexity for the MAC is n^2 and for parallel computing of MAC is thus $(n/p)^2$. However, as discussed before, we can see that the computational speed is not a monotone decreasing function of the number of processors. Theoretically, it is true that more processors can reduce the computing complexity at the initial step. However, in practice, if the data set is not sufficiently large, using more processors may not save time, as it may produce a large number of modes for the next level of HMAC. When the $n=10,000$ or $n=50,000$, including more processors provides a dramatic decrease in computing time, whereas for $n=2,000$, there is no clear decrease in time elapsed when using 4 or 8 processors instead of the

Table 1. Comparison of computing time (elapsed time in seconds) using different number of processors.

Data dimensions		Number of processors				
		1	2	4	8	12
n = 2000	d = 2	56.58	17.01	7.84	6.91	8.02
n = 2000	d = 20	323.16	128.13	112.42	190.11	250.22
n = 2000	d = 40	730.18	560.16	687.79	764.29	753.36
n = 10,000	d = 2	3849.83	871.33	276.88	145.61	131.22
n = 10,000	d = 20	8410.96	1694.82	585.33	536.32	459.88
n = 50,000	d = 2	210295.29	71152.82	23383.61	11959.24	4875.64

maximum 12 processors. For $n = 50,000$, the decrease in computing time from 1 processor to using 12 processors is more than 40 fold (see **Figure 2**), but even if the user is able to use just two processors, the computing time is reduced to 1/3 of how long a single processor would take. Even for $n = 20,000$, the advantage of using 12 processors is almost 30 fold, whereas for $n = 2,000$, the advantage is only 8 folds. In fact, the lowest time is actually clocked by 8 processors for $n = 20,000$, but using all 12 processors does not increase the time significantly. These comparisons show the potential for parallelizing the modal clustering algorithm and its inherent use for clustering high throughput data.

The R package *Modalclust* was created to implement the HMAC and PHMAC. There are also some plotting tools that give the user a comprehensive visual and understanding of the clustering result. Sources, binaries and documentation of *Modalclust* are available for download from the Comprehensive *R* Archive Network http://cran.r-project.org/ under the GNU Public License.

4. Example of Using *R* Package *Modalclust*

In this section, we demonstrate the usage of the functions and plotting tools that are available in the *Modalclust* package.

4.1. Modal Clustering

First, we provide an example of performing modal clustering to extract the subpopulations in the *logcta*20 data. The description of the dataset is given in the package. The scatter plot, along with its smooth density, is provided in **Figure 3**. First, we use the following command to download and install the package:

R > install.packages (“Modalclust”)

R > library (“Modalclust”)

Using the following command, we can get the standard (serial) HMAC and parallel HMAC using two processors for logctA20 data.

R > logcta20.hmac < −phmac(logcta20,npart=1,parallel=FALSE)

R > logcta20p2.hmac < −phmac(logcta20,npart=2,parallel=TRUE)

Both implementation results are given in **Figure 4**, which clearly identifies the three distinct subpopulations. Other model-based clustering methods, such as EM-clustering or K-means, could not capture the subpopulation structure, as the individual subpopulation is not a normal density. Distance based clustering method e.g., hierarchical clustering, with a range of linkage functions performed even worse.

By default, the function selects an interesting range of smoothing parameters with ten σ^2 values, and the final clustering only shows the results from the levels which produced merging from the previous level. For example, for the logcta20, the smoothing parameters chosen automatically are

R > logcta20.hmac$sigma

[1] 0.26 0.29 0.31 0.34 0.38 0.43 0.49 0.58 0.72 0.94,

which are chosen using the *spectral degrees of freedom* criterion introduced in [10]. Though we started with 10 different smoothing levels, the final clustering shows only 6 different levels along with a decreasing number of hierarchical cluster.

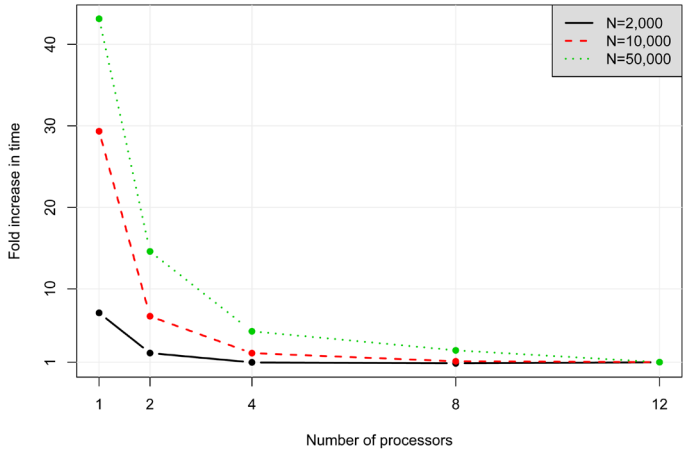


Figure 2. Comparison of fold increase in time for clustering two dimensional data of different sample sizes with respect to using 12 processors.

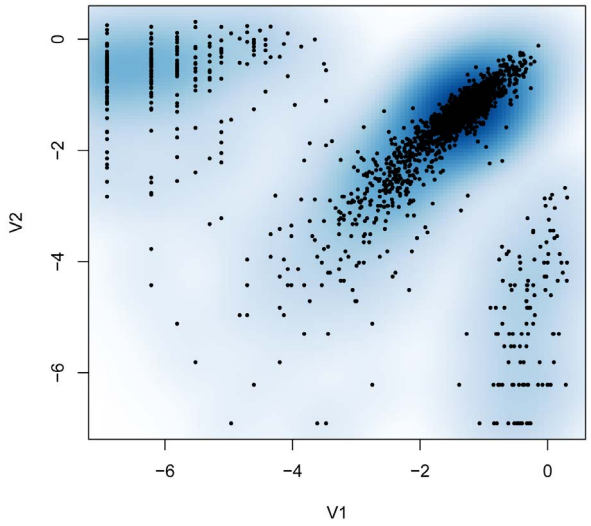


Figure 3. Smoothing scatter plot of *logctA*20 data.

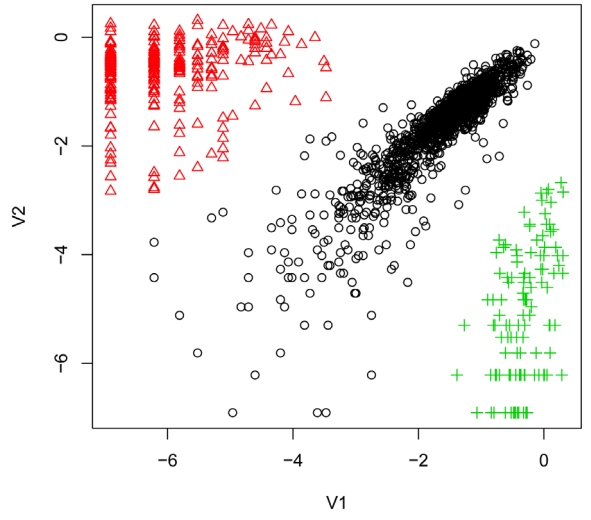


Figure 4. HMAC output of *logctA*20 data.

```
R > logcta20.hmac$level
[1] 1 2 3 3 3 4 4 4 5 6
R > logcta20.hmac$n.cluster
[1] 11 7 5 5 5 3 3 3 2 1
```

The user can also provide smoothing levels using the option *sigmaselect* in *phmac*. There is also the option of starting the algorithm from user defined modes instead of the original data points. This option becomes handy if the user wishes to merge clusters obtained from other clustering methods, e.g., EM-clustering or K-means.

4.2. Some Examples of Plotting

There are several plotting functions in *Modalclust*, which can be used to visualize the output from the function *phmac*. The plotting functions are defined on object class *hmac*, which is the default class of a *phmac* output. These plot functions will be illustrated through a data set named *disc2d*, which has 400 observations displaying the shape of two half discs. The scatter plot of *disc2d* along with its contour plot are given in **Figure 5**.

First, we introduce the standard *plot* function for an object of class "hmac". This unique and informative plot shows the hierarchical tree obtained from modal clustering. It can be obtained by

```
R > data ("disc2d.hmac")
R > plot (disc2d.hmac)
```

The dendrogram obtained from the *disc2d* data is given in **Figure 6**. The y-axis gives the different levels, and the tree displays the merging at different levels. There are several options available for drawing the tree,

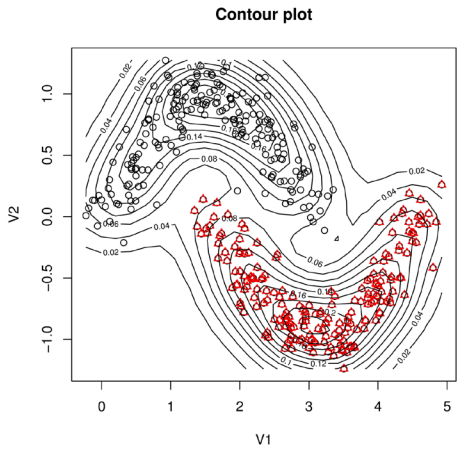


Figure 5. The scatter plot of *disc2d* data along with its probability contours.

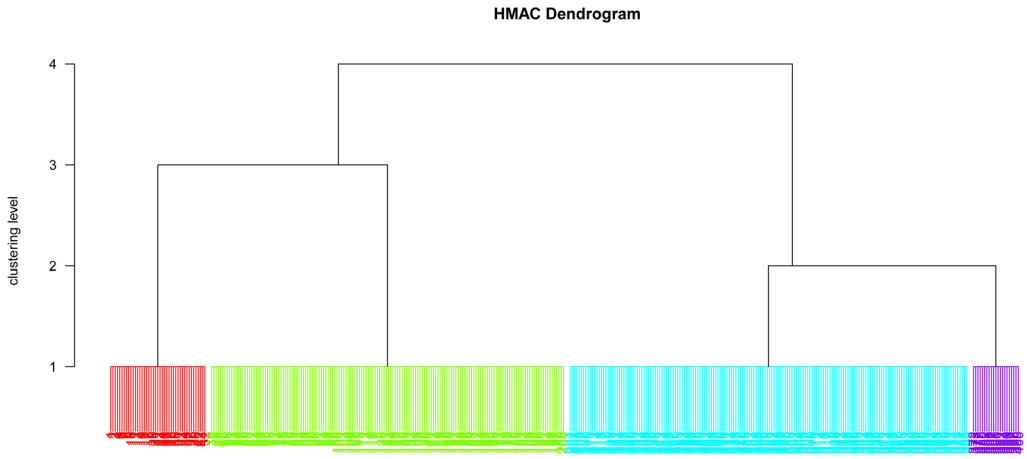


Figure 6. Hierarchical tree (Dendrogram) of *disc2d* data showing the clustering at four levels of smoothing.

including starting the tree from a specific level, drawing the tree only up to a desired number of clusters, and comparing the clustering results with user defined clusters.

There are some other plotting functions that are designed mainly for visualizing clustering results for two dimensional data, although one can provide multivariate extensions of the functions by considering all possible pairwise dimensions. One can obtain the hard clustering of the data for each level using the command

R > hard.hmac(disc2d.hmac)

Alternatively, the user can specify the hierarchical level or the number of desired clusters, and obtain the corresponding cluster membership (hard clustering) of the data. For example, the plot in **Figure 7** can be obtained by either of the following two commands:

R > hard.hmac (disc2d.hmac, n.cluster=2)

R > hard.hmac (disc2d.hmac, level=3)

Another function, which allows the user to visualize the soft clustering of the data, is based on the posterior probabilities of each observation belonging to the clusters at a specified level. For example, the plot in **Figure 8** can be obtained using

R > soft.hmac (disc2d.hmac, n.cluster=3)

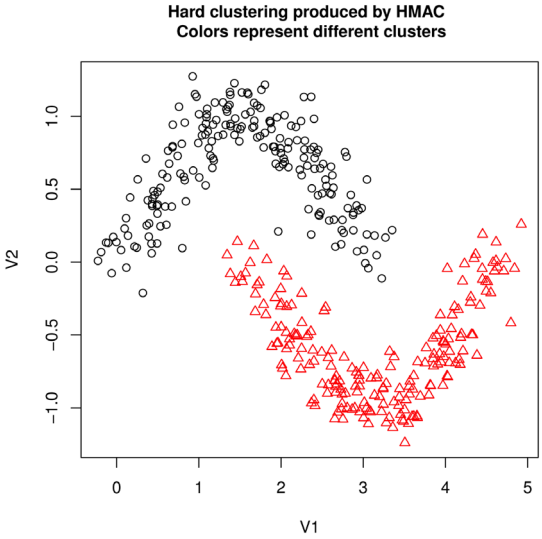


Figure 7. Hard clustering for *disc2d* data at level 3.

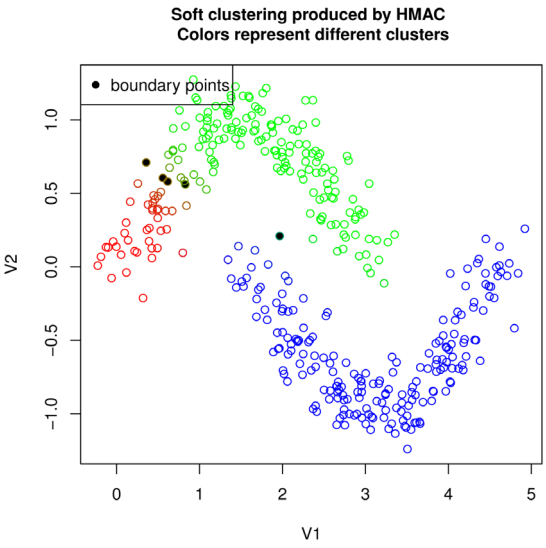


Figure 8. Soft clustering for *disc2d* data at level 2.

The plot enables us to visualize the probabilistic clustering of the three cluster model. A user can specify a probability threshold for assigning observations which clearly belong to a cluster or lie in the "boundary" of more than one cluster. Points having posterior probability below the user specified *boundlevel* (default value 0.4) are assigned as boundary points and colored in gray. In **Figure 8**, we have five boundary points among the 400 original observations. Additionally, at any specified level or cluster size, the *plot=FALSE* option in *hard.hmac* returns the cluster membership. Similarly, *plot=FALSE* option in *soft.hmac* returns a list that contains the posterior probability of each observation and boundary points.

```
R > disc2d.2clust < −hard.hmac (disc2d.hmac,n.cluster=2, plot=FALSE)
R > disc2d.2clust.soft < −soft.hmac (disc2d.hmac,n.cluster=2, plot=FALSE)
```

5. Discussion

Modalclust performs a hierarchical model based clustering allowing for arbitrary density shapes. Parallel computing can dramatically increase the computing speed by splitting the data and running the HMAC simultaneously on multi-core processors. Plotting functions give the user a comprehensive visualizing and understanding of the clustering result. One future work from this stage would be to increase computing speed, especially for large data set. From the discussion in Section 3, it is clear to see, parallel computing increases the computing speed a lot. That relies on the computing equipment. If one user has no multicore or a few multicore processors available, it will take a lot of the computing resources when clustering large data sets. One potential way to solve the computing speed problem is using *k*-means or other faster clustering techniques initially, and using the HMAC from the centers of each cluster of initial clustering results. For example, if we have a data set with 20,000 observations, we can use *k*-means clustering and choose a certain number of centers, like 200 centers and run *k*-means clustering first. And then we start from the centers of 200 clusters and clustering by HMAC. Theoretically it is a sub-optimal way compared with running HMAC for all points. In practice, it is very useful to reduce the computing costs and still obtain the right clustering.

In addition, we are currently working on an implementation of modal clustering for online or streaming data, where the goal would be to update an existing cluster with the new data without storing all the original data points and allowing for creation of new clusters and merging of existing clusters.

Sources, binaries and documentation of *Modalclust* are available for download from the Comprehensive *R* Archive Network http://cran.r-project.org/ under the GNU Public License.

References

[1] Fraley, C. (1998) Algorithms for Model-Based Gaussian Hierarchical Clustering. *SIAM Journal on Scientific Computing*, **20**, 270-281. http://dx.doi.org/10.1137/S1064827596311451

[2] Fraley, C. and Raftery, A. (1998) How Many Clusters? Which clustering method? Answers via Model-Based Cluster Analysis. *The Computer Journal*, **41**, 578-588. http://dx.doi.org/10.1093/comjnl/41.8.578

[3] Fraley, C. and Raftery, A. (1999) Mclust: Software for Model-Based Cluster Analysis. *Journal of Classification*, **16**, 297-306. http://dx.doi.org/10.1007/s003579900058

[4] Fraley, C. and Raftery, A. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, **97**, 611-631. http://dx.doi.org/10.1198/016214502760047131

[5] Fraley, C. and Raftery, A. (2002) Mclust: Software for Model-Based Clustering, Density Estimation and Discriminant Analysis. Tech. Rep., DTIC Document.

[6] Ray, S. and Lindsay, B. (2005) The Topography of Multivariate Normal Mixtures. *The Annals of Statistics*, **33**, 2042-2065. http://dx.doi.org/10.1214/009053605000000417

[7] Ray, S. and Lindsay, B. (2007) Model Selection in High Dimensions: A Quadratic-Risk-Based Approach. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**, 95-118.

[8] Baudry, J., Raftery, A., Celeux, G., Lo, K. and Gottardo, R. (2010) Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, **19**, 332-353. http://dx.doi.org/10.1198/jcgs.2010.08111

[9] Hennig, C. (2010) Methods for Merging Gaussian Mixture Components. *Advances in Data Analysis and Classification*, **4**, 3-34. http://dx.doi.org/10.1007/s11634-010-0058-3

[10] Tantrum, J., Murua, A. and Stuetzle, W. (2003) Assessment and Pruning of Hierarchical Model Based Clustering. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 197-205. http://dx.doi.org/10.1145/956750.956775

[11] Li, J., Ray, S. and Lindsay, B. (2007) A Nonparametric Statistical Approach to Clustering via Mode Identification. *Journal of Machine Learning Research*, **8**, 1687-1723.

[12] Cheng, Y. and Ray, S. (2014) Multivariate Modality Inference Using Gaussian Kernel. *Open Journal of Statistics*, **4**, 419-434. http://dx.doi.org/10.4236/ojs.2014.45041

[13] Lindsay, B., Markatou, M., Ray, S., Yang, K. and Chen, S. (2008) Quadratic Distances on Probabilities: A Unified Foundation. *The Annals of Statistics*, **36**, 983-1006. http://dx.doi.org/10.1214/009053607000000956

Open Journal of Statistics, 2014, 4, 837-850
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410079

Hierarchical Cores Applied to an Analysis of Use of Technologies Level among Higher Education Students in Mexico

Francisco Casanova-del-Angel

SEPI-ESIA, Unit ALM of the Polytechnic Institute National, Mexico City, Mexico
Email: fcasanova49@prodigy.net.mx, fcasanova@ipn.mx

Received 7 September 2014; revised 5 October 2014; accepted 1 November 2014

Abstract

Using the theory shown, Cores Optimal Criterion, three factors from which hierarchical aggregation of variables under study was built, as well as hierarchical cores showing the level of use of pocket computing technologies by students. The principal factors influencing the level of use of pocket computing technologies among higher education students are analyzed from a theoretical aggregation development based on hierarchical cores. The theoretical part includes the development of an algorithm used to obtain an interesting class or partition from a hierarchy. The experimental work carried out included design, preparation and application of a questionnaire to higher education students in Mexico. A pilot test was carried out to check timing and repetition of questions. Data was recorded, validated, and mathematically and statistically analyzed.

Keywords

Use of Technologies, Higher Education, Questionnaire, Pocket Calculators, Hierarchical Cores

1. Introduction

The purpose of this work is to statistically analyze the level of use of pocket computing technologies amongst higher education students in Mexico, in order to quantify the degree of influence of marketing and training factors on the demand of calculators with CAS (Computer Algebraic Systems) technology. Experimental work was carried out by González Meneses, M.S. [1], and included the use of a couple of questionnaires, one for students, and one for teachers, in Technological Institutes in Mexico.

The incorporation of new technologies in Middle and Higher Education is one of the principal purposes for amending syllabuses. Nowadays, there is a wide range of new technologies, from distance education to didactic

How to cite this paper: Casanova-del-Angel, F. (2014) Hierarchical Cores Applied to an Analysis of Use of Technologies Level among Higher Education Students in Mexico. *Open Journal of Statistics*, **4**, 837-850.
http://dx.doi.org/10.4236/ojs.2014.410079

software for classrooms. Particularly in teaching mathematics, there are many resources to help the teaching-learning experience. One of such resources is the use of calculators with CAS technology (Computer Algebraic Systems). The market for calculators sale is limited to three or four brands who are distributed directly from companies, and there exists the possibility to generate micro and small companies devoted to education and provision of various services such as: didactic aids, syllabus design, and training for teachers, among others, depending on the technological development and implementation of new technologies in the classroom [2].

This topic has been looked at by J. R. Rodríguez and L. F. Flores López, from the Technological Institute of Los Mochis, Sonora México by means of a didactic proposal for calculation using Texas Instruments Voyage 200 calculators, where the use of CAS technology calculators is shown to improve learning of Differential Calculus [3]. Since Latin America is highly interested in the implementation of new technologies in syllabuses, the following analysis allows us to know factors enabling the proposal of market technologies from regional to national levels, with the potential for making proposals at a Latin American scale [4] [5].

2. Theoretical Development of Hierarchy by Cores

Based on the fact that factorial correspondence analysis represents, on the same graphic, both sets comprising a tabular correspondence arrangement; sets *I* of individuals and *Q* of classes defined for each variable *J*, and that when such must be taxonomized, a rigid class system must be fixed, then the global and spatial vision provided by factorial analysis allows us establish, through some kind of aggregation method, a type of hierarchy of the data under analysis.

The method herein shown is tributary to three options: 1) calculation of distance between elements where factorial coordinates are known; 2) juxtaposition of mass or weight to each element; and 3) calculation of a distance between element classes, depending on an aggregation criterion based on cores. Since our data includes factorial values related to *Q* classes, we shall retain a small number of *A* cardinality factors, not higher than 75% of factorial data.

Let us define factorial set of values through set: $\left\{F_\alpha(q) \middle| q \in Q \text{ and } \alpha \in A\right\}$, with which it is possible to calculate many tabular arrangements for distances between elements. In our case, we shall introduce the following distance. Let q and q' be two classes of a variable $j \in J$ such that q and $q' \in Q$. Classes q and q' belong to a normed factorial space with a fixed set of coordinates. If $d: F \to \mathbb{R}$ then (F, d) is a metric space. Factorial distance between $F(q)$ and $F(q')$ is the addition of lengths of projections of line segment between factorial values on the axes system. This is mathematically expressed as follows:

$$d^2(q,q') = \|q,q'\|^2 = \sum_{\alpha \in A}\left(F_\alpha(q) - F_\alpha(q')\right)^2 \tag{1}$$

where q and q' are classes of variable $j \in J$, d is the distance between classes, α is the axis, A is the set of axes and $F_\alpha(q)$ and $F_\alpha(q')$ are factorial values of classes.

In accordance with the second option of the aggregation method defined, the distance between classes is juxtaposed by inertia λ of the set of dots along axis α, which is represented by the own value related to the corresponding axis, because of this Equation (1) may be re-expressed as follows:

$$d^2(q,q') = \|q,q'\|^2 = \sum_{\alpha \in A}\lambda_\alpha^{-1}\left(F_\alpha(q) - F_\alpha(q')\right)^2 \tag{2}$$

where q and q' are the classes of variable $j \in J$, d is the distance between classes, α is the axis, λ_α^{-1} is the inverse of distance between classes on axis α and $F_\alpha(q)$ represents factorial value of class q on axis α [6].

Once the distance between values has been defined, the diameter index of nodes of classification ν of such hierarchy must be calculated, through:

$$\nu(n) = \frac{f_a * f_b}{f_a + f_b}\left\|F_\alpha(a) - F_\alpha(b)\right\|^2 \quad \forall n \in Nodo \tag{3}$$

where a and b are barycenter's of elements of the index, f_a and f_b are the mass in a and b barycenter's, and $F_\alpha(a)$ and $F_\alpha(b)$ are factorial values of a and b barycenter's. In addition, $a \cup b = n$ and $a \cap b = \Phi$.

Every time, the distance between elements that are hierarchized must be recalculated with those to be hierarchized, because of this the following diameter index $\nu(n)$ is:

$$\nu(n)=\frac{f_a * f_b}{f_a+f_b}\left\|\lambda_\alpha^{-1}F_\alpha(a)-\lambda_\alpha^{-1}F_\alpha(b)\right\|^2 \quad \forall n\in Nodo \tag{4}$$

where $\nu(n)$ is diameter index, f_a and f_b are masses of *a* and *b* barycenter's, $F_\alpha(a)$ and $F_\alpha(b)$ are factorial values of *a* and *b* barycenter's, and λ_α^{-1} is the square root of total distance of the *A* set of dots, along axis α.

Now, from Equation (3) it may be seen that the addition of values of diameter indexes is equal to the addition of total distance λ of the set of dots along α axis, that is:

$$\sum_{n\in Nodo}\nu(n)=\sum_{\alpha\in A}\lambda_\alpha \tag{5}$$

where $\nu(n)$ diameter is index and λ_α is total distance of the set of axes. From Equation (4) it may be seen that the addition of values of diameter indexes is equal to *A*'s cardinality.

$$\sum_{n\in Nodo}\nu(n)=Card(A) \tag{6}$$

The Algorithm

Classification algorithm looks for two minimum values of the table of factors of classes to be hierarchized.

$$\delta(q,q')=\frac{f_q * f_{q'}}{f_q+f_{q'}}\left\|F_\alpha(q)-F_\alpha(q')\right\|^2 \quad \forall q,q'\in Q \tag{7}$$

From this aggregation, defined as $k=q\cup q'$, a new partition or core of the set of *Q* classes must be updated making: $\mathcal{P}=Q\cup\{k\}-\{q,q'\}$. Distances between this new element *k* and q'' are recalculated, showing the following minimum value of the factors table, through Formula (3), thus making $\nu(n)=\delta(a,b)$. The minimum of the new table is investigated, aggregated and a new partition is updated below. The above is carried out until there is no more than the two last cores to be added, taking into account that the link is the base set [7] and [8].

Theorem Cores Optimal Criterion. If aggregation cores are groups of factors with same cardinality and Ω the space of cores, optimal election criterion is:

$$d(L,P)=\sum_{i=1}^{k}d(A_i-P_i)$$

where *L* is the total set of cores, A_i is the i^{th} core containing a certain number of objects of *P* population.

Demonstration. Let $L=\{A_1,\cdots,A_h\}$, $A_i\subset\mathcal{L}$ be the i^{th} core containing *q* elements of population. $P=\{P_1,\cdots,P_h\}$ is partition of space Ωinto *k*-classes. Let $\mathcal{L}_k$ be the set of k^{th} cores and $\mathcal{P}_k$ the set of partitions of Ω cores space into classes. $d(A_i,\mathcal{P}_i)$ measures dissimilarities between core A_i and class $\mathcal{P}_i$. Based on the above, the principal problem is to look for a $L^*\subset\mathcal{L}_k$ and a population $\mathcal{P}\subset\mathcal{P}_k$ that minimize *d* dissimilarity.

Let $d(q_1,q_2)$ be a measure for dissimilarities between couples of individuals or classes. Let us suppose that:

$$d(q_1,q_2)=\sum_{q_1\in X}\sum_{q_2\in Y}d(q_1-q_2)$$

where *X* and *Y* are parts of the set of Ω individuals, then:

$$d(q_2,\{q_1\})=d(Y,q_1) \text{ and } d(\{q_1\},Y)=d(q_1,Y)$$

In case that cores are groups of individuals, the algorithm shall be specified, since such is basedon choosing two functions: assignation function and representation function.

For the assignation function, given the cores $\{A_1,\cdots,A_h\}$, partition $P=\{P_1,\cdots,P_h\}$ deducted is defined by:

$$P_i=\left\{q_1\in\Omega\middle|d(A_i,q_1)\le d(A_j,q_1)\ \forall i,j\right\}$$

In case of equality, q_1 shall be assigned to the lowest index class. Partitions *P* thus deducted from *L* are shown by $P=f(L)$, where *f* is an application of $\mathcal{L}_k$ in $\mathcal{P}_k$; that is: $f:\mathcal{L}_k\to\mathcal{P}_k$, and it is called assignation function.

For the representation function, given partition *P*, $L=\{A_1,\cdots,A_h\}$ cores are deducted as:

$$A_i = \left\{ q_1 \in \mathcal{L} \middle| q_1 \in \{q\} \text{ wich produce lowest possible dissimilarity } d\left(q_1, \mathcal{P}_i\right) \right\} \quad (8)$$

In order to ensure the unit of A_i, the set of q elements of Ω space minimizing $\sum_{q_1 \in A_i} d\left(q_1, \mathcal{P}_i\right)$ $\forall \mathcal{P}_i \subset \Omega$, exists and is unique. Therefore, the representation function exists.

QED

Observation 1. It is possible to define representation function from a given $f^{-1}: \mathcal{P}_k \to \mathcal{L}_k$, such that $f^{-1}(P) = L = \{A_1, \cdots, A_h\}$, since A_i are defined from $P = \{P_1, \cdots, P_h\}$ with (8).

Observation 2. With the *Theorem of Cores Optimal Criterion* and *Observation* 1, the algorithm implies alternatively implementing f and f^{-1} from a partition or k^{th} core randomly estimated. Every iteration implies applying function f from and $L \in \mathcal{L}_k$ element or function f^{-1} from a $P \in \mathcal{P}_k$ element.

3. Application

The attachment shows the questionnaire developed for application on the student population. The survey was partially national (center and north of the country) due, mainly, to the features of the student population (at this education level, the student population in Mexico is 10,803,868—both males and females—between 18 and 22 years old) and null financial support available for calculation of a probabilistic sample and its application (trip expenses of specialized survey personnel). The questionnaire was applied with the consent of the student, and students came from various higher education institutions (public and private) professors interested in the topic were also surveyed [9].

3.1. Data under Analysis

Data used and analyzed is a data table $I \times J$, with tabular arrangement: $k_{IJ} = \{k(i, j) \ \forall i \in I, j \in J\}$ [10], where I is the set of questionnaires with cardinality 1839 and the set of questions with cardinality 16. The definition of variables is shown in chart I.2 of the Annex, and its frequency structure is the following.

The use of the questionnaire with students of bachelor degrees of the public education system shows a log-normal distribution, the most participative students where those of mechatronics, while the less participative were those of mathematics. This is rather logical, since seeing a mathematician with a calculator is as horrible as seeing a software developer exploring a computer with a screwdriver. The semester variable shows a bimodal behavior where the most participative are freshmen. The variable grouping current type of calculator of the student, shows a leptokurtic distribution, where Casio calculators have the highest percentage, 55.07%, while Sharp calculators have the lowest percentage, 6.65%. The place of purchase of equipment variable shows the same leptokurtic distribution, where department stores have the highest percentage of sales of such equipment's. The influence on purchase by brand shows a behavior not defined. To study it, it has been defined in percentages where 50.9% of people in the survey answers that the name of the equipment influences 80% the purchase. The influence on purchase, due to its technical features, shows a distribution J, where 66.27% answers that it does influence in 80% [11].

3.2. Correlations

Since it is a well-known theory, its development is not shown here, we only mention that the calculation of correlations or *degree of association* among variables has been carried out based on ordinary Euclidian distance $d(j, j')$ among variables j and j'. Besides, it must be remembered that, if two variables are *strongly correlated*, those are near to each other $(c_{jj'} = 1)$ or, on the contrary, as far as possible from each other $(c_{jj'} = -1)$, as linear relationship linking them is direct or inverse, and that when $c_{jj'} = 0$ those are at middle distance or that j and j' are orthogonal. In box (k, j) there is $\mathrm{Cov}(x_k, x_j)$. The k^{th} diagonal term is $\mathrm{Var}(x_k)$. It should be noticed that symmetry of matrix: $\mathrm{Cov}(x_k, x_j) = \mathrm{Cov}(x_j, x_k)$. Regarding interpretation, variables with strongest correlation are brand and price, with 0.438, **Table 1**. Calculator brand and type of calculator, with −0.311, are correlated below.

Table 2 shows values obtained from the multiple correlation analysis of variables under study. Here, no variable shows a high multiple correlations. Most variables multiply correlated to 0.5 correlative values are: influence of make, price and type of calculating machine.

Table 1. Correlations between variables of use of technologies level among higher education students.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	M1	M2	M3	M4	M5	M6	M7
N1	1.000															
N2	−0.121	1.000														
N3	−0.029	−0.137	1.000													
N4	−0.032	−0.005	0.309	1.000												
N5	0.052	0.008	−0.018	−0.066	1.000											
N6	0.026	0.024	0.075	0.025	0.438	1.000										
N7	−0.005	0.030	0.019	−0.070	0.266	0.230	1.000									
N8	0.046	0.033	0.068	−0.015	0.010	0.063	0.011	1.000								
N9	−0.167	0.205	−0.311	−0.041	0.041	−0.039	0.049	−0.140	1.000							
M1	0.067	−0.089	0.050	0.001	0.088	0.052	0.143	0.027	−0.194	1.000						
M2	−0.085	−0.055	0.057	0.082	−0.031	−0.014	−0.059	−0.151	0.072	0.019	1.000					
M3	−0.107	0.018	0.005	0.011	0.092	0.058	0.093	−0.049	−0.041	0.057	0.025	1.000				
M4	−0.034	−0.012	0.080	0.069	−0.069	0.045	−0.060	0.014	−0.033	0.015	−0.000	−0.020	1.000			
M5	0.066	−0.081	0.050	−0.017	0.058	−0.008	0.019	0.027	−0.054	−0.003	0.020	−0.116	0.023	1.000		
M6	0.106	−0.123	0.053	−0.025	−0.031	0.026	−0.022	0.120	−0.087	−0.041	−0.008	−0.135	0.045	0.085	1.000	
M7	−0.010	0.065	0.053	−0.007	0.046	0.028	0.052	0.015	−0.036	−0.032	−0.064	0.007	−0.008	0.106	0.113	1.000

Table 2. Multiple correlations of variables under study.

N1	N2	N3	N4	N5	N6	N7	N8	N9	M1	M2	M3	M4	M5	M6	M7
0.275	0.301	0.361	0.157	**0.498**	**0.478**	0.351	0.249	**0.456**	0.278	0.240	0.248	0.165	0.215	0.279	0.207

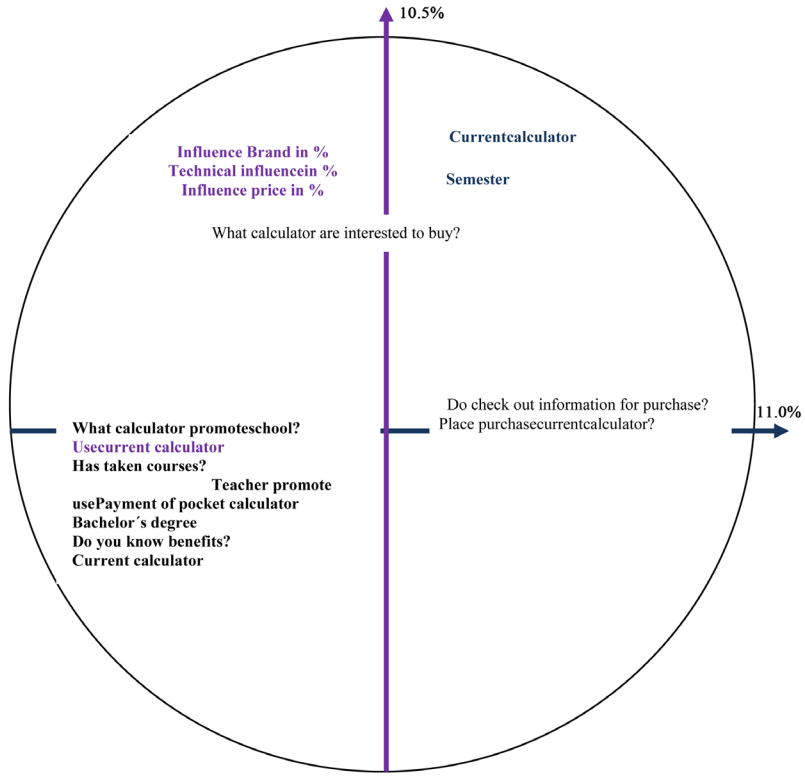
3.3. Principal Components Analysis

Let us now see the results of the Principal Component Analysis, PCA, on a tabular arrangement of gross data $I \times J$ (1839 × 16) on a correlations matrix. The theoretical description of the method is shown in [12], pp. 65-78.

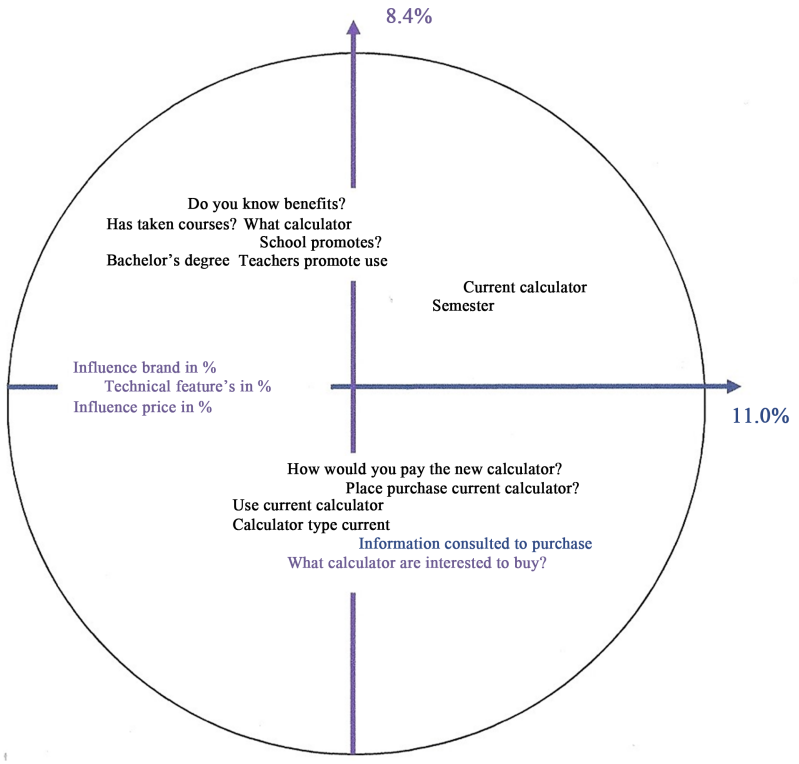
Interpretation of correlations circle 1 - 2, **Figure 1(a)** and **Figure 1(b)**, shows that the first two principal components explain 11.0% and 10.5%, respectively, that is, the first correlations circle contains 21.5% of gross data, and shows a contraposition between the type of calculator currently owned by a student (without knowing which type of calculator it is) and the semester he/she is in (without knowing in whish semester he/she is enrolled), *versus* brand, technical features of the equipment, price (every figure in percentage), and how much he/she uses the applications on his/her equipment. Regarding the second correlations circle, where the principal components 1 - 3 intervene, and which explains 18.9% of gross data (10.5% and 8.4%, respectively), it shows contraposition regarding the first component of type of calculator currently owned by the student (without knowing which type of calculator it is) and the information consulted before the purchase (without knowing if such includes brochures, recommendation or Internet), *versus* brand, technical features of the equipment, price (every figure in percentage), how much he/she uses the applications of his/her equipment, the type of calculator he/she currently owns and the calculator he/she would like to buy, as well as the knowledge he/she has about Texas Instrument calculators.

3.4. Hierarchical Ascending Classification with Euclidean Distance

The hierarchical dendrogram, built based on Euclidean distance, is composed of 3 branches, **Figure 2**. Reading and interpretation run from right to left, for hierarchical reasons [13].



(a)



(b)

Figure 1. Correlations circle. a) Principal Components 1 - 2; b) Principal Components 1 - 3.

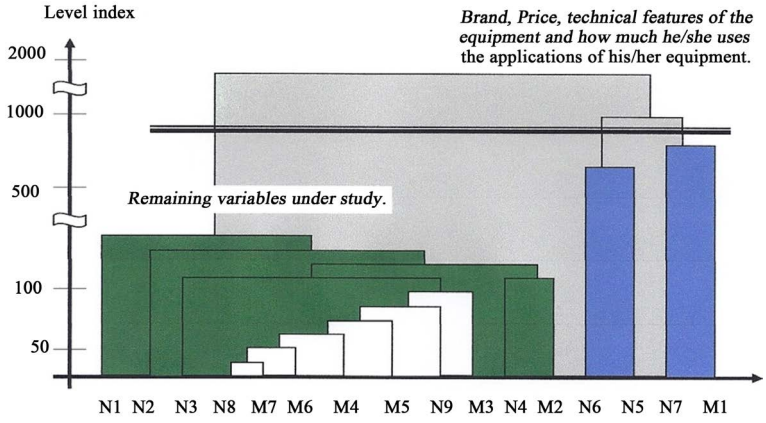


Figure 2. Hierarchical dendrogram of the use of technologies in higher education based on Euclidean distance (see **Table 2**, attachment, for definition of variables).

The dendrogram shows two aggregations of variables, the first one agglutinates variables making the first one a principal component: brand, price, technical features of the equipment and how much he/she uses the applications of his/her equipment. The second aggregation is composed by the remaining variables under study.

3.5. Factorial Analysis on Gross Data

The factorial method chosen to describe data under study is the Correspondence Analysis, CA, method. This method allows a direct search for the best simultaneous representation of sets under study; I questionnaires completed by students, and J variables describing the use of micro computing technologies in teaching practice. The CA applied to gross data K_{IJ} has the following factorial features: variances on the principal three axes or own values are: X_1 = 0.0502, X_2 = 0.0341 and X_3 = 0.0307, while percentages of habit explained by such axes are, respectively: 35.5%, 24.1% and 21.7%. The first factorial plane 1 - 2 has no defined shape and origin mass center. Variables of highest importance are brand, price, technical features of equipment and use of applications, with values ranging from 21.18 through 24.81. The first factorial axis is defined by the four variables mentioned above, of the highest importance in this study. The second factorial axis is defined by technical features in the purchase of the equipment and in its use. The third factor is defined by brand and price of the equipment.

3.6. Classes of Variables' Cut and Its Factors

Since the PCA and CA used on data do not show any relationship whatsoever between variables, it was necessary to fragment the first data table in a class table, [12], Chapter III. Let

$$k\left(i, c_r^{j_n}\right) \quad \forall i \in I \text{ and } c \text{ be classes of } I \text{ such that } j \in J \text{ and } r = 1, \cdots, m$$

that is, for every element I in the set of answers to variables determining the level of use of technologies in the practice of teaching mathematics, there is a set of variables J whose elements each contain a subset C called classes c_r, such that for each variable there are tabular arrangements $k\left(i, c_r^{j_n}\right)$ for $r = 1, \cdots, m$ with whole values between 1 and m. Ranges in which variables were fragmented are shown in Chart A.2, Annex I.

A table of generalized contingency has been created, based on the classes table *ibid* p. 28, Chapter III. The tabular arrangement created has a dimension of 1839 × 67 elements. Classes of highest importance in this study are: has not taken courses to use his/her calculator; technical features and price influence on purchase from 25% to 50%; already has a scientific calculator and is not interested in purchasing a new one. The less important ones in the study are: chemistry, materials and pure mathematics students, which is rather logical, since they are students of scientific specialty who do not need a calculator to carry out their professional studies.

The first factorial plane of the table in classes of the level of use of technologies among students of higher education has only 8% of data and has a slight parabolic structure, **Figure 3**. The first factor is composed by students of fourth semester, who use Texas Instruments symbolic calculators, students of mechatronics, who

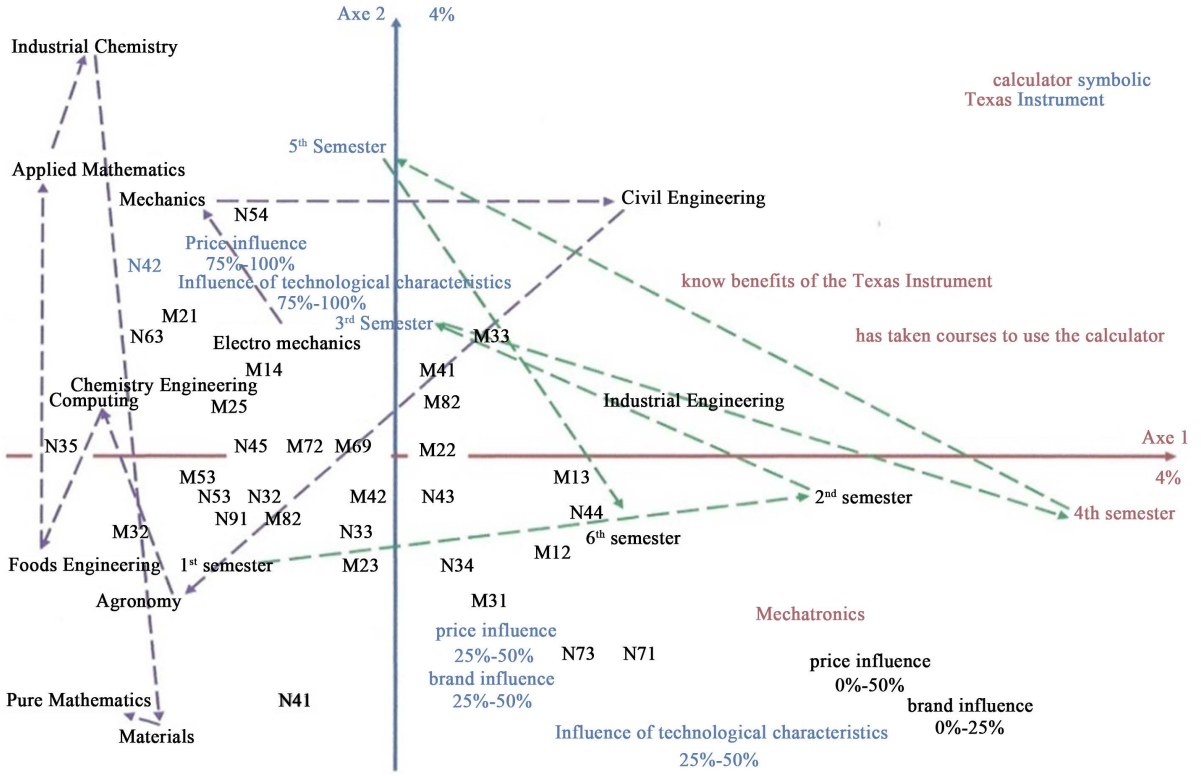


Figure 3. First factorial plane of use of technologies among higher education students in Mexico.

have taken courses to use such and know their benefits. The second factor is composed by influence of brand, weight and technical features of calculators for all percentages. The third factor is composed by students of fifth and sixth semester of all careers, who need an additional graph maker.

4. Use of Technologies Dendrogram

The hierarchical dendrogram built under the aggregation criterion of the central moment of order two, is composed by five branches, **Figure 4**. Reading and interpretation go from left to right; the hierarchical level scale has a maximum of 16 hierarchical units and the symbol near the 15th unit means a jump of scale units. In the bottom of the hierarchical structure the definition of class are briefly recorded.

The first hierarchical branch is composed of the second factor of factorial analysis, as well as some classes which do not show up in the analysis, such as the mechatronics and computing students then in the fourth and sixth semester of the career, who know how to use the equipment's under analysis. The second hierarchical branch is composed by three sub-branches: first, the most important classes in this study, that is, the chemistry and industrial engineering students who have a scientific calculator in first semester. The second sub-branch is composed by electronic, foods and civil engineering students in third semester, who know the benefits of such equipment's and are certain that the school and the teachers promote their use and purchase. The third sub-branch is composed by mechanics, applied mathematics and industrial chemistry students, who get the technical information with friends and show that the influence of price is 75%. The fourth and fifth hierarchical branches are rather a single branch, since their final aggregation comes after the cut and, put together, constitute the first factor.

5. Discussion of Results

This work is presented in accordance with its development. The theory developed on hierarchical cores is shown, where the method shown is tributary to three options: 1) calculation of distance between elements where factorial coordinates are known; 2) juxtaposition of mass or weight to each element; and 3) calculation of a distance between element classes, depending on an aggregation criterion based on hierarchical cores.

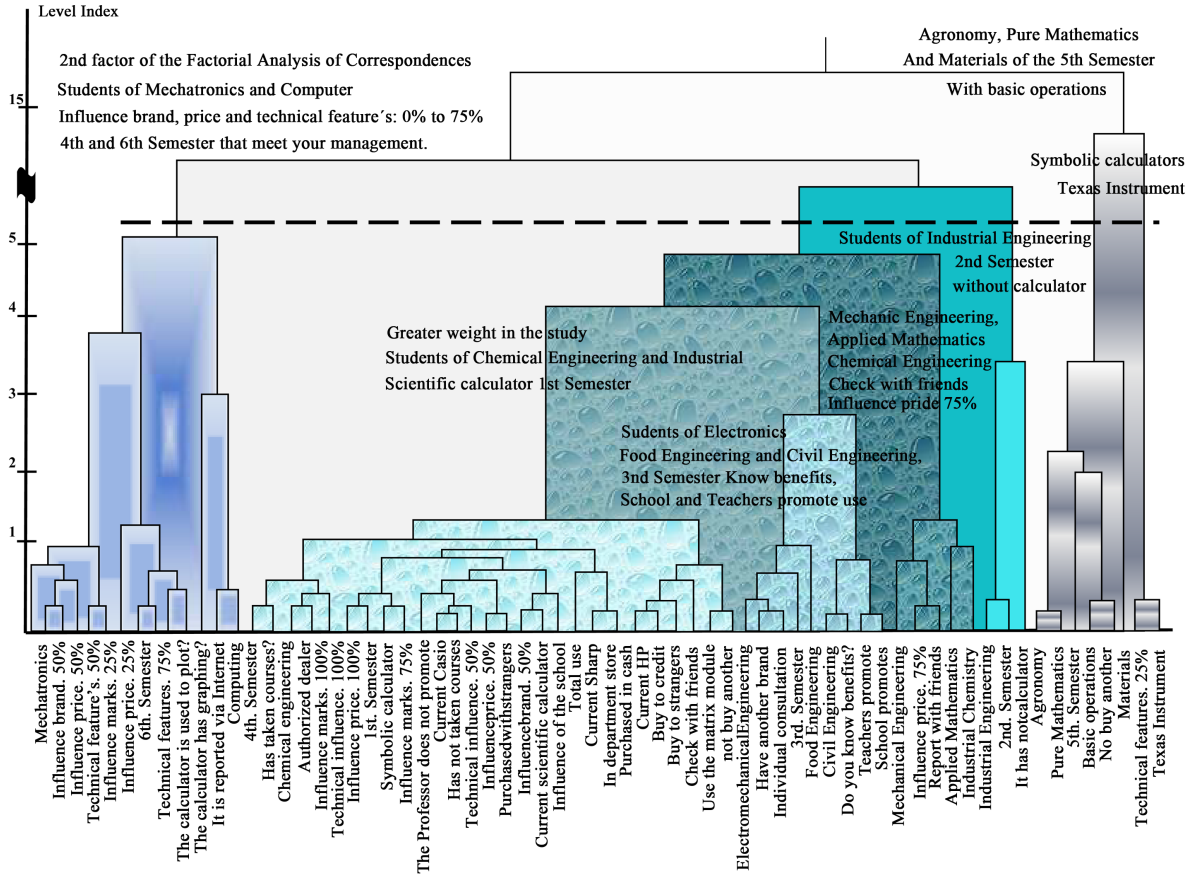


Figure 4. Dendrogram of use of technologies among higher level students.

Development of a proper data collection vehicle and its pilot test, provide enough data for national application and subsequent statistical analysis which allows constructing hierarchical cores based on an ascending hierarchical classification.

Results provided by linear statistical part are not enough to obtain conclusions on factors influencing quantification of CAS calculator's demand, basic fact influencing theoretical development. The first factorial plane of technologies use by higher education students in Mexico accounts for the path of classes making factors, which subsequently define hierarchical cores.

6. Conclusions

From the point of view of theory developed, it may be seen that from various starting points, the problem of looking for stable classes may be resolved. Starting points may be chosen by the user, with the help of a hierarchical classification.

The theorem demonstrated and called *Cores Optimal Criterion Theorem* allows implementing f *and* f^{-1} functions from a k^{th} core randomly estimated with the algorithm.

The purpose of analyzing and defining factors influencing the use of new technologies in the practice of teaching mathematical calculations in Mexico is achieved, since, as has been explained in the statistical analysis of data, it has been observed that the most important classes in this study are: 1) no courses to use the calculator; influence of technical features and price on the purchase; 2) 20% to 50% already has a scientific calculator and is not interested in purchasing a new one. The less important classes in this study are: chemistry, materials, and pure mathematics students. This is rather logical, since such are students of scientific specialty who do not need a calculator to carry out their professional studies.

The first factor is composed by mid-term engineering students using Texas Instruments symbolic calculators, who have taken courses to use them and know their benefits well. The second factor is composed by the influ-

ence of brand, weight and technical features of such calculation equipment's. The third factor is composed by students in the second half of the career, who need an additional graph maker.

From the point of view of hierarchical classification, the first branch is composed by the second factor of factorial analysis, as well as some classes which do not show up in the analysis, such as the engineering students who are halfway through their degree, who know how to use the equipment under analysis. The second hierarchical branch is composed by three sub-branches: first, the most important class in this study is the engineering students who have a scientific calculator in first semesters. The second sub-branch is composed by engineering students in third semester, who know the benefits of such equipment's and are certain that the school and the teachers promote their use and purchase. The third sub-branch is composed by engineering students, who get the technical information with friends and show that the influence of price is 75%.

Acknowledgements

I thank the disinterested collaboration and datacontribution of Myrna E. González Meneses, M.S., to carry out this multidimensional data analysis. I also acknowledge the contribution for research project recorded at the National Polytechnic Institute. Mexico with number SIP-20130585.

References

[1] González Meneses, M.E. (2007) Estudio de impacto en el uso de calculadoras con sistemas algebraicos (Tecnologías CAS) en instituciones de educación superior. Instituto Tecnológico de Apizaco. Dirección de Educación Superior Tecnológica. Secretaría de Educación Pública, Mexico.

[2] Lagrange, J.B., Artigue, M., Laborde, C. and Trouche, L. (2003) Technology and Mathematics Education: Multidimensional Overview of Recent Research and Innovation. In: Bishop, A.J., Clements, M.A., Keitel, C., Kill Patrick, J. and Leung, F.K.S., Eds., *Second International Handbook of Mathematics Education*, Vol. 1, Kluwer Academic Publishers, Dordrecht, 237-270. http://dx.doi.org/10.1007/978-94-010-0273-8_9

[3] Rodríguez, J.R. and Flores López, L.F. (2005) Propuesta didáctica para el cálculo usando calculadoras Texas Instruments. InstitutoTecnológico de los Mochis, Sonora, Mexico.

[4] Buteau, C. and Muller, E. (2006). Evolving Technologies Integrated into Undergraduate Mathematics Education. In: Son, L.H., Sinclair, N., Lagrange, J.B. and Hoyles, C., Eds., *Proceedings of the ICMI* 17 *Study Conference*: *Background Papers for the ICMI* 17 *Study*, Hanoi University of Technology, Hanoi.

[5] OECD (2004) Learning for Tomorrow's World—First Results from PISA 2003. OECD—Programme for International Student Assessment. http://www.pisa.oecd.org/dataoecd/1/60/34002216.pdf

[6] Marion, M. and Signonello, S. (2011) From Histogram Data to Model Data Analysis. In: Fichet, B., *et al*., Eds., *Classification and Multivariate Analysis for Complex Data Structures*, *Studies in Classification Data Analysis, and Knowledge Organization*, Springer-Verlag, Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-13312-1_38

[7] Diday, E. and Noirhomme, M. (2008) Symbolic Data Analysis and the SODAS Software. Wiley, Hoboken, 457 p.

[8] Diday, E. (2008) Spatial Classification. DAM (Discrete Applied Mathematics). Vol. 156.

[9] Gonzáles, P., Guzmán, J.C., Partelow, L., Pahlke, E., Jocely, L., Kastberg, D. and Williams, T. (2004) Highlights from the Trends in International Mathematics and Science Study: TIMSS 2003. National Center for Education Statistics Institute of Education Sciences, U.S. Department of Education.
http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005005

[10] Benzécri, J.P. and Benzécri F. (1980) Pratique de L´Analyse des Données. 1. Analyse des Correspondances. Exposé Élémentaire. Dunod. Bordas, Paris.

[11] Ruthven, K. and Hennessy, S. (2002) A Practitioner Model of the Use of Computer-Based Tools and Resources to Support Mathematics Teaching and Learning. *Educational Studies in Mathematics*, **49**, 47-88.
http://dx.doi.org/10.1023/A:1016052130572

[12] Casanova-del-Angel, F. (2001) Análisis multidimensional de datos. Editorial Logiciels, Mexico.

[13] Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis. 5th Edition, Wiley Series in Probability and Statistics. http://dx.doi.org/10.1002/9780470977811

Annex I Questionnaire

Data obtained from this questionnaire aims at determining the level of use of new technologies in teaching mathematics and developing a marketing proposal for some calculator models Texas Instruments.

BRAND PRODUCT

Bachelor's degree in engineering: ____ Semester: ____ Age: _____ Do you work? _____

1. What calculator do you have now? ☐ Texas Instruments ☐ Casio ☐ HP ☐ Sharp ☐ Other:

2. Where did you buy your current calculator?

PLACEMENT OF PRODUCT

☐ I don't have one ☐ Authorized distributor ☐ Department store ☐ From an acquaintance ☐ Other

3. What percentage influences you to buy a calculator?

Brand product _____ (0% - 100%) **BRAND PRODUCT**
Price _____ (0% - 100%) **PRICE**
Technical features_____ (0% - 100%) **USE OF PRODUCT**

4. Have you taken any course to use a calculator? ☐ Yes ☐ No **TRAINING**

5. Choose the type of calculator you currently have (no matter the make, only the features of the model)

USE OF PRODUCT

SCIENTIFIC
Any model
☐

GRAPHICS
Any model.
Casio, HP, TI
☐

SIMBOLIC
TI89, TI92, Voyage 200
Casio, ClassPad 300
☐

6. How many of the calculator's applications do you use? **TRAINING**

Basic operations, statistics _____ (0% - 100%)
Basic operations, statistics, graph making, programming, matrixes _____ (0% - 100%)
Basic operations, geometry, graph making, programming, differential and integral calculus, statistics, finance, word processor, simultaneous equations, polynomial roots, _____(0% - 100%)

7. When you buy a calculator, which data do you consult? **ADVERTISEMENT**

☐ Pamphlets ☐ Acquaintances ☐ School ☐ Internet ☐ Other _____

8. Which type of calculator would you like to buy (even if you already have one)? **PRICE**

GRAPHICS
Approx. price
150 USD
☐

SIMBOLIC
Approx. price
200 USD
☐

None
☐

9. If you would buy any of the above calculators, how would you pay it? ☐ Cash ☐ Credit

SELLING PLANS

10. Mark if your teachers

☐ Promote use of graph making calculators or symbolic calculation calculators in any subject.
☐ Always ☐ Sometimes ☐ Never

PROMOTION

11. Do you know the benefits offered by Texas Instruments regarding technical support?

☐ Yes ☐ No

ADVERTISEMENT

12. Does your school promote the visit or calculator promoters? ☐ Yes ☐ No

PROMOTION, SELLING PLANS

Table Annex I.1. Statistical parameters of variables under study.

Variable	Max. value	Min. value	Arithmetical mean	Standard deviation	Variation coefficient	Symmetry coefficient	Kurtosis coefficient
N1	1	13	4.381	3.155	71.967	0.7409	2.829
N2	1	6	2.568	1.751	68.163	0.496	2.135
N3	1	5	2.524	1.051	41.644	1.134	3.503
N4	1	5	3.365	0.910	27.033	0.040	2.544
N5	1	100	65.292	30.161	46.167	0.640	2.546
N6	0	100	64.450	29.968	46.471	0.425	2.460
N7	0	100	75.479	30.236	40.035	1.682	3.588
N8	1	2	1.940	0.236	12.200	13.843	14.843
N9	1	3	1.232	0.552	44.792	5.275	7.062
M1	5	100	70.380	22.949	32.588	0.371	2.494
M2	1	5	2.749	1.254	45.606	0.006	1.988
M3	1	3	2.548	0.682	26.748	1.445	3.112
M4	1	2	1.547	0.498	32.175	0.003	1.035
M5	1	3	2.108	0.645	30.594	0.011	2.371
M6	1	2	1.778	0.415	23.358	1.795	2.795
M7	1	2	1.897	0.303	15.984	6.870	7.870

Table Annex I.2. Classes' cut of variables of use of technologies in higher education.

Variable	No. of classes	Mnemonics of class	Value of class	Elements of class
N1. Bachelor's degree	13	N11	Mechatronics	189
		N12	Chemistry Mechatronics	149
		N13	Industrial Engineering	132
		N14	Electro mechanics	102
		N15	Mechanics	74
		N16	Civil Engineering	72
		N17	Agronomy	68
		N18	Computing	64
		N19	Foods	56
		N01	Applied mathematics	37
		N02	Industrial chemistry	34
		N03	Materials	29
		N04	Pure mathematics	28
N2. Semester	6	N21	1st semester	401
		N22	2nd semester	76
		N23	3rd semester	179

Continued

		N24	4th semester	83
		N25	5th semester	92
		N26	6th semester	98
N3. Current calculator	5	N31	Texas Instruments	88
		N32	Casio	477
		N33	HP	187
		N34	Sharp	70
		N35	Other	92
	5	N41	Does not have	22
N4. Place of purchase of current calculator		N42	Authorized dealer	133
		N43	Department store	401
		N44	Someone known	232
		N45	Other	126
N5. % of brand influence	4	N51	0% - 25%	143
		N52	>25% - 50%	175
		N53	>50% - 75%	139
		N54	>75% - 100%	442
N6. % of price influence	4	N61	0% - 25%	134
		N62	>25% - 50%	222
		N63	>50% to 75%	145
		N64	>-75% to 100%	398
N7. % of technical features influence	4	N71	0% - 25%	105
		N72	>25% - 50%	118
		N73	>50% - 75%	105
		N74	>75% - 100%	571
N8. Has taken courses to use the calculator	2	N81	Has taken a course	65
		N82	Has not taken courses	804
N9. Type of current calculator	**3**	N91	Scientific	712
		N92	Graph maker	104
		N93	Symbolic	68
M1. How much he/she uses his/her current calculator	4	M11	Statistical operations	54
		M12	M11 + graph making	214
		M13	M12 + matrixes	171
		M14	M13 + text editor	460
M2. Which information he/she consulted to purchase	5	M21	Brochures	200
		M22	A friend	176

Continued

		M23	School	262
		M24	Internet	186
		M25	Other	90
M3. Which additional calculator would you like to buy?	3	M31	$2000.00 graph maker	106
		M32	$3000.00 symbolic calculator	212
		M33	None	566
M4. How would you pay the new one?	2	M41	Cash	395
		M42	Credit	474
M5. Do teachers promote the use of pocket calculators?	3	M51	Always	149
		M52	Sometimes	495
		M53	Never	240
M6. Do you know its benefits?	2	M61	Does know benefits	200
		M62	Does not know benefits	667
M7. Does the school promote such equipment?	2	M71	Yes, it promotes such	101
		M72	No, it does not promote such	767

Open Journal of Statistics, 2014, 4, 851-856
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410080

Estimation of Multivariate Sample Selection Models via a Parameter-Expanded Monte Carlo EM Algorithm

Phillip Li

Department of Economics, Office of the Comptroller of the Currency, Washington, DC, USA
Email: Phillip.Li@occ.treas.gov

Received 6 September 2014; revised 5 October 2014; accepted 2 November 2014

Abstract

This paper develops a parameter-expanded Monte Carlo EM (PX-MCEM) algorithm to perform maximum likelihood estimation in a multivariate sample selection model. In contrast to the current methods of estimation, the proposed algorithm does not directly depend on the observed-data likelihood, the evaluation of which requires intractable multivariate integrations over normal densities. Moreover, the algorithm is simple to implement and involves only quantities that are easy to simulate or have closed form expressions.

Keywords

Multivariate Sample Selection, Heckman Correction, Incidental Truncation, Expectation Maximization

1. Introduction

Sample selection models, pioneered in [1]-[3], are indispensable to researchers who use observational data for statistical inference. Among the many variants of these types of models, there is a growing interest in multivariate sample selection models. These are used to model a system of two or more seemingly unrelated equations, where the outcome variable for each equation may be non-randomly missing or censored according to its own stochastic selection variable. Applications range from modeling systems of demand equations [4] [5] to household vehicle usage [6]-[8]. A common specification is to assume a correlated multivariate normal distribution underlying both the outcomes of interest and the latent variables in the system.

There are two dominant approaches in the current literature to estimate these models. One approach is to use

maximum likelihood (ML) estimation. However, as noted in the literature, a major hurdle in evaluating the likelihood is that it requires computations of multivariate integrals over normal densities, which do not generally have closed form solutions. [9] discusses the ML estimation of these models and proposes to use the popular Geweke, Hajivassiliou, and Keane (GHK) algorithm to approximate these integrals in a simulated ML framework. While this strategy works reasonably well, the GHK algorithm can be difficult to implement. Another popular approach is to use two-step estimation (see [10] for a survey). In general, there is a tradeoff in the statistical properties and the computational simplicity for these estimators. If efficiency and consistency are of primary concern, then ML estimation should be preferred over two-step estimation.

The objective of this paper is to develop a simple ML estimation algorithm for a commonly used multivariate sample selection model. In particular, this paper develops a parameter-expanded Monte Carlo expectation maximization (PX-MCEM) algorithm that differs from [9] in a few important ways. First, the PX-MCEM algorithm does not use the observed-data likelihood directly, so it avoids the aforementioned integrations. Second, the proposed iterative algorithm does not require the evaluations of gradients or Hessians, which become increasingly difficult to evaluate with more parameters and equations. Third, the algorithm is straightforward to implement. It only depends on quantities that are either easy to simulate or have closed form expressions. This last point is especially appealing when estimating the covariance matrix parameter since there are non-standard restrictions imposed onto it for identification.

This paper is organized as follows. The multivariate sample selection model (MSSM) is formulated in Section 2. Section 3 begins with a brief overview of the EM algorithm for the MSSM and continues with the development of the PX-MCEM algorithm. Methods to obtain the standard errors are discussed. Section 4 offers some concluding remarks.

2. Multivariate Sample Selection Model

The MSSM is

$$y_{i,j}^* = x_{i,j}'\beta_j + \epsilon_{i,j} \tag{1}$$

$$s_{i,j}^* = w_{i,j}'\gamma_j + \nu_{i,j} \tag{2}$$

$$s_{i,j} = \mathbb{I}\left(s_{i,j}^* > 0\right) \tag{3}$$

$$y_{i,j} = \begin{cases} y_{i,j}^* & \text{if } s_{i,j} = 1 \\ \text{missing} & \text{if } s_{i,j} = 0 \end{cases} \tag{4}$$

for observations $i = 1, \cdots, N$, and equations $j = 1, \cdots, J$. In the previous expressions, $y_{i,j}^*$ is the continuous outcome of interest for observation i and equation j. Using similar indexing notation, $s_{i,j}^*$ is the latent variable underlying the binary selection variable $s_{i,j} = \mathbb{I}\left(s_{i,j}^* > 0\right)$, where $\mathbb{I}(A)$ denotes an indicator function that equals 1 if event A is true and 0 otherwise. Sample selection is incorporated by assuming that $y_{i,j}^*$ is missing when $s_{i,j} = 0$. Otherwise, $y_{i,j}^*$ is observed and equal to $y_{i,j}$. For later use, define $s_i = \left(s_{i,1}, \cdots, s_{i,J}\right)'$ and $s_i^* = \left(s_{i,1}^*, \cdots, s_{i,J}^*\right)'$, where the prime symbol in s_i, s_i^*, and in the rest of this paper is used to denote matrix transpose.

Furthermore, $x_{i,j}$ and $w_{i,j}$ are column vectors of exogenous covariates, and β_j and γ_j are conforming vectors of parameters. Define $\beta = \left(\beta_1', \beta_2', \cdots, \beta_J'\right)'$ and $\gamma = \left(\gamma_1', \gamma_2', \cdots, \gamma_J'\right)'$. For identification, $w_{i,j}$ must contain at least one exogenous covariate that does not overlap with $x_{i,j}$ (refer to [11] for these exclusion restrictions). The unobserved errors $\epsilon_i = \left(\epsilon_{i,1}, \epsilon_{i,2}, \cdots, \epsilon_{i,J}\right)'$ and $\nu_i = \left(\nu_{i,1}, \nu_{i,2}, \cdots, \nu_{i,J}\right)'$ are jointly distributed as a $2J$-dimensional multivariate normal with a mean vector of zeros and an unknown covariance matrix of Ω. Formally, $\left(\epsilon_i', \nu_i'\right)' \overset{iid}{\sim} \mathcal{N}_{2J}(0, \Omega)$ with

$$\Omega = \begin{pmatrix} \underset{(J\times J)}{\Omega_{\epsilon\epsilon}} & \Omega_{\epsilon\nu} \\ \Omega'_{\epsilon\nu} & \underset{(J\times J)}{\Omega_{\nu\nu}} \end{pmatrix}. \tag{5}$$

The submatrix $\Omega_{\nu\nu}$ is restricted to be in correlation form to identify the parameters corresponding to the latent variables [9]. The other elements of Ω are restricted such that the matrix is symmetric and positive definite.

The covariates and binary selection variables are always observed. Without loss of generality, assume that the outcomes for any observation i are only missing for the first m_i equations, where $0 \le m_i \le J$. Define $y_{i,\text{obs}} = \left(y_{i,m_i+1}, \cdots, y_{i,J}\right)'$, and let $y_{\text{obs}} = \left\{y_{i,\text{obs}}, s_i\right\}_{i=1}^{N}$ denote the observed data. The observed-data likelihood derived from (1) through (5) is denoted as $f\left(y_{\text{obs}} \middle| \beta, \gamma, \Omega\right)$. See [9] for an exact expression of this likelihood.

3. Estimation

3.1. Overview of the EM Algorithm

The PX-MCEM algorithm is based on the EM algorithm of [12]. The basic idea behind the EM algorithm is to first augment y_{obs} with a set of "missing data" y_{mis} such that the observed-data likelihood is preserved when the missing data are integrated out of the complete-data likelihood. Formally, the missing data must satisfy

$$f\left(y_{\text{obs}} \middle| \beta, \gamma, \Omega\right) = \mathbb{E}\left[f\left(y_{\text{mis}}, y_{\text{obs}} \middle| \beta, \gamma, \Omega\right)\right], \tag{6}$$

where $f\left(y_{\text{mis}}, y_{\text{obs}} \middle| \beta, \gamma, \Omega\right)$ is the complete-data likelihood to be defined later.

The EM algorithm then proceeds iteratively between an expectation step (E-step) and a maximization step (M-step) as follows. In iteration $(t+1)$ of the algorithm, compute in the E-step

$$Q\left(\beta, \gamma, \Omega \middle| \beta^{(t)}, \gamma^{(t)}, \Omega^{(t)}\right) = \mathbb{E}\left[\log\left(f\left(y_{\text{mis}}, y_{\text{obs}} \middle| \beta, \gamma, \Omega\right)\right)\right], \tag{7}$$

where the expectation is taken with respect to the conditional predictive distribution for the missing data, $\pi\left(y_{\text{mis}}, y_{\text{obs}} \middle| \beta^{(t)}, \gamma^{(t)}, \Omega^{(t)}\right)$, and in the M-step, find

$$\underset{\beta,\gamma,\Omega}{\arg\max}\, Q\left(\beta, \gamma, \Omega \middle| \beta^{(t)}, \gamma^{(t)}, \Omega^{(t)}\right). \tag{8}$$

Denote the maximal values as $\beta^{(t+1)}$, $\gamma^{(t+1)}$, and $\Omega^{(t+1)}$, and continue on with the algorithm until convergence. The final maximal values are at least local maxima of the observed-data likelihood function.

For the MSSM, y_{mis} consists of all the missing outcomes and latent variables. Specifically, $y_{\text{mis}} = \left\{y_{i,\text{mis}}, s_i^*\right\}_{i=1}^{N}$, where $y_{i,\text{mis}} = \left(y_{i,1}^*, \cdots, y_{i,m_i}^*\right)'$. Furthermore, denote $y_{i,\text{com}} = \left(y'_{i,\text{mis}}, y'_{i,\text{obs}}, s_i^{*\prime}\right)'$ as the vector of complete data, X_i as a block-diagonal matrix with the rows of covariates corresponding to the elements of $y_{i,\text{com}}$ on its block diagonals, and $\theta = \left(\beta', \gamma'\right)'$. The complete-data likelihood for the MSSM is given by

$$f\left(y_{\text{mis}}, y_{\text{obs}} \middle| \beta, \gamma, \Omega\right) = \prod_{i=1}^{N} f\left(y_{i,\text{com}} \middle| \beta, \gamma, \Omega\right) p\left(s_i \middle| y_{i,\text{com}}\right) \tag{9}$$

with $f\left(y_{i,\text{com}} \middle| \beta, \gamma, \Omega\right) = \phi_{2J}\left(y_{i,\text{com}} \middle| X_i\theta, \Omega\right)$ which is a density function for a $2J$-dimensional multivariate normal with mean $X_i\theta$ and covariance Ω, and

$$p\left(s_i \middle| y_{i,\text{com}}\right) = \prod_{j=1}^{J}\left\{\mathbb{I}\left(s_{i,j} = 1\right)\mathbb{I}\left(s_{i,j}^* > 0\right) + \mathbb{I}\left(s_{i,j} = 0\right)\mathbb{I}\left(s_{i,j}^* \le 0\right)\right\}. \tag{10}$$

Equation (10) is a degenerate density since conditioning on s_i^* in $y_{i,\text{com}}$ determines s_i from (3). Note that the observed-data likelihood from [9] is obtained when y_{mis} is integrated out of (9), hence the condition in (6) holds.

3.2. PX-MCEM Algorithm

The standard EM algorithm using (7) and (8) is difficult to implement for the MSSM as the E-step and M-step are intractable. The PX-MCEM algorithm addresses this issue by modifying the E-step in two ways and leads to an M-step that can be evaluated with closed form quantities. Stated succinctly, the PX-MCEM algorithm is as follows.

1. Initialize $\beta^{(0)}$, $\gamma^{(0)}$, $\Omega^{(0)}$, and the number of Gibbs sampling draws G.

At iteration $t+1$:

2. Draw G sets of missing data, denoted by $y_{\text{mis}}^{(1)},\cdots,y_{\text{mis}}^{(G)}$, from $\pi\left(y_{\text{mis}}\middle|y_{\text{obs}},\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)$ using Gibbs sampling.

3. PX-MC E-step: Estimate $Q\left(\alpha,\delta,\Sigma\middle|\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)$ as

$$Q_G\left(\alpha,\delta,\Sigma\middle|\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)=\frac{1}{G}\sum_{g=1}^{G}\log\left(f\left(y_{\text{mis}}^{(g)},y_{\text{obs}}\middle|\alpha,\delta,\Sigma\right)\right). \tag{11}$$

4. PX-MC M-step: Maximize $Q_G\left(\alpha,\delta,\Sigma\middle|\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)$ with iterative generalized least squares (IGLS) to obtain the maximizing parameters $\alpha^{(t+1)}$, $\delta^{(t+1)}$, and $\Sigma^{(t+1)}$.

5. Reduction step: Apply reduction functions to $\alpha^{(t+1)}$, $\delta^{(t+1)}$, and $\Sigma^{(t+1)}$ to obtain $\beta^{(t+1)}$, $\gamma^{(t+1)}$, and $\Omega^{(t+1)}$.

6. Repeat Steps 2 through 5 until convergence. The converged values are the ML estimates $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\Omega}$.

Each step is described in more detail in the subsequent sections.

3.2.1. PX-MC E-Step

Following [13], the first modification is to expand the parameter space of the complete-data likelihood function from (β,γ,Ω) to (α,δ,Σ). The expanded parameters play similar roles as the original parameters, however Σ is expanded into a standard covariance matrix without the correlation restrictions. The parameter-expanded complete-data likelihood function is

$$f\left(y_{\text{mis}},y_{\text{obs}}\middle|\alpha,\delta,\Sigma\right)=\prod_{i=1}^{N}f\left(y_{i,\text{com}}\middle|\alpha,\delta,\Sigma\right)p\left(s_i\middle|y_{i,\text{com}}\right) \tag{12}$$

with $f\left(y_{i,\text{com}}\middle|\alpha,\delta,\Sigma\right)=\phi_{2J}\left(y_{i,\text{com}}\middle|X_i\Theta,\Sigma\right)$, where $\Theta=\left(\alpha',\delta'\right)'$, and $\alpha=\left(\alpha_1',\cdots,\alpha_J'\right)'$ and $\delta=\left(\delta_1',\cdots,\delta_J'\right)'$ are defined analogously to β and γ. The advantage of using (12) instead of (9) is that Σ is easier to work with in the PX-MC M-step.

Second, instead of computing $Q\left(\alpha,\delta,\Sigma\middle|\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)=\mathbb{E}\left[\log\left(f\left(y_{\text{mis}},y_{\text{obs}}\middle|\alpha,\delta,\Sigma\right)\right)\right]$ analytically, it is approximated as (11) with Monte Carlo methods and Gibbs sampling. To draw from $\pi\left(y_{\text{mis}}\middle|y_{\text{obs}},\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)$, simply draw $y_{i,\text{mis}}$ and s_i^* from the conditional distribution $\pi\left(y_{i,\text{mis}},s_i^*\middle|y_{i,\text{obs}},s_i,\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)$ for $i=1,\cdots,N$. From (9), we have that

$$\pi\left(y_{i,\text{mis}},s_i^*\middle|y_{i,\text{obs}},s_i,\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\right)\propto\phi_{2J}\left(y_{i,\text{com}}\middle|X_i\theta^{(t)},\Omega^{(t)}\right)p\left(s_i\middle|y_{i,\text{com}}\right), \tag{13}$$

where $\theta^{(t)}=\left(\beta^{(t)\prime},\gamma^{(t)\prime}\right)'$. For the missing outcomes, it is easy to see from (13) that

$$y_{i,j}^*\middle|y_{i,\text{mis}(-j)},s_i^*,y_{i,\text{obs}},s_i,\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\sim\mathcal{N}_1\left(\mu_{i,j|(-j)},\sigma_{i,j|(-j)}^2\right) \tag{14}$$

for $j=1,\cdots,m_i$, where $y_{i,\text{mis}(-j)}$ is equivalent to $y_{i,\text{mis}}$ with $y_{i,j}^*$ removed, and $\mu_{i,j|(-j)}$ and $\sigma_{i,j|(-j)}^2$ are respectively the conditional mean and variance of $y_{i,j}^*$ given all other elements in $y_{i,\text{com}}$ from

$$\phi_{2J}\left(y_{i,\text{com}}\middle|X_i\theta^{(t)},\Omega^{(t)}\right).$$

Similarly, for the latent variables,

$$s_{i,j}^{*}\left|s_{i,(-j)}^{*},y_{i,\text{mis}},y_{i,\text{obs}},s_{i},\beta^{(t)},\gamma^{(t)},\Omega^{(t)}\sim\mathcal{TN}_{B_{i,j}}\left(\lambda_{i,j|(-j)},\omega_{i,j|(-j)}^{2}\right)\right. \tag{15}$$

for $j=1,\cdots,J$, where $\mathcal{TN}_{A}\left(a,b^{2}\right)$ denotes a univariate normal distribution with mean a and variance b^{2} truncated to the region A. In (15), $s_{i,(-j)}^{*}$ is s_{i}^{*} with $s_{i,j}^{*}$ removed, $B_{i,j}$ is the interval $\left(-\infty,0\right]$ if $s_{i,j}=0$ and $\left(0,+\infty\right)$ otherwise, and $\lambda_{i,j|(-j)}$ and $\omega_{i,j|(-j)}^{2}$ are respectively the conditional mean and variance of $s_{i,j}^{*}$ given all other elements of $y_{i,\text{com}}$ from $\phi_{2J}\left(y_{i,\text{com}}\left|X_{i}\theta^{(t)},\Omega^{(t)}\right.\right)$.

The Gibbs sampler recursively samples from the full conditional distributions in (14) and (15) in the usual way. After a sufficient burn-in period, the last G draws are used in (11).

3.2.2. PX-MC M-Step and Reduction Step

By recognizing that (11) is proportional to the log-likelihood function of a seemingly unrelated regression model with NG observations and $2J$ equations, the maximization can be performed with IGLS. IGLS utilizes the quantities

$$\tilde{\Theta}=\left(\sum_{g=1}^{G}\sum_{i=1}^{N}X_{i}'\tilde{\Sigma}^{-1}X_{i}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{N}X_{i}'\tilde{\Sigma}^{-1}y_{i,\text{com}}^{(g)}\right) \tag{16}$$

and

$$\tilde{\Sigma}=\frac{1}{NG}\sum_{g=1}^{G}\sum_{i=1}^{N}\left(y_{i,\text{com}}^{(g)}-X_{i}\tilde{\Theta}\right)\left(y_{i,\text{com}}^{(g)}-X_{i}\tilde{\Theta}\right)', \tag{17}$$

where $y_{i,\text{com}}^{(g)}$ is equivalent to $y_{i,\text{com}}$ with $y_{i,\text{mis}}^{(g)}$ and $s_{i}^{*(g)}$. First evaluate (16) with $\tilde{\Sigma}^{-1}$ removed, which amounts to estimating Θ equation by equation, and then evaluate (17) based on $\tilde{\Theta}$. Proceed by iterating (16) and (17) recursively until convergence. Denote the converged values as $\alpha^{(t+1)}$, $\delta^{(t+1)}$, and $\Sigma^{(t+1)}$.

In the reduction step, set $\beta^{(t+1)}=\alpha^{(t+1)}$, $\gamma_{j}^{(t+1)}=\delta_{j}^{(t+1)}\big/d_{J+j}$ $\left(1\le j\le J\right)$, and $\Omega^{(t+1)}=D^{-1}\Sigma^{(t+1)}D^{-1}$, where $D=\text{diag}\left(1,\cdots,1,d_{J+1},\cdots,d_{2J}\right)$ is a $2J\times 2J$ diagonal matrix with the first J diagonals equal to 1 and the remaining J diagonals equal to the square root of the last J diagonals of $\Sigma^{(t+1)}$. The previous transformations are referred to as the reduction functions, and they are needed because (12) is used instead of (9) in the algorithm [13].

3.3. Standard Errors

The observed information matrix is

$$-\mathbb{E}\left\{\frac{\partial^{2}\log\left(f\left(y_{\text{mis}},y_{\text{obs}}\left|\beta,\gamma,\Omega\right.\right)\right)}{\partial\Psi\partial\Psi'}\right\}-\mathbb{V}\left\{\frac{\partial\log\left(f\left(y_{\text{mis}},y_{\text{obs}}\left|\beta,\gamma,\Omega\right.\right)\right)}{\partial\Psi}\right\}, \tag{18}$$

where $\Psi=\left(\beta',\gamma',\Xi'\right)'$, and Ξ is a column vector denoting the unique elements in Ω. Evaluate (18) at the ML estimates, and take the expectation and variance with respect to $\pi\left(y_{\text{mis}}\left|y_{\text{obs}},\hat{\beta},\hat{\gamma},\hat{\Omega}\right.\right)$. These moments are estimated by taking additional draws from the Gibbs sampler and constructing their Monte Carlo analogs. The standard errors are the square roots of the diagonals of the inverse estimated quantity in (18).

4. Concluding Remarks

A new and simple ML estimation algorithm is developed for multivariate sample selection models. Roughly speaking, the implementation of this algorithm only involves iteratively drawing sets of missing data from well-known distributions and using IGLS on the complete data, both of which are inexpensive to perform. By using parameter expansion and Monte Carlo methods, the algorithm only depends on quantities with closed form

expressions, even when estimating the covariance matrix parameter with correlation restrictions. This algorithm is readily extendable to other types of selection models, including extensions to various types of outcome and selection variables with an underlying normal structure, and modifications to time-series or panel data.

Acknowledgements

I would like to thank the referee, Alicia Lloro, Andrew Chang, Jonathan Cook, and Sibel Sirakaya for their helpful comments.

References

[1] Heckman, J. (1974) Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, **42**, 679-694. http://dx.doi.org/10.2307/1913937

[2] Heckman, J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **5**, 475-492.

[3] Heckman, J. (1979) Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153-161. http://dx.doi.org/10.2307/1912352

[4] Su, S.J. and Yen, S.T. (2000) A Censored System of Cigarette and Alcohol Consumption. *Applied Economics*, **32**, 729-737. http://dx.doi.org/10.1080/000368400322354

[5] Yen, S.T., Kan, K. and Su, S.J. (2002) Household Demand for Fats and Oils: Two-Step Estimation of a Censored Demand System. *Applied Economics*, **34**, 1799-1806. http://dx.doi.org/10.1080/00036840210125008

[6] Hao, A.F. (2008) A Discrete-Continuous Model of Households' Vehicle Choice and Usage, with an Application to the Effects of Residential Density. *Transportation Research Part B*: *Methodological*, **42**, 736-758. http://dx.doi.org/10.1016/j.trb.2008.01.004

[7] Li, P. (2011) Estimation of Sample Selection Models with Two Selection Mechanisms. *Computational Statistics & Data Analysis*, **55**, 1099-1108. http://dx.doi.org/10.1016/j.csda.2010.09.006

[8] Li, P. and Rahman, M.A. (2011) Bayesian Analysis of Multivariate Sample Selection Models Using Gaussian Copulas. *Advances in Econometrics*, **27**, 269-288. http://dx.doi.org/10.1108/S0731-9053(2011)000027A013

[9] Yen, S.T. (2005) A Multivariate Sample-Selection Model: Estimating Cigarette and Alcohol Demands with Zero Observations. *American Journal of Agricultural Economics*, **87**, 453-466. http://dx.doi.org/10.1111/j.1467-8276.2005.00734.x

[10] Tauchmann, H. (2010) Consistency of Heckman-Type Two-Step Estimators for the Multivariate Sample-Selection Model. *Applied Economics*, **42**, 3895-3902. http://dx.doi.org/10.1080/00036840802360179

[11] Puhani, P.A. (2000) The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys*, **14**, 53-68. http://dx.doi.org/10.1111/1467-6419.00104

[12] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1-38. http://dx.doi.org/10.2307/2984875

[13] Liu, C., Rubin, D.B. and Wu, Y.N. (1998) Parameter Expansion to Accelerate EM: The PX-EM Algorithm. *Biometrika*, **85**, 755-770. http://dx.doi.org/10.1093/biomet/85.4.755

Open Journal of Statistics, 2014, 4, 857-867
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410081

Scientific Research

Improving Model Specifications When Estimating Treatment Effects across Alternative Medical Interventions

Yawen Jiang, Jeffrey McCombs

Department of Clinical Pharmacy, Pharmaceutical Economics and Policy, School of Pharmacy, University of Southern California, USC Schaeffer Center, VERNA & PETER DAUTERIVE HALL (VPD), Los Angeles, CA, USA
Email: yawenjia@usc.edu

Received 1 October 2014; revised 20 October 2014; accepted 6 November 2014

Abstract

Objective: The purpose of this paper is to critique the list of independent variables commonly used in observational research and test the impact of variables for prior use and treatment history on estimates of treatment effects. Methods: Using data from the California Medicaid program, this study generated a series of OLS estimates of the effect of atypical antipsychotic medications on costs and duration of therapy to illustrate the impact of alternative model specifications on treatment effects. The first sequence of estimates consisted of six model specifications, the last of which included variables reflecting the type of episode defined according to prior treatment history and compliance. The second sequences repeated the specification of the first 6 models but were carried out separately by episode type to examine the heterogeneity of treatment effect. The second sequence of models documented the impact of additional drug history variables. Results: Estimates of the impact of atypical antipsychotic use on total costs and duration on initial drug were statistically significant in the first 6 models. Estimates changed significantly when dummy variables indicating prior use of inpatient service and nursing home care were included in the model specification. Estimated effects changed substantially when prior total cost was included in cost analysis, or when prior treatment duration was included in duration analysis. Significant variation also existed in estimated effects across episode types, and it was particularly pronounced before controlling for prior cost/duration. Conclusion: It is important to add prior measures of the outcome variable to control for unobserved bias in retrospective studies. Also, the accuracy and utility of results to clinicians can be improved significantly if analyses are performed by episode type.

Keywords

Treatment, Effect, Intervention, Episode, History

How to cite this paper: Jiang, Y.W. and McCombs, J. (2014) Improving Model Specifications When Estimating Treatment Effects across Alternative Medical Interventions. *Open Journal of Statistics*, **4**, 857-867.
http://dx.doi.org/10.4236/ojs.2014.410081

1. Introduction

Clinicians and policy makers require medical evidence with which to effectively integrate new technologies into real-world practice. This need is especially acute when new treatment alternatives are introduced into competition with older, well established treatments. In the case of new medications, these data come from two sources: the clinical trials required for FDA and other registry approvals and observational studies that establish the "essential need" for a new treatment alternative [1]. Both sources of medical evidence are necessary to estimate the cost-effectiveness of a new technology at product launch.

Efforts to document the essential need for a new technology actually begin very early in product development. Product innovators assess how well older, competing therapies are meeting the therapeutic needs of patients treated in the real world. Therapeutic gaps with older treatments typically arise when patient adherence to current therapies is sub-optimal or treatment efficacy for compliant patients is limited. Other sources of essential need are high treatment costs or high indirect cost to the patient and their caregivers [2]. These indirect costs may include the costs of side effects, caregiver time, reductions in the quality of life and the like. Essential need data are used in a series of "go/no go" decisions that are made as the product is developed and tested.

If the evolving data on essential need are promising and/or the new product is efficacious, the new technology will move through the required registry trials testing safety and efficacy. These studies use experimental research designs [RCTs] which maximize the internal validity through random assignment and other techniques [e.g., blinding] [1]. However, the generalizability [external validity] of results from randomized clinical trials is limited:

1. RCTs are limited to a small, homogeneous study population due to cost and patient safety concerns. Data on treatment outcomes for high risk patients may be missing or, conversely, it may be ethical to only include very high risk patients who have no remaining treatment options, as in cancer trials.

2. Patient outcomes are measured over a limited time, again due to cost and to patient burden and risk of dropout. This mis-match between study duration and time to potential treatment effect is most acute for drug therapies intended to manage chronic disease such as hypertension or hyperlipidemia.

3. By design, RCTs cannot measure patient adherence to treatment under real-world conditions. RCTs employ significant effort and resources to insure patient adherence to the study protocol.

4. Finally, FDA-registration RCTs may require only placebo-controlled trials or the list of active comparators may be constrained due to cost concerns.

Conversely, essential need studies based on retrospective data can provide CE evidence for the full range of treatment alternatives, and reflect real-world clinical practice and real-world adherence. The patients included in an essential need study also include risk groups not studied in the RCT environment. Finally, retrospective observational studies can provide evidence on long term outcomes and the [rare] clinical risk associated with existing therapies.

Drug companies combine data from real-world essential need studies and registry RCT into an initial computer-based CE model to support product marketing at launch. These models project the impact of the new technology in clinical use. However, the accuracy of the initial CE models is limited by the gaps in research on real-world adherence for the new drug, long term patient outcomes using the new drug and outcomes achieved by patient sub-groups not included in the product's clinical trials [poor external validity]. While retrospective essential need studies fill in some of these gaps, the statistical validity of observational studies can be questionable if not executed well. Of equal concern, physicians, P & T committees and government program administrators may not fully understand the complexity and pitfalls of the statistical methods use in observational research.

The purpose of this paper is to critique the statistical methods commonly used in observational research by presenting a sequence of analyses which document how statistical results can change significantly as more care is taken to maximize the use of available data. Specifically, we will present a sequence of models moving from simple models to models using explanatory variables that are rarely derived from available claims data. The paper also documents the impact of alternative estimation strategies.

2. Statistical Challenges in Observational Research

Satisfactory internal validity can only be achieved in observational research by controlling for confounding factors associated with both treatment selection and patient outcomes. For example, it is challenging to measure the impact of a new medication relative to competing older drugs if the new medication is reserved for high risk pa-

tients, or if the new medication is used initially to treat patients who failed therapy using the older alternatives [3]. This bias can be reduced using multivariate statistical methods that adjust statistically for the impact of observable factors on treatment outcomes. However, treatment selection bias will continue to exist if important factors are missing from the multivariate statistical models. In the econometrics literature, this is referred to as missing variable bias. In comparative effectiveness research, missing variable bias is referred to as unobserved treatment selection bias [UTSB].

UTSB is often a function of the data available for analysis. For example, data from a health insurance program or government program [e.g., Medicare] includes the paid claim for common laboratory tests but provides no information concerning the laboratory result itself. Fortunately, the growing availability of electronic medical records [EMR] data will provide increasing opportunities for reducing the impact of UTSB in observational studies in medicine.

The first line of defense against UTSB is to use the available data to document all factors that may impact both treatment selection and patient outcomes. Researchers often ignore episodes of drug therapy initiated following the first observed treatment episode which is concerning since patient outcomes can be radically different for the second or third treatment attempt using the same drug, or for episodes of switching therapies, episodes of augmentation therapy or episodes involving combination therapy. Moreover, the later episodes contain more information about the treatment history of the patients, such as prior compliance behavior, which could significantly impact patient outcomes. This expanded use of available pharmacy data may be particularly important when newly approved medications are significantly less likely to be used as first therapy in treatment naïve patient.

Alternative model specifications make better use of available data and will also be investigated. Both difference-in-difference models (DD) [4] and fixed effects models (FE) [4] assume that UTSB is invariant across time periods (e.g. pre-treatment and post-treatment). For example, genetic factors which affect disease severity or response to drug treatment are invariant across time, and they are usually unobservable to researchers. Diff-in-diff models are popular in analysis of panel data. By differencing out fixed effects or controlling for them using dummy variables representing clusters, these models eliminate the effects of the time-invariant UTSB. But the time-invariant assumption of UTSB does not necessarily hold in practice. Even though time-invariant effects can be removed using such techniques, potential bias caused by time-varying confounding factors is still left unresolved. For instance, some clinical symptoms and health behavior are not captured by automated data systems, yet they are unlikely to remain exactly the same across time periods.

3. Methods

3.1. Data Sources

This study conducts a series of retrospective analyses of the impact of atypical antipsychotic medications to illustrate the impact of alternative model specifications and estimation methods on treatment effects. The study uses an existing California Medicaid (Medi-Cal) data set which was derived for a string of earlier studies [5] [6] from paid claims data from the fee-for-service portion of Medi-Cal. The data cover the period of 1994-2003 during which Medi-Cal revoked its restriction on the use of typical antipsychotics to patients who had failed at least two previous treatment attempts using typical antipsychotics. This formulary restriction was lifted in October 1997, three years after the introduction of risperidone in 1994 and exactly one year after the approval of olanzapine in 1996. Quetiapine was approved by the FDA in October 1997 and was immediately available to Medi-Cal patients without restrictions. This formulary expansion resulted in an immediate increase in the diffusion of atypical antipsychotics which are now accepted as first line drug therapy for these patients [7].

Initial inclusion criteria required that patients have a paid claim with a recorded diagnosis of schizophrenia (ICD-9 code = 295.xx) or bipolar disorder (ICD-9 codes = 296.4 - 296.8) and with at least one prescription for an antipsychotic medication. Additional exclusion criteria were applied once all episodes of care were identified.

3.2. Definition of the Unit of Analysis

The “standard of practice” for the unit of analysis in a retrospective CE research design data mirrors the RCT design: The episode of treatment. In the case of observational studies, the data of randomization is replaced by an “index date” defined based on the patient’s first prescription of one of the study drugs. Like most RCTs, the

patient is typically subjected to a "wash-out" period by requiring that the patient has not filled a prescription of any study drug prior to their initial prescription. Wash-out periods vary in length and 6 months to a year are common. Most studies then limit their analysis to these "first episodes" and ignore any subsequent use of related drugs such as augmentation therapy or the switching to an alternative medication. Limiting the analysis to first episodes excludes a large majority of treatment episodes. Moreover, new medications are seldom used as the first drug of choice and are regulated to treating "treatment failures" or providing augmentation therapy.

The data set used here includes all episodes of psychotropic drug therapy initiated by patients. An episode of treatment was defined each time a patient started a drug treatment using an antipsychotic, antidepressant or mood stabilizer not used previously or restarted an earlier drug treatment after a gap that was at least 15 days. The 15-day gap was defined in collaboration with the Medi-Cal program and was to comply with earlier finding by Weiden *et al*. [8], who reported that the risk of hospitalization increased substantially after breaks in therapy as short as 10 days.

The follow up period was the 12 months after the month of initiation. The 12-month follow up period was specified for the measurement of treatment outcomes which mimics intent to treat methods implemented in clinical trials. Patient episodes were then screened for eligibility during the entire pre- and post-treatment period. The amount paid for all services were inflation adjusted to 2004 using service specific rates of fee inflation from the Medi-Cal program.

Many patients had more than one treatment episode, which is very common in schizophrenia and bipolar disorders as patients switched from one antipsychotic to another or start and stop therapy. While this approach violates the usual assumption of independence across units of analysis, excluding subsequent episodes initiated by the patient was judged to generate stronger bias than hypothetical independence of sampling units [6] [9] [10]. Excluding these follow-on episodes severely restricts the utility of the analysis to clinicians who required data on treatment effects for a wide range of treatment histories.

3.3. Covariates and Model Sequencing

The focus of the proposed study is to examine how the use of an expanded list of unconventional independent variables impacts estimates of total costs and duration of therapy using standard ordinary least squares (OLS) regressions. Specifically, the following sequence of models will be estimated:

Model 1: The basic models include only age [categories with an interval of 10], gender, county population density [urban/rural/urban-rural-mix] and Medi-Cal aid categories

Model 2: The second set of models adds dichotomous variables based on non-mental health comorbidities based on ICD-9 diagnoses at baseline.

Model 3: Mental health diagnoses were added to the model specification separately to test the impact of diagnostic mix data related directly to the disease state under study.

Model 4: The list of independent variables was extended to include two dichotomous variables indicating whether or not the patient used inpatient hospital services or nursing home services in the 6 months prior to the episode start date.

Model 5: Pre-treatment measures of the outcome variables [total costs, duration of therapy] were added in this model. This specification is mathematically equivalent to difference-in-difference modeling which re-defines the outcome variable by differencing the value of the outcome measure before and after treatment.

Models 6: This model is the first to used data on the drug history of the patient at the time of treatment. The initial drug history covariates are dichotomous variable for episode type. Five types of episodes were defined in this data set:

1. First Observed Episode: The "first" episode was defined based on the patient's first psychotropic drug therapy attempt.

2. Restart Episodes: A restart episode was defined if the patient was not on active psychotropic drug therapy for 15 days or longer and initiated therapy with the medication used in their most recent episode [intermittent use].

3. Switching Episodes: A switching episode was defined if a patient changed medication while still on active therapy or within 15 days of terminating a previous therapy, and discontinued use of all previous medications within 60 days.

4. Delayed Switching Episodes: A delayed switching episode was defined if a patient changed drug therapy after a break in therapy in excess of 15 days.

5. Augmentation Episodes: An augmentation episode was defined when a patient added a second medication while continuing to purchase one or more of their previous medications beyond 60 days.

This analysis excludes first observed treatment episodes due to the lack of data on patient treatment history. The following analyses only used restart, delayed switching, switching, and augmentation episodes. In order to facilitate comparisons to Models 1 - 6, first episodes were also excluded from the sample of episodes included in these models.

Models 7 - 12: The remaining drug treatment history variables are entered sequentially in Models 7 - 12: count of the number of prior treatment attempts, monotherapy vs. combination therapy, days off therapy (for restart and delayed switching episodes), and prior use of related drugs [typical and atypical antipsychotics, mood stabilizers, antidepressants, depot-formulated drugs]. At this point, the analyses are conducted by episode type primarily because episode type is a significant predictor of cost and duration of therapy [Model 6]. It follows that clinicians will require information on the CE of atypical vs. typical antipsychotics by episode type.

4. Results

Results for the first six models for the impact of using atypical antipsychotics that used all episodes are summarized in **Table 1**. The outcome variables used in these models are total cost over the first post-treatment year and duration of therapy on the 'initial' drug of the episode. For example, in the case of augmentation episode, the initial drug is the augmenting drug. In addition to the impacts of atypical use, we also include the estimates of the effects of episode type indicators on cost and duration in Model 6 which are also included in **Table 1**.

Estimates of the impact of atypical antipsychotic use on total costs and duration on initial drug are statistically significant in the first 6 models. In Models 1 - 3, the estimated impact of using an atypical antipsychotic range from $1230 to $1399, and the estimates of the impacts on duration range from 90.2 to 95.9 days. Estimates changed significantly when dummy variables indicating prior inpatient service use and prior nursing home use were included in the model specification. The effect of atypical use on total cost decreased to $398 whereas the effect on duration only slightly changed to 89.1 days. Equally important, the R^2 of the model for total cost increased substantially (0.182 to 0.571).

Table 1. Impact of atypical antipsychotic use on total cost and duration of therapy: all episode types (N = 731,236).

		Total Cost First Post-Treatment Year		Duration of Drug Therapy	
Model	Model Specification	OLS (SE)	R-squared In OLS	OLS (SE)	R-squared In OLS
1	Demographic variables only	1350*** (63.7)	0.095	95.9*** (0.8)	0.037
2	Add: Medical Diagnostic Mix	1399*** (62.0)	0.160	93.0*** (0.8)	0.046
3	Add: Mental Health Diagnostic Mix	1230*** (61.6)	0.182	90.2*** (0.8)	0.060
4	Add: Prior Use of Hospital and Nursing Home Care	398*** (44.6)	0.571	89.1*** (0.8)	0.063
5	Add: Pre-Treatment Costs and Duration of Therapy	615*** (36.7)	0.710	76.4*** (0.8)	0.130
6	Add: Episode Type	751*** (37.8)	0.712	55.0*** (0.8)	0.157
	Estimated Impact of Episode Type on Total Cost and Duration of Drug Therapy in Model 6				
	Estimated Impact of Episode Type on Total Cost and Duration of Drug Therapy in Model 6 (Restart as baseline)				
	Switching	1221*** (65.3)		136.9*** (1.4)	
	Delayed switching	1360*** (55.3)		73.8*** (1.2)	
	Augmentation	2237*** (58.3)		−75.9*** (1.2)	

OLS results are presented as estimate (SE). Abbreviations: N, number of episodes; OLS, ordinary least squares; SE, standard error.

Difference-in-difference modeling is frequently used in observational research testing the effect of new treatments or policy changes on patient outcomes. When prior total cost was included in cost analysis [Model 5], the estimated effect of atypical use increased from $398 to $615 and the R^2 further increased from 0.571 to 0.710. Similarly, when prior treatment duration was included in duration analysis, the estimated effect of atypical use decreased from 89 days to 76 days and the R^2 doubled from 0.063 to 0.130.

Model 6 estimates the impact of atypical use controlling for episode type. The results from this model demonstrate the importance of drug use history when estimating the impact of atypical antipsychotics on cost and duration of therapy in two ways. First, the estimated effect of atypical use changed to $751 while the estimated effect on duration decreased to 55 days. But more importantly, episode type has very significant impacts of costs and duration. Compared with restart episodes, switching episodes, delayed switching episodes, and augmentation episodes increased total cost by $1221, $1360, and $2237, respectively. However, the impacts of the episode type on duration were not uniformly positive. Switching and delayed switching episodes lasted an additional 137 days, 74 days relative to re-start episodes. Conversely, the use of the initial drug decreased by 76 days in augmentation episodes relative to re-start episodes, possibly reflecting intended short term use of augmentation therapy.

The results from Model 6 provide an estimate of the average impact of using an atypical antipsychotic on cost and duration of therapy controlling for how atypical antipsychotic drugs are used by episode type. However, clinicians need to know how these new drugs perform by episode type, not on average. This dictates that these analyses be conducted separately by episode type. Conducting analyses by episode type also allows researchers to add other treatment history variables to the analyses which can vary by episode type. Our analyses of use and cost by episode type are displayed in **Tables 2-5**. The results for the average impact of atypical use derived in Model 6 using data for all episode types is also listed in these tables as a reference.

Table 2. Impact of atypical antipsychotic use on total cost and duration of therapy: restart episodes (N = 445,258).

		Total Cost First Post-Treatment Year		Duration of Drug Therapy	
Model	Model Specification	OLS (SE)	R-squared In OLS	OLS (SE)	R-squared In OLS
1	Demographic variables	2301*** (77.3)	0.098	38.2*** (0.9)	0.018
2	Demo + Medical Diagnosis	2616*** (75.6)	0.166	33.0*** (0.9)	0.026
3	Demo + MedicalDx + MHDx	2485*** (75.7)	0.185	28.1*** (0.9)	0.039
4	+prior hospitalization +prior long term care	1077*** (53.9)	0.588	26.2*** (0.9)	0.043
5	+prior and switch total costs/prior episode duration	384*** (42.8)	0.740	24.4*** (0.9)	0.056
6	Model 6 Specification Using Data for All Episodes	751*** (37.8)	0.712	55.0*** (0.8)	0.157
7	Add: prior episode count	500*** (43.3)	0.741	22.2*** (0.9)	0.060
8	Add: mono/poly	493*** (43.4)	0.741	21.7*** (0.9)	0.060
9	Add: prior depot use indicator	497*** (43.5)	0.741	21.7*** (0.9)	0.060
10	Add: time off Rx	567*** (44.2)	0.741	20.3*** (0.9)	0.060
11	Add: prior Rx mix	563*** (111.3)	0.741	24.2*** (2.3)	0.062

OLS results are presented as estimate (SE). Abbreviations: N, number of episodes; OLS, ordinary least squares; SE, standard error.

Table 3. Impact of atypical antipsychotic use on total cost and duration of therapy: switching episodes (N = 71,917).

		Total Cost First Post-Treatment Year		Duration of Drug Therapy	
Model	Model Specification	OLS (SE)	R-squared In OLS	OLS (SE)	R-squared In OLS
1	Demographic variables	1678*** (217.4)	0.147	107.0*** (4.0)	0.037
2	Demo + Medical Diagnosis	1796*** (211)	0.203	106.1*** (4.0)	0.053
3	Demo + MedicalDx + MHDx	1901*** (208.8)	0.220	108.0*** (3.9)	0.068
4	+prior hospitalization +prior long term care	1289*** (154.9)	0.571	106.6*** (3.9)	0.071
5	+prior and switch total costs/prior episode duration	1171*** (136.4)	0.667	24.6*** (3.1)	0.453
6	Model 6 Specification Using Data for All Episodes	751*** (37.8)	0.712	55.0*** (0.8)	0.157
7	Add: prior episode count	1128*** (145.4)	0.670	26.3*** (3.2)	0.453
8	Add: mono/poly	1122*** (145.6)	0.670	25.4*** (3.2)	0.454
9	Add: prior depot use indicator	1270*** (149.1)	0.670	26.4*** (3.2)	0.454
10	Add: time off Rx				
11	Add: prior Rx mix	1262*** (152.2)	0.670	37.2*** (3.2)	0.459

OLS results are presented as estimate (SE). Abbreviations: N, number of episodes; OLS, ordinary least squares; SE, standard error.

Table 4. Impact of atypical antipsychotic use on total cost and duration of therapy: delayed switching episodes (N = 97,704).

		Total Cost First Post-Treatment Year		Duration of Drug Therapy	
Model	Model Specification	OLS (SE)	R-squared In OLS	OLS (SE)	R-squared In OLS
1	Demographic variables	1140*** (185.1)	0.161	95.9*** (2.8)	0.032
2	Demo + Medical Diagnosis	1487*** (179.7)	0.219	93.0*** (2.8)	0.045
3	Demo + MedicalDx + MHDx	1454*** (178.2)	0.236	93.4*** (2.8)	0.061
4	+prior hospitalization +prior long term care	1128*** (131.2)	0.586	92.2*** (2.8)	0.069
5	+prior and switch total costs/prior episode duration	1179*** (117.6)	0.667	88.6*** (2.85)	0.075
6	Model 6 Specification Using Data for All Episodes	751*** (37.8)	0.712	55.0*** (0.8)	0.157
7	Add: prior episode count	1217*** (121.9)	0.671	86.6*** (2.9)	0.076
8	Add: mono/poly	1231*** (123.5)	0.671	84.0*** (2.9)	0.076
9	Add: prior depot use indicator	1287*** (126.7)	0.671	84.2*** (2.9)	0.076
10	Add: time off Rx	1283*** (127)	0.671	84.3*** (2.9)	0.076
11	Add: prior Rx mix	1120*** (137.1)	0.672	91.7*** (3.1)	0.077

OLS results are presented as estimate (SE). Abbreviations: N, number of episodes; OLS, ordinary least squares; SE, standard error.

Table 5. Impact of atypical antipsychotic use on total cost and duration of therapy: augmentation episodes (N = 116,357).

		Total Cost First Post-Treatment Year		Duration of Drug Therapy	
Model	Model Specification	OLS (SE)	R-squared In OLS	OLS (SE)	R-squared In OLS
1	Demographic variables	−4936*** (185.1)	0.105	170.0*** (2.1)	0.075
2	Demo + Medical Diagnosis	−4549*** (180.8)	0.152	168.7*** (2.1)	0.086
3	Demo + MedicalDx + MHDx	−4390*** (179.9)	0.165	168.6*** (2.1)	0.095
4	+prior hospitalization +prior long term care	−1741*** (136.6)	0.520	167.6*** (2.1)	0.096
5	+prior and switch total costs/prior episode duration	−289** (112.2)	0.677	146.7*** (2.1)	0.171
6	Model 6 Specification Using Data for All Episodes	751*** (37.8)	0.712	55.0*** (0.8)	0.157
7	Add: prior episode count	66 (115.4)	0.676	142.6*** (2.1)	0.173
8	Add: mono/poly	65 (115.6)	0.676	143.3*** (2.1)	0.174
9	Add: prior depot use indicator	74 (118.7)	0.676	143.4*** (2.1)	0.174
10	Add: time off Rx				
11	Add: prior Rx mix	−156 (122.0)	0.677	155.7*** (2.2)	0.179

OLS results are presented as estimate (SE). Abbreviations: N, number of episodes; OLS, ordinary least squares; SE, standard error.

Table 2 presents the results using restart episodes starting with the original set of independent variable used in Model 1. Models 5 and 6 are equivalent when estimated using only restart episodes. In models 1 - 3 using restart episodes, the estimated effects of atypical antipsychotic use on total cost range from $2301 to $2616. Including prior inpatient services use and prior nursing home use decreased the estimated effect to $1077. It further decreased to $384 after controlling for prior total cost. The estimated effect remained stable across Models 7 - 11 ($493 - $567). The R^2 increased significantly at the stages of Model 4 (0.185 to 0.588) and Model 5 (0.588 to 0.740). The estimated effects of atypical antipsychotic use on duration is much more stable than estimated for cost across all models using restart episodes are between 20.3 and 38.2 days. Also, the increase in the R^2 was modest from Model 1 to Model 11 (0.018 to 0.062).

Table 3 presents the results of analyses using switching episodes. In Models 1 - 3 using switching episodes, the estimated effects of atypical antipsychotic use on total cost are between $1678 and $1901. Including prior inpatient services use and prior nursing home use in the model decreased the estimated effect to $1289. Including prior total cost changed the estimated effect to $1171. The estimated effects on total costs are between $1122 and $1270 in Models 7 - 11. The R^2 increased by a large amount at the stages of Model 4 (0.220 to 0.571) and Model 5 (0.571 to 0.667). The estimated effects of atypical antipsychotic use on duration in Models 1 - 4 are between 106.1 and 108.0 days. But the estimated effect of atypical use dropped significantly to 24.6 days after controlling for prior treatment duration. In Models 7 - 11, the estimated effects are between 25.4 days and 37.2 days. The R^2 increased from 0.071 to 0.453 at the stage of Model 5.

Table 4 lists the results of analyses using delayed switching episodes. Throughout the 10 models, the estimated effects of atypical antipsychotic use on total cost range from $1120 to $1487, and the estimated effects on duration range from 84.0 to 95.9 days. The R^2 in the cost analysis increased from 0.236 to 0.586 at the stage of Model 4 and increased from 0.586 to 0.667 at the stage of Model 5. However, the R^2 in duration analysis only increased modestly from 0.031 in Model 1 to 0.077 in Model 11.

Finally, the results of analyses using augmentation episodes are included in **Table 5**. In Models 1 - 3, the estimated effects of atypical antipsychotic use on total cost are between −$4936 and −$4390. The estimated effect changed to −$1741 after controlling for prior inpatient services use and prior nursing home use, and further

changed to −$289 after controlling for prior total costs. In Models 7 - 11, the estimated effects on total costs are between −$156 and $74 and are all statistically insignificant. These are the only set of insignificant estimates in all analyses in the current study. The estimated effects on duration are between 142.6 days and 170.0 days for all 10 models. The R^2 in the cost analysis increased from 0.165 to 0.520 at the stage of Model 4 and increased from 0.520 to 0.677 at the stage of Model 5. Likewise, the R^2 in duration analysis increased from 0.096 to 0.171 at the stage of Model 5.

5. Discussion and Conclusion

The purpose of this study was to investigate the changes in estimated treatment effects in response to a series of explanatory variables, some of which are rarely derived from claims databases. The results from this series of estimates are illustrated in **Figure 1** [Total Cost] and **Figure 2** [Duration of Therapy]. Two statistical effects are evident. First, controlling for prior total cost/treatment duration led to significant changes of estimates in most analyses, and the results for the impact of using atypical antipsychotics [treatment effect] tended to settle down across model specifications after that stage. This result validates the value of adding prior measures of the outcome variable which corresponds to the popular difference-in-difference estimation technique. Second, it is evident that great variation exists in estimated effects of atypical antipsychotic use across episode types which persists across model specification and is particularly pronounced before adding prior total cost/treatment duration to the model specification. But as an added bonus, conducting the analysis of treatment effects by episode type significantly increases the utility of study results to clinicians who are looking for guidance as to what works best for patients with different treatment history.

Episode type can significantly impact the estimated treatment effects because episode type has a major impact on treatment outcomes. Accordingly, comparative effectiveness research should take into account the differential treatment effects in episode-type subgroups.

A major limitation of observational result to measure treatment effects stems from the nature of claims databases. Claims databases do not usually capture important information such as disease severity and clinical symptoms. Although we controlled a long list of variables and used various model specifications in the regressions, potential bias due to unmeasured covariates could not be ruled out thoroughly. However, the future of

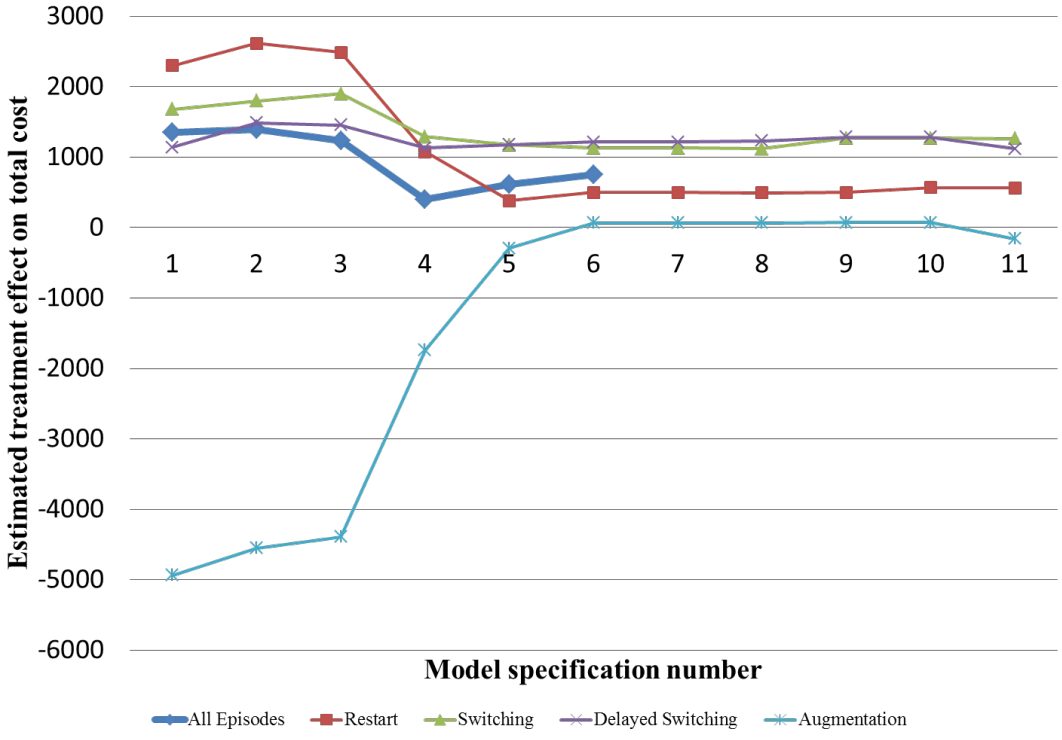


Figure 1. Impact of using atypical antipsychotics on total cost in first post-treatment year.

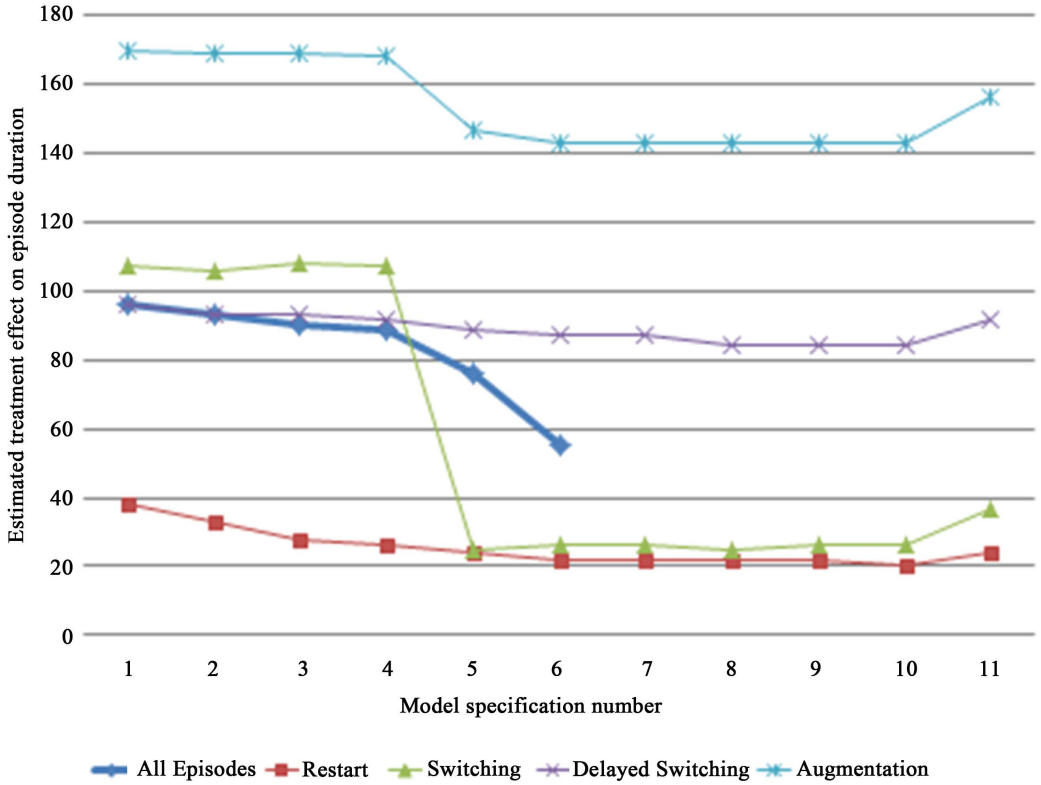


Figure 2. Impact of using atypical antipsychotics on duration of therapy.

observational research in comparative effectiveness research is bright as data from electronic medical record [EMR] systems become more available. The internal validity of estimated differences between alternative treatments will only improve as better clinically relevant data are available.

References

[1] Faries, D.E., Leon, A.C., Haro, J.M. and Obenchain, R.L. (2010) Analysis of Observational Health Care Data Using SAS. SAS Institute.

[2] Berger, M.L., Dreyer, N., Anderson, F., Towse, A., Sedrakyan, A. and Normand, S.L. (2012) Prospective Observational Studies to Assess Comparative Effectiveness: The ISPOR Good Research Practices Task Force Report. *Value in Health*, **15**, 217-230. http://dx.doi.org/10.1016/j.jval.2011.12.010

[3] Soumerai, S.B., Zhang, F., Ross-Degnan, D., Ball, D.E., LeCates, R.F., Law, M.R., Hughes, T.E., Chapman, D. and Adams, A.S. (2008) Use of Atypical Antipsychotic Drugs for Schizophrenia in Maine Medicaid Following a Policy Change. *Health Affairs*, **27**, 185-195.

[4] Greene, W.H. (2012) Econometric Analysis. Pearson Education, Inc., Upper Saddle River.

[5] Chen, L., McCombs, J.S. and Park, J. (2008) Duration of Antipsychotic Drug Therapy in Real-World Practice: A Comparison with CATIE Trial Results. *Value in Health*, **11**, 487-496. http://dx.doi.org/10.1111/j.1524-4733.2007.00262.x

[6] Chen, L., McCombs, J.S. and Park, J. (2008) The Impact of Atypical Antipsychotic Medications on the Use of Health Care by Patients with Schizophrenia. *Value in Health*, **11**, 34-43. http://dx.doi.org/10.1111/j.1524-4733.2007.00212.x

[7] Narayan, S., Sterling, K.L. and McCombs, J.S. (2006) The Impact of Open Access to Atypical Antipsychotics on Treatment Costs for Medical Patients with Bipolar Disorder. *Disease Management & Health Outcomes*, **14**, 287-301. http://dx.doi.org/10.2165/00115677-200614050-00004

[8] Weiden, P.J. (2004) Partial Compliance and Risk of Rehospitalization among California Medicaid Patients with Schizophrenia. *Psychiatric Services*, **55**, 886-891. http://dx.doi.org/10.1176/appi.ps.55.8.886

[9] Gianfrancesco, F.D., Grogg, A.L., Mahmoud, R.A., Wang, R.-H. and Nasrallah, H.A. (2002) Differential Effects of Risperidone, Olanzapine, Clozapine, and Conventional Antipsychotics on Type 2 Diabetes: Findings from a Large He-

alth Plan Database. *The Journal of Clinical Psychiatry*, **63**, 920-930. http://dx.doi.org/10.4088/JCP.v63n1010

[10] Gianfrancesco, F., Pesa, J., Wang, R.-H. and Nasrallah, H. (2006) Assessment of Antipsychotic-Related Risk of Diabetes Mellitus in a Medicaid Psychosis Population: Sensitivity to Study Design. *American Journal of Health-System Pharmacy*, **63**, 431-441. http://dx.doi.org/10.2146/ajhp050144

Open Journal of Statistics, 2014, 4, 868-878
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410082

Model Detection for Additive Models with Longitudinal Data

Jian Wu, Liugen Xue

[1]College of Applied Sciences, Beijing University of Technology, Beijing, China
[2]College of Science, Northeastern University, Shenyang, China
Email: wujian@emails.bjut.edu.cn, mbaron@utdallas.edu

Received 1 October 2014; revised 28 October 2014; accepted 15 November 2014

Abstract

In this paper, we consider the problem of variable selection and model detection in additive models with longitudinal data. Our approach is based on spline approximation for the components aided by two Smoothly Clipped Absolute Deviation (SCAD) penalty terms. It can perform model selection (finding both zero and linear components) and estimation simultaneously. With appropriate selection of the tuning parameters, we show that the proposed procedure is consistent in both variable selection and linear components selection. Besides, being theoretically justified, the proposed method is easy to understand and straightforward to implement. Extensive simulation studies as well as a real dataset are used to illustrate the performances.

Keywords

Additive Model, Model Detection, Variable Selection, SCAD Penalty

1. Introduction

Longitudinal data arise frequently in biological and economic applications. The challenge in analyzing longitudinal data is that the likelihood function is difficult to specify or formulate for non-normal responses with large cluster size. To allow richer and more flexible model structures, an effective semi-parametric regression tool is the additive model introduced by [1], which stipulates that

$$Y = \mu + \sum_{l=1}^{d} m_l\left(X^{(l)}\right) + \varepsilon, \tag{1}$$

where Y is a varaible of interest and $X = \left(X^{(1)}, \cdots, X^{(d)}\right)^T$ is a vector of predictor variables, μ is a

How to cite this paper: Wu, J. and Xue, L.G. (2014) Model Detection for Additive Models with Longitudinal Data. *Open Journal of Statistics*, **4**, 868-878. http://dx.doi.org/10.4236/ojs.2014.410082

unknown constant, and $m(X)=\sum_{l=1}^{d} m_l\left(X^{(l)}\right)$ are unknown nonparametric functions. As in most work on nonparametric smoothing, estimation of the non-parametric functions $m(X)=\sum_{l=1}^{d} m_l\left(X^{(l)}\right)$ is conducted on a compact support. Without loss of generality, let the compact set be $\mathcal{X}=[0,1]^d$ and also impose the condition $E\left[m_l\left(X^{(l)}\right)\right]=0$ which is required for identifiability of model (1.1), $l=1,\cdots,d$. We propose a penalized method for variable selection and model detection in model (1.1) and show that the proposed method can correctly select the nonzero components with probability approaching one as the sample size goes to infinity.

Statistical inference of additive models with longitudinal data has also been considered by some authors. By extending the generalized estimating equations approach, [2] studied the estimation of additive model with longitudinal data. [3] focuses on a nonparametric additive time-varying regression model for longitudinal data. [4] considered the generalized additive model when responses from the same cluster are correlated. However, in semiparametric regression modeling, it is generally difficult to determine which covariates should enter as nonparametric components and which should enter as linear components. The commonly adopted strategy in practice is just to consider continuous entering as nonparametric components and discrete covariates entering as parametric. Traditional method uses hypothesis testing to identify the linear and zero component. But this might be cumbersome to perform in practice whether there are more than just a few predictor to test. [5] proposed a penalized procedure via the LASSO penalty; [6] presented a unified variable selection method via the adaptive LASSO. But these methods are for the varying coefficient models. [7] established a model selection and semiparametric estimation method for additive quantile regression models by two-fold penalty. To our knowledge, the model selection and variable selection simultaneously with longitudinal data have not been investigated. We make several novel contributions: 1) We develop a new strategies for model selection and variable selection in additive model with longitudinal data; 2) We develop theoretical properties for our procedure.

In the next section, we will propose the two-fold SCAD penalization procedure based on QIF and computational algorithm; furthermore we present its theoretical properties. In particular, we show that the procedure can select the true model with probability approaching one, and show that newly proposed method estimates the non-zero function components in the model with the same optimal mean square convergence rate as the standard spline estimators. Simulation studies and an application of proposed methods in a real data example are included in Sections 3 and 4, respectively. Technical lemmas and proofs are given in **Appendix**.

2. Methodology and Asymptotic Properties

2.1. Additive Models with Two Fold Penalized Splines

Consider a longitudinal study with n subjects and n_i observations over time for the ith subject $(i=1,\cdots,n)$ for a total of $N=\sum_{i=1}^{n} n_i$ observation. Each observation consists of a response variable Y_{ij} and a covariate vector $X_{ij}\in R^d$ taken from the ith subject at time t_{ij}. We assume that the full data set

$$\left\{\left(X_{ij},Y_{ij}\right), i=1,\cdots,n,\ j=1,\cdots,n_i\right\}$$

is observed and can be modelled as

$$Y_{ij}=\mu+\sum_{l=1}^{d} m_l\left(X_{ij}^{(l)}\right)+\varepsilon_{ij},\quad i=1,\cdots,n,\ j=1,\cdots,n_i, \tag{2}$$

where ε_{ij} is random error with $E\left(\varepsilon_{ij}\middle|X_{ij}\right)=0$ and σ_ε^2.

At the start of the analysis, we do not know which component functions in model (1.1) are linear or actually zero. We adopt the centered B-spline basis, where $\boldsymbol{B}(\boldsymbol{X})=\left\{B_{s,l}\left(x_l\right):1\le l\le d,1\le s\le J\right\}^{\mathrm{T}}$ is a basis system $B_{s,l}\left(x_l\right)=\sqrt{K}\left[b_{s+1,l}\left(x_l\right)-\left\{E\left(b_{s+1,l}\right)/E\left(b_{1,l}\right)\right\}b_{1,l(x_l)}\right]$ and $(x)=\left(x_l\right)_{l=1}^{d}$. Equally-spaced knots are used in this article for simplicity of proof. Other regular knot sequences can also be used, with similar asymptotic results. Suppose that $m_l(\cdot)$ can be approximated well by a spline function, so that

$$m_l\left(x^{(l)}\right) \approx m_l^{sp}\left(x^{(l)}\right) = \sum_{s=1}^{J} \beta_{sl} B_{s,l}\left(x^{(l)}\right). \tag{3}$$

To simplify notation, we first assume equal cluster size $n_i = m < \infty$, and let $\beta_l = \left(\beta_{1l}, \cdots, \beta_{Jl}\right)^{\mathrm{T}}$, $\beta = \left\{\beta_{00}, \beta_1^{\mathrm{T}}, \cdots, \beta_d^{\mathrm{T}}\right\}_{Jd+1}^{\mathrm{T}}$ be the collection of the coefficients in (2.3), and $\mu = \beta_{00}$, denote $\boldsymbol{B}_{ij}^{(l)} = \left\{B_{1,l}\left(X_{ij}^{(l)}\right), \cdots, B_{J,l}\left(X_{ij}^{(l)}\right)\right\}_{J\times 1}^{\mathrm{T}}$ and $\boldsymbol{B}_{ij} = \left\{1, \boldsymbol{B}_{ij}^{(1)\mathrm{T}}, \cdots, \boldsymbol{B}_{ij}^{(d)\mathrm{T}}\right\}_{(Jd+1)\times 1}^{\mathrm{T}}$, then we have an approximation $\mu + m\left(\boldsymbol{X}_{ij}\right) = \boldsymbol{B}_{ij}^{\mathrm{T}}\beta$. We can also write the approximation of (2.1) in matrix notation as

$$\boldsymbol{Y}_i = \boldsymbol{B}_i^{\mathrm{T}}\beta + \varepsilon_i, \tag{4}$$

where $\boldsymbol{B}_i = \left\{\boldsymbol{B}_{i,1}, \cdots, \boldsymbol{B}_{im}\right\}_{n_i\times(Jd+1)}^{\mathrm{T}}$, $\boldsymbol{Y}_i = \left\{Y_{i1}, Y_{i2}, \cdots, Y_{im}\right\}^{\mathrm{T}}$ and $\varepsilon_i = \left\{\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{im}\right\}^{\mathrm{T}}$. [8] introduced the QIF that approximates the inverse of $\boldsymbol{R}$ by a linear combination of some basis matrixes, *i.e.*

$$\boldsymbol{R}^{-1} \approx a_0\boldsymbol{I} + a_1\boldsymbol{M}_1 + \cdots + a_K\boldsymbol{M}_K,$$

where $\boldsymbol{I}$ is the identity and $\boldsymbol{M}_i$ are known symmetric basis matrices and $a_0, a_1, \cdots, a_K$ are unknown constants. The advantage of the QIF approach is that it does not require the estimation of the linear coefficients a_i's associated with the working correlation matrix, which are treated as nuisance parameters here.

$$\boldsymbol{G}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\beta) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{A}_i^{-1/2}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \\ \vdots \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_K\boldsymbol{A}_i^{-1/2}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \end{pmatrix}. \tag{5}$$

The vector $\boldsymbol{G}_n(\beta)$ contains more estimating equations than parameters, but these estimating equations can be combined optimally using the generalised method of the moment. So according to [8], the QIF approach estimates β by setting G_n as close to zero as possible, in the sense of minimizing the quadratic inference function $\boldsymbol{Q}_n(\beta)$.

$$\boldsymbol{Q}_n(\beta) = n\boldsymbol{G}_n^{\mathrm{T}}(\beta)\boldsymbol{C}_n^{-1}(\beta)\boldsymbol{G}_n(\beta), \tag{6}$$

where

$$\boldsymbol{C}_n^{-1}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\beta)\boldsymbol{g}_i^{T}(\beta).$$

Our main goal is to find both zero components (*i.e.*, $m_j \equiv 0$) and linear compoents (*i.e.*, m_j is a linear function). The former can be achieved by shrinking $\left\|m_j^{sp}\right\|$ to zero. For the latter, we want to shrink the second derivative $\left\|m_j^{sp(")}\right\|$ to zero instead. This suggests the following minimization problem

$$\hat{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta) + n\sum_{l=l}^{d} p_{\lambda_1}\left(\left\|m_l^{sp}\right\|\right) + n\sum_{l=1}^{d} p_{\lambda_2}\left(\left\|m_l^{sp(")}\right\|\right), \tag{7}$$

where $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$ are two penalties used to find zero and linear coefficients respectively, with two regularization parameters λ_1 and λ_2, and $m_l^{sp} = \beta_l^{\mathrm{T}}\boldsymbol{B}^{(l)}$, $\boldsymbol{B}^{(l)} = \left\{B_{1,l}, B_{2,l}, \cdots, B_{J,l}\right\}^{\mathrm{T}}$. Note that since $\left\|m_l^{sp}\right\|^2 = \left\|\hat{\beta}_l^{\mathrm{T}}\boldsymbol{B}^{(l)}\right\|^2 = \int\left(\sum_k \beta_{kl}B_{kl}(x)\right)\left(\sum_{k'}\beta_{k'l}B_{k'l}(x)\right)\mathrm{d}x$ and

$$\left\|m_l^{sp(")}\right\|^2 = \left\|\hat{\beta}_l^{\mathrm{T}}\boldsymbol{B}^{(l)}\right\|^2 = \int\left(\sum_k \beta_{kl}B_{kl}^{(")}(x)\right)\left(\sum_{k'}\beta_{k'l}B_{k'l}^{(")}(x)\right)\mathrm{d}x,$$

$\left\|m_l^{sp}\right\|$ and $\left\|m_l^{sp(")}\right\|$ can be equivalently written as $\|\beta_l\|_{\boldsymbol{D}_l} = \sqrt{\beta_l^{\mathrm{T}}\boldsymbol{D}_l\beta_l}$ and $\|\beta_l\|_{\boldsymbol{E}_l} = \sqrt{\beta_l^{\mathrm{T}}\boldsymbol{E}_l\beta_l}$ respectively,

with (k,k') entry of $\boldsymbol{D}_l$ being $\int_0^1 B_{kl}(x)B_{k'l}(x)\mathrm{d}x$.

2.2. Asymptotic Properties

To study the rate of convergence for $\hat{\mu}$ and $\hat{\beta}$, we first introduce some notations and present regularity conditions. We assume equal cluster sizes $(n_i = m < \infty)$, and $(Y_i, X_i), i = 1,\cdots,n$ are i.i.d. from (Y, X) with $\boldsymbol{Y}_i = (Y_{i1},\cdots,Y_{im})^{\mathrm{T}}$, and $\boldsymbol{X}_i = (X_{i1}^{\mathrm{T}},\cdots,X_{im}^{T})^{\mathrm{T}}$. For convenience, we assume that $m_j(\cdot)$ is truly nonparametric for $1 \le j \le d_1$, is linear for $d_1+1 \le j \le d_1+d_2 = s$, and is zero for $s+1 \le j \le d$. The asymptotic result still hold for data with unequal cluster sizes m_i using a cluster-specific transformation as discuss in [4]. For any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes the modulus of the largest singular value of $\boldsymbol{A}$. To prove the theoretical arguments, we need the following assumptions:

(A1) The covariates $\boldsymbol{X}_i = \{X_{i1}^{\mathrm{T}},\cdots,X_{im}^{\mathrm{T}}\}^{\mathrm{T}}$ are compactly supported, and without loss of generality, we assume that each X_{ij}^{T} has support $\chi = [0,1]^d$. The density of X_{ij}^{T}, denoted by $f_j(\boldsymbol{x})$, is absolutely continuous and there exist constants C_1 and C_2 such that $0 < C_1 \le \min_{x\in\chi} f_j(\boldsymbol{x}) \le C_2 < \infty$ for all $j = 1,\cdots,m$.

(A2) Let $e = Y - \mu - m(\boldsymbol{X})$. Then $\tilde{\Sigma} = Eee^{\mathrm{T}}$ is positive definite and for some $\delta > 0$, $E\|e\|^{2+\delta} < +\infty$.

(A3) For each $l = 1,\cdots,d$, $m_l(\cdot)$ has r continuous derivatives for some $r \ge 2$.

(A4) $$\boldsymbol{G}_n(\beta) = \lim_{n\to\infty}\frac{1}{n}E\begin{pmatrix}\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i\\ \vdots \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_K\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i\end{pmatrix} = J_0. \tag{8}$$

(A5) Let $\boldsymbol{M} = (\boldsymbol{M}_0^{\mathrm{T}},\cdots,\boldsymbol{M}_k^{\mathrm{T}})^{\mathrm{T}}$. Assume the modular of the singular value of $\boldsymbol{M}$ is bounded away from 0 and infinity.

(A6) The matrix $\boldsymbol{A}$ defined in Theorem 3 is positive definite.

Theorem 1. Suppose that the regularity conditions A1-A5 hold and the number of knots $K = O_p(n^{1/(2r+1)})$, $\lambda_1, \lambda_2 \to 0$. Then there exists a local minimizer of (2.7) such that

$$|\hat{\mu} - \mu_0| = O_p(n^{-r/(2r+1)}),$$

$$\max_{1\le l\le d}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\hat{m}_l(x) - m_l(x)\}^2 = O_p(n^{-2r/(2r+1)}).$$

For $m_i = m = 1$, it reduces to a special case where the responses are i.i.d. The rate of convergence given here is the same optimal rate as that obtain for polynomial spline regression for independent data [9] [10]. The main advantage of the QIF approach is that it incorporates within-cluster correlation by optimally combing estimating equations without estimating the correlation parameters. the estimator of two fold penalized QIF achieve the same rate of convergence as un-penalized estimator. Furthermore, we prove that the penalized estimators $\{\hat{\beta}_l\}_{l=1}^{d}$ in Theorem 1 possess the sparsity property, $\hat{m}_l = 0$ almost surely for $l = s+1,\cdots,d$. The sparsity property ensures that the proposed model selection is consistent, that is, it selects the correct variables with probability tending to 1 as the sample size goes to infinity.

Theorem 2. Under the same assumptions of Theorem 1, and if the tuning parameter $n^{r/(2r+1)}\min\{\lambda_1,\lambda_2\} \to \infty$. Then with probability approaching 1.

a) $\hat{m}_j \equiv 0, s+1 \le j \le d$

b) $\hat{m}_j$ is a linear function for $d_1+1 \le j \le s$

Theorem 2 also implies that above additive model selection possesses the consistency property. The results in Theorems 2 are similar to semiparametric estimation of additive quantile regression model in [7]. However, the theoretical proof is very different from the penalized quantile loss function due to the two fold penalty and longitudinal data.

Finally, in the same spirit of that [11], we come to the question of whether the SIC can identify the true model in our setting.

Theorem 3. Suppose that the regularity conditions A1-A5 hold and the number of knots $K = O_p\left(n^{1/(2r+1)}\right)$ as assumed in Theorem 1, The parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ selected by SIC can select the true model with probability tending to 1.

3. Simulation Study

In this section, we conducted Monte Carlo studies for the following longitudinal data and additive model. the continuous responses $\{Y_{ij}\}$ are generated from

$$Y_{ij} = \sum_{l=1}^{d} m_l\left(X_{ij}^{(l)}\right) + \varepsilon_{ij}, \quad 1,\cdots,n, \quad j = 1,\cdots,5 \tag{9}$$

where $d = 10$ and the number of clusters $n = 100, 250, 500$. The additive functions are $m_1(x) = 5\sin(2\pi x)/(2-\sin(2\pi x)), m_2(x) = 8(x-0.5)^2, m_3(x) = 2x, m_4(x) = x, m_5(x) = -x$. Thus the last 5 variables in this model are null variables and do not contribute the model. The covariates $\boldsymbol{X}_{ij} = \left(X_{ij}^{(1)},\cdots,X_{ij}^{(10)}\right)^{\mathrm{T}}$ are generated independently from uniform. The error $\varepsilon_i = (\varepsilon_{i1},\cdots,\varepsilon_{i6})^{\mathrm{T}}$ follows a multivariate normal distribution with mean 0, a common marginal variance $\sigma^2 = 1$, it has first-order autoregressive (AR-1) and an compound symmetry (CS) correlation (*i.e.* exchangeable correlation) structure with different within correlation coefficient, and consider $\rho = 0.8$ and $\rho = 0.3$ representing a strong and weak within correlation structure. The predictors $\boldsymbol{X}_{ij} = \left(X_{ij}^{(1)},\cdots,X_{ij}^{(10)}\right)^{\mathrm{T}}$ are generated by $X_{ij}^{(l)} = \Phi\left(Z_{ij}^{(l)}\right)$, $\boldsymbol{Z}_{ij} = \left(Z_{ij}^{(1)},\cdots,Z_{ij}^{(l)}\right) \sim N(0,\Sigma)$, $i = 1,\cdots,n, j = 1,\cdots,m$, where Φ is the standard normal c.d.f. and $\Sigma = (1-r)\boldsymbol{I}_{(d\times d)} + r\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}$. The parameter $r\ (0 \le r < 1)$ controls the correlation between $Z_{ij}^{(l)}, 1 \le (l) \le d$.

To illustrate the effect on estimation efficiency, we compare the penalized QIF approach in [4] (PQIF) and an Oracle model (ORACLE). here the full model consists of all ten variable, and oracle model only contains the first five relevant variables and we know it's a partial additive model. The oracle model is only available in simulation studies where the true information is known. In all simulation, the number of replications is 100 and the result are summarized in **Table 1** and **Table 2**. In **Table 1**, the model selection result for both our procedure

Table 1. The estimation results for our estimator (TFPQIF) and sparse additive estimator (PQIF) and ORACLE esitmator.

n	Correlation	Method	μ	m_1	m_2	m_3	m_4	m_5
100	CS	PQIF	0.32	0.42	0.3	0.29	0.31	0.26
		TFPQIF	0.3	0.46	0.28	0.25	0.23	0.22
		ORACLE	0.14	0.14	0.15	0.15	0.13	0.12
	AR(1)	PQIF	0.36	0.39	0.32	0.3	0.29	0.25
		TFPQIF	0.29	0.39	0.35	0.2	0.25	0.22
		ORACLE	0.13	0.15	0.22	0.14	0.12	0.1
250	CS	PQIF	0.25	0.29	0.25	0.24	0.19	0.15
		TFPQIF	0.22	0.31	0.26	0.14	0.16	0.15
		ORACLE	0.12	0.11	0.19	0.097	0.098	0.09
	AR(1)	PQIF	0.28	0.24	0.31	0.33	0.28	0.19
		TFPQIF	0.20	0.2	0.27	0.24	0.14	0.15
		ORACLE	0.1	0.11	0.13	0.21	0.1	0.096
500	CS	PQIF	0.15	0.14	0.25	0.23	0.2	0.17
		TFPQIF	0.15	0.3	0.26	0.11	0.12	0.1
		ORACLE	0.09	0.13	0.12	0.07	0.07	0.07
	AR(1)	PQIF	0.18	0.3	0.26	0.13	0.12	0.14
		TFPQIF	0.16	0.23	0.25	0.09	0.09	0.09
		ORACLE	0.08	0.13	0.12	0.077	0.081	0.07

Table 2. Model selction results for our estimator (TFPQIF) and sparse additive estimator (PQIF) and ORACLE esitmator.

λ_0	λ_1	CS				AR(1)			
		NCC	NNT	NLC	NLT	NCC	NNT	NLC	NLT
100	PQIF	5.96	2	0	0	5.83	2	0	0
	TFPQIF	2.64	2	2.58	2.36	2.52	2	2.63	2.46
	ORACLE	2	2	3	3	2	2	3	3
250	PQIF	5.63	2	0	0	5.45	2	0	0
	TFPQIF	2.34	2	2.66	2.65	2.41	2	2.59	2.5
	ORACLE	2	2	3	3	2	2	3	3
500	PQIF	5.35	2	0	0	5.2	2	0	0
	TFPQIF	2.04	2	2.93	2.93	2.1	2	2.89	2.86
	ORACLE	2	2	3	3	2	2	3	3

with the one penalty QIF when the error are Gaussian, and we also list the oracle model as a benchmark, the oracle model is only available in simulation studies where the true information is known in **Table 1**, in which the column labeled "NNC" presents the average number of nonparametric components selected, the column "NNT" depicts the average number of nonparametric components selected that are truly nonparametric (truly nonzero for one penalty QIF), "NLC" presents the average number of linear components, "NLT" depicts the average number of linear components selected that are truly linear.

In **Table 2**, we conduct some simulations to evaluate finite sample performance of the proposed method. Let $\hat{m}_k(\cdot)$ be the estimator of a nonparametric function $m_k(\cdot)$ and $\{u_s\}_{s=1}^{M}$ be the grid points, the performance of estimator $\hat{m}_j(\cdot)$ will be assessed by using the square root of average square errors(RASE), we compare the performance of above estimators. On the nonparametric coponents, the errors for estimators with a single penalty and our procedure are similar, and both are qualitatively close to those of the oracle estimator. However, for the parametric components, our estimator is obviously more efficient,leading to about 40% - 50% reduction in RASE.

$$RASE = \left\{ \frac{1}{M} \sum_{s=1}^{M} \sum_{k=1}^{d} \left[\hat{m}_k(u_s) - m_k(u_s) \right]^2 \right\}^{1/2}.$$

4. Real Data Analysis

In this subsection, we analyze data from the Multi-Center AIDS Cohort Study. The dataset contains the human immunodeficiency virus, HIV, status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. All individuals were scheduled to have their measurements made during semi-annual visits. Here $t_{ij}, i = 1, \cdots, n, j = 1, \cdots, m_i$ denotes the time length in years between seroconversion and the j-th measurement of the i-th individual after the infection. [12] analyzed the dataset using partial linear models. The primary interest was to describe the trend of the mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage, and age at infection on the mean CD4 cell percentage after the infection.

In our analysis, the response variable is the CD4 cell percentage of a subject at distinct time points after HIV infection. We take four covariates for this study: X_1, the CD4 cell percentage level before HIV infection; and X_2, age at HIV infection; X_3 the individual's smoking status, which takes binary values 1 or 0, according to whether a individual is a smoker or nonsmoker; T_{ij} denote $i = 1, \cdots, n, j = 1, \cdots, m$, denotes the time length in years between seroconversion and the j-th measurement of the i-th individual after the infection. We construct the following additive model;

$$Y_{ij} = \mu + m_1\left(X_{(1)}^{ij}\right) + m_2\left(X_{ij}^{(2)}\right) + m_3\left(T_{ij}\right) + \beta_0 X_{ij}^{(3)} + \epsilon_{ij}.$$

the partially linear additive models instead of additive model because of the binaray variable $X^{(3)}$, but we not select the linear component. using our procedure, we wang to ensure which is linear component and which is zero in the non-parametirc function. For implement our procedure, linear transformation be used to the variable $X^{(1)}, X^{(2)}, T$. The result of our procedure select the m_1 is zero function and select the m_2 is a linear function, m_3 is a non-parametric. As shown in **Figure 1**, we see that the mean baseline CD4 percentage of the population

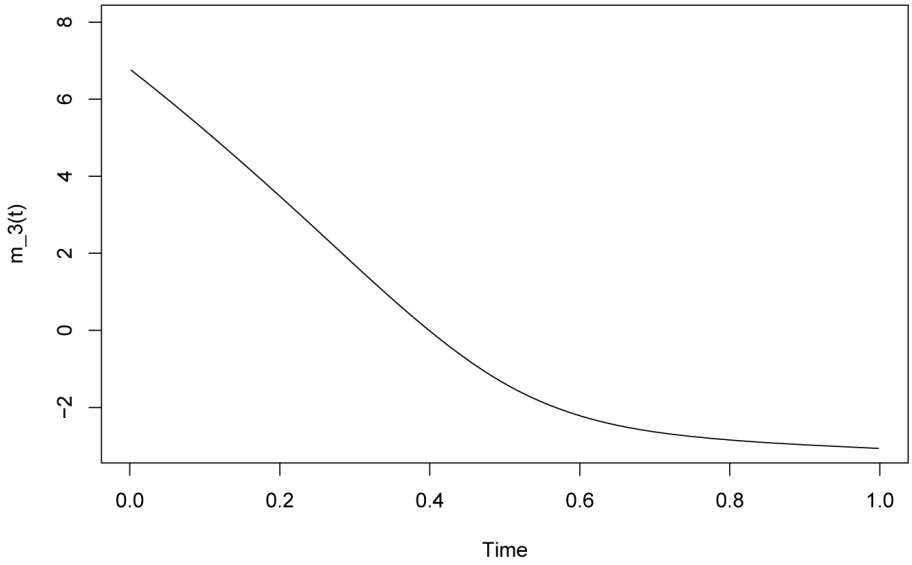


Figure 1. The estimator of $m_3(x)$.

depletes rather quickly at the beginning of HIV infection, but the rate of depletion appears to be slowing down at four years after the infection. This result is the same as before [13].

5. Concluding Remark

In summary, we present a two-fold penalty variable selection procedure in this paper, which can select linear component and significant covariate and estimate unknown coefficient function simultaneously. The simulation study shows that the proposed model selection method is consistent with both variable selection and linear components selection. Besides, being theoretically justified, the proposed method is easy to understand and straightforward to implement. Further study of the problem is how to use the multi-fold penalty to solve the model selection and variable selection in generalized additive partial linear models with longitudinal data.

Acknowledgements

Liugen Xue's research was supported by the National Nature Science Foundation of China (11171012), the Science and Technology Project of the Faculty Adviser of Excellent PhD Degree Thesis of Beijing (20111000503) and the Beijing Municipal Education Commission Foundation (KM201110005029).

References

[1] Hastie, T.J. and Tibshirani, R.J. (1990) Generalized Additive Models. Chapman and Hall, London.

[2] Berhane, K. and Tibshirani, R.J. (1998) Generalized Additive Models for Longitudinal Data. *The Canadian Journal of Statistics*, **26**, 517-535. http://dx.doi.org/10.2307/3315715

[3] Martinussen, T. and Scheike, T.H. (1999) A Semiparametric Additive Regression Model for Longitudinal Data. *Biometrika*, **86**, 691-702. http://dx.doi.org/10.1093/biomet/86.3.691

[4] Xue, L. (2010) Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data. *Journal of the American Statistical Association*, **105**, 1518-1530. http://dx.doi.org/10.1198/jasa.2010.tm10128

[5] Hu, T. and Xia, Y.C. (2012) Adaptive Semi-Varying Coefficient Model Selection. *Statistica Sinica*, **22**, 575-599. http://dx.doi.org/10.5705/ss.2010.105

[6] Tang, Y.L., Wang, H.X., Zhu, Z.Y. and Song, X.Y. (2012) A Unified Variable Selection Approach for Varying Coefficient Models. *Statistica Sinica*, **22**, 601-628. http://dx.doi.org/10.5705/ss.2010.121

[7] Lian, H. (2012) Shrinkage Estimation for Identification of Linear Components in Additive Models. *Statistics and Probability Letters*, **82**, 225-231. http://dx.doi.org/10.1016/j.spl.2011.10.009

[8] Qu, A., Lindsay, B.G. and Li, B. (2000) Improving Generalised Estimating Equations Using Quadratic Inference Func-

tions. *Biometrika*, **87**, 823-836. http://dx.doi.org/10.1093/biomet/87.4.823

[9] Huang, J.Z. (1998) Projection Estimation in Multiple Regression with Application to Functional ANOVA Models. *The Annals of Statistics*, **26**, 242-272. http://dx.doi.org/10.1214/aos/1030563984

[10] Xue, L. (2009) A Root-N Consistent Backfitting Estimator for Semiparametric Additive Modeling. *Statistica Sinica*, **19**, 1281-1296.

[11] Wang, H., Li, R. and Tsai, C.L. (2007) Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553-568. http://dx.doi.org/10.1093/biomet/asm053

[12] Xue, L.G. and Zhu, L.X. (2007) Empirical Likehood for a Varying Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **102**, 642-654.

[13] Wu, C.O., Chiang, C.T. and Hoover, D.R. (2010) Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **93**, 1388-1402.

[14] De Boor, C. (2001) A Practical Guide to Splines. Springer, New York.

Appendix: Proofs of Theorems

For convenience and simplicity, let C denote a positive constant that may be different at each appearance throughout this paper. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

Lemma 1. Under the conditions (A1)-(A6), minimizing the no penalty QIF $\tilde{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta)$. Then $|\tilde{\mu} - \mu_0| = O\left(n^{-\frac{2r}{2r+1}}\right)$, $\max_{1\le l\le d} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\tilde{m}_l(x_{ij}) - m_l(x_{ij})\right\}^2 = O_p\left(n^{-2r/(2r+1)}\right)$.

Proof: According to [14], for each $l = 1,\cdots,d$, we can get $m = \mu + \sum_{l=1}^{d} m_l$ satisfying the condition (4). There exists a constant $C > 0$ and a spline function $\tilde{m} \in \mathcal{G}_n$, such that $\|\tilde{m} - m\|_{\infty} \le CK^{-r}$. Using the triangular in equality $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\tilde{m}_l(x_{ij}) - m_l(x_{ij})\right\}^2 \le \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\boldsymbol{B}_l(x_{ij})\left(\tilde{\beta}_l - \beta_l^0\right)\right\}^2 + O\left(K^{-2r}\right)$. Therefore, it is sufficient to show that $\left\|\boldsymbol{B}_l\left(\tilde{\beta}_l - \beta_l^0\right)\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\boldsymbol{B}_l(x_{ij})\left(\tilde{\beta}_l - \beta_l^0\right)\right\}^2 = O_p\left(n^{-1}K\right)$. According to [8] entail that for any $\epsilon > 0$. exists sufficiently large $C > 0$. such that $n \to \infty$ $p\left\{\inf_{\|\boldsymbol{B}(\tilde{\beta}-\beta^0)\|_n \le C(K/n)^{1/2}} \boldsymbol{Q}_n(\beta) > \boldsymbol{Q}_n(\beta_0)\right\} > 1-\epsilon$, therefore $p\left\{\left\|\boldsymbol{B}\left(\tilde{\beta} - \beta^0\right)\right\|_n = O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)\right\}$. Furthermore, for each $l = 1,\cdots,d$. There exists a constant $C > 0$. such that $\left\|\boldsymbol{B}\left(\tilde{\beta}_l - \beta_l^0\right)\right\|_n^2 \le C\left\|\boldsymbol{B}\left(\tilde{\beta} - \beta^0\right)\right\|_n^2 = O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)$.

Proof of Theorem 1. Let

$$\boldsymbol{L}_n(\beta) = \boldsymbol{Q}_n(\beta) + n\sum_{j=1}^{d} p_{\lambda_1}\left(\left\|\tilde{m}_j\right\|\right) + n\sum_{j=1}^{d} p_{\lambda_2}\left(\left\|\tilde{m}_j''\right\|\right)$$

be the object function in (2.7), where $\|\beta_j\|_{\boldsymbol{D}_j} = \sqrt{\beta_j^{\mathrm{T}}\boldsymbol{D}_j\beta_j}$ and $\|\beta_j\|_{\boldsymbol{E}_j} = \sqrt{\beta_j^{\mathrm{T}}\boldsymbol{E}_j\beta_j}$, as a special case of no penalty QIF. Let $\tilde{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta)$, and $\tilde{m}_j = \boldsymbol{B}^{\mathrm{T}}\tilde{\beta}_j$, well known result is $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\tilde{\beta} - \beta_0\right)\right\|_n = O_p\left(\sqrt{K/n}\right)$, we want to show that for large n and any $\varepsilon > 0$, there exist a constant C large enough such that

$$p\left\{\inf_{\beta:\left\|\boldsymbol{B}^{\mathrm{T}}(\tilde{\beta}-\beta_0)\right\| = C\left(nK^{-1}\right)^{-1/2}} \boldsymbol{L}_n(\beta) > \boldsymbol{L}_n\left(\tilde{\beta}\right)\right\} \ge 1-\epsilon. \tag{A1}$$

As a result, this implies that $\boldsymbol{L}_n(\cdot)$ has a local minimum in the ball $\left\{\beta : \left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta - \tilde{\beta}\right)\right\|\right\}$. Thus, $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta - \tilde{\beta}\right)\right\|_n = O_p\left(1/\sqrt{nK^{-1}}\right)$. Further, the triangular inequality gives $\left\|\boldsymbol{B}^{\mathrm{T}}\hat{\beta} - m\right\|_n = O_p\left(\left(nK^{-1}\right)^{-1/2} + K^{-r}\right)$. To show (A1), For convenience, we assume that m_j is truly nonparametric for $1 \le j \le d_1$ is linear for $d_1 + 1 \le j \le s = d_1 + d_2$ and zero for $s+1 \le j \le d$.

$$\begin{aligned}\boldsymbol{L}_n(\beta) - \boldsymbol{L}_n\left(\tilde{\beta}_j\right) \ge \boldsymbol{Q}_n(\beta) - \boldsymbol{Q}_n\left(\tilde{\beta}_j\right) &+ \sum_{j=1}^{s} n\left\{p_{\lambda_1}\left(\|\beta_j\|_{\boldsymbol{D}_j}\right) - p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right\} \\ &+ \sum_{j=1}^{d_1} n\left\{p_{\lambda_2}\left(\|\beta_j\|_{\boldsymbol{E}_j}\right) - p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right\}.\end{aligned} \tag{A2}$$

Since $\left\|\hat{\beta}_j - \beta_{0j}\right\|_n$. Since $\lambda_1, \lambda_2 = o(1)$. We have $p\left(p_{\lambda_1}\left(\|\beta_j\|_{\boldsymbol{D}_j}\right) = p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\right) \to 1$. If $i \le s = d_1 + d_2$,

similarly, $p\left(p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)=p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right)\to 1$. If $l\le d_1$. These facts imply that $n\sum_{j=1}^{d}p_{\lambda_1}\left(\left\|\beta_j\right\|_{\boldsymbol{D}_j}\right)-n\sum_{j=1}^{d}p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\ge 0$ and $n\sum_{j=1}^{d}p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)-n\sum_{j=1}^{d}p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\ge 0$ with probability tending to 1. If $\lambda_1,\lambda_2=o(1)$, $\left\|\beta_j\right\|_{\boldsymbol{D}_j}\ge C\lambda_{\max}\left(D_j\right)$, $\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\ge C\lambda_{\max}\left(E_j\right)$, for $i=1,\cdots,d$. Therefore, when n is large enough,

$$\sum_{j=1}^{d}n\left\{p_{\lambda_1}\left(\left\|\beta_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\right\}=C_1nK^{-1/2}\lambda_1\sum_j\left\|\beta_j-\tilde{\beta}_j\right\|_n\to 0$$

$$\sum_{j=1}^{d}n\left\{p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)-p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right\}=C_2nK^{-1/2}\lambda_2\sum_j\left\|\beta_j-\tilde{\beta}_j\right\|_n\to 0$$

By the definition of SCAD penalty function, removing the regularizing terms in (A2)

$$\boldsymbol{Q}_n(\beta)-\boldsymbol{Q}_n\left(\tilde{\beta}_j\right)=\left(\beta-\tilde{\beta}\right)^{\mathrm{T}}\dot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)+\frac{1}{2}\left(\beta-\tilde{\beta}\right)^{\mathrm{T}}\ddot{\boldsymbol{Q}}_n\left(\tilde{\beta}\right)\left(\beta-\tilde{\beta}\right)\left\{1+o_p(1)\right\}\tag{A3}$$

with $\dot{\boldsymbol{Q}}_n$ and $\ddot{\boldsymbol{Q}}_n$ being the gradient vector and hessian matrix $\boldsymbol{Q}_n$, respectively. Following [8], and Lemma A1 in supplement, for any β with $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta-\tilde{\beta}\right)\right\|_n=C\left(nK^{-1}\right)^{-1/2}$, one has

$$n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)=n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{G}}_n^{\mathrm{T}}\left(\tilde{\beta}_j\right)\boldsymbol{C}_n^{-1}\left(\tilde{\beta}_j\right)\boldsymbol{G}_n\left(\tilde{\beta}_j\right)\left\{1+o_p(1)\right\}=O(K)$$

and

$$\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\ddot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)\left(\beta-\tilde{\beta}_j\right)=n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{G}}_n^{\mathrm{T}}\left(\tilde{\beta}_j\right)\boldsymbol{C}_n^{-1}\left(\tilde{\beta}_j\right)\dot{\boldsymbol{G}}_n\left(\tilde{\beta}_j\right)\left(\beta-\tilde{\beta}_j\right)\left\{1+o_p(1)\right\}=O(K)$$

where $\dot{\boldsymbol{G}}_n$ is the first order derivative of $\boldsymbol{G}_n$. Therefore, by choosing C large enough, the second term on (A3) dominates its first term. therefore (A1) holds when C and n are large enough. This completes the proof of Theorem 1. □

Proof of Theorem 2. We only show part (b), as an illustration and part (a) is similar. Suppose for some $d_1<j\le s$, $B_j^{\mathrm{T}}\hat{\beta}_j$ does not represent a linear function. Define $\hat{\beta}_j^{\star}$ to be the same as $\hat{\beta}$ except that $\hat{\beta}_j$ is replaced by its projection onto the subspace $\{\beta_j:B_j^{\mathrm{T}}\beta_j$ represents a linear function$\}$, we have

$$\boldsymbol{L}_n\left(\hat{\beta}\right)-\boldsymbol{L}_n\left(\tilde{\beta}_j\right)=\boldsymbol{Q}_n\left(\hat{\beta}\right)-\boldsymbol{Q}_n\left(\hat{\beta}_j^{\star}\right)+\sum_j n\left\{p_{\lambda_1}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right\}$$
$$+\sum_j n\left\{p_{\lambda_2}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}\right)-p_{\lambda_2}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{E}_j}\right)\right\}.$$

As in the proof of Theorem 1, we have $P\left(p_{\lambda_1}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right)\to 1$ and thus with probability approaching 1

$$0\ge\boldsymbol{L}_n\left(\hat{\beta}\right)-\boldsymbol{L}_n\left(\tilde{\beta}_j\right)=\boldsymbol{Q}_n\left(\hat{\beta}\right)-\boldsymbol{Q}_n\left(\tilde{\beta}_j\right)+n\sum_j p_{\lambda_2}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}\right).\tag{A4}$$

$\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}=\sqrt{\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}=\sqrt{\left(\hat{\beta}_j-\beta_{0j}\right)^{\mathrm{T}}\boldsymbol{E}_j\left(\hat{\beta}_j-\beta_{0j}\right)}=O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)=o(\lambda_2)$. $p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)=\lambda_2\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}$, with probability tending to 1. by the definition of SCAD penalty. $\left\|\hat{\beta}_j-\hat{\beta}_j^{\star}\right\|=O_p\left(\sqrt{K\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}\right)$, $np_{\lambda_2}\left(\sqrt{\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}\right)=n\lambda_2\sqrt{K\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}$. Therefore, similar to the proof of Theorem 1, by choosing C large enough, the second term on the right had side of (A4) dominates its first term. □

Proof of Theorem 3. For any regularization parameters $\lambda = (\lambda_1, \lambda_2)$, we denote the estimator of two fold penalty $\hat{\beta}_\lambda$, and denote by $\hat{\beta}$ the minimizer when the optimal sequence of regularization parameters is chosen. There are four separate cases to consider

CASE 1: $B_j \beta_{\lambda j}$ represents a linear component for some $j \le d_1$. Similar to the proof of Theorems 1 and 2, we have

$$\boldsymbol{Q}_n(\hat{\beta}) - \boldsymbol{Q}_n(\hat{\beta}_\lambda) = (\hat{\beta} - \hat{\beta}_\lambda)^{\mathrm{T}} \dot{\boldsymbol{Q}}_n(\hat{\beta}_\lambda) + \frac{1}{2}(\hat{\beta} - \hat{\beta}_\lambda)^{\mathrm{T}} \ddot{\boldsymbol{Q}}_n(\hat{\beta})(\hat{\beta} - \hat{\beta}_\lambda)\{1 + o_p(1)\}.$$

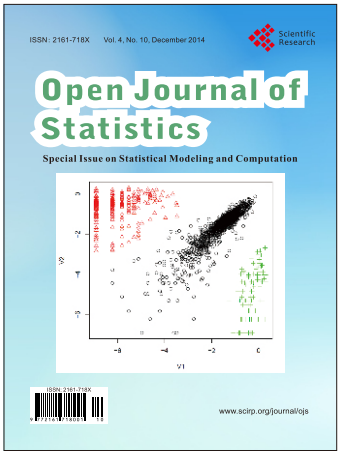
Since true m_j not linear and $\hat{\beta}_j$ is consistent in model selection, $\dfrac{\|\hat{\beta} - \hat{\beta}_\lambda\|}{K}$ is bounded away form zero, thus $\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta})\right\} - \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} > (C_1 K + C_2)\dfrac{\log(n)}{2n}$, for any $0 \le C_1, C_2 \le d$, with probability tending to 1 and the SIC cannot select such λ.

CASE 2: $\hat{\beta}_{\lambda j}$ is zero for some $1 \le j \le s$. The proof is very similar with CASE 1 and therefore omitted.

CASE 3: $\boldsymbol{B}_j \hat{\beta}_j$ represents a nonlinear component for some $d_1 < j \le s$. Here when considering CASE 3, we implicity exclude all previous cases that no underfitting cases. $\tilde{\beta}$ is the estimator of minimizing the no penalty QIF (2.6) $\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda) - \frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta}) \ge -\left|O\left(\frac{K}{n}\right)\right|$. Thus $\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} - \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta}) \ge -\left|O\left(\frac{K}{n}\right)\right|\right\}$ and

$\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} + \dfrac{K\log(n)}{2n} \ge \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta})\right\} + \dfrac{d\log(n)}{2n}$ with probability tending to 1. □

CASE 4: $\hat{\beta}_{\lambda j}$ is nonzero for $j \ge s$. The case is similar to case 3. Thus the proof is omitted.



Call for Papers

Open Journal of Statistics

ISSN 2161-718X (Print) ISSN 2161-7198 (Online)
http://www.scirp.org/journal/ojs

Open Journal of Statistics (OJS) is an international journal dedicated to the latest advancement of statistics. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of statistics.

Editor-in-Chief

Prof. Qihua Wang Chinese Academy of Sciences, China

Editorial Board

Prof. Ana M. Aguilera
Prof. Essam K. Al-Hussaini
Prof. Erniel B. Barrios
Prof. Alexander V. Bulinski
Prof. Junsoo Lee
Prof. Tae-Hwy Lee
Dr. Qizhai Li
Dr. Xuewen Lu
Prof. Claudio Morana
Prof. Thu Pham-Gia
Prof. Gengsheng Qin
Prof. Kalyan Raman
Prof. Mohammad Z. Raqab
Dr. Jose Antonio Roldán-Nofuentes
Prof. Sunil K. Sapra
Prof. Raghu Nandan Sengupta
Prof. Siva Sivaganesan
Dr. Zheng Su
Prof. Jianguo Sun
Prof. Aida Toma
Dr. Florin Vaida
Dr. Haiyan Wang
Prof. Augustine Chi Mou Wong
Dr. Peng Zeng
Dr. Hongmei Zhang
Dr. Jin-Ting Zhang
Prof. Yichuan Zhao
Dr. Wang Zhou

Subject Coverage

This journal invites original research and review papers that address the following issues in statistics. Topics of interest include, but are not limited to:

- Actuarial science
- Applied information economics
- Asymptotic statistics
- Bayesian statistics
- Biostatistics
- Business statistics
- Causal inference
- Chemometrics
- Computational statistics
- Data mining
- Decision theory
- Demography
- Descriptive statistics
- Design of experiments
- Econometrics
- Energy statistics
- Engineering statistics
- Epidemiology
- Estimation theory
- Geographic information systems
- Graphic models and related theory
- High dimensional data analysis
- Image processing
- Multivariate analysis
- Non-parametric statistics
- Parametric statistics
- Psychological statistics
- Regression analysis
- Reliability
- Reliability engineering
- Sample survey
- Sampling theory
- Semiparametric statistics
- Social statistics
- Statistical analysis with complex data
- Statistical computing
- Statistical inference
- Statistical methods
- Survival analysis
- Theoretic methods
- Time series analysis

We are also interested in short papers (letters) that clearly address a specific problem, and short survey or position papers that sketch the results or problems on a specific topic. Authors of selected short papers would be invited to write a regular paper on the same topic for future issues of the OJS.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

http://www.scirp.org/journal/ojs Email: ojs@scirp.org

What is SCIRP?

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?

All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience

Scientific Research

Website: http://www.scirp.org

Subscription: sub@scirp.org

Advertisement: service@scirp.org