ISSN: 2469-5831

Gu et al. Int J Clin Biostat Biom 2018, 4:017
DOI: 10.23937/2469-5831/1510017
Volume 4 | Issue 1
Open Access

International Journal of
Clinical Biostatistics and Biometrics

RESEARCH ARTICLE

Partial Variable Selection and its Applications in Biostatistics

***Jingwen Gu*[1]*, Ao Yuan*[1,2*]*, Chunxiao Zhou*[2]*, Leighton Chan*[2] *and Ming T Tan*[1*]**

[1]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, USA

[2]Epidemiology and Biostatistics Section, Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, USA

***Corresponding author:** *Ao Yuan, Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington DC 20057, USA; Epidemiology and Biostatistics Section, Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, Bethesda MD 20892, USA, E-mail: ay312@georgetown.edu;*
Ming T Tan, Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington DC 20057, USA, E-mail: mtt34@georgetown.edu

Abstract

We propose and study a method for partial covariates selection, which only select the covariates with values fall in their effective ranges. The coefficients estimates based on the resulting data is more interpretable based on the effective covariates. This is in contrast to the existing method of variable selection, in which some variables are selected/deleted in whole. To test the validity of the partial variable selection, we extended the Wilks theorem to handle this case. Simulation studies are conducted to evaluate the performance of the proposed method, and it is applied to a real data analysis as illustration.

Keywords

Covariate, Effective range, Partial variable selection, Linear model, Likelihood ratio test

Introduction

Variables selection is a common practice in biostatistics and there is vast literature on this topic. Commonly used methods include the likelihood ratio test [1], Akaike information criterion, AIC [2] Bayesian information criterion, BIC [3], the minimum description length [4,5] stepwise regression and Lasso [6], etc. The principal components model linear combinations of the original covariates, reduces large number of covariates to a handful of major principal components, but the result is not easy to interpret in terms of the original covariates. The stepwise regression starts from the full model and deletes the covariate one by one according to some statistical significance measure. May, et al. [7] addressed variable selection in artificial neural network models, Mehmood, et al. [8] gave a review for variable selection with partial least squares model. Wang, et al. [9] addressed variable selection in generalized additive partial linear models. Liu, et al. [10] addressed variable selection in semiparametric additive partial linear models. The Lasso [6,11] and its variation [12,13] are used to select some few significant variables in the presence of a large number of covariates.

However, existing methods only select the whole variable(s) to enter the model, which may not the most desirable in some bio-medical practice. For example, in two heart disease studies [14,15] there are more than ten risk factors identified by medical researchers in their long time investigations, with the existing variable selection methods, some of the risk factors will be deleted wholly from the investigation, this is not desirable, since risk factors will be really risky only when they fall into some risk ranges. Thus deleting the whole variable(s) in this case seems not reasonable, while a more reasonable way is to find the risk ranges of these variables, and delete the variable values in the un-risky ranges. In some other studies, some of the covariates values may just random errors which do not contribute to the influence of the responses, and remove these covariates values will make the model interpretation more accurate. In this sense we select the variables when their value falls within some range. To our knowledge, method for this kind of partial variable selection hasn't been seen in

Citation: Gu J, Yuan A, Zhou C, Chan L, Tan MT (2018) Partial Variable Selection and its Applications in Biostatistics. Int J Clin Biostat Biom 4:017. doi.org/10.23937/2469-5831/1510017

Received: January 08, 2018: **Accepted:** April 12, 2018: **Published:** April 14, 2018

the literature, which is the goal of our study here. Note that in existing method of variable selection, some variables are selected/deleted, while in our method, some variable(s) are partially selected/deleted, i.e., only some proportions of some variable observations are selected/deleted. The latter is very different from the existing methods. In summary, traditional variable selection methods, such as stepwise or Lasso, some covariate(s) will be removed either wholly or none from the analysis. This is not very reasonable, since some of the removed covariates may be partially effective, removing all their values may yield miss-leading results, or at least cost information loss; while for the variables remaining in the model, not all their values are necessarily effective for the analysis. With the proposed method, only the non-effective values of the covariates are removed, and the effective values of the covariates are kept in the analysis. This is more reasonable than the existing methods of removing all or nothing.

In the existing method of deleting whole variable(s), the validity of such selection can be justified using the Wilks result, under the null hypothesis of no effect of the deleted variable(s), the resulting two times log-likelihood ratio will be asymptotically chi-squared distributed. We extended the Wilks theorem to the case for the proposed partial variable deletion, and use it to justify the partial deletion procedure. Simulation studies are conducted to evaluate the performance of the proposed method, and it is applied to analyze a real data set as illustration.

The Proposed Method

The observed data is $(y_i, \mathbf{x}_i)(i=1,\ldots,n)$, where y_i is the response and $\mathbf{x}_i \in R^d$ is the covariates, of the i-th subject. Denote $\mathbf{y}_n = (y_1,\ldots,y_n)'$ and $\mathbf{X}_n = (\mathrm{x}_1',\ldots,\mathrm{x}_\mathrm{n}')'$. Consider the linear model

$$\mathbf{y}_\mathrm{n} = \mathbf{X_n}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathbf{n}, \qquad (1)$$

where $\beta = (\beta_1,\ldots,\beta_d)'$ is the vector of regression parameter, $\boldsymbol{\varepsilon}_\mathbf{n} = (\varepsilon_1,\ldots,\varepsilon_n)'$ is the vector of random errors, or residual departure from the linear model assumption. Without loss of generality we consider the case the ε_i's are independently identically distributed (iid), i.e. with variance matrix $Var(\varepsilon) = \sigma^2 I_n$, where I_n is the n-dimensional identity matrix. When the ε_i's are not iid, often it is assumed $Var(\varepsilon) = \Omega$ for some known positive-definite Ω, then make the transformation $\tilde{\mathbf{y}}_n = \Omega^{-1/2}\mathbf{y}_n, \tilde{\mathbf{X}}_n = \Omega^{-1/2}\mathbf{X}_n$ and $\tilde{\boldsymbol{\varepsilon}} = \Omega^{-1/2}\boldsymbol{\varepsilon}$, then we get the model $\tilde{\mathbf{y}}_\mathrm{n} = \tilde{\mathbf{X}}_n\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, and the $\tilde{\boldsymbol{\varepsilon}}_i$'s are iid with $Var(\tilde{\boldsymbol{\varepsilon}}) = I_n$. When Ω is unknown, it can be estimated by various ways. So below we only need to discuss the case the ε_i's are iid.

Summary of existing work:

We first give a brief review of the existing method of variable selection. Assume the model residual $\epsilon = y - \mathbf{x}'\boldsymbol{\beta}$ has some known density function $f(\cdot)$ (such as normal), with possibly some unknown parameter(s). For simple of discussion we assume there are no unknown parameters. Then the log-likelihood is

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(\mathrm{y}_i - \mathbf{x}_i'\beta).$$

Let $\hat{\beta}$ be the Maximum Likelihood Estimate (MLE) of $\boldsymbol{\beta}$ (when $f(\cdot)$ is the standard normal density, $\hat{\boldsymbol{\beta}}$ is just the least squares estimate). If we delete $k(\leq d)$ columns of $\mathbf{X}_n$ and the corresponding components of $\boldsymbol{\beta}$, denote the remaining covariate matrix as $\mathbf{X}_n^-$ and the resulting $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^-$, and the corresponding MLE as $\hat{\boldsymbol{\beta}}^-$. Then under the hypothesis H_0: the deleted columns of $\mathbf{X}_n$ has no effects, or equivalently the deleted components of $\boldsymbol{\beta}$ are all zeros, then asymptotically [1].

$$2\left[\ell_n\left(\hat{\boldsymbol{\beta}}\right) - \ell_n\left(\hat{\boldsymbol{\beta}}^-\right)\right] \xrightarrow{D} \chi_k^2,$$

where χ_k^2 is the chi-squared distribution with k -degrees of freedom. For a given nominal level α, let $\chi_d^2(1-\alpha)$ be the $(1-\alpha)$-th upper quantile of the χ_k^2 distribution, if $2\left[\ell_n\left(\hat{\boldsymbol{\beta}}\right) - \ell_n\left(\hat{\boldsymbol{\beta}}^-\right)\right] \geq \chi_d^2(1-\alpha)$, then H_0 is rejected at significance level α, and its not good to delete these columns of $\mathbf{X}_n$; otherwise we accept H_0 and delete these columns of $\mathbf{X}_n$.

There are some other methods to select columns of $\mathbf{X}_n$, such as AIC, BIC and their variants, as in the model selection field. In these methods, the optimal deletion of columns of $\mathbf{X}_n$ corresponds to the best model selection, which maximize the AIC or BIC. These methods are not as solid as the above one, as may sometimes depending on eye inspection to choose the model which maximize the AIC or BIC.

All the above methods require the models under consideration be nested within each other, i.e., one is a sub-model of the other. Another more general model selection criterion is the Minimum Description Length (MDL) criterion, a measure of complexity, developed by Kolmogorov [4], Wallace and Boulton (1968) [16], etc. The Kolmogorov complexity has close relationship with the entropy, it is the output of a Markov information source, normalized by the length of the output. It converges almost surely (as the length of the output goes to infinity) to the entropy of the source. Let $\mathcal{G} = \{g(\cdot,\cdot)\}$ be a finite set of candidate models under consideration, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_j : j = 1,\ldots,h\}$ be the set of parameters of interest. $\boldsymbol{\theta}_i$ may or may not be nested within some other $\boldsymbol{\theta}_j$, or $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ both in $\boldsymbol{\Theta}$ may have the same dimension but with different parametrization. Next consider a fixed density $f(.|\boldsymbol{\theta}_j)$, with parameter $\boldsymbol{\theta}_j$ running through a subset $\Gamma_j \subset R^k$, to emphasize the index of the parameter, we denote the MLE of $\boldsymbol{\theta}_j$ under model $(\cdot|\cdot)$ by $\hat{\boldsymbol{\theta}}_j$ (instead of by $\hat{\boldsymbol{\theta}}_n$ to emphasize the dependence on the sample size), $I(\theta_j)$ the Fisher information for $\boldsymbol{\theta}_j$ under $f(\cdot|\cdot)$, $|I(\theta_j)|$ its determinant, and k_j the dimension of $\boldsymbol{\theta}_j$. Then the MDL criterion (for example, Rissanen [17] and the review paper by Hansen and Yu [5], and references there) chooses $\boldsymbol{\theta}_j$ to minimize

$$-\sum_{i=1}^{n}\log f\left(Y_i|\hat{\boldsymbol{\theta}}_j\right)+\frac{k_j}{2}\log\frac{n}{2\pi}+\log\int_{\Gamma_j}\sqrt{\left|I\left(\boldsymbol{\theta}_j\right)\right|}d\boldsymbol{\theta}_j,\ (j=1,\ldots,h).\quad (3)$$

This method does not require the models be nested, but still require select/delete some whole columns. The other existing methods for variable selection, such as stepwise regression and Lasso, etc., are all for deleting/keeping some whole variables, and does not apply to our problem.

The proposed work

Now come to our question, which is non-standard and we are not aware of a formal method to address this problem. However, we think the following question is of practical meaning. Consider deleting some of the components within fixed k $(k \le d)$ columns of $\mathbf{X}_n$, the deleted proportions for these columns are $\gamma_1,\ldots,\gamma_k (0<\gamma_j<1)$. Denote X_n^- for the remaining covariate matrix, which is $\mathbf{X}_n$ with some entries replaced by 0's, corresponding to the deleted elements. Before the partial deletion, the model is

$$\mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n.$$

After the partial deletion of covariates, the model becomes

$$\mathbf{y}_n = \mathbf{X}_n^-\boldsymbol{\beta}^- + \boldsymbol{\varepsilon}_n.$$

Note that here $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^-$ have the same dimension, as no covariate is completely deleted. $\boldsymbol{\beta}$ is the effects of the original covariates, $\boldsymbol{\beta}^-$ is the effects of the covariates after some possible partial deletion. It is the effects of the effective covariates. As an over simplified example, we have individuals, with five responses $\mathbf{y}_n = (y_1, y_2, y_3, y_4, y_5)$ and covariate vectors $\mathbf{x}_1 = (1.3, 0.2, -1.5)'$, $\mathbf{x}_2 = (-0.1, 0.9, -1.3)'$, $\mathbf{x}_3 = (1.1, 1.4, -0.3)'$, $\mathbf{x}_4 = (0.8, 1.2, -1.7)'$, $\mathbf{x}_5 = (1.0, 2.1, -1.1)'$ and $\mathbf{X}_n = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)$. Then $\boldsymbol{\beta}$ is the effects of the regression of on $\mathbf{X}_n$. If we remove some seemingly insignificant covariate components, for example, let $\mathbf{x}_1^- = (1.3, 0, -1.5)'$, $\mathbf{x}_2^- = (1.1, 1.4, 0)'$, $\mathbf{x}_3^- = (1.1, 1.4, 0)'$, $\mathbf{x}_4^- = (0.8, 1.2, -1.7)'$, $\mathbf{x}_5^- = (1.0, 2.1, -1.1)'$ and $\mathbf{X}_n^- = (\mathbf{x}_1^-, \mathbf{x}_2^-, \mathbf{x}_3^-, \mathbf{x}_4^-, \mathbf{x}_5^-)$. In this case $\boldsymbol{\beta}^-$ is the effects of $\mathbf{y}_n$ regressing on $\mathbf{X}_n$. Thus, though $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^-$ have the same structure, they have different interpretations. The problem can be formulated as testing the hypothesis:

$$H_0: \boldsymbol{\beta} = \boldsymbol{\beta}^- \, vs \, H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}^-$$

If H_0 is accepted, the partial deletion is valid.

Note that different from the standard null hypothesis that some components of the parameters be zeros, the above null hypothesis is not a nested hypothesis, or $\boldsymbol{\beta}^-$ is not a subset of $\boldsymbol{\beta}$, so the existing Wilks' theorem for likelihood ratio statistic does not directly apply here.

Denote $\ell_n^-(\boldsymbol{\beta})$ be the corresponding log-likelihood based on data $(\mathbf{y}_n, \mathbf{X}_n^-)$, and the corresponding MLE as $\hat{\boldsymbol{\beta}}^-$. Since after the partial deletion, $\hat{\boldsymbol{\beta}}^-$ is the MLE of $\boldsymbol{\beta}$ under a constrained log-likelihood, while $\hat{\boldsymbol{\beta}}$ is the MLE under the full likelihood, we have $\ell_n^-(\hat{\boldsymbol{\beta}}^-) \le \ell_n(\hat{\boldsymbol{\beta}})$. Parallel to the log-likelihood ratio statistic for (whole) variable deletion, let, for our case,

$$\Lambda_n = 2\left[\ell_n(\hat{\boldsymbol{\beta}}) - \ell_n^-(\hat{\boldsymbol{\beta}}^-)\right].$$

Let $(j_1,\ldots,j_k)$ be the columns with partial deletions, $C_{j_r} = \{i : x_{j_r,i}$ is deleted $1 \le i \le n\}$ be the index set for the deleted covariates in the j_r -th column $(r=1,\ldots,k)$; $|C_{j_r}|$ be the cardinality of C_{j_r}, thus $\gamma_r = |C_{j_r}|/n(r=1,\ldots,k)$. For different j_r and j_s, C_{j_r} and C_{j_s} may or may not have some common components. We first give the following Proposition, in the simple case in which the index sets C_{j_r} 's are mutually exclusive. Then in Corollary 1 we give the result in more general case in which the index sets C_{j_s} 's are not need to be mutually exclusive.

For given $\mathbf{X}_n$, there are many different ways of partial column deletions, we may use Theorem 1 to test each of these deletions. Given a significance level α, a deletion is valid at level α if $\Lambda_n < \chi^2(1-\alpha)$, where $\chi^2(1-\alpha)$ is the $(1-\alpha)$- th upper quantile of the $\sum_{j=1}^{k}\gamma_j\chi_j^2$ distribution, which can be computed by simulation for given $(\gamma_1,\ldots,\gamma_k)$.

The following Theorem is a generalization of the Wilks Theorem [1]. Deleting some whole columns in X_n corresponds to $\gamma_j = 1$ $(j=1,\ldots,k)$ in the theorem, and then we get the existing Wilks' Theorem.

Theorem 1: *Under H_0, suppose $C_{j_r} \cap C_{j_s} = \phi$, the empty set, for all $1 \le r \ne s \le k$, then we have*

$$\Lambda_n \xrightarrow{D} \sum_{j=1}^{k}\gamma_j\chi_j^2 \quad .$$

where $\chi_1^2,\ldots,\chi_k^2$ are iid chi-squared random variable with 1-degree of freedom.

Note that in Wilks problem the null hypothesis is that, the coefficients corresponding to some variables are zero. The null hypothesis is nested within the alternative; while the null hypothesis in our problem is: The coefficients correspond to some partial variables, and the null hypothesis is not nested within the alternative. So the results of the two methods are not really comparable.

The case the C_{j_r} 's are not mutually exclusive is a bit more complicated. We first re-write the sets C_{j_r} 's such that

$$\cup_{r=1}^{k} C_{j_r} = \cup_{r=1}^{k} \cup_{j_1,\ldots,j_r} D_{j_1,\ldots,j_r},$$

where the $D_{j_1,\ldots,j_r}$'s are mutually exclusive, $D_{j_1},\ldots,D_{j_k}$ are index sets for one column of X_n only; the D_{j_1,j_2} 's are index sets common for columns j_1 and j_2 only; the D_{j_1,j_2,j_3} 's are index sets common for columns j_1, j_2 and j_3 only,.... Generally some of the $D_{j_1,\ldots,j_r}$'s are empty sets. Let $\gamma_{j_1,\ldots,j_r} = |D_{j_1,\ldots,j_r}|$

be the cardinality of $D_{j_1,\ldots,j_r}$ and $\gamma_{j_1,\ldots,j_r} = \left|D_{j_1,\ldots,j_r}\right|/n$ $(r=1,\ldots,k)$.

By examining the proof of Theorem 1, we get the following corollary which gives the result in the more general case.

Corollary 1: *Under H_0, we have*

$$\Lambda_n = 2\left[\ell_n\left(\hat{\boldsymbol{\beta}}\right)-\ell_n^-\left(\hat{\boldsymbol{\beta}}^-\right)\right]\xrightarrow{D}\sum_{r=1}^{k}\sum_{j_1,\ldots,j_r}\gamma_{j_1,\ldots,j_r}\chi^2_{j_1,\ldots,j_r},$$

where the $\chi^2_{j_1,\ldots,j_r}$'s are all independent chi-squared random variables with r-degrees of freedom $(r=1,\ldots,k)$.

Below we give two examples to illustrate the usage of Proposition.

Example 1: $n=1000$, $d=5$, $k=3$. Columns $(1,2,4)$ has some partial deletions with $C_1=\{201,202,\ldots.,299,300\}$, $C_2=\{351,352,\ldots,549,550\}$, $C_3=\{601,602,\ldots,849,850\}$, the C_j's have no overlap; $\gamma_1=1/10$, $\gamma_2=1/5$, $\gamma_3=1/4$. So by the Proposition, under H_0 we have

$$2\left[\ell_n\left(\hat{\beta}\right)-\ell_n^-\left(\hat{\beta}^-\right)\right]\xrightarrow{D}\frac{1}{10}\chi_1^2+\frac{1}{5}\chi_2^2+\frac{1}{4}\chi_3^2,$$

where all the chi-squared random variables are independent, each has 1 degree of freedom.

Example 2: $n=1000$, $d=5$, $k=3$. Columns $(1,2,4)$ has some partial deletions with $C_1=\{101,102,\ldots.,299,300;651,652,\ldots,749,750\}$, $C_2=\{201,202,\ldots,349,350\}$, $C_3=\{251,252,\ldots,299,300;701,702,\ldots,799,800\}$. In this case the C_j's have overlaps, the Proposition can not be used directly, so we use the Corollary. Then $D_1=\{101,102,\ldots,199,200\}$, $D_2=\{301,302,\ldots,349,350\}$, $D_3=\{701,702,\ldots,799,800\}$, $_{1,2}$ $\{201,202,\ldots,249,250\}$, $D_{1,3}=\{701,702,\ldots,749,750\}$, $D_{2,3}=\phi$, $D_{1,2,3}=\{251,252,\ldots,299,300\}$; $\gamma_1=1/5$, $\gamma_2=1/20$, $\gamma_3=1/10$, $\gamma_{1,2}=1/20$, $\gamma_{1,3}=1/20$, $\gamma_{2,3}=0$, $\gamma_{1,2,3}=1/20$. So by the Corollary, under H_0 we have

$$2\left[\ell_n\left(\hat{\boldsymbol{\beta}}\right)-\ell_n^-\left(\hat{\boldsymbol{\beta}}^-\right)\right]\xrightarrow{D}\frac{1}{5}\chi_1^2+\frac{1}{20}\chi_2^2+\frac{1}{10}\chi_3^2+\frac{1}{20}\chi_{1,2}^2+\frac{1}{20}\chi_{1,3}^2+\frac{1}{20}\chi_{1,2,3}^2,$$

where all the chi-squared random variables are independent, with χ_1^2, χ_2^2 and χ_3^2 are each of 1 degree of freedom, $\chi_{1,2}^2$ and $\chi_{1,3}^2$ are each of 2-degrees of freedom, and $\chi_{1,2,3}^2$ is of 3-degrees of freedom.

Next, we discuss the consistency of estimation of $\hat{\boldsymbol{\beta}}^-$ under the null hypothesis H_0. Let $\mathbf{x}^- = \mathbf{x}_r^-$ with probability γ_r $(r=0,1,\ldots,k)$, where x_r^- is an i.i.d. copy of the $x_{i,r}^-$'s, whose components with index in C_{jr}, in particular C_{j0} is the index set for those covariates without partial deletion.

Theorem 2: *Under conditions of Theorem 1,*

$$\hat{\boldsymbol{\beta}}^- \to \boldsymbol{\beta}_0\,(a.s.).$$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^- - \boldsymbol{\beta}_0\right)\xrightarrow{D}N(0,\Omega),$$

where

$$\Omega = \mathrm{E}_{\beta_0}\left[\dot{\ell}(\boldsymbol{\beta}_0)\dot{\ell}'(\boldsymbol{\beta}_0)\right] = E\left[\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)'\right]\int\frac{\dot{f}^2(\epsilon)}{f(\epsilon)}d\epsilon.$$

To extend the results of Theorem 2 to the general case, we need the following more notations. Let $x_{(j_1,\ldots,j_k)}$ be an i.i.d. copy of data in the set $D_{j_1,\ldots,j_k}$. Let $\mathbf{x}^- = \mathbf{x}^-_{j_1,\ldots,j_r}$ with probability $\gamma_{j_1,\ldots,j_r}$ $(r=0,1,\ldots,k)$, where $\mathbf{x}^-_{j_1,\ldots,j_r}$ is an i.i.d. copy of the $x^-_{i,j_1,\ldots,j_r}$'s, whose components with index in $C_{j_1,\ldots,j_r}$.

Corollary 2: *Under conditions of Corollary 1, results of Theorem 2 hold with $\mathbf{x}^-$ given above.*

Computationally $E\left[\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)'\right]$ is well approximated by

$$E\left[\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)\left(\mathbf{x}^- - \boldsymbol{\mu}^-\right)'\right]\approx\sum_{r=0}^{k}\frac{\left|D_{j_1,\ldots,j_r}\right|}{n}\frac{1}{\left|D_{j_1,\ldots,j_r}\right|}\sum_{(i,j)\in D_{j_1,\ldots,j_r}}\left(\mathbf{x}^-_{i,j}-\hat{\boldsymbol{\mu}}^-_{j_1,\ldots,j_r}\right)\left(\mathbf{x}^-_{i,j}-\hat{\boldsymbol{\mu}}^-_{j_1,\ldots,j_r}\right)',$$

where the notation $_{(i,j)\in D_{j\ ,\ldots,j}}$ means summation over those $\mathbf{x}^-_{i,j}$'s with deletion index in $D_{j_1,\ldots,j_r}$, and $\left(\hat{\mu}^-_{j_1,\ldots,j_r}\right)=\frac{1}{\left|D_{j_1,\ldots,j_r}\right|}\Sigma_{(i,j)\in D_{j_1,\ldots,j_r}}x^-_{i,j}$.

Simulation Study and Application

Simulation study

We illustrate the proposed method with two examples, Examples 3 and 4 below. The former rejects the null hypothesis H_0 while the latter accepts. In each case we simulate $n=1000$ i.i.d. data with response y_i and with covariates $\mathbf{x}_i=(x_{i1},x_{i2},x_{i3},x_{i4},x_{i5})(i=1,\ldots,n)$. We first generate the covariates, sample the $\mathbf{x}_i$'s from the 5-dimensional normal distribution with mean vector $\boldsymbol{\mu}=(3.1,1.8,-0.5,0.7,1.5)'$ and a given covariance matrix Γ.

Then we generate the response data, which, given the covariates. The y_i's are generated as

$$y_i=\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}_0+\epsilon_i,(i=1,\ldots,n)$$

$\boldsymbol{\beta}_0=(0.42,0.11,0.65,0.83,0.72)'$, the ϵ_i's are i.i.d. $N(0,1)$.

Hypothesis test is conducted to examine if the partial deletion is valid or not. Significant level is set as $\alpha=0.05$. The experiment repeated 1000 times, $Prop$ represents the proportion $\Lambda_n>Q(1-\alpha)$, where $Q(1-\alpha)$ is the $(1-\alpha)$ -th upper quantile of the distribution $\sum_{j=1}^{k}\gamma_j\chi_j^2$ given in Theorem 1, computed via simulation.

Example 3: In this example, five data sets are generated according to the mentioned method, with five different values of $\boldsymbol{\beta}_0$. We are interested to know whether covariates with $|x_{ij}|<\frac{1}{10}$ can be deleted. Five data set with different $\boldsymbol{\beta}_0$ values are simulated. The proportion $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_k)$ of x_{ij}'s with $|x_{ij}|<\frac{1}{10}$ are shown for each data set, the results are shown in Table 1. The five rows in Table 1 are the results for the five data sets. For each data, the parameter $\boldsymbol{\beta}$ is estimated, a and test is conducted using the given $\boldsymbol{\gamma}$, the Λ_n is computed, $Q(1-\alpha)$ is given, and the corresponding p-value is provided. Note that for our problem, a p-value smaller than α

Table 1: The simulation result of γ, Λ_n, $Q(1-\alpha)$ and p-value according to $\boldsymbol{\beta}_0$.

No.	$\boldsymbol{\beta}_0$	γ	Λ_n	$Q(1-\alpha)$	p-value
1	(0.42, 0.11, 0.65, 0.83, 0.72)	(0.008, 0.022, 0.043, 0.037, 0.030)	14492.91	4.5767	0.006
2	(0.12, 0.85, 0.44, 0.73, 0.62)	(0.004, 0.020, 0.041, 0.040, 0.020)	13010.97	4.5748	0.016
3	(0.59, 0.27, 0.73, 0.35, 0.66)	(0.008, 0.032, 0.031, 0.048, 0.025)	13505.90	4.5786	0.000
4	(0.21, 0.45, 0.78, 0.56, 0.63)	(0.007, 0.022, 0.039, 0.053, 0.033)	12487.58	4.5281	0.005
5	(0.77, 0.51, 0.48, 0.89, 0.32)	(0.01, 0.022, 0.042, 0.045, 0.026)	15437.66	4.5317	0.000

Table 2: The simulation result of γ, Λ_n, $Q(1-\alpha)$ and p-value according to $\boldsymbol{\beta}_0$.

No.	$\boldsymbol{\beta}_0$	γ	Λ_n	$Q(1-\alpha)$	p-value
1	(0.42, 0.11, 0.65, 0.83, 0.72)	(0.1, 0.1, 0.1)	1.0146	4.6034	0.998
2	(0.12, 0.85, 0.44, 0.73, 0.62)	(0.1, 0.1, 0.1)	0.3576	4.6414	0.977
3	(0.59, 0.27, 0.73, 0.35, 0.66)	(0.1, 0.1, 0.1)	3.2480	4.6756	0.965
4	(0.21, 0.45, 0.78, 0.56, 0.63)	(0.1, 0.1, 0.1)	3.3003	4.6306	0.972
5	(0.77, 0.51, 0.48, 0.89, 0.32)	(0.1, 0.1, 0.1)	3.3531	4.6326	0.955

means a significant value of Λ_n, or significant difference between the regression coefficients of original covariates and those of the covariates after partial deletion, which implies in turn that the null hypothesis should be rejected, or the partial deletion should not be conducted (Table 1).

We see that the p-values of rejecting H_0, are all smaller than 0.05 in the five set of $\boldsymbol{\beta}_0$. This suggests that covariates with $|x_{ij}| < \frac{1}{10}$ should not be deleted at significance level $\alpha = 0.05$.

Example 4: In this example, the original $\mathbf{X}$ as in Example 3, but now we replace the entries in first 100 rows and first three columns by noise $\in$, where $\epsilon\ N\left(0, \frac{1}{9}\right)$. The delete proportion $\gamma = (0.1, 0.1, 0.1)$ is fixed with x_{ij} 's having absolute values smaller than the lower 0.1 percent being deleted. We are interested to see in this case whether these noises can be deleted, i.e. H_0 can be rejected or not. The results are shown in the following (Table 2).

We see that the p-values of rejecting H_0 are all greater than 0.95 for the five sets of $\boldsymbol{\beta_0}$. It suggests that the data provided strong evidence to conclude that the deleted values are noises and they are not necessary to the data set at 0.05 significance level.

Application to real data problem.

We analyze a data set from the Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism, which is obtained from The National Institutes of Health (NIH). (For detailed description and data link, https://www.ncbi.nlm.nih.gov/pubmed/2515723). It is a multi-center, placebo-controlled clinical trial that aimed to determine a treatment for early Parkinson's disease patient to prolong their time requiring levodopa therapy. The number of patients enrolled was 800. The selected object were untreated patients with Parkinson's disease (stage I or II) for less than five years and met other eligible criteria. They were randomly assigned according to a two-by-two factorial design to one of four treatment groups: 1) Placebo 2) Active tocopherol 3) Active deprenyl 4) Active deprenyl and tocopherol. The observation continued for 14 ± 6 months and reevaluated every 3 months. At each visit, Unified Parkinson's Disease Rating Scale (UPDRS) including its motor, mental and activities of daily living components were evaluated. Statistical analysis result was based on 800 subjects. The result revealed that no beneficial effect of tocopherol. Deprenyl effect was found significantly prolong the time requiring levodopa therapy which reduced the risk of disability by 50 percent according to the measurement of UPDRS.

Our goal is to examine whether some of the covariates can be partially deleted. If traditional variable selection methods are used, such as stepwise or Lasso, it will end up with some covariate(s) been removed wholly from the analysis. This is not very reasonable, since some of the removed covariates may be partially effective, removing all their values may yield miss-leading results, or at least cost information loss. We use the proposed method to examine three of the response variables, PDRS, TREMOR and PIGD, and three covariates, Age, Motor and ADL for all these responses. The deleted covariates are the ones with values below the γ -th data quantile, with $\gamma = 0.01, 0.02, 0.03$ and 0.05. We examine each response and covariate one by one. The results are shown in Table 3, Table 4 and Table 5 below.

In Table 3, response TREMOR is examined. For covariable Age, the likelihood ratio Λ_n is larger than the cut-off point $Q(1-\alpha)$ at all the deletion proportions, it suggests that for Age, no partial deletions with these proportions should be removed. For covariable Motor, Λ_n is smaller than the cutoff point $Q(1-\alpha)$ at the 0.01 proportion, this covariable can be partially deleted at this proportion. In other words, the covariate Motor with values smaller than 1%-th of its quantile have no impact on the analysis, or can be treated as noise and

Table 3: Response TREMOR: Λ_n values and estimated regression coefficients.

	Age		**Motor**		**ADL**	
Estimated coefficient	**0.0240456**		**0.1801616**		**0.00451205**	
Delete proportion	Λ_n	$Q(1-\alpha)$	Λ_n	$(\quad)$	Λ_n	$Q(1-\alpha)$
0.01	11.5171	0.0593	0.35929	0.8787	0.00425	0.1897
0.03	20.0485	0.1245	6.2598	0.6924	0.00425	0.1861
0.05	14.0114	0.2937	8.7075	0.9034	0.00425	0.1496
0.1					0.0238	0.3841

Table 4: Response PIGD: Λ_n values and estimated regression coefficients.

	Age		**Motor**		**ADL**	
Estimated coefficient	**-0.0049032**		**0.02467423**		**0.2084862**	
Delete proportion	Λ_n	$Q(1-\alpha)$	Λ_n	$Q(1-\alpha)$	Λ_n	$Q(1-\alpha)$
0.02	0.4956	0.0849	0.0031	0.0972	3.5607	0.2210
0.03	1.3908	0.1306	0.0166	0.1513	3.5607	0.2256
0.05	0.4816	0.2596	1.2607	0.2188	3.6607	0.1925

Table 5: Response PDRS: Λ_n values and estimated regression coefficients.

	Age		**Motor**		**ADL**	
Estimated coefficient	**1.563389**		**0.0476914**		**-1.139864**	
Delete proportion	Λ_n	$Q(1-\alpha)$	Λ_n	$Q(1-\alpha)$	Λ_n	$Q(1-\alpha)$
0.01	1.0073	0.0497	80.7411	0.0944	6.3392	0.0453
0.02	1.0475	0.0669	142.3528	0.0841	57.5051	0.2216
0.03	0.8609	0.1486	321.5332	0.1906	57.5051	0.2111
0.05	0.8650	0.2379	397.6481	0.2227	57.5051	0.2199

should be removed from the analysis. For covariable ADL, with deletion proportions $0.01\text{-}0.1$, the likelihood ratio Λ_n is smaller than $Q(1-\alpha)$ which suggest that the lower percentage of $1\%\text{-}10\%$ of this covariate have no impact on the analysis and should be deleted. After removing the corresponding proportions of Motor and ADL, the model is re-fitted to get the parameter estimates shown there. These estimates have better meaning than the ones based on the whole covariates data, since now the noise values of covariates are removed, and only the effective covariates entered the analysis. However, if traditional variable methods are used, such as stepwise regression or Lasso, it may end up with the whole covariate Motor, ADL, or both to be removed, and leads loss of information or even miss-leading results.

In Table 4, response PIGD is investigated. For covariable age, Λ_n is larger than the cut-off point $Q(1-\alpha)$ at the 0.02, 0.03 and 0.05 proportions, suggests that partial deletion with these proportions are not appropriate. For covariate Motor, Λ_n is smaller than cut-off point $Q(1-\alpha)$ at the deletion proportions 0.02 and 0.03, suggests that the lower percentage of $2\text{-}3\%$ should be deleted from the analysis. For the variable ADL, Λ_n is larger than the cut-off point $Q(1-\alpha)$ at the delete proportions 0.02, 0.03 and 0.05, hence partial deletion at these proportions are not valid. After deleting 3% of the smallest values of Motor, the model is re-fit to get the parameter estimates shown in the Table. The new estimates are more meaning full since the on-effective values of covariate Motor are removed from the analysis.

In Table 5, the response is PDRS. The likelihood ratios Λ_n of Age, Motor and ADL all are larger than $\chi^2(1-\alpha)$ at the deletion proportions of 0.01, 0.02, 0.03 and 0.05. Thus the null hypothesis are rejected at all these proportions, or no deletion is valid at these proportions, and the analysis should be based on the original full data, with the parameter estimates shown in the Table (Table 3, Table 4, and Table 5).

Note that the coefficient for Age is insignificant, and hence the corresponding Λ_n values with deleted proportions are senseless.

Concluding Remarks

We proposed a method for partial variable deletion, in which only some proportion(s) of covariate(s) values are to be deleted. This is in contrast to the existing methods either select or delete the entire variable(s). Thus this method is new and is a generalization of the existing variable selection. The question is motivated from practical problems. It can used to find the effective ranges of the covariates, or to remove possible noises in the covariates, and thus the corresponding estimated effects are more interpretable. The proposed test statistic is a generalization of the Wilks likelihood ratio statistic, the asymptotic distribution of the proposed statistic is generally a chi-squared mixture distribution, the corresponding cut-off point can be computed by simulation. Simulation studies are conducted to evaluate the performance of the method, and it is applied to analyze a real Parkinson disease data as illustration. A drawback of the current version of the method is that it needs to specify the proportions of possible deletions for the variables, this makes the optimal proportions are not easy to find. In our next step research we will try to implement an algorithm which finds the optimal proportions automatically, and more easy to use. As suggested from a reviewer, simulation studies should be performed for statistical significance test between the proposed method and existing variable selection method(s) to address the contribution of the proposed method. This will be potential for our future research work (Appendix).

Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9: 60-62.
2. Akaike H (1974) A new look at the statistical identification model. IEEE Transaction on Automatic Control 19: 716-723.
3. Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6: 461-464.
4. Kolmogorov A (1963) On tables of random numbers. Sankhya 25: 369-375.
5. Hansen M, Yu B (2001) Model selection and the principle of minimum description length. Journal of American Statistical Association 96: 746-774.
6. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58: 267-288.
7. May RJ, Maier HR, Dandy GC, Fernando TG (2008) Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modelling and Software 23: 1312-1326.
8. Mehmood T, Liland KH, Snipen L, Saebo S (2012) A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems 118: 62-69.
9. Wang L, Liu X, Linag H, Carroll R (2011) Estimation and variable selection for generalized additive partial linear models. Annals of Statistics 39: 1827-1851.
10. Liu X, Wang L, Liang H (2011) Estimation and variable selection for semiparametric additive partial linear models. Stat Sin 21: 1225-1248.
11. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Statistics in Medicine 16: 385-395.
12. Fan J, Li R (2001) Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96: 1348-1360.
13. Fan J, Li R (2002) Variable selection for Cox's proportional hazards model and frailty model. Annals of Statistics 30: 74-99.
14. Wang HX, Leineweber C, Kirkeeide R, Svane B, Schenck-Gustafsson K, et al. (2007) Psychosocial stress and atherosclerosis: family and work stress accelerate progression of coronary disease in women. The Stockholm Female Coronary Angiography Study. J Intern Med 261: 245-254.
15. Shara NM, Wang H, Valaitis E, Pehlivanova M, Carter EA, et al. (2011) Comparison of estimated glomerular filtration rates and albuminuria in predicting risk of coronary heart disease in a population with high prevalence of diabetes mellitus and renal disease. Am J Cardiol 107: 399-405.
16. Wallace CS, Boulton DM (1968) An information measure for classification. Computer Journal 11: 185-194.
17. Rissanen J (1996) Fisher information and stochastic complexity. IEEE Transactions on Information Theory 42: 40-47.
18. Stat 701 (2002) Proof of Wilks' Theorem on LRT.
19. Bickel PJ, Klaassen CA, Ritov Y, Wellner JA (1993) Efficient and Adaptive Estimation for Semiparametric Models. The Indian Journal of Statistics 62: 157-160.

Appendix

Proof of Theorem 1: Let $\boldsymbol{\beta}_0$ be the true parameter value generating the observed data, $\dot{\ell}_n(\boldsymbol{\beta})=\partial\ell_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\ddot{\ell}_n(\boldsymbol{\beta})=\partial^2\ell_n(\boldsymbol{\beta})/[\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}']$. Since $\hat{\boldsymbol{\beta}}$ is the MLE, $\dot{\ell}_n(\boldsymbol{\beta})=0$, so we have

$$-\dot{\ell}_n(\boldsymbol{\beta}_0)=\dot{\ell}_n(\hat{\boldsymbol{\beta}})-\dot{\ell}_n(\boldsymbol{\beta}_0)=\ddot{\ell}_n(\boldsymbol{\beta}_n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_n$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Since $\hat{\boldsymbol{\beta}}\to\boldsymbol{\beta}_0$ (a.s.), we have $\boldsymbol{\beta}_n\to\boldsymbol{\beta}_0$ (a.s.). We get

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=\sqrt{n}\left(-n^{-1}\ddot{\ell}_n(\boldsymbol{\beta})\right)^{-1}n^{-1}\dot{\ell}_n(\boldsymbol{\beta}_0)$$

Note that $-n^{-1}\ddot{\ell}_n(\boldsymbol{\beta})\xrightarrow{P}I(\boldsymbol{\beta}_0)$, and that $\dot{\ell}_n(\boldsymbol{\beta}_0)=\sum_{i=1}^{n}v_i$, $v_i=\left[\partial f\left(y_i-\frac{\mathbf{x}_i\boldsymbol{\beta}_0}{\partial\boldsymbol{\beta}}\right)\right]/f(y_i-\mathbf{x}_i\boldsymbol{\beta}_0)$. The v_i 's are iid with $E(v_i)=0$ and $Var(v_i)=E(v_iv_i^T)=I(\boldsymbol{\beta}_0)$, consequently,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{D}N(0,I^{-1}(\boldsymbol{\beta}_0)).\ (A.1)$$

To simplify notation, assume the columns with partial deletions are $(1,\ldots,k)$. Denote $(n_1,\ldots,n_k)$ be the numbers of x_{ij} 's deleted from columns $(1,\ldots,k)$ of $\mathbf{X}_n$, so $(n_1,\ldots,n_k)/n=(\gamma_1,\ldots,\gamma_k)$, denote $n_0=n-(n_1+\cdots+n_k)$ and $\gamma_0=n_0/n$. Let $\ell_n^-(\boldsymbol{\beta}_0)$ be the likelihood with proportions $(\gamma_1,\ldots,\gamma_k)$, be deleted from columns $(1,\ldots,k)$ in $\mathbf{X}_n$ and denote $\dot{\ell}_n^-(\boldsymbol{\beta})$ and $\ddot{\ell}_n^-(\boldsymbol{\beta})$ be the partial derivatives accordingly. Write

$$\ell_n^-(\boldsymbol{\beta})=\ell_{0,n_0}(\boldsymbol{\beta})+\sum_{j=1}^{k}\ell_{-j,n_j}(\boldsymbol{\beta}),$$

where $\ell_{0,n_0}(\boldsymbol{\beta})$ is the part of log-likelihood for the n_0 data without covariate deletion, and $\ell_{-j,n_j}(\boldsymbol{\beta})$ is that for all the n_j data with the j -th covariate deleted. Denote $\ddot{\ell}_n(\boldsymbol{\beta})=\ddot{\ell}_{0,n_0}(\boldsymbol{\beta})+\sum_{j=1}^{k}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta})$ accordingly.

For the log-likelihood without data deletion we make the similar decomposition as

$$\ell_n(\boldsymbol{\beta})=\ell_{0,n_0}(\boldsymbol{\beta})+\sum_{j=1}^{k}\ell_{j,n_j}(\boldsymbol{\beta}),$$

where $\ell_{j,n_j}(\boldsymbol{\beta})$ is the part of log-likelihood using data from the same individuals as those in $\ell_n(\boldsymbol{\beta})$, but without covariate deletion, and denote $\ddot{\ell}_n(\boldsymbol{\beta})=\ddot{\ell}_{0,n_0}(\boldsymbol{\beta})+\sum_{j=1}^{k}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta})$ accordingly.

Note that the same term $\ell_{0,n_0}(\boldsymbol{\beta})$ appears in both the decompositions of $\ell_n(\boldsymbol{\beta})$ and $\ell_n^-(\boldsymbol{\beta})$, the same term $\ddot{\ell}_{0,n_0}(\boldsymbol{\beta})$ appears in both the decompositions of $\ddot{\ell}_n(\boldsymbol{\beta})$ and $\ddot{\ell}_n^-(\boldsymbol{\beta})$, and that for $j=1,\ldots,k$, $\ell_{j,n_j}(\boldsymbol{\beta})$ is different from $\ell_{-j,n_j}(\boldsymbol{\beta})$ in that there is no deletion in $\ell_{j,n_j}(\boldsymbol{\beta})$, although both partial log-likelihoods use data from the same set.

We have

$$\ell_n(\hat{\boldsymbol{\beta}})=\ell_n(\boldsymbol{\beta}_0)-\dot{\ell}_n(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}})-\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\ddot{\ell}_n(\boldsymbol{\beta}_n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$$

$$=\ell_n(\boldsymbol{\beta}_0)-\frac{1}{2}\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\ddot{\ell}_{0,n_0}(\boldsymbol{\beta}_n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\sum_{j=1}^{k}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right).$$

Also, let $Z = (Z_1,...,Z_d)$ with Z_j 's iid $N(0,1)$, and note $-n_j^{-1}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)\xrightarrow{P} I(\boldsymbol{\beta}_0)$ for $(j=0,\ldots,k)$, so we have

$$-(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=-\gamma_j\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'n_j^{-1}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$$

$$\xrightarrow{D}\gamma_j Z'Z.$$

So we get

$$\ell_n(\hat{\boldsymbol{\beta}})=\ell_n(\boldsymbol{\beta}_0)+\frac{1}{2}\sum_{j=0}^{k}\gamma_j Z'Z+o_p(1).$$

Similarly to (A.1) we have, with $\boldsymbol{\beta}_n^-$ lies between $\hat{\boldsymbol{\beta}}^-$ and $\hat{\boldsymbol{\beta}}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)=\sqrt{n}\left(-n^{-1}\ddot{\ell}_n^-(\boldsymbol{\beta}_n^-)\right)^{-1}n^{-1}\ell_n^-(\dot{\boldsymbol{\beta}}_0).$$

Consequently,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)\xrightarrow{D}N(0,I_-^{-1}(\boldsymbol{\beta}_0)) \qquad \text{(A.2)}$$

Where $I_-(\boldsymbol{\beta}_0)=E_{H_0}(v_i^- v_i^-{}')$, and $v_i^-{}'$ is v_i with the j -th covariate being removed with probability $\gamma_j\,(j=1,\ldots,k)$.

Also

$$\ell_n^-(\hat{\boldsymbol{\beta}}^-)=\ell_n^-(\boldsymbol{\beta}_0)-\frac{1}{2}\left((\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)'\ddot{\ell}_{0,n_0}(\boldsymbol{\beta}_n^-)(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)+\sum_{j=1}^{k}(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)'\ddot{\ell}_{-j,n_j}(\boldsymbol{\beta}_n^-)(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)\right).$$

and we have

$$-(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)'\ddot{\ell}_{0,n_0}(\boldsymbol{\beta}_n^-)(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)=-\gamma_0\sqrt{n}(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)'n_0^{-1}\ddot{\ell}_{0,n_0}(\boldsymbol{\beta}_n^-)\sqrt{n}(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)$$

$$\xrightarrow{D}\gamma_0 Z'Z$$

Note that $\ddot{\ell}_{-j,n_j}(\boldsymbol{\beta}_n^-)$ is a $d\times d$ matrix with the j -th row and j -th column be zeros, so

$$(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)'\ddot{\ell}_{-j,n_j}(\boldsymbol{\beta}_n^-)(\hat{\boldsymbol{\beta}}^- -\boldsymbol{\beta}_0)=(\hat{\boldsymbol{\beta}}_{-j}-\boldsymbol{\beta}_{0,-j})'\ddot{\tilde{\ell}}_{-j,n_j}(\boldsymbol{\beta}_n^-)(\hat{\boldsymbol{\beta}}_{-j}-\boldsymbol{\beta}_{0,-j})$$

where $\hat{\beta}_{-j}$ is the $(d-1)$ -dimensional vector with the j -th element removed from $\hat{\boldsymbol{\beta}}^-$, $\boldsymbol{\beta}_{0,-j}$ is the $(d-1)$ -dimensional vector with the j -th element removed from $\boldsymbol{\beta}_0$, and $\ddot{\tilde{\ell}}_{-j,n_j}(\boldsymbol{\beta}_n^-)$ is the $(d-1)\times(d-1)$ matrix with the j -th column and j -th row removed from $\ddot{\ell}_{-j,n_j}(\boldsymbol{\beta}_n)$.

Since under H_0, $l_n(\boldsymbol{\beta}_0)=l_n^-(\boldsymbol{\beta}_0)$, now we have, under H_0,

$$2\left[\ell_n(\hat{\boldsymbol{\beta}})-\ell_n^-(\hat{\boldsymbol{\beta}}^-)\right]=\sum_{j=0}^{k}\gamma_j\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\left(-n_j^{-1}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)\right)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\sqrt{n}(\hat{\boldsymbol{\beta}}_{-j}-\boldsymbol{\beta}_{0,-j})^T\left(-n_0^{-1}\ddot{\tilde{\ell}}_{-j,n_j}(\boldsymbol{\beta}_n^-)\right)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right)$$

$$=\sum_{\mathrm{j=1}}^{\mathrm{k}}\gamma_j\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\left(-n_j^{-1}\ddot{\ell}_{j,n_j}(\boldsymbol{\beta}_n)\right)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\sqrt{n}(\hat{\boldsymbol{\beta}}_{-j}-\boldsymbol{\beta}_{0,-j})^T\left(-n_0^{-1}\ddot{\tilde{\ell}}_{-j,n_j}(\boldsymbol{\beta}_n^-)\right)\sqrt{n}(\hat{\boldsymbol{\beta}}_{-\boldsymbol{j}}-\boldsymbol{\beta}_{0,-j})\right)+o_p(1),$$

note that the first term in the above bracket is asymptotically a χ^2 random variable with d -degrees of freedom, while the second term is asymptotically a χ^2 random variable with $(d-1)$ -degrees of freedom. As in the proof of Wilks' Theorem (or some more recent proofs, such as in Stat701, 2002 [18]), for each j we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)^{'}\left(-n_j^{-1}\ddot{\ell}_{j,n_j}\left(\boldsymbol{\beta}_n\right)\right)\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)-\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{-j}-\boldsymbol{\beta}_{0,-j}\right)^{T}\left(-n_0^{-1}\ddot{\tilde{\ell}}_{-j,n_j}\left(\boldsymbol{\beta}_n^{-}\right)\right)\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{-\boldsymbol{j}}-\boldsymbol{\beta}_{0,-j}\right)\xrightarrow{D}\chi_j^2$$

where χ_j^2 is a chi-squared distribution with 1-degree of freedom, and the χ_j^2 's are independent for different j , hence we get

$$2\left[\ell_n\left(\hat{\boldsymbol{\beta}}\right)-\ell_n^{-}\left(\hat{\boldsymbol{\beta}}^{-}\right)\right]\xrightarrow{D}\sum_{j=1}^{k}\gamma_j\chi_j^2.$$

Proof of Theorem 2: i) Is from standard argument for the consistency of MLE.

ii) After deleting the irrelevant covariates, the model is

$$y_i=\mathbf{x}_i^{-}\boldsymbol{\beta}+\varepsilon_i,\quad \varepsilon_i\sim f\left(.\right),$$

where the $\mathbf{x}_i^{-}$ are i.i.d. $\mathbf{x}^{-}$, and $\mathbf{x}^{-}=\mathbf{x}_r^{-}$ with probability $\gamma_r\left(r=0,1,\ldots,k\right)$, where $\mathbf{x}_r^{-}$ is an i.i.d. copy of the $x_{i,r}^{-}$'s, whose components with index in C_{jr} , in particular C_{j0} is the index set for those covariates without partial deletion. The log-likelihood is

$$\ell_n^{-}\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\log f\left(y_i-\mathbf{x}_i\boldsymbol{\beta}\right).$$

By the standard result on regression parameter estimation (eg; Proposition 4.3.1 D and Example 4.3.1 in Bickel, et al.) [19], the efficient score for $\boldsymbol{\beta}$ based on $\ell^{-}\left(\boldsymbol{\beta}\right)$ is

$$\dot{\ell}\left(\boldsymbol{\beta}\right)=\partial\log\frac{f\left(y-\mathbf{x}^{-}\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}=\frac{\dot{f}\left(y-\mathbf{x}^{-}\boldsymbol{\beta}\right)}{f\left(y-\mathbf{x}^{-}\boldsymbol{\beta}\right)}\left(\mathbf{x}^{-}-\boldsymbol{\mu}^{-}\right)$$

Where $\boldsymbol{\mu}_j^{-}=E\left(\mathbf{x}^{-}\right)$. Under the common assumption that $\epsilon=y-\mathbf{x}^{-}\boldsymbol{\beta}_0$ is independent of $\mathbf{x}^{-}$ (or just conditioning on $\mathbf{x}^{-}$), it follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{-}-\boldsymbol{\beta}_0\right)\xrightarrow{D}N\left(0,\Omega\right),$$

Where

$$\Omega=E_{\boldsymbol{\beta}_0}\left[\dot{\ell}(\boldsymbol{\beta}_0)\hat{\ell}^{'}\left(\boldsymbol{\beta}_0\right)\right]=E\left[\left(\mathbf{x}^{-}-\boldsymbol{\mu}^{-}\right)\left(\mathbf{x}^{-}-\boldsymbol{\mu}^{-}\right)^{'}\right]\int\frac{\dot{f}^2\left(\varepsilon\right)}{f\left(\varepsilon\right)}d\varepsilon.$$