SMOOTHLY CLIPPED ABSOLUTE DEVIATION IN ANALYSIS OF SURVEY DATA

Chen Xu[1], Jiahua Chen[1], Harold Mantel[2]

ABSTRACT

The penalized likelihood approach with smoothly clipped absolute deviation (SCAD) penalty has been demonstrated to be an attractive technique for variable selection. Under certain regularity conditions such as the independence assumption on the data structure, the SCAD has been shown to be able to consistently select the important variables and efficiently estimate the coefficients. In this paper, we study the use of SCAD in the analysis of complex survey data, where the structure of data is intrinsically dependent. Under a two-stage sampling framework, we prove that the sample-based SCAD estimator enjoys desired consistency properties in both variable selection and parameter estimation. The approach is further illustrated using data from the Hypertension component of the 2009 Survey on Living with Chronic Diseases in Canada.

KEY WORDS: Consistency, Penalized Likelihood, Regularization, SCAD, Super-Population Model, Variable Selection.

RÉSUMÉ

L'approche de vraisemblance pénalisée ayant un écart absolu avec coupure lisse (smoothly clipped absolute deviation, SCAD) a montré qu'elle était une technique attrayante pour la sélection de variables. Sous certaines conditions de régularité, comme l'hypothèse d'indépendance sur la structure des données, le SCAD s'est révélé capable de sélectionner de façon constante les variables importantes et d'estimer efficacement les coefficients. Dans cet article, nous étudierons l'utilisation du SCAD dans l'analyse de données d'enquête complexes où la structure des données est intrinsèquement dépendante. Dans le cadre d'un échantillonnage à deux degrés, nous prouvons que l'estimateur SCAD fondé sur l'échantillon tire parti des propriétés de constance désirées tant dans la sélection des variables que dans l'estimation des paramètres. L'approche est illustrée ensuite par l'utilisation de données provenant de la composante hypertension de l'Enquête sur les personnes ayant une maladie chronique au Canada (EPMCC).

MOTS-CLÉS : Cohérence; modèle de super-population; régularisation; SCAD; sélection des variables; vraisemblance pénalisée.

1. INTRODUCTION

Variable selection is fundamental in statistical modeling process. Traditional selection procedures, such as the best subset selection and the stepwise regression, are practically useful when there are not too many covariates under consideration. However, for the cases where the number of covariates is large, these methods can be computationally expensive or unstable in the selecting process (see, e.g., Breiman 1996). Instead, the penalized likelihood methods (PLM) are now being used as computationally feasible alternatives for the variable selection. Such examples include the bridge regression (Frank and Friedman 1993), the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) and the smoothly clipped absolute deviation (SCAD; Fan and Li 2001). These approaches exclude variables from the model by estimating their coefficients at zero, and shrink other coefficients accordingly. Compared with the traditional methods, the PLM requires lower computational cost and provides more stable selection results. These advantages merit the PLM to be an attractive approach for the variable selection. Among these approaches, the SCAD is advocated by Fan and Li 2001, as it enjoys three desirable properties: sparsity, unbiasedness and continuity.

[1] Chen Xu and Jiahua Chen, Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, Canada, V6T 1Z2, chen.xu@stat.ubc.ca, jhchen@stat.ubc.ca

[2] Harold Mantel, Statistical Research and Innovation Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6, Harold.Mantel@statcan.gc.ca

Under the classical settings for statistical inference (i.e. observations are independently drawn from an infinite population), the SCAD estimator can consistently select the true variables and accurately estimate their corresponding coefficients simultaneously (see, e.g. Fan and Li 2001, Fan and Peng 2004). These results are encouraging as they provide a deeper understanding on the large sample behaviours of the SCAD estimator under the ideal independence data structures. However, in survey sampling, observations are obtained from a finite population and hence they have intrinsic dependence structure. Therefore, the consistency results of SCAD developed for independent observations are not applicable for the analysis of survey data.

In this article, we investigate the asymptotic behaviours of SCAD when observations are collected through complex survey designs, where the data structure is potentially distorted by varied inclusion probabilities across the sampling units. A framework of sample-based penalized likelihood method is proposed for variable selection in analysis of survey data. Under a two-stage sampling scheme, we show that the sample-based SCAD estimator is consistent in both variable selection and parameter estimation.

The article is organized as follows. In Section 2, we introduce the model settings and propose the framework of sample-based PLM (SCAD) in the context of survey. In Section 3, we describe the two-stage sampling scheme and investigate the asymptotic behaviours of sample-based SCAD estimator. We use numerical studies in Section 4 to further assess the performance of proposed approach and place some concluding remarks in Section 5.

2. THE FRAMEWORK OF SAMPLE-BASED PLM

2.1 Super-population Model

Let D be a finite population consists of N units. Let y_i be the response variable of the unit i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ be its p covariates, for $i = 1, \ldots, N$. We assume that $(y_i, \mathbf{x}_i)$ are independent realizations of random vector $(Y, \mathbf{X})$ from a super-population $\widetilde{D}$, which hypothetically mimics the generation of $(y_i, \mathbf{x}_i)$ appearing in the finite population *D*. We consider the case where Y and $\mathbf{X}$ follow a generalized linear model (GLM) within $\widetilde{D}$. That is, conditioning on $\mathbf{X}$, the distribution of Y follows a natural exponential family, the density of which takes the form

$$f(Y;\theta) = c(Y)\exp(\theta Y - b(\theta)). \tag{1}$$

The parameter θ is linked to the covariates $\mathbf{X}$ via $b'(\theta) = g^{-1}(\mathbf{X}^T\beta)$, where $g(.)$ is a pre-specified link function and vector $\beta = \{\beta_1, \ldots, \beta_p\}^T$ is the p-dimensional regression coefficient. Suppose a probability sample $\{i_1, \ldots, i_n\}$ is taken and their corresponding $(y_i, \mathbf{x}_i)$ are observed. Based on this sample, we investigate the relationship between $\mathbf{X}$ and Y under the above model assumption. In this paper, we focus only on the canonical link $g(b'(\theta)) = \theta$, such that $\theta = \mathbf{X}^T\beta$.

We are interested in the situation where only a few covariates are influential on the outcome *Y*. Without loss of generality, we assume the first p coefficients are non-zero and denote the true value of β by $\beta_0 = \{\beta_{01}, \beta_{02}\}$ with $\beta_{02} = 0$. Because of the sparsity, the analysis benefits from a variable selection procedure, where the goal is to correctly identify the covariates with non-zero coefficients.

2.2 The sample-based PLM and SCAD

A number of penalized likelihood methods have been demonstrated to be an attractive technique for variable selection. They shrink the fitted regression coefficients toward 0 and automatically set some exactly at 0. This key property merits some PLM as variable selection operators.

Suppose we have observed every unit in the finite population *D*. For some penalty function $\phi_\lambda(.)$, we define the penalized likelihood function

$$Q_N(\beta) = \sum_{i=1}^{N} \log f(y_i; \mathbf{x}_i^T \beta) - N \sum_{j=1}^{p} \phi_\lambda(|\beta_j|). \tag{2}$$

The PLM estimates β by $\hat{\beta}_N$ that maximizes $Q_N(\beta)$. When $\phi_\lambda(|\beta|) = \lambda |\beta|^\gamma$ for some $\gamma > 0$, we obtain the bridge estimator, which includes the Lasso as a special case when $\gamma = 1$. When the penalty function is defined by

$$\phi_\lambda'(|\beta|) = \lambda \left\{ I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\}, \tag{3}$$

we obtain the SCAD estimator. In order for PLM to have a variable selection property, the penalty function must have a spike at the origin. Under mild conditions including the observations are independent with each other, the PLM can consistently select covariates with non-zero regression coefficients. In particular, the SCAD estimator enjoys three desirable properties, sparsity, unbiasedness and continuity (Fan and Li, 2001). Because the census data discussed above have the independence structure, most existing results of PLM remain applicable to $\hat{\beta}_N$.

In practical situations, only a probability sample is available. Let $I_i = 1$ if the i th unit is in the sample and $I_i = 0$ otherwise. We propose the sample-based penalized likelihood estimator $\hat{\beta}_n$ by maximizing

$$Q_n(\beta) = \sum_{i=1}^{N} I_i w_i \log f(y_i; \mathbf{x}_i^T \beta) - N \sum_{j=1}^{p} \phi_\lambda(|\beta_j|), \tag{4}$$

where w_i is the survey weight of the i th unit. Denote $P(I_i = 1) = \pi_i$, then in general, w_i is chosen proportional to π_i^{-1} such that $\sum_{i=1}^{N} I_i w_i y_i$ is an unbiased estimator for the population total $\sum_{i=1}^{N} y_i$ under the probability randomization distribution induced by the sampling plan.

Compared with $\hat{\beta}_N$, the distribution of $\hat{\beta}_n$ involves randomness from both super-population model (1) and the sampling design. Under the joint framework, the asymptotic behaviors of $\hat{\beta}_n$ must be re-investigated.

3. ASYMPTOTIC PROPERTIES OF $\hat{\beta}_n$

3.1 A two-stage sampling scheme

The properties of the $\hat{\beta}_n$ rely on sampling schemes. In this paper, we consider a two-stage sampling design commonly used in household surveys. Suppose the finite population D is stratified into H clusters according to some auxiliary variable known in advance to the sampling process. Let N_h be the number of units in cluster h and assume that $N_h < L$ for $h = 1, \ldots, H$ and for some constant L. The population size is given by $\sum_{h=1}^{H} N_h = N$. In the first stage of the sampling, n_1 clusters are selected through simple random sampling without replacement at the cluster level; then within each selected cluster n_2 units are randomly chosen as representatives for that cluster and values of $(Y, \mathbf{X})$ are measured. In household surveys, such clusters are often referred as dwellings in a region and the secondary sampling units are the people living in the dwellings.

We consider a sequence of populations D_r and samples d_r drawn according to the above sampling scheme. We assume that the number of clusters H increases to infinity as $r \to \infty$ with L remaining a constant. Also, we assume that the

primary sampling fraction $n_1 / H = \gamma$ $(0 < \gamma < 1)$ would not change with r. Under these settings, we first give a useful lemma as follows

Lemma 1. *Consider a finite population with H clusters of cluster size N_h. Let $\{u_{hl}\}$ be the measurement of the l th unit in the h th cluster. For a random sample obtained through the two-stage sampling design described above, we denote by U_{ij} the measurement of the j th sampled units in the i th sampled cluster. Assume $N^{-1}\sum_{h=1}^{H}\sum_{l=1}^{N_h}|u_{hl}|^{1+\eta} < C$ for some positive C and η. Then*

$$\frac{1}{N}\left|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} W_{ij}U_{ij} - \sum_{h=1}^{H}\sum_{l=1}^{N_h} u_{hl}\right| \to_p 0 \quad \text{as} \quad r \to \infty,$$

here $W_{ij} = N_h H / n_1 n_2$ when the i th cluster in the sample is the h th cluster in the population.

Lemma 1 states that as $r \to \infty$ the weighted sample mean converges to the population mean in probability. This result is an extension of the weak law of large numbers established by Chen and Rao (2007) for simple random samples.

3.2 The estimation and selection consistency

With Lemma 1, we now establish the consistency of sample-based SCAD estimator in both parameter estimation and variable selection. For simplicity of notation, we use a single subscript $t = 1, \ldots, N$ for the index of units in the population D and use subscript $j = 1, \ldots, p$ to denote the j th component of $\mathbf{X}$. Some regularity conditions on the finite population are required as follows:

C1. There exists a positive constant M_1 such that

$$\frac{1}{N}\sum_{t=1}^{N} x_{tj}^{2+\eta} \le M_1 \quad \text{for } j = 1, \ldots, p.$$

C2. Let $l_N(\beta) = \sum_{t=1}^{N} \log f(y_t; \mathbf{x}_t^T\beta)$, and $H_N(\beta) = -\left[\frac{\partial^2 l_N(\beta)}{\partial\beta\partial\beta^T}\right] = \sum_{t=1}^{N} \mathbf{x}_t b''[\mathbf{x}_t^T\beta]\mathbf{x}_t^T$. There exists a positive constant M_2 such that for r large enough,

$$\lambda_{min}[N^{-1}H_N(\beta_0)] \ge M_2,$$

where λ_{min} denotes the smallest eigenvalue. Furthermore, we assume that $b''(.)$ is bounded.

C3. For any $\varepsilon > 0$, there exists a constant ξ such that, when r is large enough,

$$H_N(\beta) \ge (1-\varepsilon)H_N(\beta_0)$$

for $\beta \in \{\beta : \| \beta - \beta_0 \| \le \xi\}$.

Condition 1 requires $(2+\eta)$ th moment for each X_j exists such that Lemma 1 is applicable to the corresponding sample-based second moments. Conditions 2 and 3 require that, when population size N is sufficiently large and β is close enough to β_0, $H_N(\beta)$ performs similarly as $H_N(\beta_0)$, so that the minimum eigenvalue of $H_N(\beta)$ is bounded below. We state our main result as the following theorem.

Theorem 1. *Under conditions C1-C3, if $\lambda \to 0$ and $n^{1/2-\delta}\lambda \to \infty$ for some $\delta > 0$, as $r \to \infty$, then there exists a local maximizer $\hat{\beta}_n = (\hat{\beta}_{n1}, \hat{\beta}_{n2})$ of the sample-based penalized likelihood function (4) with the SCAD penalty, such that*

$$\|\hat{\beta}_n - \beta_0\| = O_p(n^{-1/2+\delta}) \text{ and } P\{\hat{\beta}_{n2} = 0\} \to 1.$$

Remark 1. Theorem 1 shows that with some regularity requirements on the finite population, the sample-based SCAD can consistently identify the important variables and efficiently estimate the coefficients. These results also imply that as $r \to \infty$ the sample-based SCAD estimate $\hat{\beta}_n$ converges to its population-based version $\hat{\beta}_N$ with probability tending to 1.

Remark 2. The consistency result of sample-based SCAD is obtained under a two-stage sampling framework. It is not hard to extend these results to more general designs, where the sampling process could be muti-stage and stratified. One straightforward extension is to consider a stratified sampling over the entire population and then followed by the proposed two-stage cluster sampling within each stratum.

Corollary 1. Under conditions required in Theorem 1, the sample-based SCAD estimate of the true coefficient β_1 is asymptotically equivalent to the sample-based MLE on the true model, i.e.

$$P(\hat{\beta}_{n1} = \hat{\beta}_{MLE}) \to 1.$$

Remark 3. Findings in Corollary 1 together with the selection consistency are referred as the "oracle" property of the SCAD estimator (Fan and Li, 2001).

4. NUMERICAL STUDIES

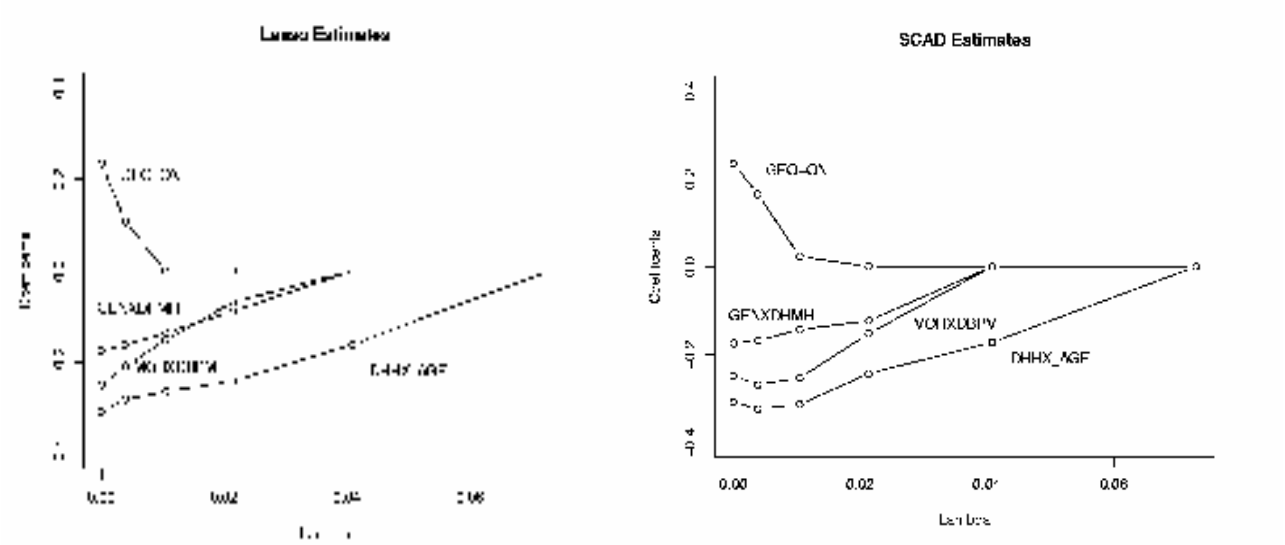
To further investigate the performance of sample-based SCAD, in this section we provide several numerical results through analyzing a health survey data set. In particular, we compare the sample-based SCAD and sample-based Lasso in terms of both estimation accuracy and selection stability.

The data we use is from the hypertension part of the Survey on Living with Chronic diseases in Canada (SLCDC) 2009. It is a cross-sectional survey sponsored by the Public Health Agency of Canada that collects information related to the experience of Canadians with chronic health conditions. The target population for the hypertension component of SLCDC is Canadians aged 20 years or older who have been diagnosed with hypertension living in private dwellings in the ten provinces. The samples for SLCDC were drawn from respondents of the 2008 Canadian Community Health Survey (CCHS) through a stratified (by age and sex) sampling scheme. An overall sample of 9,005 was selected from 17,437 CCHS respondents, and finally there are 6142 respondents who completed the SLCDC survey.

In this project, we choose to study the problem of identifying the health behaviours that affect the control of blood pressure. We are considering a logistic regression model of the blood control status (well controlled/ not) on 39 candidate covariates derived from the SLCDC data, and then use the sample-based PLM approach to select the influential ones.

Figure 1 shows the solution path of the Lasso and SCAD estimates of the four most significant coefficients according to various values of λ s. Compared with Lasso estimates, the estimates of SCAD are slightly larger in absolute value. This reflects the fact that the SCAD places fewer penalties to coefficients with large absolute fitted values. The BIC criterion

Figure 1: Solution path for Lasso and SCAD estimates



is then used to choose the optimal value of λ. Finally, the SCAD suggests 19 important covariates (with $\lambda = 0.016$) in the model, while the Lasso only selects 16 significant ones (with $\lambda = 0.011$).

To further investigate the stability of the selection results for sample-based PLM, we repeat above selection procedure on 500 independent sets of bootstrap samples drawn from the original SLCDC data. For each bootstrap sample, the weights are re-adjusted according to number of times that the unit is select in that bootstrap sample. In Table 1, we list the

Table 1: Selection rate for significant variables

	GEO_ON	DHHX_AGE	GENXDHMH	SMHXDSLT	HWTDBMI	MOHXDBPM	Ave.Size
SCAD	.782	.996	.894	.734	.882	.794	22.06
Lasso	.734	.986	.864	.670	.870	.776	20.02

bootstrap selection rate (number of times that a variable is selected divided by 500) for the six most significant covariates supported by the MLE estimates on the original SLCDC data. We observe that most of the significant variables supported by the original data are consistently selected by SCAD in the bootstrap replications, which illustrates the stability of sample-based SCAD selection procedure on the important variables. Compared with the SCAD, the Lasso selection has relatively lower selection rates on ``SMHXDSLT", which may be due to the excessive shrinkage caused by the L_1 penalty. In Table 1, we also include the average number of selected variables for both SCAD and Lasso. From our results, it seems that the SCAD tends to keep more variables than Lasso, as the shrinkage of SCAD is relatively gentle for significant coefficients.

The precision of the sample-based PLM estimates was also investigated through empirical variance estimations. The high precision of the SCAD estimates is supported by our numerical results.

5. CONCLUDING REMARKS

In this paper, we study the penalized likelihood method for variable selection in the context of survey sampling, where observations are intrinsically dependent. We propose a framework of using the PLM in analysis of survey data via introducing the sample-based penalized likelihood estimators under the super-population modelization. Under certain regularity conditions on the finite population, we show that the sample-based SCAD estimator consistently identifies the influential variables and efficiently estimates the corresponding coefficients. The good behaviours of sample-based SCAD are further supported by numerical studies.

The desired features of sample-based PLM rely on an appropriate choice of tuning parameter, which controls the amount of regularization. For the non-survey cases, Wang et al. (2007) have shown that the SCAD with tuning parameter chosen by BIC could consistently select the true model. However, their result is not directly applicable to the context of survey

where additional complications of data structure are introduced by survey designs. How should we choose such a good tuning parameter for sample-based PLM is an important issue.

ACKNOWLEDGEMENT

This project was supported by Statistics Canada and MITACS. The authors would like to thank Dr. Georgia Roberts in Statistics Canada and Prof. J.N.K Rao in Carleton University for valuable suggestions.

REFERENCES

Frank, I.E. and Friedman, J.H. (1993). "A statistical view of some chemonetrics regression tools". *Technometrics*., **35**, 109-148.

Breiman, L. (1996). "Heuristics of instability and stabilization in model selection". *Ann. of Statist.,* **24**, 2350-2383.

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso". *J. R. Statist. Soc. B*, **58**, 267-288.

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". *J. Amer. Statist. Assoc*, **96,** 1348-1360.

Fan, J. and Peng, H. (2004). "Nonconcave penalized likelihood with a diverging number of parameters". *Ann. of Statist.,* **32**, 781-813.

Chen, J. and Rao, J.N.K. (2007). "Asymptotic normality under two-phase sampling designs". *Statistica Sinica*, **17**, 1047-1064.

Wang, H., Li, R. and Tsai, C. (2007). "Tuning parameter selectors for the smoothly clipped absolute deviation Method". *Biometrika*, **94**, 553-568.