

Inconsistency of Classical Penalized Likelihood Approaches under Endogeneity

Yawei He

Department of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing, China

Email: yaweihh@163.com

How to cite this paper: He, Y.W. (2020) Inconsistency of Classical Penalized Likelihood Approaches under Endogeneity. *Journal of Applied Mathematics and Physics*, 8, 2335-2343.
<https://doi.org/10.4236/jamp.2020.810172>

Received: September 29, 2020

Accepted: October 27, 2020

Published: October 30, 2020

Abstract

With the high speed development of information technology, contemporary data from a variety of fields becomes extremely large. The number of features in many datasets is well above the sample size and is called high dimensional data. In statistics, variable selection approaches are required to extract the efficacious information from high dimensional data. The most popular approach is to add a penalty function coupled with a tuning parameter to the log likelihood function, which is called penalized likelihood method. However, almost all of penalized likelihood approaches only consider noise accumulation and supurious correlation whereas ignoring the endogeneity which also appeared frequently in high dimensional space. In this paper, we explore the cause of endogeneity and its influence on penalized likelihood approaches. Simulations based on five classical penalized approaches are provided to vindicate their inconsistency under endogeneity. The results show that the positive selection rate of all five approaches increased gradually but the false selection rate does not consistently decrease when endogenous variables exist, that is, they do not satisfy the selection consistency.

Keywords

High Dimension, Endogeneity, Feature Selection, Penalized Likelihood

1. Introduction

Along with the rapid progress of information technology and electronics industry, more and more data have been obtained from biomedical, econometrics and other fields. Therefore, in order to extract valid information from mass data, high-dimensional variable selection has been set off in statistics. Variables selection refers to the selection of important variables from the suspicious feature space and the elimination of redundant variables. High dimension indexes the

number of variables (features) is much higher than the sample size, and can even reach its exponential order. Compared with traditional data analysis, variable selection in high-dimensional space not only increases the computational burden, but also leads to noise accumulation, spurious correlation and endogeneity [1]. The noise accumulation is mainly due to the accumulation of estimation errors caused by the need to estimate a large number of unknown parameters at the same time during feature selection. To avoid noise accumulation, variable selection often makes a reasonable sparse assumption for the parameters to be evaluated [2]. The suspicious correlation is mainly due to the high sample correlation between high-dimensional variables. When important variables are highly correlated with some redundant variables, these redundant variables are easily selected and make suspicious variables. In this case, penalty function is usually added. This method of adding a penalty function after a log-likelihood function, called the penalized likelihood method, is the most common method for high-dimensional variable selection. Unfortunately, most penalized likelihood methods consider noise accumulation and spurious correlation, but ignore another important factor—endogeneity [3]. This paper studies the influence of endogeneity on the classical penalized likelihood methods, which is divided into three parts. Firstly, it introduces the origin and causes of endogeneity; secondly, it summarizes the classical penalized likelihood method and its development process; finally, comparative analysis is carried out to show the inconsistency of various penalized likelihood approaches under endogeneity.

2. The Origin and Cause of Endogeneity

The concept of endogeneity originated from economics. Under the linear regression model $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \dots + X_p\beta_p + \varepsilon$, it means that some explanatory variables correlates with the residual, namely $\text{cov}(X_j, \varepsilon) \neq 0$. The causes of endogeneity in variable selection can be roughly divided into three categories: omitted variables, measurement errors and simultaneous bias. These will be elaborate in detail under the most commonly used linear regression model. Omitted variables mean that some important variables that can affect the response variable Y are omitted in the explanatory variable. If these omitted variables were related to the pre-existing explanatory variables, endogeneity would occur. To be more specific, assuming that the true regression model is $Y = \beta_0 + X_1\beta_1 + \dots + X_k\beta_k + X_*\beta_* + \varepsilon$, but variable X_* is omitted, and the regression model is mistakenly set as $Y = \beta_0 + X_1\beta_1 + \dots + X_k\beta_k + \varepsilon$. Therefore the omitted variable actually goes into the error term u , that is, $u = X_*\beta_* + \varepsilon$. If X_* is related to X_j , then u is related to X_j , and it would lead to endogeneity. When the measurement of a variable is incomplete, the measurement bias will be included in the error term of the regression equation as a part of the regression bias. The measurement bias comes not only from the error records of variables, but also from the inevitable conceptual differences between the commonly used proxy variable and the real variable, which can be obtained from the explanatory variables and the response variable. For example, suppose the real regression mod-

el is $Y = \beta_0 + X_1\beta_1 + \dots + X_k\beta_k + \varepsilon$ and the equation to be estimated is $Y^* = \beta_0 + X_1\beta_1 + \dots + X_k\beta_k + u$, $Y^* - Y = t$ is the measurement error. If the measurement bias t is related to the explanatory variable, endogeneity will occur. In addition to the omitted variables and measurement biases leading to endogeneity, explanatory variables and response variables may also affect each other. That is not a one-way causality, leading to causal correlation bias but also endogeneity. Take resident income X and resident consumption Y as an example. In general, the interaction between income and consumption, and the process of mutual influence cannot be observed. At this time, the information about X and Y is essentially mixed up. More precisely, $Y = \beta_0 + X\beta_1 + \varepsilon$, $X = \gamma_0 + Y\gamma_1 + u$, so $\text{cov}(X, \varepsilon) \neq 0$ and endogeneity occurs.

In the analysis of high-dimensional data, endogeneity is almost inevitable. That is mainly because researchers tend to collect as many potential relevant explanatory variables as possible to avoid omission of important variables when we do not know the real model while these high-dimensional variables are usually aggregated from multiple data sources. Unintentionally, some explanatory variables may be associated with residuals, leading to endogeneity. It can also be said that the more variables, the higher the data dimension, the greater the probability of endogeneity.

3. Penalized Likelihood Method and Its Development

One of the most popular techniques in statistics for extracting information from large volumes of complex data is the high dimensional variable selection. There are two main goals in variable selection: selection consistency, that is, selecting of important variables accurately with a probability close to 1; prediction accuracy, that is, estimating coefficients as accurately as knowing in advance. An Oracle property is defined if these two goals can be satisfied simultaneously. However, due to the occurrence of over-fitting in high-dimensional space, it is difficult combine the two goals, and the selection consistency is usually considered to be more important. For example, in disease gene mapping, the main concern is which genes are the pathogenic genes and not others.

In the high dimension linear model, the penalized likelihood method, which adds a penalty function to the log-likelihood function to shrink estimates to trade between variance and bias, is the most common method of variable selection. More specifically, we consider a linear regression model with main effects only, by minimizing the penalized likelihood function $\|Y - X\beta\|^2 + \sum p_\lambda(\beta_j)$, and it's going to produce a certain amount of non-zero coefficients. And their corresponding variables will be the candidate variables. In the penalized likelihood approaches, a variety of penalized functions were selected, including Lasso [4], SCAD [5], Adaptive Lasso (ALasso) [6], MCP [7], Sequential Lasso (SLasso) [8], etc.

3.1. Lasso and Improvements

Lasso was the first to choose the most basic penalized function $p_\lambda(\beta) = \lambda|\beta|$

and has been widely cited. It is convenient and easy to compute since its entire regularization path is computed under the complexity of a single linear regression. In a high-dimensional space, the estimation of Lasso is biased, but it satisfies model's selection consistency under conditions like neighborhood stability condition [9], non-representable condition [10], and Mutual Incoherence Condition [11]. However, all of these conditions require weak correlations between non-significant variables and significant variables, which is difficult to achieve in practice. That is, Lasso performs poorly when there is a high correlation between variables. In fact, for a set of variables with a high two-way correlation, Lasso is more likely to select a variable from this set regardless of which one is selected.

Many classical feature selection methods have been proposed by on the basis of Lasso. Elastic net [12] integrated Lasso with ridge regression by defining $p_\lambda(\beta) = \lambda_1 |\beta| + \lambda_2 |\beta|^2$ and it outperforms Lasso in high correlation and prediction accuracy. However, it is easy to cause grouping effect, that is, highly correlated variables are often selected into the model or excluded at the same time. ALasso [6] considers the weighted penalized function $p_\lambda(\beta_j) = \lambda w_j |\beta_j|$ and is proved to satisfy both the selection consistency and the prediction accuracy under a reasonable initial estimator. Another significant improvement of Lasso, S-Lasso [8], takes a stepwise approach to variables selection, but only adds a L1 penalized function to variables which are not selected in previous stage. This can ensure that variables selected in the early stage are not omitted in the subsequent selection process. S-Lasso also owns the oracle property and is more computationally attractive than approaches like elastic net.

3.2. SCAD and Related

Compared with Lasso, SCAD [5] takes a different approach, resulting in a successful nonconcave penalized function

$$P'_\lambda(\beta) = \lambda I(\beta \leq \lambda) + (a\lambda - \beta) + I(\beta > \lambda) / (a - 1)$$

which has desirable properties on many occasions [13] [14] [15]. MCP [7] makes $P'_\lambda(\beta) = (a\lambda - \beta) / a$ to be similar to the penalized function used for SCAD and translates the flat part of the derivative of SCAD penalty to the origin. However, due to the nature of the nonconcave penalized function, they are both computationally unadvantageous if compared to the Lasso family.

3.3. Tuning Parameter

In addition to the chosen of penalized function, the determination of tuning parameter λ is also one of the key points of penalized likelihood approaches. If set λ to a set of values, a series of candidate models are generated. Therefore, the penalized likelihood method should be used in conjunction with the model selection criteria. The former generates candidate models; the latter decides the optimal model. Classical model selection criteria include AIC [16], BIC [17]. However, these traditional criteria are no longer suitable for high-dimensional space due to the selection of too many useless variables. In order to adapt to the

high-dimensional situation, researchers add additional penalized terms after AIC [18] or replace factor 2 with a constant term C [19]. For BIC, more efforts have been devoted to the prior probability modifications, such as modified BIC (mBIC) [20] and extended BIC (EBIC) [21]. By assigning different values to parameters γ , EBIC is essentially a set of criteria. The BIC and mBIC can be regarded as special cases of EBIC by letting $\gamma = 0$ and $\gamma = 1$. The properties of EBIC under different high-dimensional models have been extensively studied. It is consistent for the linear model [21], the generalized linear model [22], the cox model [23], etc.

4. Inconsistency under Endogeneity

When using the penalized likelihood method for variables selection, some basic conditions must be met to achieve the desired properties. This includes restrictions on explanatory variables [8] or focus on the explanatory variables and regression coefficients [24] or the restrictions on likelihood function [25]. However, when endogeneity exists, even if there's only one endogenous variable left, the above necessary conditions are hard to meet. In this case, there will be an insurmountable difference between the estimated value of regression coefficient and the true value, which will affect the selection consistency of these features. Next, we will use a simulation to show the effect of endogeneity.

4.1. Specification of Model

Consider the model $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, I)$. let sample size be $n = 50, 100$ and 200 respectively. Define the number of variables $p = \lfloor n^{1.2} \rfloor$ and $\beta_j = (-1)^j (0.8 + 0.05u)$, where u follows the two-point distribution with a parameter of 0.5 , for $j = 1, 2, \dots, 6$; $\beta_j = 0$ for $j = 7, 8, \dots, p$. Consider two different Settings:

Setting 1: $X_j = Z_j, j = 1, 2, \dots, 6$; $X_j = Z_j(1 + 2\varepsilon), j = 7, 8, \dots, p$.

Setting 2: $X_j = Z_j(1 + 2\varepsilon), j = 1, 2, \dots, p - 6$; $X_j = Z_j, j = p - 5, \dots, p$.

The difference between these two settings is that the former only has insignificant variables that are endogenous while the latter are all important variables that are endogenous. Both of them will be compared respectively with the exogenous case that $X_j = Z_j$ for all j to reflect the impact of endogeneity. The $Z \sim N(0, \Sigma)$ and is independent of ε . The setting of the covariance matrix Σ considers only two common structures: $\Sigma_{ij} = 0.5, i \neq j, \Sigma_{ij} = 1, i = j$ and $\Sigma_{ij} = 0.5^{|i-j|}$, which can be called S1 and S2 respectively. The extended Bayesian model selection criterion EBIC is used to select the tuning parameter and determine the optimal model by letting $\gamma = 1 - \log n / 4 \log p$.

4.2. Results and Interpretation

In the measurement of selection consistency, PDR (number of true selected variables/total number of true variables), FDR (number of false selected variables/total number of selected variables) and Msize (total number of selected variables) are

used. Due to the randomness of the explanatory variables and the error term, the above simulation process will be repeated 200 times to take the average value of the measures, and the results are shown in **Tables 1-4**.

It can be seen from **Tables 1-4** that when there is no endogeneity, PDR tended to 1 with an upward trend, while FDR tended to 0 with a downward trend, and the number of selected variables tended to the number of real variables, although the initial performance of various feature selection methods is different. In other words, the asymptotic consistency of these classical penalized likelihood approaches satisfied. However, when endogeneity exists, either the unimportant

Table 1. Results under setting 1 with s1.

endogeneity	PDR, FDR (Msize)		
	n = 50	n = 100	n = 200
Lasso	0.29, 0.71 (4.6)	0.38, 0.67 (5.8)	0.51, 0.64 (8.2)
S_Lasso	0.29, 0.73 (6.4)	0.45, 0.71 (9.0)	0.79, 0.70 (16.4)
SCAD	0.96, 0.55 (14.0)	0.99, 0.60 (17.5)	1.00, 0.64 (22.6)
A_Lasso	0.36, 0.70 (6.2)	0.47, 0.68 (8.8)	0.82, 0.69 (16.2)
MCP	0.97, 0.56 (14.5)	0.99, 0.64 (18.1)	1.00, 0.70 (23.6)
exogeneity	n = 50	n = 100	n = 200
Lasso	0.41, 0.25 (3.5)	0.81, 0.23 (6.8)	0.97, 0.16 (7.4)
S_Lasso	0.64, 0.46 (10.3)	0.97, 0.14 (7.0)	1.00, 0.05 (6.4)
SCAD	0.74, 0.53 (14.6)	0.99, 0.20 (11.7)	1.00, 0.04 (6.3)
A_Lasso	0.62, 0.44 (10.2)	0.95, 0.12 (6.9)	1.00, 0.05 (6.4)
MCP	0.75, 0.58 (15.2)	0.99, 0.25 (14.4)	1.00, 0.05 (6.4)

Table 2. Results under setting 1 with s2.

endogeneity	PDR, FDR (Msize)		
	n = 50	n = 100	n = 200
Lasso	0.23, 0.73 (3.9)	0.33, 0.64 (5.4)	0.44, 0.52 (6.3)
S_Lasso	0.20, 0.75 (4.8)	0.29, 0.70 (6.3)	0.44, 0.67 (9.9)
SCAD	0.93, 0.58 (15.0)	0.99, 0.67 (20.3)	0.99, 0.56 (18.1)
A_Lasso	0.31, 0.72 (4.6)	0.39, 0.62 (6.6)	0.46, 0.64 (8.6)
MCP	0.94, 0.59 (15.2)	1.00, 0.69 (20.7)	1.00, 0.75 (27.3)
exogeneity	n = 50	n = 100	n = 200
Lasso	0.49, 0.13 (3.6)	0.78, 0.14 (5.9)	0.96, 0.13 (7.0)
S_Lasso	0.58, 0.33 (7.7)	0.90, 0.17 (6.9)	1.00, 0.06 (6.4)
SCAD	0.71, 0.56 (11.3)	0.88, 0.25 (10.1)	1.00, 0.09 (6.7)
A_Lasso	0.56, 0.28 (6.8)	0.89, 0.15 (6.8)	1.00, 0.06 (6.4)
MCP	0.73, 0.66 (18.1)	0.96, 0.27 (12.9)	1.00, 0.06 (6.7)

Table 3. Results under setting 2 with s_1 .

endogeneity	PDR, FDR (Msize)		
	n = 50	n = 100	n = 200
Lasso	0.78, 0.58 (11.4)	0.89, 0.54 (12.0)	0.95, 0.42 (10.7)
SLasso	0.77, 0.70 (16.3)	0.99, 0.62 (16.4)	1.00, 0.64 (17.3)
SCAD	0.78, 0.47 (9.6)	0.99, 0.37 (10.6)	1.00, 0.38 (12.0)
ALasso	0.72, 0.65 (14.7)	0.94, 0.56 (13.7)	1.00, 0.52 (14.8)
MCP	0.74, 0.50 (9.6)	0.99, 0.40 (11.0)	1.00, 0.41 (14.8)
exogeneity	n = 50	n = 100	n = 200
Lasso	0.41, 0.25 (3.5)	0.81, 0.23 (6.8)	0.97, 0.16 (7.4)
SLasso	0.64, 0.46 (10.3)	0.97, 0.14 (7.0)	1.00, 0.05 (6.4)
SCAD	0.74, 0.53 (14.6)	0.99, 0.20 (11.7)	1.00, 0.04 (6.3)
ALasso	0.62, 0.41 (10.2)	0.95, 0.12 (6.9)	1.00, 0.06 (6.4)
MCP	0.75, 0.60 (16.6)	0.99, 0.25 (14.4)	1.00, 0.05 (6.3)

Table 4. Results under setting 2 with s_2 .

endogeneity	PDR, FDR (Msize)		
	n = 50	n = 100	n = 200
Lasso	0.75, 0.54 (10.9)	0.85, 0.51 (11.7)	0.96, 0.42 (11.0)
SLasso	0.64, 0.71 (14.8)	0.90, 0.65 (16.1)	0.99, 0.62 (16.4)
SCAD	0.76, 0.52 (10.6)	0.99, 0.40 (11.1)	1.00, 0.40 (10.7)
ALasso	0.69, 0.67 (13.3)	0.88, 0.62 (15.2)	0.98, 0.58 (15.8)
MCP	0.75, 0.55 (10.7)	0.98, 0.50 (12.6)	1.00, 0.54 (14.7)
exogeneity	n = 50	n = 100	n = 200
Lasso	0.50, 0.13 (3.6)	0.78, 0.14 (5.9)	0.97, 0.13 (7.0)
SLasso	0.58, 0.33 (7.7)	0.90, 0.17 (6.9)	1.00, 0.06 (6.4)
SCAD	0.71, 0.56 (16.3)	0.88, 0.25 (10.1)	1.00, 0.08 (6.7)
ALasso	0.57, 0.29 (6.9)	0.90, 0.15 (6.9)	1.00, 0.06 (6.5)
MCP	0.74, 0.66 (18.1)	0.96, 0.27 (13.0)	1.00, 0.06 (6.7)

endogenous variables or important endogenous variables, as the sample size increases, all approaches are selected to rate though there is a rising trend for PDR but not necessarily obvious. The performance of FDR and number of selected variables is not as expected by their asymptotic nature; it's still picking the wrong variables, which means it is no longer valid in the presence of endogeneity. In addition, these tables showed the difference in the robustness between the above penalized likelihood methods. When switching from exogenous to endogenous, SCAD is the most robust and SLasso is the lowest robust, which suggests some implications for subsequent endogenous feature selection studies.

Acknowledgements

This project is supported by National Natural Science Foundation of China (Grant No: 11701058).

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- [1] Fan, J. (2014) Challenges of Big Data analysis. *National Science Review*, **1**, 293-314. <https://doi.org/10.1093/nsr/nwt032>
- [2] Donoho, D. (2000) High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *The American Mathematical Society Conference*, Los Angeles, CA, United States, 7-12 August 2000.
- [3] Engle, R., Hendry, D. and Richard, J.-F. (1983) Exogeneity. *Econometrica*, **51**, 277-304. <https://doi.org/10.2307/1911990>
- [4] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, **58**, 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- [5] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. <https://doi.org/10.1198/016214501753382273>
- [6] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. <https://doi.org/10.1198/016214506000000735>
- [7] Zhang, C.H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. <https://doi.org/10.1214/09-AOS729>
- [8] Luo, S. and Chen, Z. (2011) Sequential Lasso for Feature Selection with Ultra-High Dimensional Feature Space. *Journal of the American Statistical Association*, **109**, 1229-1240. <https://doi.org/10.1080/01621459.2013.877275>
- [9] Meinshausen, N. and Bühlmann, P. (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462. <https://doi.org/10.1214/009053606000000281>
- [10] Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, **7**, 2541-2563.
- [11] Wainwright, M. (2009) Sharp Thresholds for High Dimensional and Noisy Sparsity Recovery Using ℓ_1 -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, **55**, 2183-2202. <https://doi.org/10.1109/TIT.2009.2016018>
- [12] 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. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- [13] Fan, J. and Li, R. (2004) New Estimation and Model Selection Procedures for Semi-parametric Modeling in Longitudinal Data Analysis. *Journal of the American Statistical Association*, **99**, 710-723. <https://doi.org/10.1198/016214504000001060>
- [14] Fan, J., Peng, H., *et al.* (2004) Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, **32**, 928-961.

- <https://doi.org/10.1214/009053604000000256>
- [15] Xie, H. and Huang, J. (2009) SCAD-Penalized Regression in High-Dimensional Partially Linear Models. *The Annals of Statistics*, **37**, 673-696. <https://doi.org/10.1214/07-AOS580>
- [16] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. *Second International Symposium on Information Theory*, 267-281.
- [17] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. <https://doi.org/10.1214/aos/1176344136>
- [18] Barron, A., Birge, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, **113**, 301-413. <https://doi.org/10.1007/s004400050210>
- [19] Baraud, Y. (2000) Model Selection for Regression on a Fixed Design. *Probability Theory and Related Fields*, **117**, 467-493. <https://doi.org/10.1007/PL00008731>
- [20] Bogdan, M., Ghosh, J.K. and Doerge, R. (2004) Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. *Genetics*, **167**, 989-999. <https://doi.org/10.1534/genetics.103.021683>
- [21] Chen, J. and Chen, Z. (2008) Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, **95**, 759-771. <https://doi.org/10.1093/biomet/asn034>
- [22] Chen, J. and Chen, Z. (2012) Extended BIC for Small- n -Large- P Sparse GLM. *Statistica Sinica*, **22**, 555-574. <https://doi.org/10.5705/ss.2010.216>
- [23] Luo, S., Xu, J. and Chen, Z. (2015) Extended Bayesian Information Criterion in the Cox Model with a High Dimensional Feature Space. *Annals of the Institute of Statistical Mathematics*, **67**, 287-311. <https://doi.org/10.1007/s10463-014-0448-y>
- [24] Lu, W., Goldberg, Y. and Fine, J.P. (2012) On the Robustness of the Adaptive Lasso to Model Misspecification. *Biometrika*, **99**, 717-731. <https://doi.org/10.1093/biomet/ass027>
- [25] Fan, J. and Liao, Y. (2012) Endogeneity in High Dimensions. *Annals of Stats*, **42**, 872-917. <https://doi.org/10.1214/13-AOS1202>