Variable selection: From full data to missing data

Yongchan Kwon

Department of Statistics, Seoul National University, Seoul, Korea

December 8, 2014

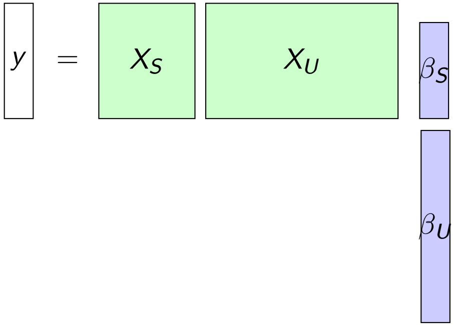
Contents

Introduction

$$y = X \beta$$

$\rightarrow$ Is variable selection important ??

Introduction



$\rightarrow$ Yes! variable selection is so important !

Classical variable selection

The best subset selection

[step1] Define $\mathcal{M}_k$ be the set of all linear functions with k nonzero coefficients.
[step2] For $k = 0, \cdots, p$, choose $m_k \in \mathcal{M}_k$ such that m_k has the minimum of $\text{RSS}(\beta) = (y - X\beta)^T(y - X\beta)$ among $\mathcal{M}_k$.
[step3] Among $m_0, \cdots, m_p$, choose one model using cross-validation, AIC, or BIC.

$\rightarrow$ Can be costly in computation.

Penalized least squares method

Let y be an $n \times 1$ vector and X be an $n \times d$ matrix. Then, a form of the penalized least squares is for $\lambda > 0$

$$\mathrm{argmin}_\beta \left(\frac{1}{2}(\mathrm{y} - \mathrm{X}\beta)^{\mathrm{T}}(\mathrm{y} - \mathrm{X}\beta) + \sum_{\mathrm{j}=1}^{\mathrm{d}} \mathrm{p}_\lambda(|\beta_{\mathrm{j}}|) \right)$$

where $p_\lambda(\cdot)$ is called a penalty function indexed up to penalty parameter λ. Penalty parameter λ can be chosen by Generalized Cross-Validation(GCV).

L_q penalty : $p_{\lambda j}(|\beta_j|) = \lambda|\beta_j|^q$
$q = 2$: Ridge regression $\rightarrow$ No variable selection features
$q = 1$: LASSO

Penalized least squares method

Fan and Li (2001) suggest that a good penalty function should result in an estimator with three properties.

To be a good estimator.....

1.**Unbiasedness:** The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.
2.**Sparsity:** The resulting estimator is a thresholding rule, which automatically sets small estimated coefficient to zero to reduce model complexity.
3.**Continuity:** The resulting estimator is continuous in the data to avoid instability in model prediction.

Penalized least squares method

Fan and Li (2001) proposed Smoothly Clipped Absolute Deviation (SCAD) penalty defined by

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}$$

for some $a > 2$ and $\beta > 0$.

Penalized least squares method

penalty.png

Figure: Comparing L_1, L_2, and SCAD penalty functions

Penalized least squares method

Oracle property

Let β^* be the true regression coefficient and $A = \{j : \beta_j^* \neq 0\}$. We will say β^o be the oracle estimator defined as

$$\beta^o = \mathrm{argmin}_{\beta,\beta_j=0,j\in A^c} \frac{1}{2}(y - X\beta)^T(y - X\beta)$$

$\hat{\beta}$ is said to possess **the oracle property** if there exists a sequence of λ_n such that with $\lambda = \lambda_n$

$$\lim_n \Pr(\hat{\beta} = \beta^0) = 1.$$

A slightly weaker definition is that if estimator satisfies
(1) $\lim_n \Pr(\hat{A} = A^*) = 1$
(2) $\sqrt{n}(\hat{\beta} - \beta^*) \overset{d}{=} \sqrt{n}(\beta^o - \beta^*)$.

Variable selection in missing data

Setting

$(X_1, z_1, y_1), \cdots, (X_n, z_n, y_n)$: n independent observations
y_i : the response variable
X_i : a completely observed covariates.
z_i : a partially observed covariates.
$(z_{m.i}, z_{o.i})$: missing and observed component of z_i.
r_i : response indicator for z_i.
$D_{f,i}$ and $D_{o,i}$: full and observed data of subject i
D_f and D_o : the entire full and observed data
D_m : missing part.

Setting2

Then,

$$f(D_c) = \prod_{i=1}^{n} f(y_i, z_i, r_i \mid x_i, \eta)$$

Where η is a parameter. According to the EM algorithm, we define Q-function given by

$$Q(\eta \mid \eta^{(s)}) = E[\log f(D_f; \eta) \mid D_o; \eta^{(s)}].$$

By definition, we can write

$$Q(\eta \mid \eta^{(s)}) = \log f(D_o; \eta) + H(\eta \mid \eta^{(s)})$$

Where $H(\eta \mid \eta^{(s)}) = E[\log f(D_m \mid D_o; \eta) \mid D_o; \eta^{(s)}]$.

Variable selection in missing data

Ibrahim, Zhu, and Tang (2008) give an idea to calculate observed likelihood by approximating H-function using a truncated Hermite expansion.(One of orthogonal series expansion.)
In the same paper, they define two new information criteria given by

$$IC_{H,Q} = -2\log f(D_o;\hat{\eta}) + c_n(\hat{\eta}) = -2Q(\hat{\eta} \mid \hat{\eta}) + 2H(\hat{\eta} \mid \hat{\eta}) + c_n(\hat{\eta})$$

$$IC_Q = -2Q(\hat{\eta} \mid \hat{\eta}) + c_n(\hat{\eta})$$

where $c_n(\hat{\eta})$ is a function of the data and the fitted model. By choosing small $IC_{H,Q}$, we can select the model(variable selection). For instance, if $c_n(\hat{\eta}) = dim(\eta) \times 2$ is an AIC-type criterion.

Variable selection in missing data

Thus, penalized idea is revisited!! Garcia, Ibrahim, and Zhu (2010) proposed the method to develop variable selection with penalty function for missing data problems.

Idea!!

The idea is that
(1) parameter is estimated by penalized likelihood method
(2) penalty parameter is chosen by minimizing IC_Q.

Variable selection in missing data

Assumptions

(A1) η^* is unique and an interior point of the compact parameter space Θ.
(A2) $\hat{\eta}_o \to \eta^*$ in probability.
(A3) For all $i, l_i(\eta)$ is three-times continuously differentiable on Θ and $l_i(\eta), |\partial_j l_i(\eta)|^2$ and $|\partial_j \partial_k \partial_l l_i(\eta)|$ are dominated by $B_i(D_{o,i})$ for all $j, k, l = 1, \cdots, d$. where d is a number of candidate covariates and $\partial_j = \partial / \partial_j$.
(A4) For each $\epsilon > 0$, there exists a finite K such that

$$sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^{n} E[B_i(D_{o,i}) 1_{B_i(D_{o,i}) > K}] < \epsilon$$

for all n.

Variable selection in missing data

Assumptions

(A5)

$$\lim_n -\frac{1}{n}\sum_{i=1}^n \partial_\eta^2 l_i(\eta^*) = A(\eta^*)$$

$$\lim_n \frac{1}{n}\sum_{i=1}^n \partial_\eta l_i(\eta^*)\partial_\eta l_i(\eta^*)^T = B(\eta^*)$$

$$\lim_n -\frac{1}{n}\sum_{i=1}^n D^{20}Q(\eta_S^*|\eta^*) = C(\eta_S^*|\eta^*)$$

$$\lim_n \frac{1}{n}\sum_{i=1}^n D^{10}Q(\eta_S^*|\eta^*)D^{10}Q(\eta_S^*|\eta^*)^T = D(\eta_S^*|\eta^*)$$

where $A(\eta^*)$ and $C(\eta_S^*|\eta^*)$ are positive definite and D^{ij} denotes the i-th and j-th derivatives of the first and second component of the Q function.

Variable selection in missing data

Assumptions

(A6) Define $a_n = \mathrm{max_j}\{\mathrm{p'}_{\lambda_{\mathrm{jn}}}(|\beta^*_{\mathrm{j}}|) : \beta^*_{\mathrm{j}} \neq 0\}$, and
$b_n = \mathrm{max_j}\{\mathrm{p''}_{\lambda_{\mathrm{jn}}}(|\beta^*_{\mathrm{j}}|) : \beta^*_{\mathrm{j}} \neq 0\}$
1. $\mathrm{max_j}\{\lambda_{\mathrm{j_n}} : \beta^*_{\mathrm{j}} \neq 0\} = \mathrm{o_p}(1)$
2. $a_n = O_p(n^{-1/2})$.
3. $b_n = o_p(1)$.

(A7) Define $d_n = \mathrm{min_j}\{\lambda_{\mathrm{j_n}} : \beta^*_{\mathrm{j}} = 0\}$.
1. For all j such that $\beta^*_j = 0$, $\lim_n \lambda^{-1}_{j_n} \liminf_{\beta \to 0+} p'_{\lambda_{j_n}}(\beta) > 0$ in probability.
2. $n^{1/2} d_n \xrightarrow{p} \infty$.

Variable selection in missing data

Theorem

Under assumptions (A1)-(A7), we have
(1) Unbiasedness: $\hat{\eta}_\lambda - \eta^* = O_p(n^{-1/2})$ *as* $n \to \infty$.
(2) Sparsity: $P(\hat{\beta}_{(2)\lambda} = 0) \to 1$.
(3) Asymptotic normality: $(\hat{\beta}_{(1)\lambda}, \hat{\tau}_\lambda, \hat{\alpha}_\lambda, \hat{\zeta}_\lambda)$ *is asymptotically normal.*

proof of (1). Given assumptions, then it follows from White (1994) that

$$n^{-1/2}\sum_{i=1}^{n}\partial_\eta l_i(\eta^*) \xrightarrow{D} N(0, B(\eta^*)).$$

and

$$n^{1/2}(\hat{\eta}_o - \eta^*) \xrightarrow{D} N(0, A(\eta^*)^{-1}B(\eta^*)A(\eta^*)^{-1})$$

To show $\hat{\eta}_\lambda$ is a $\sqrt{n}$-consistent maximizer of η^*, it is enough to show that for large C

$$P\Big(\sup_{|u|=C}\left\{l(\eta^* + n^{-1/2}u) - n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\beta_j^* + n^{-1/2}u_j|)\right\}$$
$$< l(\eta^*) + n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\beta_j^*|)\Big) \to 1$$

Since this implies there exists a local maximizer in the ball $\{\eta^* + n^{-1/2}u; |u| < C\}$ and thus unbiasedness is proved. Taking a Taylor's expansion of the penalized likelihood function, we have

$$
\begin{aligned}
& l(\eta^* + n^{-1/2}u) - l(\eta^*) + n\sum_{j=1}^{p} p_{\lambda_{jn}}(|\beta_j^*|) - n\sum_{j=1}^{p} p_{\lambda_{jn}}(|\beta_j^* + n^{-1/2}u_j|) \\
& \leq n^{-1/2}u^T\partial_\eta l(\eta^*) - \frac{1}{2}u^T A(\eta^*)u + \sqrt{p_1}n^{1/2}a_n|u| - \frac{1}{2}|b_n||u|^2 + o_p(1) \\
& \leq n^{-1/2}u^T\partial_\eta l(\eta^*) - \frac{1}{2}u^T A(\eta^*)u + \sqrt{p_1}n^{1/2}a_n|u| + o_p(1)
\end{aligned}
$$

Note that except the second term of last equation is $O_p(1)$ and $u^T A(\eta^*)u$ is bounded below by $|u|^2\times$ the smallest eigenvalue of $A(\eta^*)$, then this dominates other three terms. Thus, results can be made negative for enough large C.

References

- Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American Statistical Association, Dec 2001, Vol. 96, No. 456.
- Zou, H. (2006), "The Adaptive Lasso and its oracle properties", Journal of the American Statistical Association, Dec 2006, Vol. 101, No. 476.
- Ibrahim, J. G., Zhu, H., and Tang, N. (2008), "Model selection Criteria for missing data problems using the EM algorithm", Journal of the American Statistical Association, Dec 2008, Vol. 103, No. 484.

References

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010), "Variable selection for regression models with missing data", Statistica Sinica, 20 (2010), 149-465.

J. Fan, and J. Lv. (2010), "A selective overview of variable selection in High Dimensional Feature Space", Stat Sin. 2010 January; 20(1): 101-148.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso", Journal of the Royal Statistical Society, Series B, Volume 58, Issue 1 (1996), 267-288.