Smoothly Clipped Absolute Deviation (SCAD) for Correlated Variables

SIDI ZAKARI Ibrahim

LIB-MA, FSSM
Cadi Ayyad University (Morocco)

COMPSTAT'2010
Paris, August 22-27, 2010

Motivations

- Fan and Li (2001), Zou and Li (2008) works
- Convex penalties (e.g quadratic penalties) : make trade-off between bias and variance, can create unnecessary biases when the true parameters are large and cannot produce parsimonious models.
- Nonconcave penalties (e.g: SCAD penalty,Fan 1997 and hard thresholding penalty, Antoniadis 1997)
- Variables selection in high dimension (correlated variables)
- Penalized likelihood framework

Ideal procedure for variable selection

- Unbiasedness: The resulting estimator is nearly unbiasedness when the true unkwown parameter is large to avoid excessive estimation bias.
- Sparsity: Estimating a small coefficient as zero, to reduce model complexity.
- Continuity: The resulting estimator is continuous in the data to avoid instability in model prediction.

The Smoothly Clipped Absolute Deviation (SCAD) Penalty

The SCAD penalty noted $J_\lambda(.)$ satisfies all three requirements (unbiasedness,sparsity,continuity) and is defined by $J_\lambda(0) = 0$ and for $|\beta_j| > 0$

$$J'_\lambda(|\beta_j|) = \lambda \mathbf{I}(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{a-1} \mathbf{I}(|\beta_j| > \lambda), \qquad (1)$$

where $(z)_+ = \max(z, 0)$, $a > 2$ and $\lambda > 0$.
SCAD possesses oracle properties.

Generalities

Let $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ an i.i.d random variables sample where $\mathbf{x}_i \in R^p, y_i \in R$.
The conditional log-likelihood function knowing $\mathbf{x}_i$ is:

$$\ell_i(\boldsymbol{\beta}) = \ell_i(\boldsymbol{\beta}, \phi) = \ell_i(\mathbf{x}_i^t \boldsymbol{\beta}, y_i, \phi) \tag{2}$$

where ϕ is the dispersion parameter, supposed known.
We want to estimate β maximizing:

$$P\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) - n \sum_{j=1}^{p} J_\lambda(|\beta_j|), \tag{3}$$

- The penalized likelihood is nonconcave and nondifferentiable
- Maximization problem
- Alternative: Approximation of the SCAD penalty by convex functions
- Iterative algorithms

LQA Algorithm: Fan and Li (2001)

$$\beta^{(k+1)} = \text{argmax}_{\beta} \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} \frac{J'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|} \beta_j^2 \right\}. \quad (4)$$

- When $|\beta_j^{(k)}| < \epsilon_0$ put $\hat{\beta}_j = 0$
- Two drawbacks: Choice of ϵ_0 and definitive exclusion of variables.

LLA Algorithm: Zou and Li (2008)

$$\beta^{(k+1)} = \text{argmax}_\beta \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} J'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right\}. \qquad (5)$$

- The one step LLA estimations are good as estimations obtained after the fully iterative LLA.
- The well known LARS algorithm is used when computing the solution.
- Therefore, as with LASSO (Tibshirani, 1996) there is a problem of selection in the case $p >> n$.

Our contribution: MLLQA Algorithm

$$\beta^{(k+1)} = \text{argmax}_{\beta} \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} \omega_j^1 |\beta_j| - \frac{n}{2} \sum_{j=1}^{p} \omega_{j,\tau}^2 \beta_j^2 \right\}. \tag{6}$$

Where ω_j^1 and $\omega_{j,\tau}^2$ depend on $J_\lambda'(|\beta_j^{(0)}|)$, $|\beta_j^{(0)}|$ and eventually $\tau > 0$.

- $\beta^{(0)}$ is the Maximum Likelihood Estimator.
- The second term is for selection.
- The third one guarantees grouping effect as with the elastic net (Zou and Hastie, 2005).
- For the convergence we prove that MLLQA is an instance of MM algorithms (Hunter and Li 2005).

Augmented data problem

We show that solving problem (6) is equivalent to find:

$$\widehat{\beta} = \text{argmin}_{\beta} \left\{ \frac{1}{2} \parallel Y^{*} - X^{*}\beta \parallel^{2} + n \sum_{j=1}^{p} \omega_{j}^{1} |\beta j|. \right\} \qquad (7)$$

$Y^{*} \in R^{n+p}$, X^{*} of dimension $(n+p) * p$ and (Y^{*}, X^{*}) depend on data (Y, X).

Proposition

Solving the problem (3) via one-step MLLQA algorithm is equivalent to One-step LLA on augmented data.

Oracle and Statistical Properties of the one step MLLQA estimator

Let $\widehat{\beta}(ose)$ be the one-step estimator $\beta^{(1)}$ and β_0 the true model parameter.
Assume $\beta_0 = (\beta_{01}, ..., \beta_{0p})^T = (\beta_{10}^T, \beta_{20}^T)^T$ and $\beta_{20} = 0$. Under some regularity conditions we have the following theorem:

Theorem
If $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$, $\widehat{\beta}(ose)$ is
Sparse: with probability tending to 1, $\widehat{\beta}(ose)_2 = 0$.
Asymptotically normal: $\sqrt{n}(\widehat{\beta}(ose)_1 - \beta_{10}) \to N(0, I_1^{-1}(\beta_{10}))$

- Continuity: the minimum of the function $\mid \beta \mid + J_\lambda'(\mid \beta \mid)$ must be attained at zero (Fan and Li 2001).In the case of one-step it suffices that $J_\lambda'(\mid \beta \mid)$ be continuous for $\mid \beta \mid > 0$ to have the continuity of $\widehat{\beta}(ose)$.

Grouping effect: case of correlated variables

Assume that the response variable is centered and the predictors are standardized. If $|\beta_i^{(0)}| = |\beta_j^{(0)}| \neq 0 \; i, j \in \{1, ..., p\}$ we then have:

1. $D_{\lambda,\tau,\beta^{(0)}}(i,j) \leq \frac{|\beta_j^{(0)}|+\tau}{nJ'_\lambda(|\beta_j^{(0)}|)}\sqrt{2(1-\rho)}$
2. $x_i = x_j \Rightarrow \widehat{\beta}_i = \widehat{\beta}_j$

Where $\rho = x_i^t x_j$ and $D_{\lambda,\tau,\beta^{(0)}}(i,j) = \frac{|\widehat{\beta}_i - \widehat{\beta}_j|}{|Y|_1}$

Linear Model

In this example, simulation data were generated from the linear regression model,

$$y = x^T\beta + \epsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T$, $\epsilon \sim \mathcal{N}(0, 1)$ and x is multivariate normal distribution with zero mean and covariance between the *i*th and *j*th elements being $\rho^{|i-j|}$ with $\rho \in \{0.5, 0.7, 0.9\}$.The sample size is set to be 50 and 100.For each case we repeated the simulation 500 times.

$n = 50$

Method	MRME	No. of C	Zeros IC	Underfit	Proportion of Correctfit	Overfit
			$\rho = .5$			
LLA	0.357	3	2.712	0	0.412	0.588
MLLQA	0.331	3	2.488	0	0.492	0.508
			$\rho = .7$			
LLA	0.437	2.998	2.794	0.002	0.362	0.636
MLLQA	0.383	2.994	2.654	0.006	0.410	0.584
			$\rho = .9$			
LLA	0.616	2.884	2.676	0.116	0.282	0.606
MLLQA	0.579	2.876	2.556	0.124	0.302	0.578

$n = 100$

Method	MRME	No. of C	Zeros IC	Underfit	Proportion of Correctfit	Overfit
			$\rho = .5$			
LLA	0.492	2.998	3.154	0.002	0.460	0.538
MLLQA	0.455	2.998	3.114	0.002	0.482	0.516
			$\rho = .7$			
LLA	0.486	2.998	2.828	0.002	0.480	0.518
MLLQA	0.451	2.998	2.872	0.002	0.490	0.508
			$\rho = .9$			
LLA	0.539	2.946	2.490	0.054	0.394	0.552
MLLQA	0.491	2.944	2.516	0.056	0.412	0.532

Conclusion

- Using convexe approximation of SCAD penalty, we've transformed our initial problem in one-step LLA on augmented data.
- This approach is adapted in the high dimensional setting ($p >> n$).So, allows the selection of more than n variables.
- We considered one-step estimator as final estimation because it's naturally adopt sparse representation and has oracle properties.
- Our approach improves one-step LLA results in the case ($p < n$).

- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R.(2004): Least angle regression. *The annals of statistics (32), 407-499*
- 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*.
- HUNTER, D. and Li, R.(2005): Variable selection using MM algorithms. *The annals of statistics, volume 33, 1617-1642*
- TIBSHIRANI, R.(1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (58), 267-288*.
- ZOU, H. and HASTIE, T.(2005): Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society (67), 301-320*.
- ZOU, H. and LI, R.(2008): One-step sparse estimates in

Thank you for your attention!!!

MERCI DE VOTRE ATTENTION!!!