The Sparse High-Dimensional Linear Discriminant Analysis with Moderately Clipped LASSO

THESIS

Chang Jae Ho

KONKUK UNIVERSITY
Graduate School
Department of Applied Statistics

Supervisor: Dr. Kyusang Yu, Professor

November 28, 2019

Contents

Contents

LDA I

Classical model

Consider a classification problem where

- Predictor vector: $\mathbf{x} = (x_1, \cdots, x_p)'$
- Class label: $C = 1, 2$
- $(\mathbf{x}|C = c) \sim N_p(\mu_c, \Sigma)$ with
  $\mu_c, \Sigma :=$ mean vector of the class c and $p \times p$ covariance matrix
  $P(C = 1) = \pi_1, P(C = 2) = \pi_2$ such that $\pi_1, \pi_2 \in [0, 1]$.

Bayes rule: optimal classifier minimizing the $0 - 1$ loss
Set $\beta^{Bayes} := \Sigma^{-1}\partial$ ($\partial := \mu_2 - \mu_1$) and classify $\mathbf{x}$ to class 2 iff

$$\log \frac{P(C = 2|\mathbf{x})}{P(C = 1|\mathbf{x})} = \log P(\mathbf{x}|C = 2)\pi_2 - \log P(\mathbf{x}|C = 1)\pi_1 > 0$$

$$\Leftrightarrow \{\mathbf{x} - (\mu_1 + \mu_2)/2\}' \beta^{Bayes} + \log(\pi_2/\pi_1) > 0. \tag{1}$$

LDA II

Classical model

- Estimation[1]

  $\hat{\mu}_1 := \sum_{i=1}^{n_1} \mathbf{x}_{1i}/n_1$

  $\hat{\mu}_2 := \sum_{i=1}^{n_2} \mathbf{x}_{2i}/n_2$

  $\hat{\Sigma}$:= the pooled sample covariance estimate of Σ

LDA obtains an estimator $\hat{\beta}^{LDA} := \hat{\Sigma}^{-1}\hat{\partial}$ with $\hat{\partial} = \hat{\mu}_2 - \hat{\mu}_1$.

[1] $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ denote the sample from each class

LDA

Least Squares(LS)

LDA is analogous to **Least Sqaures** problem when $p < n$.
Put $y_i = -n/n_1$ for $C_i = 1$ and $y_i = n/n_2$ for $C_i = 2$, then

$$(\hat{\beta}^{ols}, \hat{\beta}_0^{ols}) = \arg\min_{\beta,\beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i'\beta)^2$$

where $\hat{\beta}^{ols} = constant \times \hat{\beta}^{LDA}$. This indicates that the direction of $\hat{\beta}^{ols}$ is equal to that of $\hat{\beta}^{LDA}$.

High-dimensional LDA

Limitation and Penalized LS

- $\nexists \hat{\Sigma}^{-1}$ When $n \ll p$

Adopt **Penalized Least Squares** estimation;

$$(\hat{\beta}, \hat{\beta}_0) = \underset{\beta, \beta_0}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i' \beta)^2 / n + \sum_{j=1}^{p} J(|\beta_j|) \right\} \tag{2}$$

where J can be selected among sparsity-inducing penalties such as Least Absolute Shrinkage and Selection Operator(LASSO), Minimax Concave Penalty(MCP) and Moderately Clipped LASSO(MCL; Kwon et al., 2015). For $\hat{\partial}' \hat{\beta} > 0$, assign a new observation $\mathbf{x}$ to class 2 if

$$\left(\mathbf{x} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right)' \hat{\beta} + \frac{\hat{\beta}' \hat{\Sigma} \hat{\beta}}{\hat{\partial}' \hat{\beta}} \log(n_2 / n_1) > 0, \tag{3}$$

which was verified as a rule with the closed-form optimal intercept by Mai et al. (2012).

High-dimensional LDA

Why PLS?

Q. Why did they adopt PLS for this problem?

- Does not depend on the invertibility of Σ
- Many optimal computation algorithms
- Theoretical properties (e.g. oracle property, variable selection consistency, and asymptotic normality)
- Computationally NOT sophisticated (compared to other methods)

Contents

Mai et al. (2012) I

Example

LASSOed Sparse LDA

- Low prediction error and variable selection consistency
- Emulates the Bayes rule and outperforms penaliezed Fisher's LDA(Witten and Tibshirani, 2011), ℓ_1-Constrained Quadratic Porgramming with LASSO(Wainwright, 2009), and Wu et al. (2009)
- Simulations + prostate and colon cancer data analysis

Mai et al. (2012) II

The LASSOed estimators asymptotically approach the **oracle LASSO estimator**;

$$\hat{\beta}^{oL,\gamma} := \underset{\beta_A;\ A:=\{j;\beta_j\neq 0\}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i'\beta)^2/n + \sum_{j=1}^{p} |\beta_j|\gamma \right\}$$

The sparse high-dimensional LDA with MCL estimators shows that;

- Close enough to the oracle LASSO estimator
- Performs better than the LASSOed LDA

Contents

Background I

Moderately Clipped LASSO

- Precedent research on the nonconvex penalized estimators' *oracle property*; the asymptotic equivalence with the oracle estimator
- Convex penalties(LASSO) and their derivatives lost variable selection consistency under the violation of strong irrepresentability (SI) condition
- Smoothly Clipped Absolute Deviation (SCAD; Fan et al., 2001, Kim et al., 2008); asymptotically attains an **oracle LSE** and **nearly unbiasedness** by smoothly clipping LASSO
- Minimax Concave Penalty (MCP; Zhang, 2010) preserves convexity better than SCAD while achieving an oracle LSE

Background II

Moderately Clipped LASSO

- Prediction accuracy: LSE $<$ LASSO
- Non-convex penalized estimators' predictions are mostly inferior to those of LASSO for the finite-size samples.
- When the sample size is small, the non-convex penalties fail when true β is small.
- The MCL penalty was devised in this context to combine both LASSO and MCP seeking to achieve both advantages of LASSO and MCP.

Definition I

Moderately Clipped LASSO

The MCL-penalized estimator $(\hat{\beta}^{cL}, \hat{\beta}_0^{cL})$, proposed by Kwon et al. (2015), is a minimizer defined as (2) where J is a penalty function induced by its derivative:

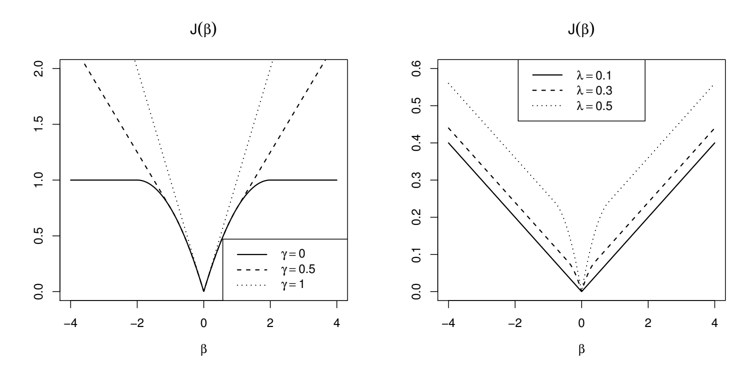
$$\begin{cases} \frac{d}{dt} J_{\gamma,\lambda}(|t|) & = \max(\gamma, \lambda - |t|/\tau) \qquad \forall t \in \mathbb{R} \sim \{0\} \\ J_{\gamma,\lambda}(0) & = 0 \end{cases}$$

for $\tau > 1, 0 \leq \gamma \leq \lambda$.

Definition II

Moderately Clipped LASSO

Figure: The penalty functions. $(\lambda, \tau) = (1, 2)$ and $(\gamma, \tau) = (0.1, 2)$ respectively.



Definition III

Moderately Clipped LASSO

There are two different regularization parameters γ and λ.

- λ controls the concavity of the MCL near the origin and the overall sparsity of the estimators.
- γ regularizes a shrinkage of nonzero estimators so and also controls the sparsity(Kwon et al., 2015).
- The MCL penalty with $\gamma = 0$ and $\gamma = \lambda$ each obtains an MCP and a LASSO penalty.
- Fix γ properly with a reasonable estimator and choose the optimal λ for the given γ in the estimation process.

Computation Algorithm[2]

CCCP

The MCL penalty can be decomposed as $J_{\gamma,\lambda}(|t|) = L_{\gamma,\lambda}(t) + \lambda|t|$ where $L_{\gamma,\lambda}$ is a continuously differentiable concave function:

$$\frac{d}{dt}L_{\gamma,\lambda}(t) = \frac{d}{dt}J_{\gamma,\lambda}(t) - \lambda = -\max(\lambda - \gamma, t/\tau) < 0$$

for $t > 0$ and $L_{\gamma,\lambda}(0) = J_{\gamma,\lambda}(0)$ (the image of $J_{\gamma,\lambda}$ for $t \neq 0$ is symmetric about $t = 0$).

- Convex-ConCave Procedure (CCCP; Yuille and Rangarajan, 2003) and an iterative way such as pathwise coordinate descent algorithm (Friedman et al., 2007) can find a local minimizer
- The package `ncpen` programmed in `R` (Kim et al., 2018) was used in the numerical study section

[2]The efficient computation algorithm for the MCL estimation had already been discussed by Kwon et al. (2015).

Contents

Terms I

Terms and Definitions

For an arbitrary $m \times n$ matrix $\mathbf{P}$ and some vector $\mathbf{a} \in \mathbb{R}^p$, define

$$|\mathbf{P}|_\infty = \max_{i=1,\cdots,m} \sum_{j=1}^{n} |P_{ij}| \tag{4}$$

$$|\mathbf{a}|_\infty = \max_{j=1,\cdots,p} |a_j| \tag{5}$$

$$|\mathbf{a}|_{\min} = \min_{j=1,\cdots,p} |a_j|$$

and define $A = \{j; \beta_j^{Bayes} \neq 0\}$, $s = |A|$, $N = A^c$. Call β_A^{Bayes} as discriminative variable(s) since it is a Bayes classification direction.

Terms II

Terms and Definitions

Let the marginal covariance matrix of $\mathbf{x}$ be $\Omega := V(\mathbf{x}) = \Sigma + V(\mu_y)$ and partition it as $\Omega = \begin{pmatrix} \Omega_{AA} & \Omega_{AN} \\ \Omega_{NA} & \Omega_{NN} \end{pmatrix}$ and define

$$\kappa = |\Omega_{NA}(\Omega_{AA})^{-1}|_\infty$$
$$\varphi = |(\Omega_{AA})^{-1}|_\infty$$
$$\psi = |\partial_A|_\infty.$$

Terms III

Terms and Definitions

Define $\mathbf{Z} := (\mathbf{I}_n - \Pi_{\mathbf{1}_n})\mathbf{X}$ and $\Pi_{\mathbf{a}} := \mathbf{a}(\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}'$ and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$ for $\mathbf{X}_j = (x_{1j}, \cdots, x_{nj})'$ $(j = 1, \cdots, p)$.
Then, the MCL estimator is defined as ;

$$(\hat{\beta}^{\gamma,\lambda}, \hat{\beta}_0^{\gamma,\lambda}) = \arg\min_{\beta,\beta_0} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i'\beta)^2/n + \sum_{j=1}^{p} J_{\gamma,\lambda}(|\beta_j|) \right\}$$

which can derive a partial estimation while plugging in the centered design matrix;

$$\hat{\beta}^{\gamma,\lambda} = \arg\min_{\beta}\{Q^{\gamma,\lambda}(\beta)\} \text{ where}$$

$$Q^{\gamma,\lambda}(\beta) := \frac{\beta'\mathbf{Z}'\mathbf{Z}\beta}{n} - 2\hat{\partial}'\beta + \sum_{j=1}^{p} J_{\gamma,\lambda}(|\beta_j|). \tag{6}$$

Terms IV

Terms and Definitions

Let $\hat{\Omega} := \mathbf{Z}'\mathbf{Z}/n$ so $\hat{\Omega}_{AA} = \mathbf{Z}'_A\mathbf{Z}_A/n$ and $\hat{\Omega}_{NA} = \mathbf{Z}'_N\mathbf{Z}_A/n$.
Denote $(\Omega_{AA})^{-1}\partial_A$ by $\tilde{\beta}_A^{Bayes}$ so that $\tilde{\beta}_N^{Bayes} := \mathbf{0}$.

Verifying the claim that suggested sparse discriminant anaysis constructed on $\hat{\beta}^{\gamma,\hat{\partial}}$ can successfully recover the support of this quantity and estimating $\tilde{\beta}^{Bayes}$ are sufficient for us in the sense that $\tilde{\beta}^{Bayes} = c\beta^{Bayes}$ for some constant c as mentioned in Mai et al. (2012).

Put $\hat{\Delta}_{AA} = \hat{\Omega}_{AA} - \Omega_{AA}$, $\hat{\Lambda}_{NA} = \hat{\Omega}_{NA}(\hat{\Omega}_{AA})^{-1} - \Omega_{NA}(\Omega_{AA})^{-1}$, $\hat{\delta}_A = \hat{\partial}_A - \partial_A$ and $\hat{\delta} = \hat{\partial} - \partial$ for the following lemmas.

Non-asymptotic properties I

Tail probabilities and Karush Kuhn Tucker conditions

Lemma

There exist positive constants ϵ_0 and c_1, c_2 such that for any $\epsilon \leq \epsilon_0$ we have

$$P\left(|\hat{\Delta}_{AA}|_\infty \geq \epsilon\right) \leq 2s^2 \exp(-nc_1\epsilon^2/s^2)$$
$$P(|\hat{\delta}_A|_\infty \geq \epsilon) \leq 2s \exp(-nc_2\epsilon^2)$$
$$P(|\hat{\delta}|_\infty \geq \epsilon) \leq 2p \exp(-nc_2\epsilon^2)$$

Non-asymptotic properties II

Tail probabilities and Karush Kuhn Tucker conditions

Lemma

There exist constants ϵ_0, c_1 such that for any $\epsilon \leq \min(\epsilon_0, 1/\varphi)$, we have

$$P\left[|\hat{\Lambda}_{NA}|_\infty \geq \frac{\epsilon\varphi(\kappa+1)}{1-\varphi\epsilon}\right] \leq 2ps\exp(-nc_1\epsilon^2/s^2).$$

ε_0 depends on $\zeta \leq \lambda_{\max}(\Omega)^{-1}$ only.

The detailed proof of Lemma 2 and 3 are given in Mai et al. (2012).

An oracle lasso estimator $\hat{\beta}^{oL,\gamma}$ is the oracle MCL estimator with $\gamma = \lambda$.

Let me claim that $\hat{\beta}^{oL,\gamma}$ is a member of (6).

Non-asymptotic properties III

Tail probabilities and Karush Kuhn Tucker conditions

Lemma

Define $\Xi^{\gamma,\lambda}$ as the set of all local minimizers of (6). Then, $\hat{\beta} \in \Xi^{\gamma,\lambda}$ if $\hat{\beta}$ satisfies

$$\hat{\partial}_S - \mathbf{Z}_S'\mathbf{Z}\hat{\beta}/n = \frac{\gamma}{2}\mathbf{sign}(\hat{\beta}_S), \tag{7}$$

$$|\hat{\beta}_S|_{\min} > \tau(\lambda - \gamma),$$

$$|\hat{\partial}_{S^c} - \mathbf{Z}_{S^c}'\mathbf{Z}\hat{\beta}/n|_\infty \leq \frac{\lambda}{2} \tag{8}$$

where $S := \{j; \hat{\beta}_j \neq 0\}$.

Non-asymptotic properties IV

Tail probabilities and Karush Kuhn Tucker conditions

Remark In addition, conditions (7) and (8) are equivalent with

$$\hat{\partial}_S - \hat{\Omega}_{SS}\hat{\beta}_S = \frac{\gamma}{2}\textbf{sign}(\hat{\beta}_S),$$
$$|\hat{\partial}_{S^c} - \hat{\Omega}_{S^cS}\hat{\beta}_S|_\infty \leq \frac{\jmath}{2}.$$

Since Mai et al. (2012) proved that $\hat{\beta}_A^{oL,\gamma}$ is asymptotically close to $\tilde{\beta}_A^{Bayes}$ and (3) can recover the support of $\tilde{\beta}^{Bayes}$, I will show that $\hat{\beta}^{\gamma,\jmath}$ can asymptotically obtain $\hat{\beta}^{oL,\gamma}$.

Non-asymptotic properties V

Tail probabilities and Karush Kuhn Tucker conditions

Theorem

Let $m_A := |\tilde{\beta}_A^{Bayes}|_{min}$. *Pick any* τ, γ *and* $\daleth$ *so* $\daleth > \gamma \times \max(1, \kappa) \geq 0$ *and* $m_A > \tau(\daleth - \gamma) + \varphi\gamma/2$. *then,*

$$P\left(\hat{\beta}^{oL,\gamma} \in \Xi^{\gamma,\daleth}\right) \geq 1 - 2p \exp\left[-nc_2 \left\{\frac{\daleth(1 - \varphi\epsilon_1) - \gamma(\kappa + \varphi\epsilon_1)}{4(1+\kappa)}\right\}^2\right]$$
$$- 2ps \exp(-c_1 n\epsilon_1^2/s^2) - 2s^3 \exp(-c_1 n\epsilon_2^2/s^2) - 2s^2 \exp(-c_2 n\epsilon_2^2)$$

for any positive $\epsilon_1 < \min\left[\epsilon_0, 1/\varphi, \frac{\daleth - \kappa\gamma}{\varphi\{4\psi(1+\kappa)+\daleth+\gamma\}}\right]$ *and*
$\epsilon_2 < \min\left[\epsilon_0, \frac{m_A - \tau(\daleth-\gamma)-\varphi\gamma/2}{\varphi\{1+\psi\varphi+m_A-\tau(\daleth-\gamma)\}}\right]$.

Asymptotic property I

with specified orders

Corollary

For $\lambda = \lambda_n$ *and* $\gamma = \gamma_n$, $P(\hat{\beta}^{oL,\gamma} \in \Xi^{\gamma,\lambda}) \to 1$ *under;*

$$\begin{aligned} &\textit{Condition } 1.\ \ n, p \to \infty \ \ \textit{and} \ \ \log(ps)s^2 = o(n), \\ &\textit{Condition } 2.\ \ \sqrt{\frac{\log(ps)s^2}{n}} \ll \lambda \ll m_A \ \ \textit{and} \ \ \gamma = o(\lambda), \\ &\textit{Condition } 3.\ \ \psi, \kappa \ \ \textit{and} \ \ \varphi \ \ \textit{are} \ \ \textit{constants}. \end{aligned}$$

First condition poses a joint restriction on n and p.
For example, as mentioned in Mai et al. (2012), it holds as long as $p \ll \exp(n^{2a})$ for $a < 1/2$ so p is allowed to grow faster than any polynomial order of n (nonpolynomial-dimension asymptotics).

Asymptotic property II

with specified orders

In second condition, since $\gamma \leq \daleth$ by the definition of the MCL, the order of m_A is specified to consistently seperate the discriminative set from the non-discriminative one.

- Mai et al. (2012) included an irrepresentability condition[3]; the value of κ to should be less than one.
- This was relaxed in here: we can allow $\kappa > 1$ under the values of γ and $\daleth$ satisfying $\kappa \leq \daleth/\gamma$.

[3] Irrepresentability condition is a sufficient condition for the asymptotic properties of $\hat{\beta}$ (oracle property & variable selection consistency) to satisfy (Meinshausen and Bühlmann, 2006; Zou, 2006; Zhao and Yu, 2006; Wainwright, 2009; Mai et al., 2012).

Contents

Simulation I

Design and results

$n_j \overset{iid}{\sim} Bin(n, \pi_1)$ $(j = 1, \cdots 6)$ given prior probabilities $\pi_1 = \pi_2 = 1/2$.
$\mathbf{x}_1|c_1, \cdots, \mathbf{x}_n|c_n \overset{iid}{\sim} N_p(\mu_c, \Sigma)$.
WLOG, set $\mu_1 = \mathbf{0}$ so that $\mu_2 = \Sigma\beta^{Bayes}$ where β^{Bayes} represents the Bayes classification direction. The values of β^{Bayes} and Σ satisfies $\kappa > 1$, **violating the irrepresentability condition**.

Table: Simulation Settings

Model number	n_j	p	Σ	β^{Bayes}
$j = 1, \cdots, 6$	$100 \times 2^{j-1}$	400	$\Sigma_{rc} = \begin{cases} 0.5^{I(r \neq c)} & r \in A, c \in A \\ 0.5^{\lvert r-c \rvert} & o.w. \end{cases}$ $(r, c = 1, \cdots p)$	$0.556 \begin{pmatrix} 3 \\ 1.5 \\ 0 \\ 0 \\ 2 \\ \mathbf{0}_{p-5} \end{pmatrix}$

Simulation II

Design and results

- LASSO and the MCP (with $\tau = 2$) estimators obtained from training data of size n_j
- Tuning parameters were selected by 5-fold cross-validation based on Root Mean Squared Error (RMSE).
- For the MCL, $\tau = 2$ and estimators for γ: $1.3^k \hat{\gamma}$ with $k = -4, -3, \cdots, 3, 4$, where $\hat{\gamma}$ is the optimally selected value of λ for the LASSO estimator.
- Tuned λ through 5-fold cross-validation to ensure a better prediction accuracy.

For the prediction assessment, I generated new test data of size $2n$. 200 replications were conducted using medians and standard errors of each measure.

Simulation III

Design and results

Table: Simulation Results

Model		MCP					MCL$_\gamma$					LASSO
			$1.3^{-4}\hat{\gamma}$	$1.3^{-3}\hat{\gamma}$	$1.3^{-2}\hat{\gamma}$	$1.3^{-1}\hat{\gamma}$	$\hat{\gamma}$	$1.3^{1}\hat{\gamma}$	$1.3^{2}\hat{\gamma}$	$1.3^{3}\hat{\gamma}$	$1.3^{4}\hat{\gamma}$	
1	Error (%)	11	11	10.5	10	9	8	8	7.5	7.5	8.5	7.5
	s.e.	(0.029)	(0.031)	(0.031)	(0.028)	(0.027)	(0.024)	(0.023)	(0.024)	(0.024)	(0.025)	(0.022)
	True.Sel.	1	1	1	1	2	2	2	2	2	2	3
	s.e.	(0.071)	(0.57)	(0.636)	(0.746)	(0.734)	(0.73)	(0.688)	(0.668)	(0.704)	(0.741)	(0.353)
	False.Sel.	0	0	1	1	1	1	0	0	0	0	5.5
	s.e.	(1.692)	(11.062)	(12.483)	(12.775)	(10.725)	(8.873)	(6.952)	(3.579)	(1.942)	(0.676)	(11.504)
2	Error (%)	9.25	8.75	8.375	8	8	7.5	7.25	7.25	7.25	7.5	7.25
	s.e.	(0.021)	(0.02)	(0.019)	(0.021)	(0.019)	(0.017)	(0.016)	(0.014)	(0.015)	(0.015)	(0.014)
	True.Sel.	1	2	2	2	2	3	3	3	3	3	3
	s.e.	(0.419)	(0.523)	(0.565)	(0.605)	(0.649)	(0.56)	(0.476)	(0.461)	(0.456)	(0.524)	(0.1)
	False.Sel.	1	1	1	1	2	2	1	0	0	0	8
	s.e.	(2.835)	(5.523)	(6.837)	(9.003)	(9.699)	(8.573)	(4.504)	(1.906)	(0.73)	(0.198)	(11.804)
3	Error (%)	8.062	7.625	7.5	7.375	7.125	7	7	6.875	6.875	7	7
	s.e.	(0.014)	(0.011)	(0.011)	(0.011)	(0.01)	(0.01)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
	True.Sel.	2	2	2	2	3	3	3	3	3	3	3
	s.e.	(0.474)	(0.472)	(0.549)	(0.587)	(0.495)	(0.371)	(0.32)	(0.232)	(0.186)	(0.196)	(0)
	False.Sel.	2	2	2	2	1	1	0	0	0	0	10
	s.e.	(3.555)	(3.852)	(4.721)	(4.678)	(8.041)	(9.289)	(3.464)	(0.962)	(0.3)	(0.157)	(11.062)

Simulation IV

Design and results

Model		MCP	MCL_γ									LASSO
			$1.3^{-4}\hat{\gamma}$	$1.3^{-3}\hat{\gamma}$	$1.3^{-2}\hat{\gamma}$	$1.3^{-1}\hat{\gamma}$	$\hat{\gamma}$	$1.3^{1}\hat{\gamma}$	$1.3^{2}\hat{\gamma}$	$1.3^{3}\hat{\gamma}$	$1.3^{4}\hat{\gamma}$	
4	Error (%)	7.469	7.125	7.063	7	6.937	6.937	6.937	6.937	6.937	6.937	7
	s.e.	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.007)
	True.Sel.	2	3	3	3	3	3	3	3	3	3	3
	s.e.	(0.476)	(0.511)	(0.453)	(0.397)	(0.332)	(0.238)	(0.157)	(0.071)	(0)	(0.071)	(0)
	False.Sel.	3	3	2	1	1	0.5	0	0	0	0	10
	s.e.	(4.102)	(3.387)	(3.143)	(3.308)	(3.906)	(6.904)	(3.491)	(0.908)	(0.258)	(0.071)	(10.367)
5	Error (%)	7.094	6.906	6.875	6.875	6.844	6.844	6.844	6.828	6.812	6.812	6.844
	s.e.	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
	True.Sel.	3	3	3	3	3	3	3	3	3	3	3
	s.e.	(0.491)	(0.401)	(0.368)	(0.332)	(0.264)	(0.196)	(0.157)	(0.071)	(0)	(0)	(0)
	False.Sel.	2	0	0	0	0	0	0	0	0	0	16
	s.e.	(2.941)	(1.813)	(1.579)	(1.785)	(4.906)	(8.862)	(4.028)	(0.943)	(0.198)	(0)	(12.005)
6	Error (%)	6.93	6.859	6.859	6.859	6.844	6.844	6.812	6.812	6.812	6.797	6.844
	s.e.	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	True.Sel.	3	3	3	3	3	3	3	3	3	3	3
	s.e.	(0.448)	(0.372)	(0.337)	(0.32)	(0.264)	(0.208)	(0.122)	(0.071)	(0)	(0)	(0)
	False.Sel.	0	0	0	0	0	0	1	0	0	0	21.5
	s.e.	(1.495)	(2.375)	(4.047)	(4.71)	(7.728)	(9.494)	(4.968)	(1.109)	(0.196)	(0)	(11.578)

Simulation V

Design and results

- The LASSO has a great accuracy overall yet shows the largest counts of irrelevant variable selection.
- MCL: selected the true variables and excluded the non-discriminant variables from the model, and even better prediction accuracy than LASSO
- The MCL with $\gamma = 1.3^3\hat{\gamma} \approx 2.2\hat{\gamma}$: the best accuracy except model 6 in which ranked the second accuracy by the difference of 0.015%.
- The LASSO sensed all three discriminative variables but numbers of the false selection tended to increase; 5.5, 8, 10, 10, 16 and 21.5 false selections.
- MCP also succeeded to show variable selection consistency.

Simulation VI

Design and results

- Boxplots of estimators and $\tilde{\beta}_A^{Bayes}$ (dotted lines)

Estimation

- Sample size $\uparrow \;\Rightarrow\;$ center-concentrated estimators
- $\hat{\beta}_A^{\hat{\gamma},\hat{\lambda}^{CV}} \approx \hat{\beta}_A^{\hat{\gamma},\hat{\gamma}}$ for all models
- The MCL tended to underestimate as the value of γ increases because the estimation became sparser.

Contents

Prostate cancer data analysis I

Table: Prostate data analysis, $\tau = 2$ for MCP and MCL.

	MCP	MCL$_\gamma$									LASSO
		$1.3^{-4}\hat{\gamma}$	$1.3^{-3}\hat{\gamma}$	$1.3^{-2}\hat{\gamma}$	$1.3^{-1}\hat{\gamma}$	$\hat{\gamma}$	$1.3^{1}\hat{\gamma}$	$1.3^{2}\hat{\gamma}$	$1.3^{3}\hat{\gamma}$	$1.3^{4}\hat{\gamma}$	
Error (%)	22.857	8.571	8.571	8.571	8.571	11.429	8.571	8.571	11.429	11.429	8.571
s.e.	(0.076)	(0.044)	(0.046)	(0.047)	(0.048)	(0.048)	(0.051)	(0.052)	(0.06)	(0.064)	(0.044)
fit size	3	25	23	21	19	17	15	14	13	12	50
s.e.	(2.119)	(5.293)	(4.977)	(4.655)	(4.576)	(4.672)	(4.609)	(4.566)	(5.778)	(6.19)	(5.573)

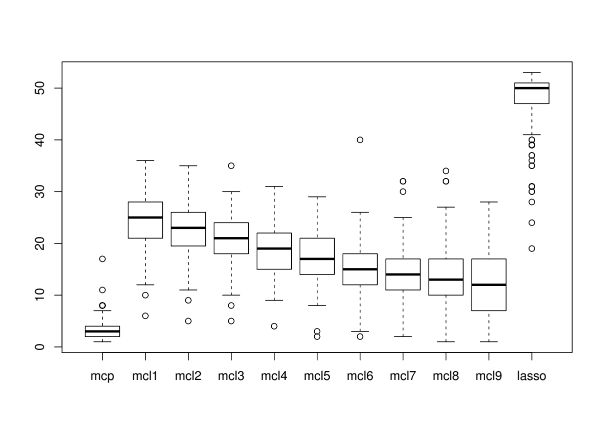
Prostate cancer data analysis II

- Prostate tumor gene expression dataset (Singh et al., 2002); available on `R` package `sda`
- train to test data split $= 2:1$
- The MCL with $\gamma = 1.3^2\hat{\gamma} \approx 1.7\hat{\gamma}$ showed the best performance.
- Even though the larger value of γ can make a simpler model possible, it will cost a large amount of accuracy loss.
- LASSO also predicted accurately but failed to narrow the pool of variables and simplify the model.

Prostate cancer data analysis III

Figure: Boxplots of the fit size

`mcl1` through `mcl9` indicate MCL with $\gamma = 1.3^{-4}\hat{\gamma}, \cdots, 1.3^{4}\hat{\gamma}$



Conclusions

- With the control of MCL regularization parameter added with the tuning parameter, it is possible searching for the optimal set of estimators.
- MCL is successful in the application of sparse LDA; the concave penalty is the breakthrough in relaxing the irrepresentability condition.
- MCL can simultaneously obtain great prediction accuracy and select true variables.
- Numerical studies showed that the MCL can select the value of γ properly using the optimal tuning parameter for the LASSO.

Conclusions

End of the presentation

Thank You

References

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). *Pathwise coordinate optimization*. The annals of applied statistics, 1(2), 302-332.

Kim, D., Lee, S., & Kwon, S. (2018). A unified algorithm for the non-convex penalized estimation: The ncpen package. *arXiv preprint arXiv:1811.05061.*

Kim, Y., Choi, H., & Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484), 1665-1673.

Kwon, S., Lee, S., & Kim, Y. (2015). Moderately clipped LASSO. *Computational Statistics & Data Analysis*, 92, 53-67.

Lee, S. & Kwon, S. (2015) Moderately Clipped LASSO for the Sparse High-Dimensional Logistic Regression Models. *Journal of the Korean Data Analysis Society*, 17, 1145-1154.

Mai, Q., Zou, H., & Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1), 29-42.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3), 1436-1462.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2), 203-209.

Sun, X., Choi, H., & Kwon, S. (2014) A Sparse Ridge Estimation for the Sparse Logistic Regression Model. *Journal of the Korean Data Analysis Society*, 16(4), 1715-1725.

References

Wainwright, M. J. (2009). Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5), 2183-2202.

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 753-772.

Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C., & Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9), 1145-1151.

Yuille, A. L., & Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15(4), 915-936.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2), 894-942.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov), 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.