High-dimensional Linear Discriminant Analysis with Moderately Clipped LASSO [1]

Jaeho Chang*, Haeseong Moon, and Sunghoon Kwon

Department of Applied Statistics, Konkuk University

**jaehochang@konkuk.ac.kr*

May 12, 2021

[1]Publication available at https://doi.org/10.29220/CSAM.2021.28.1.021.

Overview

1. Introduction
   - Linear discriminant analysis
   - Literature review
2. Main Results
   - Clipped LASSO
   - Theory
   - Simulation studies
   - Real Data Analysis
3. Concluding remarks

LDA: Bayes discriminant rule

- Bayes classifier

$$\phi^{\text{Bayes}}(\mathbf{x}) = \arg\max_{c\in\{1,2\}} \mathbf{P}(C = c|\mathbf{X} = \mathbf{x})$$

  - $\mathbf{X} \in \mathbb{R}^p$, $C \in \{1, 2\}$ are random
- The linear discriminant analysis (LDA) (Fisher, 1936)

$$\left(\mathbf{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\right)^T \boldsymbol{\beta}^{\text{Bayes}} + \log(\pi_2/\pi_1) > 0,$$

  - $\mathbf{X}|C = c \sim \mathcal{N}_p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}),\ c \in \{1, 2\}$
  - $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}$ are mean and covariance
  - $\pi_1 + \pi_2 = 1$; fixed class probabilities
  - $\boldsymbol{\beta}^{\text{Bayes}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}$ ($\boldsymbol{\theta} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$) is the Bayes direction vector

LDA: Estimation

- LDA direction vector

$$\hat{\beta}^{\mathrm{LDA}} = \hat{\mathbf{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}, \tag{1}$$

  - $\hat{\mathbf{\Sigma}}$ is the pooled sample covariance matrix and
  - $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$

Linear discriminant rule: assign $C|\mathbf{X} = \mathbf{x}$ to the class 2 if $\mathbf{x}$ satisfies

$$\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^T \hat{\beta}^{\mathrm{LDA}} + \log(n_2/n_1) > 0, \tag{2}$$

where $n_1 = |\{c_i; c_i = 1\}|$ and $n_2 = |\{c_i; c_i = 2\}|$ for $i = 1, \cdots, n$

What if $p > n$?

- Direct modification of the covariance (Krzanowski et al., 1995), (Bickel et al., 2004)
- Constructing relevant shrunken centroid means Guo et al. (2006); Fan and Fan (2008); Wu et al. (2009); Cai and Liu (2011); Witten and Tibshirani (2011); Clemmensen et al. (2011); Shao et al. (2011)
- Penalized LSE (Mai et al., 2012; Tibshirani, 1996) motivated by Hastie et al. (2009)

Connection between LS and LDA

- LSE for the LDA problem (Hastie et al., 2009)

$$(\hat{\alpha}^{\mathrm{LSE}}, \hat{\boldsymbol{\beta}}^{\mathrm{LSE}}) = \underset{\alpha,\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^T \boldsymbol{\beta})^2 / 2n \qquad (3)$$

  - $\hat{\beta}^{\mathrm{LSE}} = (\mathbf{Z}^T \mathbf{Z}/n)^{-1} \hat{\boldsymbol{\theta}}$ for the centered desgin matrix $\mathbf{Z}$
  - $y_i = (-1)^{c_i} n/n_{c_i}, i \leqslant n$.
- relation

$$\hat{\boldsymbol{\beta}}^{\mathrm{LSE}} = c \hat{\boldsymbol{\beta}}^{\mathrm{LDA}} \qquad (4)$$

- ... and the discriminant rule becomes

$$\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2} \right)^T \hat{\boldsymbol{\beta}}^{\mathrm{LSE}} + c \log(n_2/n_1) > 0 \qquad (5)$$

Penalized LSE

(1) and (4) fail to hold when $p > n$ and $\nexists \hat{\Sigma}^{-1}$

Q. How to estimate β^{Bayes}?

Mai et al. (2012) used the least absolute shrinkage and selection operator (LASSO):

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

for some $\lambda > 0$, where the solution can be well defined even when $p > n$.

Moderately Clipped LASSO

- Clipped LASSO (Kwon et al., 2015)

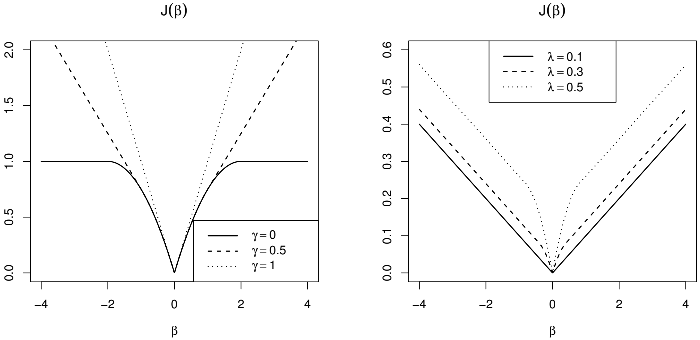
$$\begin{cases} J_{\gamma,\lambda}(0) = 0 \\ \nabla J_{\gamma,\lambda}(t) = \lambda - t/a, & t < a(\lambda - \gamma) \\ \nabla J_{\gamma,\lambda}(t) = \gamma, & o.w. \end{cases}$$

for $0 \leqslant \gamma \leqslant \lambda$

- property
  1. $J_{\lambda,\lambda}$ attains LASSO & $J_{0,\lambda}$ attains MCP
  2. λ mainly controls the concavity of the penalty near the origin for like the MCP
  3. γ regularizes the amount of shrinkage non-zero regression coefficients like the LASSO

Behavior

Figure: Various shapes of the MCL with $a = 2$: $\lambda = 1$ for the left panel and $\gamma = 0.1$ for the right panel.



What can we expect from the MCL?

- Kwon et al. (2015)
- Correct model selection + good prediction accuracy
- Does not perfectly outperform LASSO or MCP
  BUT enjoys the advantages of the both
- The MCL estimator for β^{Bayes}=oracle[2] LASSO estimator with probability tending to one

[2] a theoretically optimal estimator obtained by using the signal variables only

Oracle Property I

$Q_{\gamma,\lambda}(\beta)$: LS-type loss penalized with $J_{\gamma,\lambda}$
$\Xi^{\kappa}_{\gamma,\lambda} := \{\beta \in \{\text{all local minimizers of } Q_{\gamma,\lambda}\}; |\mathrm{supp}(\beta)| \leqslant \kappa\}.$

Theorem

Under some conditions,

$$\lim_{n\to\infty} \mathbf{P}\left(\{\hat{\beta}^{\text{Oracle LASSO},\gamma}\} = \Xi^{\kappa}_{\gamma,\lambda}\right) = 1.$$

That is, the oracle LASSO estimator is the unique minimizer of $Q_{\gamma,\lambda}$ with probability tending to one.

Oracle Property II

- This also holds for $J_{0,\lambda}$ (MCP)
  So the oracle LSE becomes the unique minimizer of $Q_{\gamma,\lambda}$
- linear regression

$$m_{\mathcal{A}} \gg \lambda \gg \sqrt{\log p/n}$$
$$for \;\; m_{\mathcal{A}} := \min_{j\in\mathcal{A}} |\beta_j^{\text{Bayes}}|,$$

- LDA

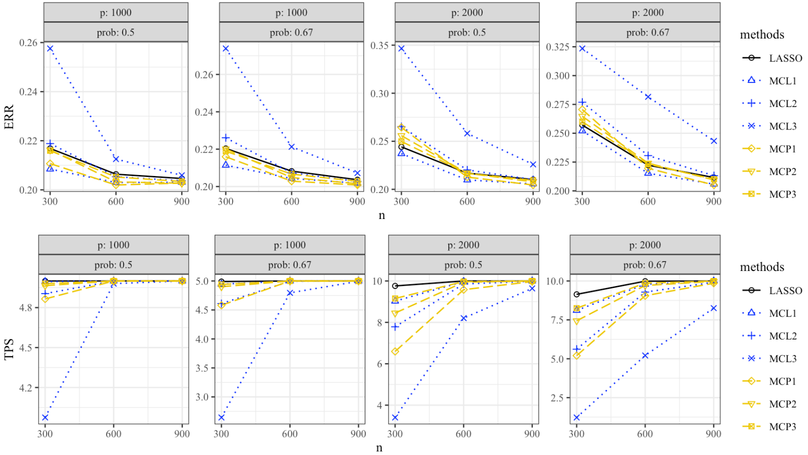
$$m_{\mathcal{A}} \gg \lambda \gg q\sqrt{\log p/n}$$

if $n \gg \kappa^3 \log p$, where $a \gg b$ implies $a/b \to \infty$ as $n \to \infty$.

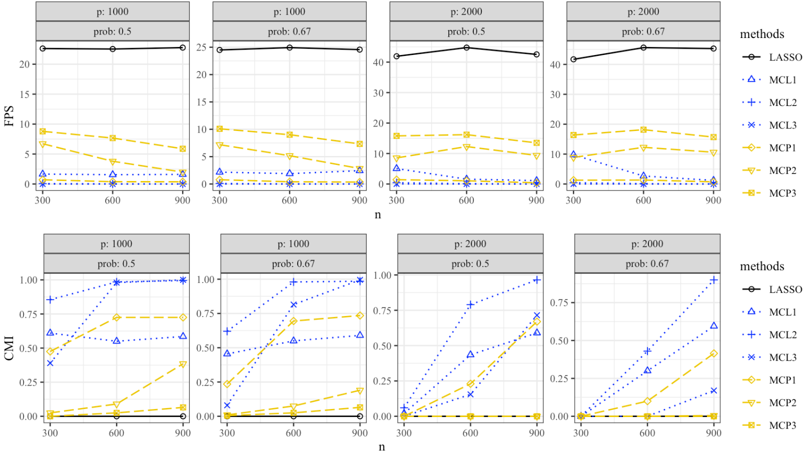
Simulations

- simulation design
  - $\pi_1 = 0.5$ and $\pi_1 = 0.67$
  - $\mathbf{\Sigma} = \mathbf{\Sigma}^{(i)}, i \leqslant 2$:
    1. $\mathbf{\Sigma}^{(1)} = \mathbf{I}_p$ (independent)
    2. $\Sigma^{(2)}_{jk} = 0.5^{|j-k|}$ for $j, k \leqslant p$ (power decaying correlation)
  - $\beta_j^{\text{Bayes}} = \alpha(-1)^j I(j \leqslant q), j \leqslant p$
    α is calculated so that all Bayes error rates to be 0.2
    $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = \mathbf{\Sigma}\boldsymbol{\beta}^{\text{Bayes}}$
  - $n \in \{300, 600, 900\}$, $(p, q) \in \{(1000, 5), (2000, 10)\}$
- estimators: for $k \in \{1, 2, 3\}$,
  - MCL_k with $a = 2.1$ and $\gamma = k\hat{\lambda}^{\text{LASSO}}$ (Kwon et al., 2015)
  - MCP_k with $a = k + 0.1$, and $\gamma = 0$.
  - LASSO

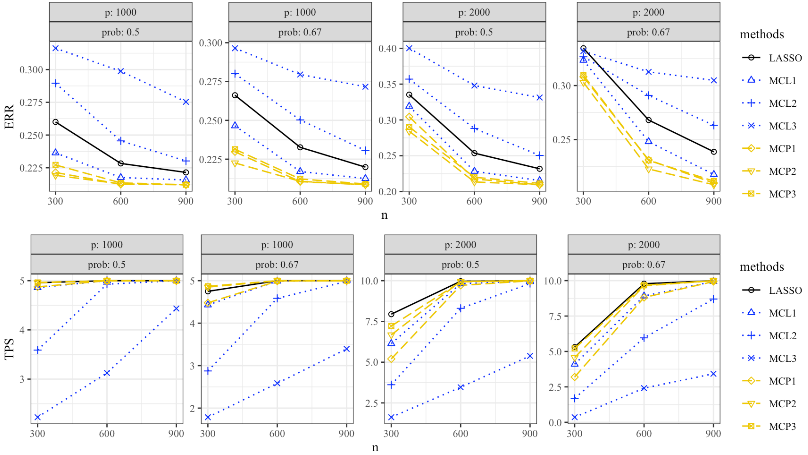
$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)}$: ERR, TPS



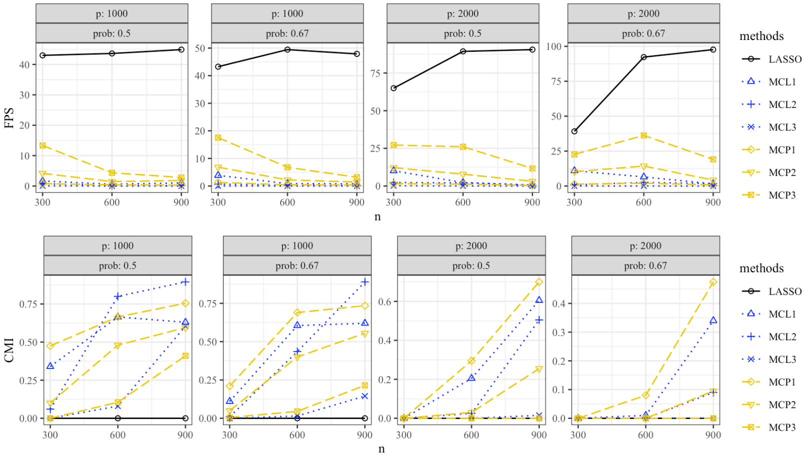
$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)}$: FPS, CMI



$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(2)}$: ERR, TPS



$\mathbf{\Sigma} = \mathbf{\Sigma}^{(2)}$: FPS, CMI



Summary

The MCL can be a nice alternative to the LASSO for the high-dimensional penalized LDA

- can correctly identify the sparse Bayes direction vector
- keeps almost the same prediction accuracy as the LASSO

Remark

MCL_1 performed quite well regardless of the simulation designs considered; this aligns with the recommendation by Kwon et al. (2015)

Analysis of micro-array samples I

HD microarray data sets in R (John, 2016): Burczynski et al. (2006); Chin et al. (2006); Chowdary et al. (2006); Gordon et al. (2002)

Sure independence screening procedure (Fan et al., 2010): used first top d predictive variables with the largest marginal regression coefficients

For comparison,

- tuning via the 10-fold CV & prediction via the leave-one-out CV
- counted the number of incorrectly classified samples (error)
- calculated the average of fit sizes (sizes)

Analysis of micro-array samples II

	d	LASSO	MCL_1	MCL_2	MCL_3	MCP_1	MCP_2	MCP_3
				Chowdary et al. (2006), $n = 104$, $p = 22283$				
errors	400	1	1	2	2	4	4	5
sizes	400	36.28	28.50	24.78	21.84	11.44	11.60	12.33
				Gordon et al. (2002), $n = 181$, $p = 12533$				
errors	400	2	2	2	1	3	3	3
sizes	400	41.28	37.33	24.41	16.71	7.17	6.79	7.03
				Burczynski et al. (2006), $n = 127$, $p = 22283$				
errors	400	6	6	7	9	15	13	15
sizes	400	47.67	41.50	26.19	24.06	13.70	13.81	13.13
				Chin et al. (2006), $n = 118$, $p = 22215$				
errors	400	12	11	13	15	21	22	22
sizes	400	33.62	27.28	21.53	15.16	9.74	9.69	9.61

Analysis of micro-array samples III

- LASSO: highest prediction accuracy BUT largest fit sizes
- MCPs: lowest prediction accuracy BUT smallest fit sizes
- $d = 400$: MCL_1 was similar to LASSO in terms of the prediction accuracy while having small fit sizes
- MCL_2: much smaller fit sizes than the LASSO + similar prediction accuracy

Conclusion

High-dimensional LDA with MCL:

- predicts similarly or better than the LASSO while recovering the sparsity of the direction vector
- get the variable selection consistency under reasonable regularity conditions as supported by numerical studies
- An additional tuning parameter γ seems inattractive; however, the heuristic choice of $\gamma = \hat{\gamma}^{\text{opt}}$ or $\gamma = 2\hat{\gamma}^{\text{opt}}$ worked practically
- Further research may focus on the choice of γ

Bickel, P. J., Levina, E., et al. (2004). Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., et al. (2006). Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The journal of molecular diagnostics*, 8(1):51–61.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association*, 106(496):1566–1577.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6):529–541.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., and Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. *The journal of molecular diagnostics*, 8(1):31–39.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Efron, B. and Morris, C. (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605.

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

Fan, J., Song, R., et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967.

Guo, Y., Hastie, T., and Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

John A. Ramey (2016). datamicroarray: Collection of Data Sets for Classification. https://github.com/ramhiser/datamicroarray.

Kim, D., Lee, S., and Kwon, S. (2020). A unified algorithm for the non-convex penalized estimation: The ncpen package. *The R Journal*, Accepted.

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

Kim, Y., Jeon, J.-J., and Han, S. (2016). A necessary condition for the strong oracle property. *Scandinavian Journal of Statistics*, 43(2):610–624.

Kim, Y. and Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, 99(2):315–325.

Krzanowski, W., Jonathan, P., McCarthy, W., and Thomas, M. (1995). Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(1):101–115.

Kwon, S., Lee, S., and Kim, Y. (2015). Moderately clipped lasso. *Computational Statistics & Data Analysis*, 92:53–67.

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

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Witten, D. M. and 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., and 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. and Rangarajan, A. (2002). The concave-convex procedure (cccp). In *Advances in neural information processing systems*, pages 1033–1040.

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

Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and 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.