On the ultrahigh dimensional linear discriminant analysis problem with a diverging number of classes

Rui Pan[1], Hansheng Wang[1], and Runze Li[2]

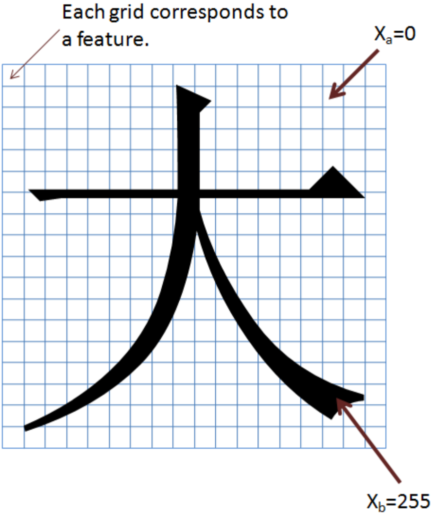
1 : Department of Business Statistics and Econometrics
Guanghua School of Management, Peking University
2 : Department of Statistics and the Methodology Center
The Pennsylvania State University

May 13, 2013

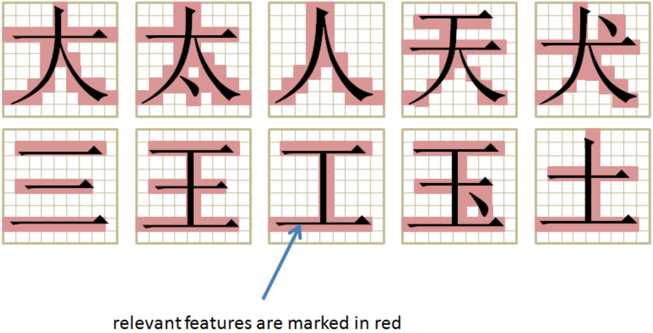
Outline

- A Motivating Example
- Introduction
- Pairwise Sure Independence Screening
  - Pairwise LDA
  - Theoretical Properties
  - Post Screening Estimation
  - Tunning Parameter Selection
- Numerical Studies
- Concluding Remarks

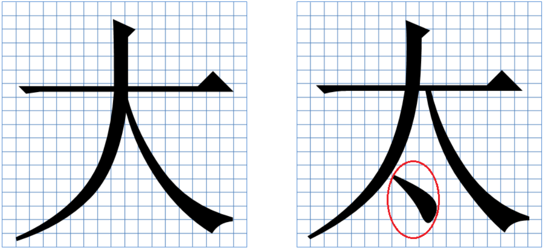
An Example: Chinese Character Recognition



An Example: Ten Chinese Characters



An Example: Pairwise Comparison



Introduction: Linear Discriminant Analysis (LDA)

- Categorical response (class label) $Y = 1, 2$ with equal prior probability, and continuous predictors (features) $X \in \mathbb{R}^p$
- Given the class label k ($k = 1, 2$), $X \sim N_p(\mu_k, \Sigma)$
- LDA Rule

$$\{X_0 - (\mu_1 + \mu_2)/2\}^\top \Sigma^{-1} (\mu_1 - \mu_2) > 0, \qquad (1)$$

which can be estimated by

$$\{X_0 - (\hat{\mu}_1 + \hat{\mu}_2)/2\}^\top \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) > 0 \qquad (2)$$

Ultrahigh Dimensional LDA: Literature Review

$$\{X_0 - (\hat{\mu}_1 + \hat{\mu}_2)/2\}^\top \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) > 0$$

- Bickel and Levina (2004):

  Independence Classification Rule, $\hat{D} = \text{diag}(\hat{\Sigma})$
- Fan and Fan (2008):

  Feature Annealed Independence Rule
- Shao et al. (2011): Sparse LDA
- Mai et al. (2012): Direct Approach

Introduction: Two Challenges

- Challenge One: LDA with high-dimensional predictors,

  $p \to \infty$

- Challenge Two: LDA with a diverging number of classes,

  $K \to \infty$

- Solution: Pairwise Feature Screening Method

Introduction: Contributions

Major Contributions:

(a) We propose to decompose the ultrahigh dimensional LDA problem with a diverging number of classes into many low dimensional ones.

(b) We propose a new pairwise feature screening of LDA, and establish the strong screening consistency of the proposed procedure.

Pairwise LDA: Notation

Let (Y_i, X_i) be the observation collected from the ith $(1 \le i \le n)$ subject.

$Y_i \in \{1, 2, \cdots, K\}$ with probability $P(Y_i = k) = \pi_k > 0$, where $\pi_k = 1/K$ for simplicity.

$X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the associated feature. Conditional on $Y_i = k$, X_i follows a multivariate normal distribution with mean $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p \times p}$.

Pairwise LDA: Pairwise Comparison

- Let (Y_0, X_0) be an independent observation. Suppose X_0 is known and we want to predict Y_0.
- $\{X_0 - (\mu_1 + \mu_2)/2\}^\top \Sigma^{-1}(\mu_1 - \mu_2) > 0$.
- For any pair (k_1, k_2) with $1 \leq k_1 \neq k_2 \leq K$, k^* can be equivalently defined as

$$k^* = \text{argmax}_{1 \leq k_1 \leq K} \sum_{k_2 \neq k_1} I\left(\left\{X_0 - (\mu_{k_1} + \mu_{k_2})/2\right\}^\top \beta_{k_1 k_2} > 0\right),$$

where $\beta_{k_1 k_2} = (\beta_{k_1 k_2, 1}, \cdots, \beta_{k_1 k_2, p})^\top = \Sigma^{-1}(\mu_{k_1} - \mu_{k_2}) \in \mathbb{R}^p$.

Pairwise LDA: Pairwise Screening

Define the notation $\mathcal{M}_{k_1 k_2} = \{j : \beta_{k_1 k_2, j} \neq 0\}$ to collect those indices associated with nonzero coefficients, and $|\mathcal{M}_{k_1 k_2}|$ the size of $\mathcal{M}_{k_1 k_2}$.

Accordingly, the original classification function can be re-written as $k^* = \text{argmax}_{k_1} \sum_{k_1 \neq k_2}$

$$I\left(\left\{X_{0(\mathcal{M}_{k_1 k_2})} - (\mu_{k_1(\mathcal{M}_{k_1 k_2})} + \mu_{k_2(\mathcal{M}_{k_1 k_2})})/2\right\}^\top \beta_{k_1 k_2(\mathcal{M}_{k_1 k_2})} > 0\right),$$

where $X_{0(\mathcal{M}_{k_1 k_2})} = (X_{0j} : j \in \mathcal{M}_{k_1 k_2})^\top \in \mathbb{R}^{|\mathcal{M}_{k_1 k_2}|}$ is the subvector of X_0 according to $\mathcal{M}_{k_1 k_2}$.

Pairwise LDA: Screening Method

Write $\beta_{k_1k_2} = \Sigma^{-1}\gamma_{k_1k_2}$ with $\gamma_{k_1k_2} = \mu_{k_1} - \mu_{k_2}$. And then investigate $\hat{\gamma}_{k_1k_2} = \hat{\mu}_{k_1} - \hat{\mu}_{k_2}$, where $\hat{\mu}_k = n_k^{-1}\sum_i X_i I(Y_i = k)$ and $n_k = \sum_i I(Y_i = k)$. For a given constant $c_{k_1k_2}$, we then estimate $\mathcal{M}_{k_1k_2}$ by

$$\widehat{\mathcal{M}}_{k_1k_2} = \left\{ j : |\hat{\gamma}_{k_1k_2,j}| > c_{k_1k_2} \right\}.$$

Theoretical Properties: Conditions

(C1) (*Pairwise Sparsity*) Assume that for any $1 \leq k_1, k_2 \leq K$, $1 \leq |\mathcal{M}_{k_1 k_2}| \leq s_{\max}$, where $s_{\max}$ is a fixed positive constant.

(C2) (*Coefficient Regularity*) (a) Assume that there exist finite positive constants $\beta_{\min}$, $\beta_{\max}$ such that $\beta_{\min} < \min_{k_1,k_2} \min_{j \in \mathcal{M}_{k_1 k_2}} |\beta_{k_1 k_2, j}| \leq \max_{k_1,k_2} \max_{j \in \mathcal{M}_{k_1 k_2}} |\beta_{k_1 k_2, j}| < \beta_{\max}$. (b) Furthermore, assume that there exists a constant $\gamma_{\min} > 0$ such that $\min_{k_1,k_2} \min_{j \in \mathcal{M}_{k_1 k_2}} |\gamma_{k_1 k_2, j}| > \gamma_{\min}$.

Theoretical Properties: Conditions

(C3) (*Covariance Matrix*) Assume that $0 < \tau_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < \tau_{\max} < \infty$ for some positive constants $\tau_{\min}$ and $\tau_{\max}$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest absolute eigenvalues of a symmetric matrix A.

(C4) (*Divergence Speed*) (a) Assume that $\log p \leq \nu_1 n^{\xi_1}$ for some constant $\nu_1 > 0$ and $0 < \xi_1 < 1$; (b) Furthermore, we assume that the number of classes $K \leq \nu_2 n^{\xi_2}$ for some constants $\nu_2 > 0$ and $0 < \xi_2 < 1$ with $\xi_1 + \xi_2 < 1$.

Theoretical Properties: Strong Screening Consistency

Theorem

Under Conditions (C1) to (C4), as $n \to \infty$, there exists a set of constants $c_{k_1 k_2}$ for every $1 \leq k_1, k_2 \leq K$, such that

$$P\left(\widehat{\mathcal{M}}_{k_1 k_2} \supset \mathcal{M}_{k_1 k_2} \text{ for every } 1 \leq k_1, k_2 \leq K\right) \to 1,$$

$$P\left(\max_{k_1, k_2} |\widehat{\mathcal{M}}_{k_1 k_2}| \leq m_{\max}\right) \to 1,$$

where $m_{\max} = 16\tau_{\max}^2 s_{\max} \beta_{\max}^2 \gamma_{\min}^{-2}$.

Post Screening Estimation: Notation

$\hat{\beta}_{k_1 k_2} = (\hat{\beta}_{k_1 k_2, j} : 1 \le j \le p)^\top \in \mathbb{R}^p$ can be obtained as,

- $\hat{\beta}_{k_1 k_2, j} = 0$ for any $j \notin \widehat{\mathcal{M}}_{k_1 k_2}$,
- $\hat{\beta}_{k_1 k_2 (\widehat{\mathcal{M}}_{k_1 k_2})} = (\hat{\beta}_{k_1 k_2, j} : j \in \widehat{\mathcal{M}}_{k_1 k_2})^\top = \hat{\Sigma}^{-1}_{(\widehat{\mathcal{M}}_{k_1 k_2})} \hat{\gamma}_{k_1 k_2 (\widehat{\mathcal{M}}_{k_1 k_2})}$.

Post Screening Estimation: Theorem 2

Theorem

Assume (C1) to (C4), as $n \to \infty$, then

$\max_{k_1 k_2} \|\hat{\beta}_{k_1 k_2} - \beta_{k_1 k_2}\| = o_p(1)$.

Post Screening Estimation: Prediction

$$k^* = \text{argmax}_{k_1} \sum_{k_1 \neq k_2}$$

$$I\left(\left\{X_{0(\mathcal{M}_{k_1 k_2})} - (\mu_{k_1(\mathcal{M}_{k_1 k_2})} + \mu_{k_2(\mathcal{M}_{k_1 k_2})})/2\right\}^\top \beta_{k_1 k_2(\mathcal{M}_{k_1 k_2})} > 0\right),$$

For a new observation X_0, we can predict Y_0 by

$$\hat{k} = \text{argmax}_{k_1} \sum_{k_2 \neq k_1}$$

$$I\left(\left\{X_{0(\widehat{\mathcal{M}}_{k_1 k_2})} - (\hat{\mu}_{k_1(\widehat{\mathcal{M}}_{k_1 k_2})} + \hat{\mu}_{k_2(\widehat{\mathcal{M}}_{k_1 k_2})})/2\right\}^\top \hat{\beta}_{k_1 k_2(\widehat{\mathcal{M}}_{k_1 k_2})} > 0\right).$$

Tuning Parameter Selection

$$\hat{\mathcal{M}}_{k_1k_2} = \left\{ j : |\hat{\gamma}_{k_1k_2,j}| > c_{k_1k_2} \right\}$$

Follow Mai et al. (2012), we construct a least squares objective function $Q_{k_1k_2} = E[(Y_{k_1k_2,i} - \beta_{k_1k_20} - \beta_{k_1k_2}^{\top} X_i)^2 | Y_i \in \{k_1, k_2\}]$, where $Y_{k_1k_2,i}$ is defined as $I(Y_i = k_1)/\pi_{k_1} - I(Y_i = k_2)/\pi_{k_2}$.

Information criteria,

$$\begin{aligned}
\text{AIC} &= \log \hat{Q}_{k_1k_2} + 2 \times n_{k_1k_2}^{-1} |\widehat{\mathcal{M}}_{k_1k_2}|, \\
\text{BIC} &= \log \hat{Q}_{k_1k_2} + \log n_{k_1k_2} \times n_{k_1k_2}^{-1} |\widehat{\mathcal{M}}_{k_1k_2}|, \\
\text{EBIC} &= \log \hat{Q}_{k_1k_2} + (\log n_{k_1k_2} + 2 \log p) \times n_{k_1k_2}^{-1} |\widehat{\mathcal{M}}_{k_1k_2}|.
\end{aligned}$$

Simulation Study: 4 Models

	1	2	3	$\cdots$	$K-1$	K	$K+1$	$\cdots$	p
1	μ	0	0	$\cdots$	0	0	0	$\cdots$	0
2	0	μ	0	$\cdots$	0	0	0	$\cdots$	0
3	0	0	μ	$\cdots$	0	0	0	$\cdots$	0
$\vdots$									
$K-1$	0	0	0	$\cdots$	μ	0	0	$\cdots$	0
K	0	0	0	$\cdots$	0	μ	0	$\cdots$	0

Remark: $\mathcal{M}_{k_1 k_2} = \{k_1, k_2\}$ and $|\mathcal{M}_{k_1 k_2}| = 2$.

Example 1. (Independent Covariance Structure) Generate $Y_i \in \{1, \cdots, K\}$ according to $P(Y_i = k) = 1/K$. Given $Y_i = k$, X_i is generated from a multivariate normal distribution with $E(X_i | Y_i = k) = \mu_k$, where $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top \in \mathbb{R}^p$ is a p-dimensional vector with $\mu_{kk} = \mu$ but $\mu_{k_1 k_2} = 0$ for any $k_1 \neq k_2$. Furthermore, the conditional covariance is given by $\mathrm{cov}(X_i | Y_i = k) = \Sigma = I_p$, where I_p is a $p \times p$ identity matrix. It is easily verified that $\mathcal{M}_{k_1 k_2} = \{k_1, k_2\}$, and thus $\mathcal{M}_T = \bigcup \mathcal{M}_{k_1 k_2} = \{j : 1 \leq j \leq K\}$.

Example 2. (Autoregressive Covariance Structure) The data are generated in a similar manner as Example 1 but with two differences. The first difference is that Σ is set to $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$. Note that $\Sigma^{-1} = (\omega_{j_1 j_2}) \in \mathbb{R}^{p \times p}$ is very sparse with $\omega_{11} = \omega_{pp} = 4/3$, $\omega_{jj} = 5/3$ for $1 < j < p$, $\omega_{j(j+1)} = \omega_{(j+1)j} = -2/3$ for $1 \leq j < p$, and $\omega_{j_1 j_2} = 0$ whenever $|j_1 - j_2| > 1$. The second difference is that the mean vector μ_k is set be $\mu_k = \mu \Sigma_{(k)}$, where $\Sigma_{(k)}$ stands for the kth column vector of Σ. Accordingly, it follows that $\mathcal{M}_{k_1 k_2} = \{k_1, k_2\}$ with $\mathcal{M}_T = \bigcup \mathcal{M}_{k_1 k_2} = \{j, 1 \leq j \leq K\}$.

Example 3. (Compound Symmetric Covariance Structure) The data are generated in a similar manner as in Example 1. However, the covariance is changed to $\sigma_{j_1 j_2} = 0.5 + 0.5 I(j_1 = j_2)$, which is a compound symmetric structure with diagonal components being 1 but all others being 0.5. One can verify that $\mathcal{M}_{k_1 k_2} = \{k_1, k_2\}$ with $\mathcal{M}_T = \{j : 1 \le j \le K\}$. Furthermore, we know that the largest eigenvalue of Σ is $\lambda_{\max}(\Sigma) = 0.5(p+1)$ while the smallest is $\lambda_{\min}(\Sigma) = 0.5$. Consequently, the technical condition (C3) is seriously violated. It is then of great interest to examine how the proposed procedure is sensitive to this violation.

Example 4. (Normality Assumption) $Y_i \in \{1, \cdots, K\}$ is generated according to $P(Y_i = k) = 1/K$. Given $Y_i = k$, we then generate the predictors as $X_i = \mu_k I(Y_i = k) + Z_i$, where $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top \in \mathbb{R}^p$ with $\mu_{kk} = \mu$ but $\mu_{k_1 k_2} = 0$ for every $k_1 \neq k_2$. Furthermore, the random vector $Z_i \in \mathbb{R}^p$ is generated as $Z_i = (Z_{i1}, \cdots, Z_{ip})^\top \in \mathbb{R}^p$ with each Z_{ij} independently simulated from a centralized standard exponential distribution, that is $\exp(1) - 1$. Once again, we have $\mathcal{M}_{k_1 k_2} = \{k_1, k_2\}$ and $\mathcal{M}_T = \{j : 1 \leq j \leq K\}$.

Simulation Study: Results

Table 1: Simulation Results for Model 1 with 1,000 Replications

Criterion	Signal μ	(n, K)	n_k	MS	MMS	CZ(%)	IZ(%)	CP(%)	UCP(%)	RSSE	MRSSE	CA(%) $\hat{k}$	k^*	RCA(%)
EBIC	3	(100,10)	10	1.0	1.2	100.0	48.5	3.5	0.0	3.1	4.0	72.1	90.3	79.9
		(400,20)	20	1.3	1.5	100.0	34.0	31.9	0.0	2.3	3.3	65.8	84.7	77.7
		(1600,40)	40	1.9	2.0	100.0	3.2	93.4	0.0	0.5	3.1	76.0	78.1	97.3
	5	(100,10)	10	1.1	1.2	100.0	45.6	8.7	0.0	4.9	5.9	97.3	99.8	97.5
		(400,20)	20	1.5	1.7	100.0	23.3	53.3	0.0	2.8	5.3	97.5	99.7	97.8
		(1600,40)	40	2.0	2.0	100.0	0.8	98.4	0.0	0.5	5.1	99.4	99.4	99.9
AIC	3	(100,10)	10	4.3	7.4	100.0	2.1	96.1	36.1	3.7	7.1	80.6	90.3	89.3
		(400,20)	20	9.2	12.3	99.9	0.0	100.0	98.9	3.5	5.6	75.5	84.7	89.2
		(1600,40)	40	20.4	24.9	99.8	0.0	100.0	100.0	3.4	4.9	70.3	78.1	90.0
	5	(100,10)	10	3.6	6.0	100.0	0.8	98.5	54.8	3.6	7.9	99.5	99.8	99.6
		(400,20)	20	6.8	9.5	100.0	0.0	100.0	96.9	2.9	6.0	99.3	99.7	99.6
		(1600,40)	40	12.6	15.9	99.9	0.0	100.0	100.0	2.5	5.1	99.0	99.4	99.6
BIC	3	(100,10)	10	3.7	6.0	100.0	3.1	94.0	16.1	3.3	6.9	81.5	90.3	90.3
		(400,20)	20	5.0	6.4	100.0	0.1	99.8	69.7	2.2	5.3	79.0	84.7	93.3
		(1600,40)	40	4.0	4.5	100.0	0.0	100.0	99.5	1.1	4.2	76.6	78.1	98.2
	5	(100,10)	10	3.0	4.4	100.0	1.6	96.7	26.0	3.1	7.6	99.5	99.8	99.7
		(400,20)	20	3.1	3.8	100.0	0.1	99.9	75.0	1.5	5.4	99.6	99.7	99.9
		(1600,40)	40	2.5	2.6	100.0	0.0	100.0	99.8	0.7	3.3	99.5	99.4	100.1

Real Example: Handwritten Chinese Characters

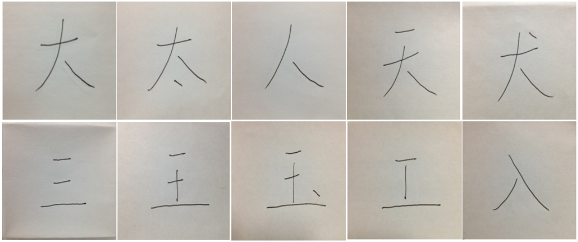


Figure: The Ten Handwritten Characters used for Real Data Analysis

Table: Detailed Results for the Real Example

Method	MS	MMS	Total No. of selected features	CA(%)
AIC	19.6	30.0	222.9	70.4
BIC	7.5	26.8	112.8	67.2
EBIC	1.3	3.4	32.2	58.4
NM	625	625	625	46.1

Concluding Remarks

- Pairwise Feature Screening
- Strong Screening Consistency Property
- Future Study

Reference

Akaike, H. (1973), “Information theory and an extension of the maximum likelihood principle,” *In 2nd International Symposium on Information Theory, Ed. B. N. Petrov & F. Csaki*, 267–281. Budapest: Akademia Kiado.

Bickel, P. J. and Levina, E. (2004), “Some theory for Fisher’s linear discriminant function, “naïve Bayes”, and some alternatives when there are many more variables than observations,” *Bernoulli*, 10, 989–1010.

Bickel, P. J. and Levina, E. (2008), “Regularized estimation of large covariance matrices,” *The Annals of Statistics*, 36, 199–227.

Chen, J. and Chen, Z. (2008), “Extended Bayesian information criterion for model selection with large model spaces,” *Biometrika*, 95, 759–771.

Clemmensen, L., Hastie, T., and Ersbøll (2011), “Sparse discriminant analysis,” *Technometrics*, 53(4), 406–413.

Fan, J. and Fan, Y. (2008), “High dimensional classification using features annealed independence rules,” *The Annals of Statistics*, 36, 2605–2637.

Fan, J., Fan, Y., and Lv, J. (2008), “High dimensional covariance matrix estimation using a factor model,” *Journal of Econometrics*, 147, 186–197.

Fan, J., Feng, Y., and Song, R. (2011), “Nonparametric independence screening in sparse ultra-high dimensional additive models,” *Journal of the American Statistical Association*, 116, 544–557.

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.

Fan, J. and Lv, J. (2008), “Sure independence screening for ultra-high dimensional feature space (with discussion),” *Journal of the Royal Statistical Society, Series B*, 70, 849–911. Fan, J. and Song, R. (2010),“Sure independent screening in generalized linear models with NP-dimensionality,” *The Annals of Statistics*, 38, 3567–3604.

Guo, Y., Hastie, T., and Tibshirani, R. (2007),“Regularized discriminant analysis and its application in microarrays,” *Biostatistics*, 1, 86–100.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer.

Huang, J., Ma, S., and Zhang, C. H. (2007), “Adaptive LASSO for sparse high dimensional regression,” *Statistica Sinica*, 18, 1603–1618.

Johnson, R. A. and Wichern, D. W. (2003), *Applied Multivariate Statistical Analysis (5th Ed.)*, New York: Pearson Education.

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

McQuarrie, D. R. and Tsai, C. L. (1998), *Regression and Time Series Model Selection*, World Scientific, Singapore.

Schwarz, G. (1978), “Estimating the dimension of a model,” *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), “An asymptotic theory for linear model selection,” *Statistica Sinica*, 7, 221–264.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011), “Sparse linear disciminant analysis by thresholding for high dimensional data,” *The Annals of Statistics*, 39, 1241–1265.

Tibshirani, R., Hastie, T., Narashimhan, B., and Chu, G. (2003), “Class prediction by nearest shrunken centriods with applications to DNA microarrays,” *Statistical Science*, 18, 104–117.

Tibshirani, R. J. (1996), “Regression shrinkage and selection via the LASSO,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H. (2009), “Forward regression for ultra-high dimensional variable screening,” *Journal of the American Statistical Association*, 104, 1512–1524.

— (2012), “Factor profiled independence screening,” *Biometrika*, 99, 15–28.

Wang, H., Li, R., and Tsai, C. L. (2007), “Tuning parameter selectors for the smoothly clipped absolute deviation method,” *Biometrika*, 94, 553–568.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. (2005), “Text Mining: Predictive Methods for Analyzing Unstructured Information,” New York: Springer.

Zhang, H. H. and Lu, W. (2007), “Adaptive lasso for Cox’s proportional hazard model,” *Biometrika*, 94, 691–703.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011), “Model-free feature screening for ultrahigh dimensional data,” *Journal of the American Statistical Association*, 106, 1464–1475.

Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

Thanks!