Comparisons of discriminant analysis techniques for high-dimensional correlated data

Line H. Clemmensen
DTU Informatics
lhc@imm.dtu.dk

Overview

- Linear discriminant analysis (notation)
- Issues for high-dimensional data
- Assumptions about variables - independent or correlated?
- Within-class covariance estimates in a range of recently proposed methods
- Simulations
- Results and discussion

Linear discriminant analysis

- We model *K* classes by Gaussian normals
- k^{th} class has distribution $C_k \sim N(\mu_k, \Sigma)$
- Maximum-likelihood estimate of within-class covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

Linear discriminant analysis

- A new observation $\boldsymbol{x}_{new}$ is classified using the following rule

$$\max_{C_k}\{\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\mathbf{x}_{new}^T - \frac{1}{2}\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k^T\}$$

Issues and fixes for high dimensions (*p*>>*n*)

- Within-class covariance matrix becomes singular
- Regularize within-class covariance matrix to have full rank
- Introduce sparseness in feature-space (dimension reduction)
- So far papers have focused on sparseness criterion, cost function and speed.

Focus here

- The estimate of the within-class covariance matrix is crucial

Assuming independence

- Use a diagonal estimate of the within-class covariance matrix
- Similar to a univariate regression approach

Nearest shrunken centroids

- Diagonal estimate of within-class covariance matrix

$$\hat{\mathbf{\Sigma}}_{NSC} = \mathrm{diag}(\hat{\mathbf{\Sigma}})$$

- Soft-thresholding to perform feature selection

- $$\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k^* = \mathrm{sign}(\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k)(|\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k| - \Delta)_+$$

Penalized linear discriminant analysis

- Diagonal estimate of within-class covariance

$$\tilde{\mathbf{\Sigma}}_{PLDA} = \mathrm{diag}(\hat{\mathbf{\Sigma}})$$

- Using L_1-norm to introduce sparsity in Fisher's criterion and a maximization-minorization algorithm for optimization.

Assuming correlations exist

- Estimate off-diagonal in within-class covariance matrix
- Should preferably exploit high correlations in data and “average out noise”

Regularized discriminant analysis

- Trade-off diagonal estimate and full estimate of within-class covariance matrix

$$\hat{\mathbf{\Sigma}}_{RDA}(\alpha) = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathrm{diag}(\hat{\mathbf{\Sigma}})$$

- Use soft-thresholding to obtain sparseness

$$\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k^* = \mathrm{sign}(\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k)(|\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k| - \Delta)_+$$

Sparse discriminant analysis

- Full estimate of covariance matrix based on a L_1- and L_2-penalized feature-space

$$\hat{\boldsymbol{\Sigma}}_{SDA} = 1/n \sum_{k=1}^{K} \sum_{i \in C_k} (\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_k)(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_k)^T$$

- Where $\tilde{\mathbf{x}}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$, and β are the estimated sparse and regularized discriminant directions in SDA.

Sparse linear discriminant analysis by thresholding

- Using thresholding to obtain sparsity in the within-class covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{ij,SLDAT} = \hat{s}_{ij} I(|\hat{s}_{ij}| > t_1), \text{ with } t_1 = M_1 \sqrt{\log p}/\sqrt{n}$$

- As well as in the feature-space

$$\tilde{\delta}_{i,kl} = \hat{\delta}_{i,kl} I(|\hat{\delta}_{i,kl}| > t_2)$$

- where $\delta_{kl} = \mu_k - \mu_l$

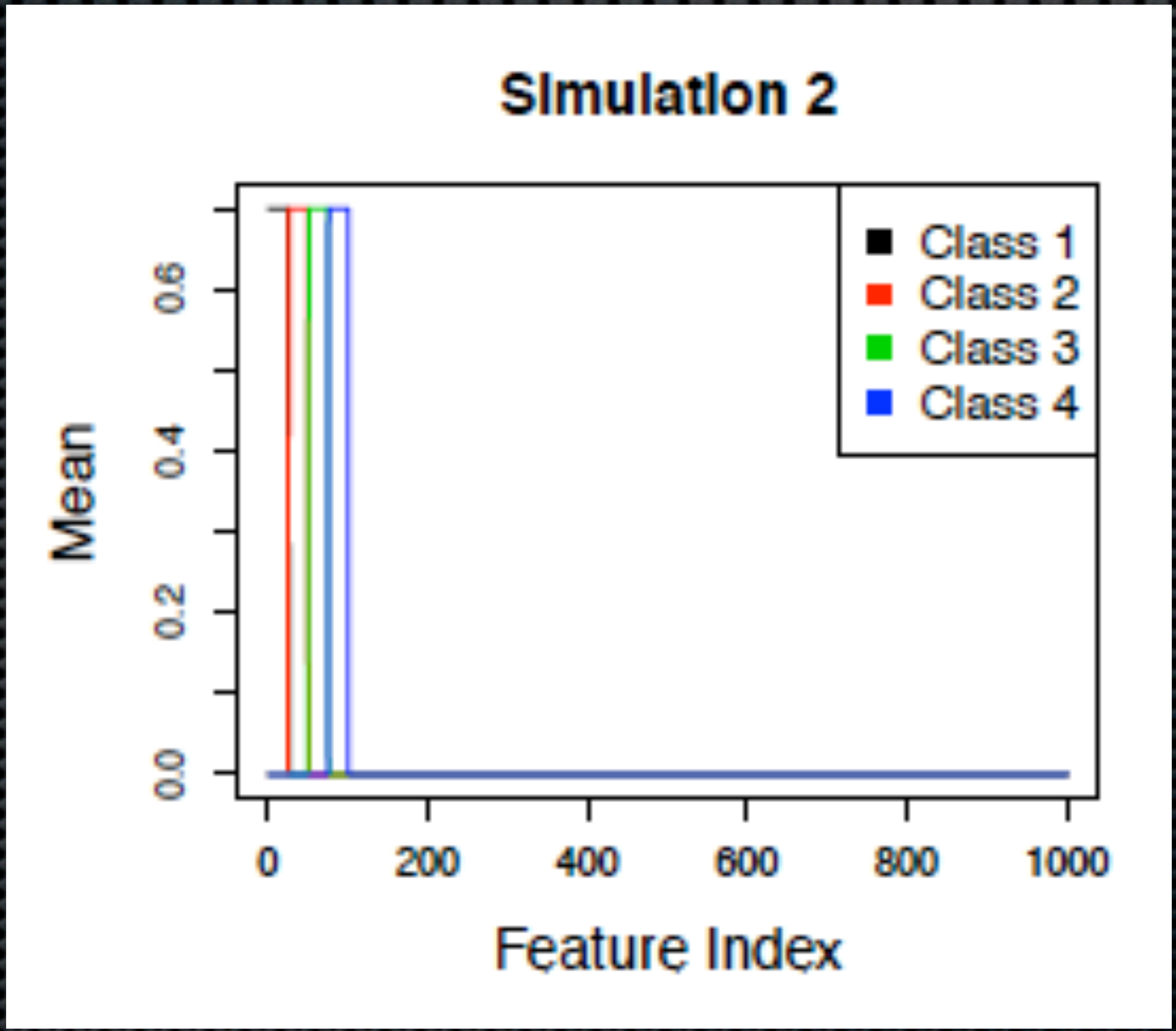
Simulations S

- Four classes of Gaussian distributions C_k: $x_i \sim N(\mu_k, \Sigma)$ with means

$$\boldsymbol{\mu_{jk}} = 0.7 \times 1_{((k-1)\times 100+1 \leq j \leq k\times 100)}$$

- And within-class covariance matrix is block-diagonal with 100 variables in each block and the $(j, j')^{th}$ element of each block equal to $r^{\text{abs}(j-j')}$ *where* $0 \leq r \leq 1$.

Simulation means of four classes



Simulations S

- S1: Independent variables *r=0, p=500*
- S2: Correlated variables *r=0.99, p=500*
- S3: Correlated variables *r=0.99, p=1000*
- S4: Correlated variables *r=0.9, p=1000*
- S5: Correlated variables *r=0.8, p=1000*
- S6: Correlated variables *r=0.6, p=1000*

Simulations X

- Four Gaussian classes with means as in S simulations
- Off-diagonal of within-class covariance matrix equal to ρ (diagonal equals one)

Simulations X

- X1: Correlated variables with $\rho=0.8$, $p=1000$
- X2: Correlated variables with $\rho=0.6$, $p=1000$
- X3: Correlated variables with $\rho=0.4$, $p=1000$
- X4: Correlated variables with $\rho=0.2$, $p=1000$
- X5: Correlated variables with $\rho=0.1$, $p=1000$
- X6: Independent variables with $\rho=0$, $p=1000$

Procedure

- 1200 observations were simulated for each case
- 100 observations were used to train the model
- another 100 to validate and tune parameters
- 1000 observations were used to report test errors
- 25 repetitions were performed and mean and standard deviations reported

Results

		PLDA	NSC	SDA	RDA	SLDAT
S1:	#errors	116.6(4.3)	**88.5**(2)	124.4(4.6)	**90.9**(2.4)	141.2(5.9)
	#features	348(18.8)	276(17.1)	261.7(18.1)	218.1(12.3)	292(23.1)
S2:	#errors	539.72(23.9)	424.84(26.6)	**0**(0)	**0.36**(0.3)	13.2(10.8)
	#features	264.32(34)	143.92(12.3)	500(0)	449.52(14.7)	473.28(14.8)
S3:	#errors	602.1(18.8)	449.2(24.9)	**0**(0)	**0.04**(0)	18.6(6.5)
	#features	444.4(69.5)	170.2(27.3)	847.6(1.6)	715.9(39.2)	890.8(43.5)
S4:	#errors	622.4(18.2)	440.2(21)	**0.12**(0.1)	3.1(0.8)	256.9(24.7)
	#features	566.9(66.6)	153.5(23)	841.4(10.8)	955.7(35.8)	711(76.9)
S5:	#errors	550.7(22.7)	412.9(26.9)	**2.2**(0.4)	5(1.4)	397.4(21.9)
	#features	436.2(68)	161.6(21.1)	814.3(18.2)	867.7(62.4)	585(85.2)
S6:	#errors	540.7(20.1)	398.9(18)	44.1(4.4)	**39.2**(5.7)	463.8(22.5)
	#features	457.5(60.1)	143.6(16.7)	406.5(30.8)	260.1(58.5)	365.6(72.1)
X1:	#errors	166.9(10.1)	58.4(10.4)	**0**(0)	2.2(0.6)	12.5(1.5)
	#features	133.7(16.6)	125.6(24.8)	857.4(1.7)	376.4(86.7)	725.8(73.3)
X2:	#errors	134.7(7.9)	29(6.2)	**0**(0)	6.72(2.1)	42.4(6.6)
	#features	155.2(6.6)	141(14.3)	857.3(2.1)	293(81.1)	218.3(53.9)
X3:	#errors	106.3(7.8)	17.4(3.4)	**0.04**(0)	7.12(1.5)	21.4(6.1)
	#features	192.2(6.5)	161.6(30.6)	858.3(1.8)	477.4(94.2)	125.6(6.3)
X4:	#errors	36(4.3)	5.6(1.1)	**0.08**(0.1)	6.4(1.4)	5(1.5)
	#features	245.2(36.4)	363.5(47.8)	862.4(1.7)	594.9(93)	181.2(16.8)
X5:	#errors	11.1(1.7)	6(1.5)	**0.4**(0.1)	2.8(0.7)	4.3(1.5)
	#features	208.2(15)	650.7(47.7)	861.3(1.5)	797.4(73.7)	366.2(51.9)
X6:	#errors	166.3(6.7)	**116.7**(3.3)	174.6(4.2)	**120**(5.1)	211.7(6.2)
	#features	418.2(45.1)	320.6(33.4)	339.6(27)	296(22.3)	357.4(50.8)

Discussion

- Assuming independence works best when variables are independent
- Assuming correlations exist works best when variables are correlated
- An illustration of a part of the correlation matrix may reveal the structure of data
- Interpretability - low dimensional projections of data

References

- Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**(4): 406-413 (2011)
- CRAN: The comprehensive r archive network (2009). URL http://cran.r-project.org/
- Fisher, R.: The use of multiple measurements in axonomic problems. Annals of Eugenics **7**:179-188 (1936)
- Guo, Y., Hastie, T., Tibshirani, R.: Regularized linear discriminant analysis and itsapplications in microarrays. Biostatistics **8**(1), 86-100 (2007)
- Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. The Annals of Statistics **23**(1), 73-102 (1995)
- Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2 edn. Springer (2009)
- Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**, 55-67 (1970)
- Shao, J., Wang, G., Deng, X., Wang, S.: Sparse linear discriminant analysis by thresholding for high dimensional data. The Annals of Statistics 39(2), 1241-1265 (2011)
- Sjöstrand, K., Carden, V.A., Larsen, R., Studholme, C.: A generalization of voxel-wise procedures for highdimensional statistical inference using ridge regression. In: J.M.Reinhardt, J.P.W. Pluim (eds.) SPIE, SPIE 6914, Medical Imaging (2008)
- Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statistical Science **18**, 104-11 (2003)
- Tibshirani, R., Saunders, M.: Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society - Series B **67**(1), 91{108 (2005)
- Witten, D., Tibshirani, R.: Penalized classication using Fisher's linear discriminant, Journal of the Royal Statistical Society, Series B (2011)
- Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of Royal Statistical Society - Series B **67**(Part 2), 301-320 (2005)