Compressive Regularized Discriminant Analysis (CRDA)

High-dimensional Classification and Feature Selection – Applications to Microarray Studies

Muhammad Naveed Tabassum and Esa Ollila

Dept. of Signal Processing and Acoustics, Aalto University, Finland

Aalto University

Motivation

Goals:

1 Given G possible classes (or populations), classify a p-dimensional observation $\mathbf{x}$ accurately to its correct class.

2 Reduce the number of variables (or features) without sacrificing the accuracy.

Challenge: High-dimension (HD) low-sample size settings, where p is often several magnitudes larger than the number of observations, n (i.e., $p \gg n$, for example *microarray data*.

Sparsity facilitates interpretation and stabilizes estimation in the HD situations.

Problem Formulation

⋆ Following *rule* assigns $\mathbf{x}$ to one of the G classes

$$\mathbf{x} \in \texttt{group}\,\big[\,\tilde{g} = \arg\max_g\; d_g(\mathbf{x})\,\big], \tag{1}$$

where $g \in \{1,\ldots,G\}$ and $d_g(\mathbf{x})$ is called the *discriminant function*.

Linear discriminant analysis (LDA) uses the rule (1) with,

$$d_g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_g + c_g$$

for $g = 1,\ldots,G$, where:

$$\boldsymbol{\beta}_g = \boldsymbol{\beta}_g(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g \quad \in \mathbb{R}^p \tag{2}$$

$$c_g = -\frac{1}{2}\boldsymbol{\mu}_g^\top \boldsymbol{\beta}_g + \ln p_g \quad \in \mathbb{R} \tag{3}$$

where $\boldsymbol{\Sigma}$ is common covariance matrix of the classes, μ_g denotes the class mean vector ($g = 1,\ldots,G$) and p_g is a prior probability that $\mathbf{x}$ is from class g.

If the i^{th} entry of $\boldsymbol{\beta}_g$ is zero, then the i^{th} feature does not contribute in the classification to g^{th} population.

Regularized LDA

⋆ Training dataset $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n) \in \mathbb{R}^{p\times n}$ is given with associated class labels $c(i) \in \{1,\ldots,G\}$.

⋆ Unknowns, p_g, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}$, are estimated from $\mathbf{X}$.

⋆ $\hat{p}_g = \pi_g = n_g/n$, where $(n_g = \sum_{i=1}^n \mathbf{I}(c(i) = g)$.

For $g = 1,\ldots,G$, assuming observations in $\mathbf{X}$ are centered by the sample mean vectors of the classes

$$\hat{\boldsymbol{\mu}}_g = \overline{\mathbf{x}}_g = \frac{1}{n_g}\sum_{c(i)=g} \mathbf{x}_i, \tag{4}$$

the pooled sample covariance matrix (SCM) is given as:

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top.$$

In practice, the rule (1) uses $\hat{d}_g(\mathbf{x})$ with $\hat{\boldsymbol{\beta}}_g = \mathbf{S}^{-1}\hat{\boldsymbol{\mu}}_g$ in (2). However, $\mathbf{S}$ is singular and is no longer invertible in the HD settings.

⋆ Thus, a *regularized SCM (RSCM)* $\hat{\boldsymbol{\Sigma}}$ is used to avoid the singularity and to construct the empirical LDA rule.

⋆ Such approaches are referred to as regularized LDA (see e.g., [1,2,3]) which we refer shortly as RDA.

As RSCM we use

$$\hat{\boldsymbol{\Sigma}} = \alpha\,\mathbf{S} + (1-\alpha)\,\eta\mathbf{I} \tag{5}$$

where $\eta = \mathrm{Tr}(\mathbf{S})/p$ and $\alpha \in [0,1)$ is a regularisation parameter that is calculated using the method proposed in [4] or using cross-validation (CV).

Next, the computational complexity of matrix inversion is reduced from $\mathcal{O}(p^3)$ to $\mathcal{O}(pn^2)$ using the SVD-trick [1].

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{U}\left[\left(\frac{\alpha}{n}\mathbf{D}^2 + (1-\alpha)\eta\mathbf{I}_m\right)^{-1} - \frac{1}{(1-\alpha)\eta}\mathbf{I}_m\right]\mathbf{U}^\top + \frac{1}{(1-\alpha)\eta}\mathbf{I}_p, \tag{6}$$

where $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and $\eta = \mathrm{Tr}(\mathbf{S})/p = \mathrm{Tr}(\mathbf{D}^2)/np$.

Compressive RDA (CRDA)

We express LDA discriminant rule in vector form:

$$\mathbf{d}(\mathbf{x}) = (d_1(\mathbf{x}),\ldots,d_G(\mathbf{x})) = \mathbf{x}^\top\mathcal{B} - \frac{1}{2}\mathrm{diag}\big(\mathbf{M}^\top\mathcal{B}\big) + \ln\mathbf{p}, \tag{7}$$

where $\ln\mathbf{p} = (\ln p_1,\ldots,\ln p_G)$, $\mathbf{M} = \big(\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_G\big)$, $\mathcal{B} = \boldsymbol{\Sigma}^{-1}\mathbf{M}$ and $\mathrm{diag}(\mathbf{A}) = (a_{11},\ldots,a_{GG})$ for some matrix $\mathbf{A}$.

K-rowsparsity of $\mathcal{B} \in \mathbb{R}^{p\times G} \rightsquigarrow p-K$ features (genes) do not contribute in the classification procedure.

⋆ The simultaneous feature selection (**SFS**) is obtained by using *hard-thresholding operator* $H_K(\cdot, q)$.

⋆ It is defined as a transform $H_K(\mathcal{B}, q)$.

⋆ It retains the elements of the K rows of $\mathcal{B}$ that possess largest ℓ_q norm and set elements of the other rows to zero, as illustrated in Figure 1.

⋆ We use $q = 1, 2, \infty$.

Our *CRDA* uses estimated discriminant function

$$\hat{\mathbf{d}}(\mathbf{x}) = \mathbf{x}^\top\hat{\mathcal{B}} - \frac{1}{2}\mathrm{diag}\big(\hat{\mathbf{M}}^\top\hat{\mathcal{B}}\big) + \ln\boldsymbol{\pi}, \tag{8}$$

where $\ln\boldsymbol{\pi} = (\ln\pi_1,\ldots,\ln\pi_G)$, $\hat{\mathbf{M}} = \big(\hat{\boldsymbol{\mu}}_1 \ldots \hat{\boldsymbol{\mu}}_G\big)$ and

$$\hat{\mathcal{B}} = H_K(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{M}}, q)$$

having K non-zero rows, e.g., as shown in Figure 2.

Results and Discussions

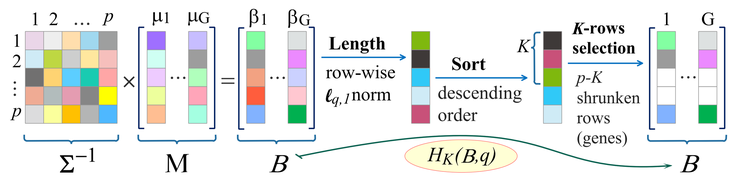

Figure 1: The simultaneous feature selection (SFS) algorithm.

CRDA uses easy to tune joint-sparsity level $K \in \{1,2,\ldots,p\}$ and benefits from SFS approach, where features are eliminated across all groups instead of group-wise.

⋆ Optimal pair $(\hat{\alpha}_{\text{cv}}, \hat{K}_{\text{cv}})$ using $\varepsilon_{\text{cv}}(\alpha, K) \le \max(0.15\,n,\, \varepsilon_{\text{cv}})$, which has the smallest number of features selected (NFS).

⋆ CV employs grids with 100 K-values and 25 α-values, and comparison using estimated $\hat{\alpha}_{ell}$ from [4] is provided in paranthesis (see Tables).

⋆ CRDA based classifiers are compared against the NSC [2], SCRDA [1] and PLDA [3].

⋆ The simulation setup mimics real microarray data and generates $n = 200$ training and 1000 test observations each having $p = 10,000$ features.

⋆ It has 200 differential features for $G = 3$ groups. Table lists average results over 10 Monte-Carlo trials using 10-fold CV.

Test error (TE), NFS, false positive (FP) and detection rate (DR) for the simulation setup.

Methods	TE/1000	NFS	DR	FP
CRDA$^{\ell_1}$	**46** (50)	**205** (259)	90 (**94**)	**12** (27)
CRDA$^{\ell_2}$	49 (**46**)	240 (**203**)	**92** (92)	23 (**10**)
CRDA$^{\ell_\infty}$	50 (52)	238 (252)	89 (92)	27 (27)
SCRDA	108	282	69	51

A summary of the used real microarray datasets.

Dataset	N	p	G	Disease
Ramaswamy *et al.*	190	16,063	14	Cancer
Yeoh *et al.*	248	12,625	6	Leukemia
Sun *et al.*	180	54,613	4	Glioma
Nakayama *et al.*	105	22,283	10	Sarcoma

Average comparison results for 10 training-test (75%-25%) set splits using 5-fold CV. Values in bold-face indicate the best results and in parenthesis are obtained using $(\hat{\alpha}_{ell}, \hat{K}_{\text{cv}})$ instead of $(\hat{\alpha}_{\text{cv}}, \hat{K}_{\text{cv}})$.

Methods	Ramaswamy *et al.* dataset		Yeoh *et al.* dataset		Sun *et al.* dataset		Nakayama *et al.* dataset	
	TE/47	NFS	TE/62	NFS	TE/45	NFS	TE/26	NFS
CRDA$^{\ell_1}$	10.6 (**9.9**)	2634 (4899)	9.6 (7.5)	2525 (4697)	12.5 (**12.9**)	23320 (27416)	8.3 (7.9)	2941 (6952)
CRDA$^{\ell_2}$	10.4 (10.3)	2683 (3968)	9.7 (**6.0**)	2273 (**4659**)	12.9 (13.3)	**20589** (23484)	7.9 (7.6)	3142 (7755)
CRDA$^{\ell_\infty}$	**10.3** (10.3)	3405 (4530)	**9.3** (6.5)	**846** (4697)	**12.4** (13.5)	21354 (**20207**)	7.6 (7.6)	**2719 (2340)**
PLDA	18.8	5023	NA	NA	15.2	21635	4.4	10479
SCRDA	24	14874	NA	NA	15.7	54183	**2.8**	22283
NSC	16.3	**2337**	NA	NA	15	30005	4.2	5908

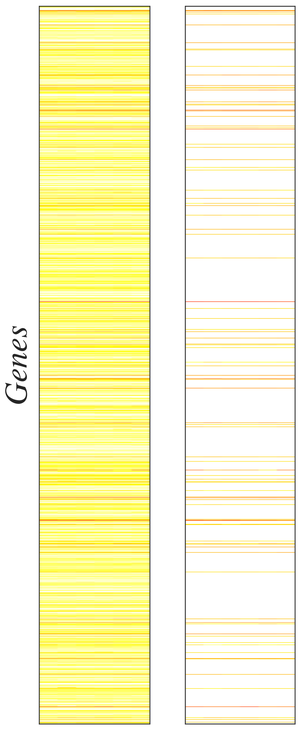

Figure 2: $\mathcal{B}$ in Yeoh *et al.* dataset: before (left) and after (right) the transformation $H_K(\mathcal{B} = \boldsymbol{\Sigma}^{-1}\mathbf{M}, q)$.

Conclusions

⋆ *Proposed CRDA of data in high-dimension low-sample size situations was shown to outperform competing methods in most of the cases.*

⋆ *It can be a useful tool for accurate selection of (differentially expressed) features, i.e., genes in microarray studies.*

*See our paper for more detailed results and discussions.

Email: {muhammad.tabassum, esa.ollila}@aalto.fi

References

[1] Yaqian Guo, Trevor Hastie, and Robert Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2006.

[2] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu, "Class prediction by nearest shrunken centroids, with applications to dna microarrays," *Statistical Science*, pp. 104–117, 2003.

[3] Daniela M Witten and Robert Tibshirani, "Penalized classification using fisher's linear discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.

[4] Esa Ollila, "Optimal high-dimensional shrinkage covariance estimation for elliptical distributions," in *European Signal Processing Conference (EUSIPCO 2017)*, Kos, Greece, pp. 1689–1693.

IEEE ICASSP 2018 || Calgary, Canada