Motivation

Goals
- Given \( G \) possible classes (or populations), classify a \( p \)-dimensional observation \( x \) accurately to its correct class.
- 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 \( x \) to one of the \( G \) classes
  \[ x \in \{ \overline{G}_1, \ldots, G_G \}, \]
  where \( g \in \{ 1, \ldots, G \} \) and \( d_g(x) \) is called the discriminant function.

Linear discriminant analysis (LDA) uses the rule (1) with
- \( d_g(x) = x^T \beta_g + c_g \) for \( g = 1, \ldots, G \), where
  \[ \beta_g = \beta_g(\Sigma_g) = \Sigma_g^{-1}\mu_g \in \mathbb{R}^p \] (2)
- \( c_g = \frac{1}{2} \mu_g^T \beta_g + \ln p_g \in \mathbb{R} \) (3)

where \( \Sigma_g \) is common covariance matrix of the classes, \( \mu_g \) denotes the class mean vector (\( g = 1, \ldots, G \)) and \( p_g \) is a prior probability that \( x \) is from class \( g \).

If the \( i \)-th entry of \( \beta_g \) is zero, then the \( i \)-th feature does not contribute in the classification to \( g \)-th population.

Regularized LDA

- Training dataset \( X = \{ x_1, \ldots, x_n \} \in \mathbb{R}^{p \times n} \) is given with associated class labels \( c \in \{ 1, \ldots, G \} \).
- Unknown \( \mu_g, \Sigma_g \), and \( \alpha_g \) are estimated from \( X \).
- \( \beta_g = \beta_g(\Sigma_g) = \Sigma_g^{-1}\mu_g \) where \( (n_g = \sum_{c_i = g} 1) \).

For \( g = 1, \ldots, G \), assuming observations in \( X \) are centered by the sample mean vectors of the classes
- \( \mu_g = \mu_g = \frac{1}{n_g} \sum_{x_i} x_i \) (4)
- the pooled sample covariance matrix (SCM) is given as:
  \[ S = \frac{1}{n} XX^T \]

In practice, the rule (1) uses \( d_g(x) \) with \( \beta_g = S^{-1}\mu_g \) in (2). However, \( S \) is singular and is no longer invertible in the HD settings.

Thus, a regularized SCM (RSCM) \( \Sigma \) 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
  \[ \Sigma = \alpha + (1 - \alpha) q \mathbf{I} \] (5)
where \( q = \text{Tr}(S)/p \) 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 \( \mathcal{O}(p^2) \) using the SVD-trick [1].

\[ \Sigma^{-1} = U \left( \frac{\alpha}{p} D^2 + (1 - \alpha) q \mathbf{I} \right)^{-1} \frac{1}{(1 - \alpha) q} \mathbf{I} U^T + \frac{1}{(1 - \alpha) q} \] (6)
where \( \mathbf{X} = UDV^T \) and \( q = \text{Tr}(S)/p = \text{Tr}(D^2)/np \).

Compressive RDA (CRDA)

We express LDA discriminant rule in vector form:
- \( d(x) = (d_1(x), \ldots, d_G(x)) \)
- \( x^T \mathbf{B} = -\frac{1}{2} \text{diag}(\mathbf{M}^T \mathbf{B}) + \ln p \) (7)

where \( \mathbf{M} = [\mu_1, \ldots, \mu_G] \) and \( \mathbf{B} = \Sigma^{-1} \mathbf{M} \) and \( \text{diag}(\mathbf{A}) = [a_1, \ldots, a_{G-1}] \) for some matrix \( \mathbf{A} \).

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

- It is a transformation \( H_k(B, \beta) \).
- It retains the elements of the \( K \) rows of \( B \) that possess largest \( \ell_1 \) norm and set elements of the other rows to zero.

Therefore, the rule (1) uses \( d_g(x) \) with \( \beta_g = \beta_g(\Sigma_g) = \Sigma_g^{-1}\mu_g \) in (2).

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

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 feature selection of (different) expressed features, i.e., genes in microarray studies.

References


Figures

- Figure 1: The simultaneous feature selection (SFS) algorithm.
- Figure 2: In Yodh et al. dataset (left) and after (right) the transformation \( H_k(B; \beta) \) instead of \( \beta_g \).

IEEE ICASSP 2018 • Calgary, Canada