

# Optimal Nearest Shrunken Centroids Method for High-dimensional Classification

Tiejun (Ty) Tong  
University of Colorado at Boulder

(Joint work with Herbert Pang)

December 11, 2009  
The 7th Annual Rocky Mountain Bioinformatics Conference  
Snowmass, CO

# High-dimensional Data Classification

- Let  $p$  denote the total number of covariates (e.g., genes), and  $n$  the total number of samples. This “**large  $p$  small  $n$** ” paradigm has posed challenges to traditional statistical methods.
- **Linear Discriminant Analysis (LDA)**: Assign a new subject  $\mathbf{y}$  to a class that minimizes the following discriminant score,

$$d_k(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{x}}_k)^T \hat{\Sigma}^{-1} (\mathbf{y} - \bar{\mathbf{x}}_k) - 2 \ln \hat{\pi}_k, \quad k = 1, \dots, K.$$

LDA requires that  $n \geq p$  to ensure  $\hat{\Sigma}$  non-singular, which usually doesn't hold for high-dimensional data.

- **Diagonal LDA** (Dudoit et al, 2002): By assuming a diagonal covariance matrix  $\Sigma$ ,

$$d_k(\mathbf{y}) = \sum_{i=1}^p (y_i - \bar{x}_{ik})^2 / s_i^2 - 2 \ln \hat{\pi}_k,$$

where  $\bar{x}_{ik}$  is the  $i$ th component of the centroid for class  $k$ . It is also called the Nearest Centroids classification.

## Nearest Shrunken Centroids (NSC) Classification

- The **Nearest Shrunken Centroids** (NSC) method (Tibshirani et al, 2002) is a modification to the Nearest Centroids classification,

$$d'_k(\mathbf{y}) = \sum_{i=1}^p (y_i - \bar{x}'_{ik})^2 / s_i^2 - 2 \ln \hat{\pi}_k.$$

where  $\bar{x}'_{ik}$  are the shrunken centroids that shrink the standard centroids  $\bar{x}_{ik}$  toward the overall centroid  $\bar{x}_i$ .

- Let  $d_{ik} = (\bar{x}_{ik} - \bar{x}_i) / (m_k s_i)$  be the standardized distance. The NSC method shrinks each  $d_{ik}$  toward zero (hard thresholding),

$$d'_{ik} = I(|d_{ik}| > \Delta) d_{ik}.$$

where  $\Delta$  is a shrinkage parameter. This leads to

$$\bar{x}'_{ik} = \bar{x}_i + m_k s_i d'_{ik}.$$

- Attractive Property: When a sufficiently large  $\Delta$  is used, many non-differentially expressed features will be **eliminated** from the study.

## New Algorithm

- Motivation: Though the NSC method works well in high-dimensional classification, it has the following limitation:

*When the sample size is small, the estimation of  $\Delta$  by cross-validation is fairly unstable which leads to a large variation in the number of features selected. As a consequence, the performance of the NSC method is unsatisfactory.*

- The data set of SRBCTs used in Tibshirani et al (2002) has **63** training samples and **25** test samples.
- We propose to estimate  $\Delta$  by minimizing the following risk function,

$$R(d', \tau; \Delta) = \frac{1}{pK} \sum_{i=1}^p \sum_{k=1}^K E\{L(d'_{ik}, \tau_{ik}; \Delta)\},$$

where

$L(\cdot)$  be a loss function,

$\tau_{ik} = (\mu_{ik} - \mu_i)/(m_k \sigma_i)$  be the true standardized distances,

$d' = (d'_{ik})$  are the shrunken estimates of  $\tau = (\tau_{ik})$ .

## New Algorithm

- Under the squared loss, the risk function is

$$R(d', \tau; \Delta) = 1 + \frac{1}{pK} \sum_{i=1}^p \sum_{k=1}^K \int_{-\Delta - \tau_{ik}}^{\Delta - \tau_{ik}} (\Delta^2 - x^2) \phi(x) dx.$$

The optimal shrinkage estimator is defined as

$$\Delta_{opt} = \operatorname{argmin}_{\Delta \geq 0} R(d', \tau; \Delta)$$

- We propose to estimate  $\Delta_{opt}$  by the following procedure:

$$\hat{R}(d', \tau; \Delta) = 1 + \frac{1}{pK} \sum_{i=1}^p \sum_{k=1}^K \int_{-\Delta - d_{ik}}^{\Delta - d_{ik}} (\Delta^2 - x^2) \phi(x) dx.$$

which leads to  $\hat{\Delta}_{opt} = \operatorname{argmin}_{\Delta \geq 0} \hat{R}(d', \tau; \Delta)$ .

- For any fixed  $n$ , under some mild conditions,

$$\begin{aligned} \hat{R}(d', \tau; \Delta) &\xrightarrow{a.s.} R(d', \tau; \Delta), \quad \text{as } p \rightarrow \infty; \\ \hat{\Delta}_{opt} &\xrightarrow{a.s.} \Delta_{opt}, \quad \text{as } p \rightarrow \infty. \end{aligned}$$

## Performance of optNSC – Binary Classification

- Let  $p = 1000$ ,  $\mu_1 = (0, 0, \dots, 0)$  and  $\mu_2 = (0, \dots, 0, 0.5, \dots, 0.5)$  with  $\pi_0 = 0.8$  or  $0.95$ .
- Simulate  $2n$  samples ( $n$  training samples and  $n$  testing samples) from  $N(\mu_k, I)$  for each class.
- We repeat the procedure 200 times and report their average misclassification error rates.

