On Some Fast And Robust Classifiers For HDLSS Data

Sarbojit Roy [1] Jyotishka Ray Choudhury [2] Subhajit Dutta [1]

[1]Indian Institute of Technology Kanpur [2]Indian Statistical Institute Kolkata

Classification in High Dimension, Low Sample Size Settings

Suppose $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ and $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jp})^\top$ are i.i.d. random vectors from distribution functions $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively, for $1 \le i \le n_1$ and $1 \le j \le n_2$. Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ denote the training sample of size $n = n_1 + n_2$, where $\mathcal{X}_1 = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\mathcal{X}_2 = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$. The class prior probabilities $0 < \pi_1, \pi_2 < 1$ satisfy $\pi_1 + \pi_2 = 1$.

Given the training sample $\mathcal{X}$, our aim is to develop a classifier δ such that its misclassification probability Δ goes to zero under fairly general conditions in the high dimension, low sample size (HDLSS) regime, where n is held fixed while $p \to \infty$.

Limitations of Existing Classifiers

- $\mu_1 = \mathrm{E}[\mathbf{X}]$, $\mu_2 = \mathrm{E}[\mathbf{Y}]$, $\Sigma_1 = \mathrm{Cov}[\mathbf{X}]$ and $\Sigma_2 = \mathrm{Cov}[\mathbf{Y}]$.
- Define $\nu^2 = \lim_{p\to\infty} \frac{1}{p}\|\mu_1 - \mu_2\|^2$ and $\sigma_j^2 = \lim_{p\to\infty} \frac{1}{p}\mathrm{trace}(\Sigma_j)$ for $j = 1, 2$.
- Existing classifiers yield *perfect classification* (i.e., $\Delta \to 0$ as $p \to \infty$) if
  - $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ for the nearest neighbor (NN) classifier, average distance (AVG) classifier, support vector machines (SVM) [2].
  - $\nu^2 > 0$, or $\sigma_1^2 \neq \sigma_2^2$ for the scale adjusted AVG (SAVG) classifier [1].
- In HDLSS settings, behavior of the existing classifiers is governed by the constants ν^2, σ_1^2 and σ_2^2.

Our Contribution

We propose classifiers that are **robust**, **computationally fast**, **free from tuning parameters** and have **strong theoretical properties**.

A Robust and Tuning-free Classifier

Define $h(u, v) = \sin^{-1}\left((1+uv)/\sqrt{(1+u^2)(1+v^2)}\right)/2\pi$ for $u, v \in \mathbb{R}$ and

$$\bar{h}(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\sum_{k=1}^{p} h(u_k, v_k) \text{ for } \mathbf{u} = (u_1, \ldots, u_p)^\top, \mathbf{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p.$$

$$\bar{T}_{11} = \sum_{i<j} \frac{\bar{h}(\mathbf{X}_i, \mathbf{X}_j)}{n_1(n_1-1)},\ \bar{T}_{12} = \sum_{i,j}\sum \frac{\bar{h}(\mathbf{X}_i, \mathbf{Y}_j)}{n_1 n_2},\ \bar{T}_{22} = \sum_{i<j} \frac{\bar{h}(\mathbf{Y}_i, \mathbf{Y}_j)}{n_2(n_2-1)},$$

$$\bar{T}_1(\mathbf{z}) = \frac{1}{n_1}\sum_{i=1}^{n_1} \bar{h}(\mathbf{X}_i, \mathbf{z}),\ \bar{T}_2(\mathbf{z}) = \frac{1}{n_2}\sum_{i=1}^{n_2} \bar{h}(\mathbf{Y}_i, \mathbf{z}),\ \bar{L}_j(\mathbf{z}) = \bar{T}_{jj} - 2\bar{T}_j(\mathbf{z}) \text{ for } j = 1, 2.$$

- Discriminant: $\bar{L}(\mathbf{z}) = \bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z})$.
- Classifier: $\delta_1(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{L}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases}$

A measure of distance between $\mathbf{F}_1$ and $\mathbf{F}_2$

- h is a bounded function and free of parameters.
- Define $\bar{\tau}_p = \mathrm{E}[\bar{h}(\mathbf{X}_1, \mathbf{X}_2)] + E[\bar{h}(\mathbf{Y}_1, \mathbf{Y}_2)] - 2E[\bar{h}(\mathbf{X}_1, \mathbf{Y}_1)]$ for $p \ge 1$.
- Clearly, $\bar{\tau}_p \ge 0$. Equality holds iff $F_{1k} = F_{2k}$ for all $1 \le k \le p$ where F_{1k} and F_{2k} are one-dimensional marginals of $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively (see [3]).
- $\bar{\tau}_p$ is a measure of distance between $\mathbf{F}_1$ and $\mathbf{F}_2$.
- Now, $\mathrm{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] = \bar{\tau}_p \ge 0$, while $\mathrm{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_2] = -\bar{\tau}_p \le 0$.

Limitations of $\bar{\tau}_p$

Define $\bar{\tau}_p(1,1) = \mathrm{E}[\bar{h}(\mathbf{X}_1, \mathbf{X}_2)]$, $\bar{\tau}_p(2,2) = E[\bar{h}(\mathbf{Y}_1, \mathbf{Y}_2)]$ and $\bar{\tau}_p(1,2) = E[\bar{h}(\mathbf{X}_1, \mathbf{Y}_1)]$.

$$\text{Observe that } \bar{\tau}_p = \{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\} + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}.$$

- Suppose $\bar{\tau}_p(1,1) < \bar{\tau}_p(1,2) < \bar{\tau}_p(2,2)$. Then, the value of $\bar{\tau}_p$ may become small.
- An improved dissimilarity index: $\bar{\psi}_p = \{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\}^2 + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}^2$.
  - Squaring the terms before addition eliminates the possibility of cancellations.
  - Further, $\bar{\psi}_p = 0$ iff $F_k = G_k$ for all $1 \le k \le p$.

A Classifier Based on $\bar{\psi}_p$

Define $\bar{T} = \bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}$.

- Discriminant: $\bar{\theta}(\mathbf{z}) = \bar{T}\bar{L}(\mathbf{z})/2 + \left\{\bar{T}_{22} - \bar{T}_{11}\right\}\left\{\bar{L}_1(\mathbf{z}) + \bar{L}_2(\mathbf{z}) + 2\bar{T}_{12}\right\}/2.$
- Classifier: $\delta_2(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases}$

Asymptotic Behavior in HDLSS Settings

Suppose $\mathbf{U} = (U_1, \ldots, U_p)^\top$ and $\mathbf{V} = (V_1, \ldots, V_p)^\top$ are two independent vectors such that $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$ for $j, j' \in \{1, 2\}$. Let us assume the following:

- Weak dependence among the component variables:

$$\text{(A1)} \sum_{1 \le k < k' \le p}\sum \mathrm{Corr}\Big(h(U_k, V_k), h(U_{k'}, V_{k'})\Big) = o(p^2).$$

  - Assumption (A1) is trivially satisfied if the component variables of the underlying distributions are independently distributed.
  - It continues to hold when the components have weak dependence among them. For example, (A1) is satisfied when $\{h(U_k, V_k), k \ge 1\}$ has the ρ-mixing property.
  - If assumption (A1) is satisfied, then we have the following:

$\mathbf{Z} \sim \mathbf{F}_1$	$\lvert\bar{L}(\mathbf{Z}) - \bar{\tau}_p\rvert \xrightarrow{\mathrm{P}} 0$ and $\lvert\bar{\theta}(\mathbf{Z}) - \bar{\psi}_p\rvert \xrightarrow{\mathrm{P}} 0$ as $p \to \infty$
$\mathbf{Z} \sim \mathbf{F}_2$	$\lvert\bar{L}(\mathbf{Z}) + \bar{\tau}_p\rvert \xrightarrow{\mathrm{P}} 0$ and $\lvert\bar{\theta}(\mathbf{Z}) + \bar{\psi}_p\rvert \xrightarrow{\mathrm{P}} 0$ as $p \to \infty$.

- Asymptotic separability of $\mathbf{F}_1$ and $\mathbf{F}_2$:

$$\text{(A2)} \liminf_{p\to\infty} \bar{\tau}_p > 0.$$

  - If the component variables are identically distributed, then (A2) is satisfied.
  - (A2) also implies that $\liminf_{p\to\infty} \bar{\psi}_p > 0$.

Theorem 1: Perfect Classification

If (A1) and (A2) are satisfied, then for any $\pi_1 > 0$, we have $\Delta_1 \to 0$ and $\Delta_2 \to 0$ as $p \to \infty$.

Relative Performance of δ_1 and δ_2

- Both δ_1 and δ_2 yield *perfect classification* under the same set of assumptions.
- We now provide a set of sufficient conditions under which one classifier outperforms the other. First, let us assume the following:

(A3) There exists a $p_0 \in \mathbb{N}$ such that $\bar{\tau}_p(1,2) > \min\{\bar{\tau}_p(1,1), \bar{\tau}_p(2,2)\}$ for all $p \ge p_0$.

- If assumption (A3) is satisfied, then either of $\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)$ and $\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)$ is positive, while the other quantity is negative.

Theorem 2: Ordering Between Δ_1 and Δ_2

If assumptions (A1) – (A3) are satisfied, then there exists an integer p_0' such that $\Delta_2 \le \Delta_1$ for all $p \ge p_0'$.

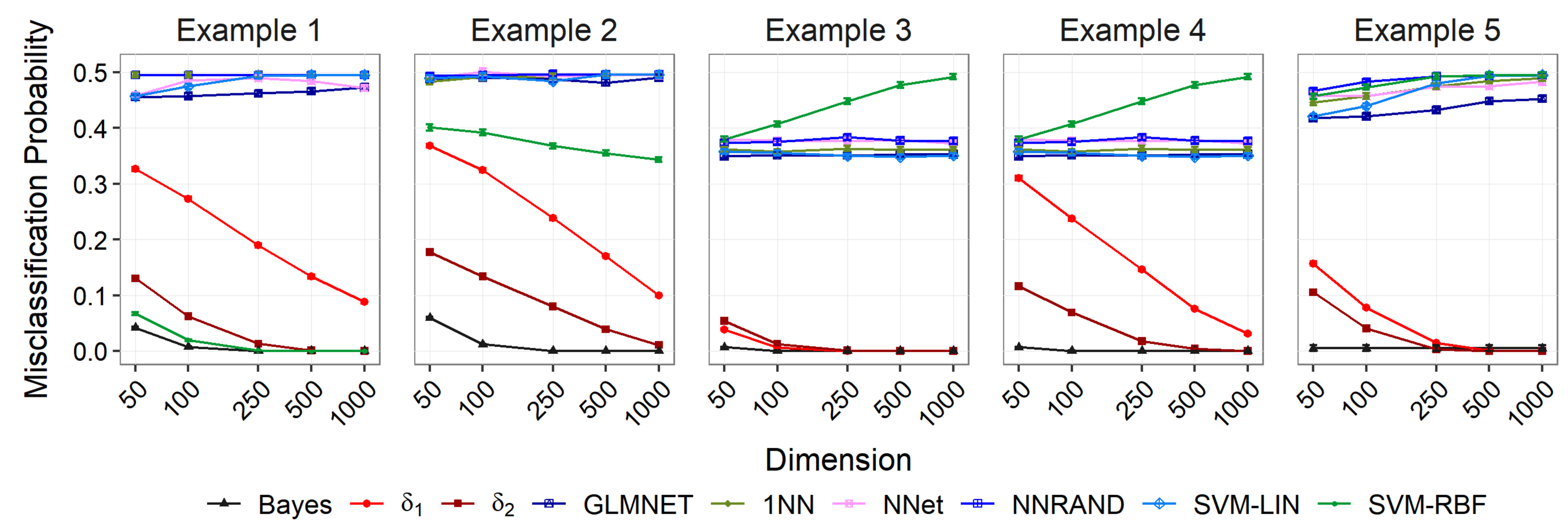
If the inequality in (A3) is inverted, then the ordering of Δ_1 and Δ_2 is reversed.

Simulation Study

Fix $1 \le k \le p$. Now, consider the following examples:

Example	$(\nu, \sigma_1^2, \sigma_2^2)$	$\bar{T}_{12} > \min\{\bar{T}_{11}, \bar{T}_{22}\}$
1. $X_{1k} \overset{i.i.d.}{\sim} N(1,1)$ and $Y_{1k} \overset{i.i.d.}{\sim} N(1,2)$	$\nu^2 < \lvert\sigma_1^2 - \sigma_2^2\rvert$	True
2. $X_{1k} \overset{i.i.d.}{\sim} N(0,3)$ and $Y_{1k} \overset{i.i.d.}{\sim} t_3$	$\nu^2 = \sigma_1^2 - \sigma_2^2 = 0$	True
3. $X_{1k} \overset{i.i.d.}{\sim} C(0,1)$ and $Y_{1k} \overset{i.i.d.}{\sim} C(1,1)$	do not exist	False
4. $X_{1k} \overset{i.i.d.}{\sim} C(0,1)$ and $Y_{1k} \overset{i.i.d.}{\sim} C(0,2)$	do not exist	True
5. $X_{1k} \overset{i.i.d.}{\sim} \mathrm{Par}(1,1)$ and $Y_{1k} \overset{i.i.d.}{\sim} \mathrm{Par}(1.25,1)$	do not exist	True

$N(\mu, \sigma)$: Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$.
t_α: the Student's t-distribution with $\alpha > 0$ degrees of freedom.
$C(\mu, \sigma)$: Cauchy distribution with location $\mu \in \mathbb{R}$ and scale $\sigma > 0$.
$\mathrm{Par}(\theta, s)$: Pareto distribution with $\theta > 0$ and scale $s > 0$.



References

[1] Yao-Ban Chan and Peter Hall. Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96(2):469–478, 2009.

[2] Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444, 2005.

[3] Zhimei Li and Yaowu Zhang. On a projective ensemble approach to two sample test for equality of distributions. In *International Conference on Machine Learning*, pages 6020–6027. PMLR, 2020.