Discriminant Analysis in High-Dimensional Gaussian Latent Mixtures

Marten Wegkamp

Department of Mathematics
Department of Statistics and Data Science
Cornell University
Ithaca, New York

Joint work with Xin (Mike) Bing,
University of Toronto

References

Based on

- *Xin Bing and Marten Wegkamp.* Interpolating Discriminant Functions in High-Dimensional Gaussian Latent Mixtures. *Biometrika* (2023)
- *Xin Bing and Marten Wegkamp.* Optimal Discriminant Analysis in High-Dimensional Latent Factor Models. *Annals of Statistics* (2023)

Outline

Introduction

Latent Factor Model

We observe independent copies of the pair (X, Y) with features $X \in \mathbb{R}^p$ according to

$$X = AZ + W$$

and labels $Y \in \{0, 1\}$.

- Only X is observed
- A is a deterministic, unknown $p \times K$ loading matrix
- $Z \in \mathbb{R}^K$ are unobserved, latent factors
- W is unobserved, random noise

Assumptions

(i) W is independent of both Z and Y

(ii) $\mathbb{E}[Z] = \mathbf{0}_K$, $\mathbb{E}[W] = \mathbf{0}_p$

(iii) A has rank K

(iv) $Z \mid Y = k \sim N_K(\alpha_k, \Sigma_{Z|Y})$ with $\alpha_k := \mathbb{E}[Z|Y = k]$ and

$$\Sigma_{Z|Y} := \mathrm{Cov}(Z|Y = 0) = \mathrm{Cov}(Z|Y = 1) > 0$$

(v) $W = \Sigma_W^{1/2} V$ with $\mathbb{E}[V] = \mathbf{0}_p$, $\mathbb{E}[VV^\top] = \boldsymbol{I}_p$ and

$$\sup_{\|u\|_2=1} \mathbb{E}[\exp(u^\top V)] \le \exp(\gamma^2/2)$$

(vi) For some absolute constant $c \in (0,1)$, $\min\{\pi_0, \pi_1\} \ge c$ with $\pi_k := \mathbb{P}\{Y = k\}$, $k = 0, 1$

Basic inequality

Lemma

Under (i), (ii), (iii), we have

$$R_x^* := \inf_g \mathbb{P}\{g(X) \neq Y\} \;\geq\; R_z^* := \inf_h \mathbb{P}\{h(Z) \neq Y\}$$

Oracle Benchmark

We have the explicit expression

$$R_z^* = 1 - \pi_1 \Phi\left(\frac{\Delta}{2} + \frac{\log \frac{\pi_1}{\pi_0}}{\Delta}\right) - \pi_0 \Phi\left(\frac{\Delta}{2} - \frac{\log \frac{\pi_1}{\pi_0}}{\Delta}\right).$$

Here

$$\Delta^2 := (\alpha_0 - \alpha_1)^\top \Sigma_{Z|Y}^{-1} (\alpha_0 - \alpha_1)$$

is the Mahalanobis distance between the conditional means $\alpha_0 = \mathbb{E}[Z \mid Y = 0]$ and $\alpha_1 = \mathbb{E}[Z \mid Y = 1]$.

Oracle Benchmark

- If $\Delta \to \infty$, then $R_z^* \to 0$. Trivial asymptotic Bayes error - Expect fast rates
- If $\Delta \to 0$ and $\pi_0 > \pi_1$, then $R_z^* \to \pi_1$. Trivial asymptotic Bayes rule votes 0 all the time - Expect fast rates
- If $\Delta \to 0$ and $\pi_0 = \pi_1 = 1/2$, then $R_z^* \to 1/2$. Asymptotic random guessing - Expect slow rates

Conclusion:

In a way, the most interesting case is $\Delta \asymp 1$.

Oracle Benchmark

$$\Sigma_{X|Y} = A\Sigma_{Z|Y}A^\top + \Sigma_W$$

If the signal-to-noise ratio

$$\xi := \frac{\lambda_K(A\Sigma_{Z|Y}A^\top)}{\lambda_1(\Sigma_W)}$$

for predicting Z from X given Y is large, the gap between R_x^* and R_z^* is small.

Minimax Lower Bounds

Minimax Lower Bound

We establish minimax-optimal rates of convergence of the excess risk

$$R_x(\widehat{g}) - R_z^* := \mathbb{P}\{\widehat{g}(X) \neq Y\} - \inf_h \mathbb{P}\{h(Z) \neq Y\}$$

for any classification rule $\widehat{g} : \mathbb{R}^p \to \{0, 1\}$ based on independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ from our factor model (i)–(iv).

Minimax Lower Bound

- Define the parameter space of $\theta := (A, \Sigma_{Z|Y}, \Sigma_W, \alpha)$ as

$$\pi_0 = \pi_1 = 1/2$$
$$\lambda_1(\Sigma_W) \asymp \lambda_p(\Sigma_W) \asymp \sigma^2$$
$$\lambda_1(A\Sigma_{Z|Y}A^\top) \asymp \lambda_K(A\Sigma_{Z|Y}A^\top) \asymp \lambda$$

Set

$$\omega^2 := \frac{K}{n} + \frac{\sigma^2}{\lambda}\Delta + \frac{\sigma^2 p}{\lambda n}\frac{\sigma^2}{\lambda}\Delta.$$

Theorem

Assume (i) – (vi), $K \geq 2$, $K/(n \wedge p) \leq c_1$, $\sigma^2/\lambda \leq c_2$ and $\sigma^2 p/(\lambda n) \leq c_3$ for some small constants $c_1, c_2, c_3 > 0$.

1. If $\Delta \asymp 1$, then there exists some constants $c_0 \in (0,1)$ and $C > 0$ such that
$$\inf_{\widehat{g}} \sup_{\theta} \mathbb{P}_\theta \left\{ R_x(\widehat{g}) - R_z^* \geq C\omega^2 \right\} \geq c_0.$$
2. If $\Delta \to \infty$ and $\sigma^2/\lambda \to 0$, as $n \to \infty$, then there exists some constants $c_0 \in (0,1)$ and $C > 0$ such that
$$\inf_{\widehat{g}} \sup_{\theta} \mathbb{P}_\theta \left\{ R_x(\widehat{g}) - R_z^* \geq C\omega^2 e^{-\frac{1}{8}\Delta^2 + o(\Delta^2)} \right\} \geq c_0.$$
3. If $\Delta \to 0$, as $n \to \infty$, then there exists some constants $c_0 \in (0,1)$ and $C > 0$ such that
$$\inf_{\widehat{g}} \sup_{\theta} \mathbb{P}_\theta \left\{ R_x(\widehat{g}) - R_z^* \geq C\omega \min\left(\frac{\omega}{\Delta}, 1\right) \right\} \geq c_0.$$

Minimax lower bound

$$\omega^2 := \frac{K}{n} + \frac{\sigma^2}{\lambda}\Delta + \frac{\sigma^2 p}{\lambda n}\frac{\sigma^2}{\lambda}\Delta.$$

The lower bounds consist of three terms:

- the one related with K/n is the optimal rate of the excess risk even when Z were observable;
- the second one related with σ^2/λ is the irreducible error for not observing Z;
- the last one involving $\sigma^2 p/(\lambda n) \times (\sigma^2/\lambda)$ is the price to pay for estimating the column space of A.
- The third term can be absorbed by the second term as $\sigma^2 p/(\lambda n) \leq c_3$.
- The lower bounds are tight (later).

Methodology

Methodology

To motivate our approach, suppose that we observe Z.
The optimal Bayes rule to classify a new point $z \in \mathbb{R}^K$ is

$$g_z^*(z) = \mathbb{1}\{z^\top \eta + \eta_0 \geq 0\}$$

where

$$\eta = \Sigma_{Z|Y}^{-1}(\alpha_1 - \alpha_0), \qquad \eta_0 = -\frac{1}{2}(\alpha_0 + \alpha_1)^\top \eta + \log \frac{\pi_1}{\pi_0}.$$

This rule is optimal in the sense that it has the smallest possible misclassification error $\mathbb{P}\{Y \neq g(Z)\}$.

- Modern efficient empirical LDA in the high-dimensional setting exploit potential sparsity of $\Sigma_{X|Y}^{-1}(\mu_1 - \mu_0)$.
  See, e.g., Tibshirani et al (2002), Fan and Fan (2008), Witten and Tibshirani (2011), Shao, Wang, Deng, Wang (2011), Cai and Liu (2011), Mai, Zou, Yuan (2012), Cai and Zhang (2019ab).
- In the high-dimensional regime, many features are highly correlated and any sparsity assumption becomes questionable.
- Instead: assume low-dimensional structure and "classify projections".

Connection between LDA and Regression

Let $\Sigma_Z = \mathbb{E}[ZZ^\top]$ be the unconditional covariance matrix of Z. Define

$$\beta = \pi_0\pi_1\Sigma_Z^{-1}(\alpha_1 - \alpha_0),$$
$$\beta_0 = -\frac{1}{2}(\alpha_0 + \alpha_1)^\top\beta + \pi_0\pi_1\left[1 - (\alpha_1 - \alpha_0)^\top\beta\right]\log\frac{\pi_1}{\pi_0}.$$

Proposition

Under Assumptions (ii) and (iv), we have

$$z^\top\eta + \eta_0 \geq 0 \quad \Longleftrightarrow \quad z^\top\beta + \beta_0 \geq 0.$$

Furthermore,

$$\beta = \Sigma_Z^{-1}\mathbb{E}[ZY].$$

Methodology

- The key difference is the use of the unconditional Σ_Z, as opposed to the conditional $\Sigma_{Z|Y}$.
- We can interpret β as the regression coefficient of Y on Z. This suggests to estimate β via least squares.
- We only have access to $x \in \mathbb{R}^p$, $\boldsymbol{X} = [X_1 \cdots X_n]^\top \in \mathbb{R}^{n \times p}$, and $\boldsymbol{y} = (Y_1, \ldots, Y_n)^\top \in \{0,1\}^n$.
- Since $\boldsymbol{X} = \mathbf{Z}A^\top + \mathbf{W}$, we need to find some appropriate matrix $B \approx A(A^\top A)^{-1}$ so that $\boldsymbol{X}B \approx \mathbf{Z} + \mathbf{W}A(A^\top A)^{-1}$.

Methodology

Estimate the inner-product $z^\top \beta$ by

$$x^\top \widehat{\theta} \;:=\; x^\top B(\boldsymbol{X}B)^+ \boldsymbol{y} \;=\; x^\top B(B^\top \boldsymbol{X}^\top \boldsymbol{X}B)^+ B^\top \boldsymbol{X}^\top \boldsymbol{y}$$

for some appropriate matrix B.

Estimate β_0 by

$$\widehat{\beta}_0 := -\frac{1}{2}(\widehat{\mu}_0 + \widehat{\mu}_1)^\top \widehat{\theta} + \widehat{\pi}_0 \widehat{\pi}_1 \left[1 - (\widehat{\mu}_1 - \widehat{\mu}_0)^\top \widehat{\theta}\right] \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0}$$

based on standard non-parametric estimates

$$n_k = \sum_{i=1}^{n} \mathbb{1}\{Y_i = k\}, \quad \widehat{\pi}_k = \frac{n_k}{n}, \quad \widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} X_i \mathbb{1}\{Y_i = k\}.$$

Our proposed classifier (Bing and W. 2023) is

$$\widehat{g}_x(x) := \mathbb{1}\{x^\top \widehat{\theta} + \widehat{\beta}_0 \geq 0\}.$$

The estimates $\widehat{\theta}$ and $\widehat{\beta}_0$ depend on B.

- We investigate $B = \boldsymbol{U}_r \in \mathbb{R}^{p \times r}$, where $\boldsymbol{U}_r$ consists of the first r right-singular vectors of $\widetilde{\boldsymbol{X}}$.
- $\widetilde{\boldsymbol{X}}$ is an auxiliary $n \times p$ data matrix (unlabelled observations only), independent of the training data $(\boldsymbol{X}, \boldsymbol{y})$. If not available, split the data in two equal parts.
- What if we use $\boldsymbol{X}$ instead?

PCR-based LDA

Bing and W. (2019, 2023) propose to use $r = \widehat{K}$ with

$$\widehat{K} := \arg \min_{0 \le k \le \bar{K}} \frac{\sum_{j>k} \sigma_j^2}{np - 2.1(n+p)k}$$

based on the singular-values σ_j of $\widetilde{\boldsymbol{X}}$, with $\bar{K} < \lfloor \frac{1}{4.2}(n \wedge p) \rfloor$.

Real data analysis

- We analyze three popular gene expression datasets (leukemia data, colon data and lung cancer data).
- For all three data sets, the features are standardized to zero mean and unit standard deviations.
- For each dataset, we randomly split the data, within each category, into 70% training set and 30% test set.
- We compare our proposed algorithm, PCLDA-$\widehat{K}$, with the
  - Nearest Shrunken Centroids classifier (PAMR) of Tibshirani, Hastie, Narasimhan, Chu (2002),
  - ℓ_1-Penalized Linear Discriminant (PenalizedLDA) of Witten and Tibshirani (2011),
  - Direct Sparse Discriminant (DSDA) of Mai, Zou, Yuan (2012).

Data name	p	n	n_0 (category)	n_1 (category)
Leukemia	7129	72	47 (acute lymphoblastic leukemia)	25 (acute myeloid leukemia)
Colon	2000	62	22 (normal)	40 (tumor)
Lung cancer	12533	181	150 (adenocarcinoma)	31 (malignant pleural mesothelioma)

Summary of three data sets.

	PCLDA-$\widehat{K}$	DSDA	PenalizedLDA	PAMR
Leukemia	**3.57** (0.036)	5.52 (0.044)	3.91 (0.043)	4.61 (0.039)
Colon	**16.37** (0.077)	18.11 (0.07)	33.95 (0.086)	19.00 (0.089)
Lung cancer	**0.55** (0.008)	1.69 (0.017)	1.80 (0.026)	0.91 (0.011)

The averaged misclassification errors (in percentage). The numbers in parentheses are the standard deviations over 100 repetitions.

Rates of convergence for the excess risk

General method for deriving upper bounds

We view R_z^* as an oracle risk since the Z_i aren't observed.

Our proposed classifier is designed to estimate the Bayes classifier g_z^* in $\mathbb{R}^K$ and to adapt to the underlying low-dimensional structure.

We define

$$\widehat{G}_x(x) := x^\top \widehat{\theta} + \widehat{\beta}_0, \qquad G_z(z) := z^\top \beta + \beta_0$$

so that $\widehat{g}_x(x) = \mathbb{1}\{\widehat{G}_x(x) \geq 0\}$ and $g_z^*(z) = \mathbb{1}\{G_z(z) \geq 0\}$.

Theorem

Set $c_* = (1 + \pi_0\pi_1\Delta^2)/(\pi_0\pi_1)$. For all $t > 0$,

$$R_x(\widehat{g}_x) - R_z^* \;\leq\; \mathbb{P}\{|\widehat{G}_x(X) - G_z(Z)| > t\} \;+ c_* t P(t),$$

with

$$\begin{aligned} P(t) := \;& \pi_0 \mathbb{P}\{-c_* t < G_z(Z) < 0 \mid Y = 0\} + \\ & \pi_1 \mathbb{P}\{0 < G_z(Z) < c_* t \mid Y = 1\}. \end{aligned}$$

Rate depends on

- estimate of optimal half space
- behavior around the decision boundary

Explicit expression for $P(t)$

Since Z is Gaussian, $P(t)$ can be simplified.

Proposition

Assume (i) – (iv). For all $\omega_n \to 0$, the exists $0 < c < 1/8$,

$$P(\omega_n) \lesssim \begin{cases} \omega_n & \text{if } \Delta \asymp 1 \\ \omega_n \exp(-c\Delta^2) & \text{if } \Delta \to \infty \\ \omega_n \exp(-c/\Delta^2) & \text{if } \Delta \to 0 \text{ and } \pi_0 \neq \pi_1 \\ \min(1, \omega_n/\Delta) & \text{if } \Delta \to 0 \text{ and } \pi_0 = \pi_1 = 1/2 \end{cases}$$

Estimation of optimal boundary

Since

$$\widehat{G}_x(X) - G_z(Z) = Z^\top (A^\top \widehat{\theta} - \beta) + W^\top \widehat{\theta} + \widehat{\beta}_0 - \beta_0$$

the key quantities to bound are

- $\|\widehat{\theta}\|_2$
- $\|\Sigma_Z^{1/2}(A^\top \widehat{\theta} - \beta)\|_2$.

Rates of Convergence

Theorem - simplified case

Let $\theta \in \Theta(\lambda, \sigma, \Delta)$ with $\Delta \asymp 1$ and $\kappa(A\Sigma_Z A^\top) \asymp 1$. With probability $1 - \mathcal{O}(n^{-1})$,

$$R_x(\widehat{g}_x) - R_z^* \lesssim \left[\frac{K \log n}{n} + \frac{\sigma^2}{\lambda} + \left(\frac{p}{n}\frac{\sigma^2}{\lambda}\right)^2\right] \log n, \quad \text{if } B = \boldsymbol{U}_K;$$

$$R_x(\widehat{g}_x) - R_z^* \lesssim \left[\frac{K \log n}{n} + \frac{\sigma^2}{\lambda}\right] \log n, \quad \text{if } B = \widetilde{\boldsymbol{U}}_K.$$

Rates of Convergence

(1) If $p < n$, the two rates coincide and consistency of both PC-based classifiers requires that $K \log^2 n/n \to 0$ and $\sigma^2 \log n/\lambda \to 0$.

(2) If $p > n$, and

$$\frac{\lambda}{\sigma^2} \gtrsim \min\left\{\left(\frac{p}{n}\right)^2, \frac{p}{\sqrt{nK \log n}}\right\},$$

the two rates coincide.

(3) If $p > n$ and λ/σ^2 is relatively small, the effect of using $B = \widetilde{\boldsymbol{U}}_K$ based on an independent data set $\tilde{\boldsymbol{X}}$ is real as evidenced on the next slide where we keep λ/σ^2, n and K fixed but let p grow.

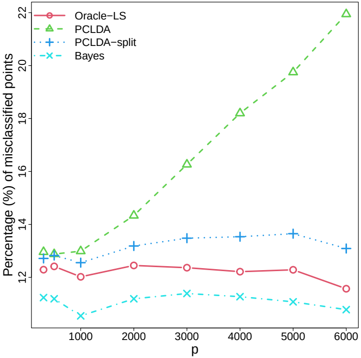


Illustration of the advantage of constructing $\widetilde{\boldsymbol{U}}_K$ from an independent dataset: PCLDA represents the PC-based classifier based on $B = \boldsymbol{U}_K$ while PCLDA-split uses $B = \widetilde{\boldsymbol{U}}_K$ that is constructed from an independent $\tilde{\boldsymbol{X}}$. Oracle-LS is the oracle benchmark that uses both Z and $\mathbf{Z}$ while Bayes represents the risk of using the oracle Bayes rule. We fix $n = 100$ and $K = 5$ and keep λ/σ^2 fixed, while we let p grow.

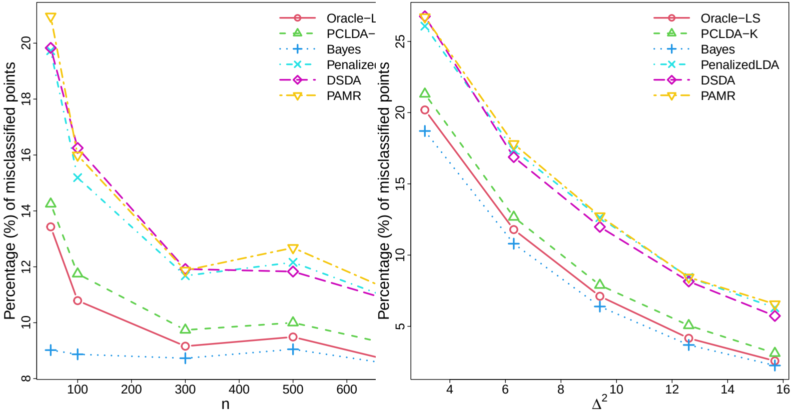
Simulations

Simulations

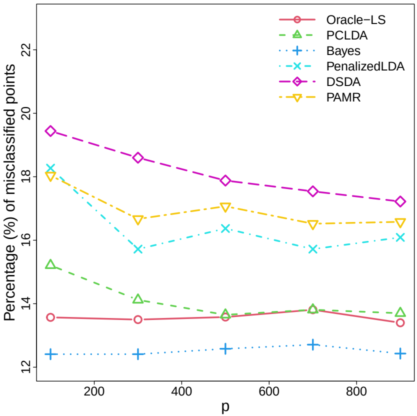
- We set $\pi_0 = \pi_1 = 1/2$, $\alpha_0 = -\alpha_1 = -(\frac{1}{2}\sqrt{\eta/K})\,\mathbf{1}_K$.
- The parameter η controls the signal strength Δ.
- We generate $\Sigma_{Z|Y}$ as follows:
  - $[\Sigma_{Z|Y}]_{ii}$ are iid Unif(1,3)
  - $[\Sigma_{Z|Y}]_{ij} = \sqrt{[\Sigma_{Z|Y}]_{ii}[\Sigma_{Z|Y}]_{jj}}(-1)^{i+j}(0.5)^{|i-j|}$ for each $i \neq j$.
- We generate Σ_W in the same way, except $\mathrm{diag}(\Sigma_W) = \mathbf{1}_p$.
- Rows of $\mathbf{W} \in \mathbb{R}^{n\times p}$ are iid $N_p(0, \Sigma_W)$.
- Entries of A are iid $N(0, 0.3^2)$.

$\eta = 5,\ K = 10,\ p = 300$ and $n \in \{50, 100, 300, 500, 700\}$

$K = 5,\ n = 100,\ p = 300$ and $\eta \in \{2, 4, 6, 8, 10\} \Longrightarrow \Delta^2 \in \{3.1, 6.3, 9.4, 12.6, 15.7\}$



$K = 5,\ \eta = 5,\ n = 100$ and $p \in \{100, 300, 500, 700, 900\}$.



Interpolation

Question:

What happens if $B = \boldsymbol{I}_p$, hence $\widehat{\theta} = \boldsymbol{X}^+\boldsymbol{y}$ (generalized least squares)?

Interpolation

Phenomenon: deep neural networks

It is possible to achieve good generalization error despite zero training error (overfitting)!

- In regression context: Bartlett et al (2020), Belkin et al (2018), Hastie et al (2022)
  For this model: Bing, Bunea, Strimas-Mackey, W (2021), Bunea, Strimas-Mackey, W (2022)
- For binary classification: Cao et al (2021), Chatterji and Long (2021), Hsu et al (2021), Minsker et al (2021), Muthukumar et al (2019), Wang and Thrampoulidis (2021)

- Current literature on classification considers
  - Decision boundaries are hyperplanes through origin
  - Misclassification risk, not excess risk, is bounded.
- These interpolation methods without intercept actually fail when the mixture probabilities are asymmetric and the Bayes error does not vanish.

Our results:

- We will show that $\widehat{g}(x) = x^\top \widehat{\theta} + \widehat{\beta}_0$ has zero training error, but is inconsistent due to plug-in estimate $\widehat{\beta}_0$.
- We need to use an independent hold-out sample to estimate intercept β_0 to obtain consistency and sometimes even minimax optimality.
- The interpolation property may be destroyed. However, if we encode the labels differently, e.g., via ± 1, interpolation is preserved (if one cares).
- We provide a concrete instance of the interesting phenomenon that overfitting and minimax-optimal generalization performance can coexist in a latent low-dimensional statistical model, against traditional statistical belief.

Interpolation

Proposition (Bunea, Strimas-Mackey, W 2022)

Assume $n \geq K$. Then, there exist finite, positive constants C, c depending on σ only, such that, provided $r_e(\Sigma_W) = tr(\Sigma_W)/\|\Sigma_W\|_{op} \geq Cn$,

$$\mathbb{P}\left\{\sigma_n^2(\boldsymbol{X}) \geq \frac{1}{8}\mathrm{tr}(\Sigma_W)\right\} \geq 1 - 3\exp(-c\ n)$$

Corollary: interpolation is common

Assume $p \geq n \geq K$, $\|\Sigma_W\|_{op} \asymp 1$ and $tr(\Sigma_W) \asymp p$. Then the GLS $\widehat{\theta} = \boldsymbol{X}^+\boldsymbol{y}$ interpolates the data

$$\lim_{n\to\infty} \mathbb{P}\{\boldsymbol{X}\widehat{\theta} = \boldsymbol{y}\} = 1.$$

Interpolation

Observation: zero training error if intercept in $(-1, 0]$

If $\widehat{\theta} = \boldsymbol{X}^{+}\boldsymbol{y}$ interpolates, then the classifier

$$\mathbb{1}\{x^{\top}\widehat{\theta} + \bar{\beta}_0 > 0\}$$

perfectly classifies the training data for *any* $\bar{\beta}_0 \in (-1, 0]$ (including zero intercept).

Simply note that, as long as $\bar{\beta}_0 \in (-1, 0]$,

$$X_i^{\top}\widehat{\theta} + \bar{\beta}_0 = Y_i + \bar{\beta}_0 > 0 \iff Y_i = 1, \text{ for all } i \in [n]$$

We will argue that interpolation depends on how we encode labels

Interpolation

Question:

Does the classifier $\mathbb{1}\{x^\top \widehat{\theta} + \beta_0 > 0\}$ that uses the true intercept β_0 yield zero training error ?

This is equivalent with verifying if $\beta_0 \in (-1, 0]$.

Answer:

It depends! Only if we encode the majority class as 0.

Lemma

The true intercept β_0 satisfies

$$\operatorname{sgn}(\beta_0) = \operatorname{sgn}\left(\frac{1}{2} - \pi_0\right), \qquad |\beta_0| \le \left|\frac{1}{2} - \pi_0\right|.$$

Observation

- The optimal decision boundary in the latent space is independent of the particular encoding.
- Interpolation property crucially depends on the way we encode the labels.
- For instance, if we encode Y as $\{-1, 1\}$, the classifier

$$2\mathbb{1}\{x^\top \widehat{\theta} + 2\beta_0 > 0\} - 1$$

always has zero training error (as $|\beta_0| \leq 1/2$).

Interpolation leads to inconsistency

The following lemma shows that $\widehat{\beta}_0 = -1/2$, irrespective of the true value of β_0, whenever $\widehat{\theta}$ interpolates.

Proposition

Let $\widehat{\beta}_0$ be the plug-in estimate. On the event $\{\boldsymbol{X}\widehat{\theta} = \boldsymbol{y}\}$ where $\widehat{\theta}$ interpolates, we have $\widehat{\beta}_0 = -1/2$.

- $\widehat{g}(x) = \mathbb{1}\{x^\top \widehat{\theta} + \widehat{\beta}_0 > 0\}$ always interpolates as $\widehat{\beta}_0 \in (-1, 0]$.
- $\widehat{\beta}_0$ is an inconsistent estimate of β_0 in general.
- Confirmed in simulations: classifier is inconsistent.

What can we do?

1. $\pi_0 = \pi_1 = 1/2$. In this case $\beta_0 = 0$, no need to estimate β_0 (current literature).
2. $\pi_0 \neq \pi_1$. Estimate β_0 by

$$\widetilde{\beta}_0 := -\frac{1}{2}(\widetilde{\mu}_0 + \widetilde{\mu}_1)^\top \widehat{\theta} + \left[1 - (\widetilde{\mu}_1 - \widetilde{\mu}_0)^\top \widehat{\theta}\,\right] \widehat{\pi}_0 \widehat{\pi}_1 \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0}$$

with $\widehat{\theta}$ and $\widehat{\pi}_k$ as before, but

$$\widetilde{\mu}_k = \frac{1}{\widetilde{n}_k} \sum_{i=1}^{n'} X_i' \mathbb{1}\{Y_i' = k\}, \quad \widetilde{n}_k = \sum_{i=1}^{n'} \mathbb{1}\{Y_i' = k\}$$

are based on an independent hold-out sample of size $n' \asymp n$.

Modified classifier

Theorem: Simplified rates of convergence

Suppose

$$\theta \in \Theta(\lambda, \sigma, A), \;\; p \gg n \gg K, \;\; \Delta \asymp 1, \;\; n \asymp n', \;\; \kappa \asymp 1$$

Then $\widetilde{g}(x) = \mathbb{1}\{x^\top \widehat{\theta} + \widetilde{\beta}_0 > 0\}$ satisfies

$$R_x(\widetilde{g}) - R_z^* \;\lesssim\; \left[\frac{K \log(n)}{n} + \frac{n}{p} + \left(\frac{p}{n\,\xi} \right)^2 + \frac{1}{\xi} \right] \log(n).$$

Simplified rates of convergence

Summary

- If $\xi \gg p/n$, then $\widetilde{g}$ is consistent
- If, furthermore, $\xi \gtrsim (p/n) \cdot (n/K)^{1/2}$, then

$$\mathbb{P}\{\widetilde{g}(X) \neq Y\} - R_z^* \lesssim \frac{K}{n} \log^2(n) + \frac{n}{p} \log(n).$$

- If, in addition, $p \gtrsim n^2/K$, then $\widetilde{g}$ is minimax-optimal.

Simulations

Simulations

We generated the data as follows:

- $\pi_0 = \pi_1 = 0.5$
- $\alpha_0 = -\alpha_1$, $\alpha_1 = \mathbf{1}_K \sqrt{2/K}$
- $\Sigma_{Z|Y} = \boldsymbol{I}_K$ (This implies $\Delta^2 = 8$).
- Entries of $\mathbf{W}$ and A are independent realizations of $N(0, 1)$ and $N(0, 0.3^2)$, respectively.

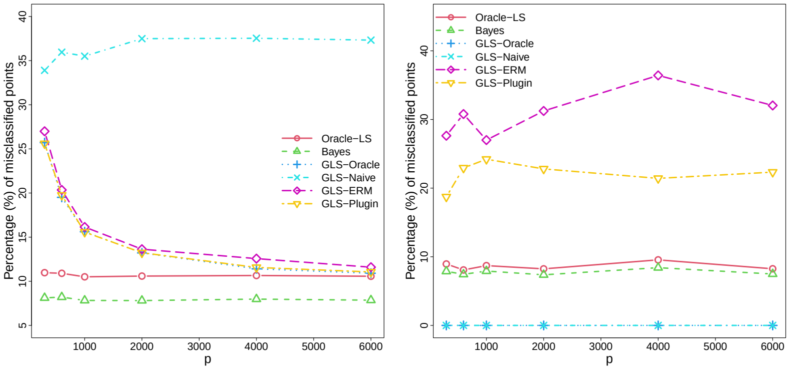
Simulations

We first verify the inconsistency of the naive classifier that uses the naive plug-in estimator of β_0 and contrast with other consistent classifiers.

- GLS-Naive: classifier $\widehat{g}(x) = \mathbb{1}\{x^\top \widehat{\theta} + \widehat{\beta}_0 > 0\}$ with $\widehat{\beta}_0$ being the naive plug-in estimator
- GLS-Oracle, GLS-Plugin and GLS-ERM represent $\mathbb{1}\{x^\top \widehat{\theta} + \bar{\beta}_0 > 0\}$ with $\bar{\beta}_0$ chosen as the true β_0, the plug-in estimate based on data splitting, and the estimate based on empirical risk minimization, respectively.
- Besides the optimal Bayes classifier (Bayes), we also choose the oracle procedure (Oracle-LS) that uses both $\mathbf{Z}$ and Z as our benchmark.

Simulations

The performance of all classifiers on 200 test data points, averaged over 100 simulations, for $K = 5$ and $n = 100$, and $p \in \{300, 600, 1000, 2000, 4000, 6000\}$.



Simulations

- We evaluate the performance of our proposed classifier and examine its dependence on p, K and ξ.
- We consider the misclassification error on 200 test data points, the estimation error $\|\beta - A^\top \widehat{\theta}\|_{\Sigma_Z}$ of β, and the estimation error $|\widetilde{\beta}_0 - \beta_0|$ of β_0.
- The sample size is fixed as $n = 100$ and we use a validation set with 100 data points to compute $\widetilde{\beta}_0$.

Simulations

Setting	Misclassification errors	Errors of estimating β	Errors of estimating β_0
$K = 5,\ \sigma_A = 0.3$			
$p = 300$	0.256 (0.046)	0.144 (0.052)	0.040 (0.031)
$p = 600$	0.198 (0.037)	0.127 (0.046)	0.034 (0.023)
$p = 1000$	0.156 (0.032)	0.117 (0.041)	0.029 (0.021)
$p = 2000$	0.132 (0.034)	0.115 (0.039)	0.029 (0.024)
$p = 4000$	0.116 (0.027)	0.112 (0.032)	0.027 (0.020)
$p = 1000,\ \sigma_A = 0.3$			
$K = 3$	0.152 (0.033)	0.091 (0.039)	0.028 (0.020)
$K = 5$	0.161 (0.029)	0.117 (0.039)	0.032 (0.022)
$K = 10$	0.178 (0.036)	0.180 (0.036)	0.033 (0.027)
$K = 15$	0.186 (0.038)	0.219 (0.040)	0.030 (0.022)
$p = 1000,\ K = 5$			
$\sigma_A = 0.01$	0.479 (0.038)	0.397 (0.004)	0.048 (0.039)
$\sigma_A = 0.05$	0.282 (0.039)	0.239 (0.024)	0.034 (0.026)
$\sigma_A = 0.1$	0.187 (0.035)	0.124 (0.037)	0.029 (0.019)
$\sigma_A = 0.24$	0.161 (0.033)	0.109 (0.034)	0.029 (0.022)

Thank you!