Classical vs high-dimensional theory

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Sur, Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences, 116, 14516–14525

Classical theory

- It concerns the behaviour when the *sample size* $n \to \infty$
- Suppose $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \underset{p\times 1}{Y}$ with mean $\mu = \mathbb{E}(Y)$ and finite variance $\Sigma = \mathbb{V}\mathrm{ar}(Y)$
- *Law of large numbers*: the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ converges in probability to μ
- *Central limit theorem*: the rescaled deviation $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix Σ
- *Consistency of maximum likelihood estimation*
- Etc.

Suppose that we are given $n = 1000$ samples from a statistical model in $p = 500$ dimensions

Will theory that requires $n \to \infty$ with the dimension p remaining fixed provide useful predictions?

High-dimensional data

- The data sets arising in many parts of modern science have a "high-dimensional flavor", with p on the same order as, or possibly larger than n

$$p \gg n$$

- Classical "large n, fixed p" theory fails to provide useful predictions
- Classical methods can break down dramatically in high-dimensional regimes

Linear discriminant analysis in high-dimensions

Classification problem

- Let's turn to the classification problem involving the allocation of the observed unit x to one of two classes A and B
- For a Bayesian analysis suppose that the prior probabilities are $\pi_A \equiv \mathrm{P}(Y = A)$ and $\pi_B \equiv \mathrm{P}(Y = B)$ with $\pi_A + \pi_B = 1$. Then the posterior probabilities satisfy

$$\frac{\mathrm{P}(Y = B | X = x)}{\mathrm{P}(Y = A | X = x)} = \frac{\pi_B}{\pi_A} \frac{f_B(x)}{f_A(x)}$$

  giving the class with the higher posterior probability
- As a special case, suppose that the two classes are distributed as multivariate Gaussians $X_A \sim N(\mu_A, I_p)$ and $X_B \sim N(\mu_B, I_p)$, with $\pi_A = \pi_B = 1/2$

Optimal decision

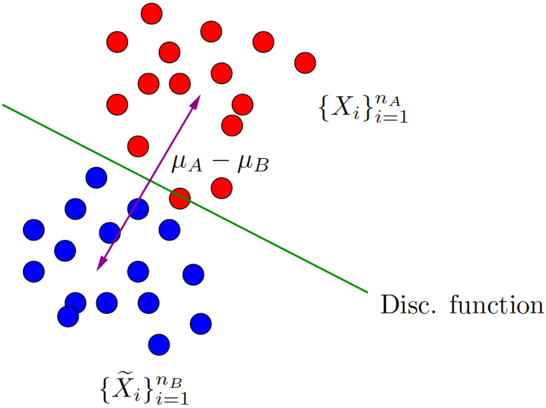
- The optimal decision rule is to threshold the log-likelihood ratio

$$\Psi(x) = \langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2}\right) \rangle$$

  where $\langle x, z \rangle = x^t z = \sum_{j=1}^{p} x_j z_j$ denotes the Euclidean inner product in $\mathbb{R}^p$
- If $\Psi(x) > 0$ then classify A, otherwise B
- Error probability of the optimal rule:

$$\mathrm{Err}(\Psi) = \frac{1}{2}\mathrm{P}(\Psi(X_A) < 0) + \frac{1}{2}\mathrm{P}(\Psi(X_B) \geq 0) = \Phi\left(-\frac{\gamma}{2}\right)$$

  where $\gamma = \|\mu_A - \mu_B\|$, $\|\mu\| = \sqrt{\mu^t \mu}$, and Φ is the cdf of a standard normal variable



$$\langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2}\right)\rangle = 0$$

source: Wainwright

Linear Discriminant Analysis

- Fisher's LDA: uses the plug-in principle based on n_A samples from class A and n_B samples from class B

$$\hat{\Psi}(x) = \langle \hat{\mu}_A - \hat{\mu}_B, x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \rangle$$

- Error probability of LDA (is itself a random variable)

$$\mathrm{Err}(\hat{\Psi}) = \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_A) < 0) + \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_B) \geq 0)$$

- Classical theory: if $(n_A, n_B) \to \infty$ and p remains fixed, then $\hat{\mu}_A \overset{prob.}{\to} \mu_A$, $\hat{\mu}_B \overset{prob.}{\to} \mu_B$ and the asymptotic error probability is $\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \mathrm{Err}(\Psi) = \Phi(-\gamma/2)$

High-Dimensional Theory

- What happens if $(n_A, n_B, p) \to \infty$ with
  - $p/n_A \to \delta$ with $\delta \geq 0$
  - $p/n_B \to \delta$
  - $\|\mu_A - \mu_B\|_2 \to \gamma > 0$
- Kolmogorov (1960) showed that

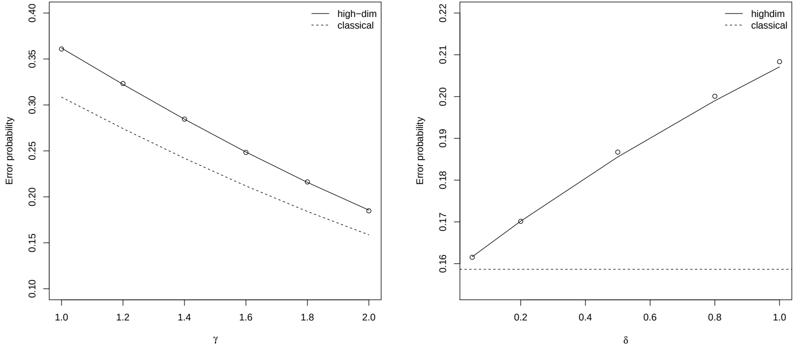
$$\mathrm{Err}(\hat{\Psi}) \overset{\textit{prob.}}{\to} \Phi\Big(- \frac{\gamma^2}{2\sqrt{\gamma^2 + 2\delta}}\Big)$$

- If $p/n \to 0$, then the asymptotic error probability is $\Phi(-\gamma/2)$ as is predicted by classical theory
- If $p/n \to \delta > 0$, then the asymptotic error probability is strictly larger than $\Phi(-\gamma/2)$

The error probability of $\hat{\Phi}$, for the finite triple

$$(p, n_A, n_B) = (400, 800, 800)$$

is better described by the classical $\Phi(-\gamma/2)$, or the high-dimensional analog $\Phi(-\gamma^2/(2\sqrt{\gamma^2 + 2\delta}))$?



circles correspond to the empirical error probabilities, averaged over 10 trials

What can help us in high dimensions?

- An important fact is that high-dimensional phenomena are unavoidable
- If the ratio p/n stays bounded strictly above zero, then it is not possible to achieve the optimal classification rate
- Our only hope is that the data is endowed with some form of *low-dimensional structure*

- What is the underlying cause of the inaccuracy of the prediction for the LDA in high-dimensions?
- The squared Euclidean error

$$\|\hat{\mu} - \mu\|^2 = \sum_{j=1}^{p} (\hat{\mu}_j - \mu_j)^2$$

concentrates sharply around p/n, i.e. for $t \in (0, 1)$

$$\mathrm{P}\left(\left| \|\hat{\mu} - \mu\|^2 - \frac{m}{n} \right| \geq \frac{p}{n} t \right) = \mathrm{P}\left(\left| \frac{1}{p} \sum_{j=1}^{p} Z_j^2 - 1 \right| \geq t \right) \leq 2e^{-\frac{pt^2}{8}}$$

where $Z_j = \sqrt{n}(\hat{\mu}_j - \mu_j) \sim N(0, 1)$; for the upper bound see Wainwright (2019), Example 2.11

Sparsity

- Suppose that the p-vector μ is *sparse*, with only s of its p entries being nonzero, for some sparsity parameter $s \ll p$
- If sparsity holds, we can obtain a better estimator by thresholding the sample means

$$\tilde{\mu}_j = \hat{\mu}_j \mathbb{1}\{|\hat{\mu}_j| > \lambda\}$$

  where

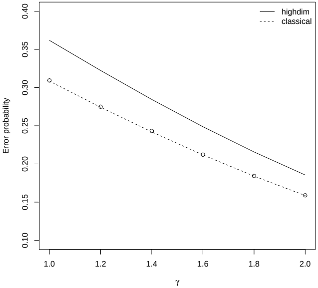
$$\lambda = \sqrt{\frac{2 \log p}{n}}$$

Thresholded mean

Suppose to replace $\hat{\mu}$ by the thresholded mean $\tilde{\mu}$, then

$$\tilde{\Psi}(x) = \langle \tilde{\mu}_A - \tilde{\mu}_B, x - \frac{\tilde{\mu}_A + \tilde{\mu}_B}{2} \rangle$$

approaches the optimal $\mathrm{Err}(\Psi)$ if $\log \binom{p}{s} / n \to 0$. For $s = 5$:



circles correspond to the empirical error probabilities, averaged over 10 trials