Classical versus high-dimensional theory

Kaveh S. Nobari

Reading Sessions in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Classical framework

- Classical theory provides statement for a fixed class of models, parameterised by an index n (interpreted as sample size) that is allowed to increase.
- An instance of the earlier statement is the simplest setup of the law of large numbers: given restrictions on dependence, heterogeneity, and moments of sequence of random variables $\{X_i\}_{i=1}^n$, $\bar{X}_n - \mu \xrightarrow{a.s.} 0$, where $\bar{X}_n \equiv n^{-1} \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X_n]$ [see White (2014)]. Consequently, the sample mean $\bar{X}_n$ is a consistent estimator of the unknown population mean.
- A more refined statement is provided by the central limit theorem that guarantees the rescaled deviation $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to a Gaussian distribution with zero mean and covariance matrix Σ.
- In classical framework, the dimension d of the data space is typically fixed, while $d \to \infty$.

High-dimensional framework

- In a high-dimensional framework, the dimension d is either of the same order as the sample size n - i.e. $d \asymp n$ or the degree of dimensionality is much larger than the sample size - i.e. $d \gg n$.
- In such framework, the classical large n, fixed d theory fails to provide useful predictions.
- Classical methods can break down dramatically in high-dimensional regimes.
- In this session, we follow chapter 1 of Wainwright (2019) to motivate the study of high-dimensional statistics.

Example: linear discriminant analysis

- Suppose we wish to determine whether an observation vector $x \in \mathbb{R}^d$ has been drawn from one of two possible distributions, say P_1 v.s. P_2.
- When the two distributions are known, then a natural decision rule is based on thresholding the log-likelihood ratio

$$\psi(x) = \log \frac{P_2[X]}{P_1[X]}$$

where we reject the null hypothesis that x comes from the distribution $P_1[x]$ when

$$\psi(x) > c$$

such that c is the smallest constant that satisfies $P[TS > c] \leq \alpha$, for $0 \leq \alpha \leq 1$, where α is an arbitrary significance level.

- The Neyman-Pearson Lemma guarantees that these family of decision rules are optimal (maximize power for a given testing problem).

Example: linear discriminant analysis

Now suppose P_1 and P_2 are distributed as multivariate Gaussian, say $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, with differing only in their mean vectors. Furthermore recall, that for a multivariate Gaussian random variable, the density $P_i[X]$ for $i = 1, 2$ is defined as follows

$$P_i[x] = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right)$$

therefore,

$$\begin{aligned}\psi(x) &= \log \frac{P_2[X]}{P_1[X]} \\ &= \left\{\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right\}\end{aligned}$$

Example: linear discriminant analysis

Noting that $B(x, y) := \frac{1}{2} x' \Sigma^{-1} y$ is a symmetric bilinear form, we have

$$
\begin{aligned}
\psi(x) &= \log \frac{P_2[X]}{P_1[X]} \\
&= B(x - \mu_1, x - \mu_1) - B(x - \mu_2, x - \mu_2) \\
&= B(x, x) - B(x, \mu_1) - B(\mu_1, x) + B(\mu_1, \mu_1) \\
&\quad - B(x, x) + B(x, \mu_2) + B(\mu_2, x) - B(\mu_2, \mu_2) \\
&= 2B(x, \mu_2) - 2B(x, \mu_1) + B(\mu_1, \mu_1) - B(\mu_2, \mu_2) \\
&= 2B(x, \mu_2 - \mu_1) + B(\mu_1 - \mu_2, \mu_1 + \mu_2) \\
&= 2B(x, \mu_2 - \mu_1) - B(\mu_2 - \mu_1, \mu_2 + \mu_1) \\
&= x' \Sigma^{-1} (\mu_2 - \mu_1) - \frac{1}{2} (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 + \mu_1)
\end{aligned}
$$

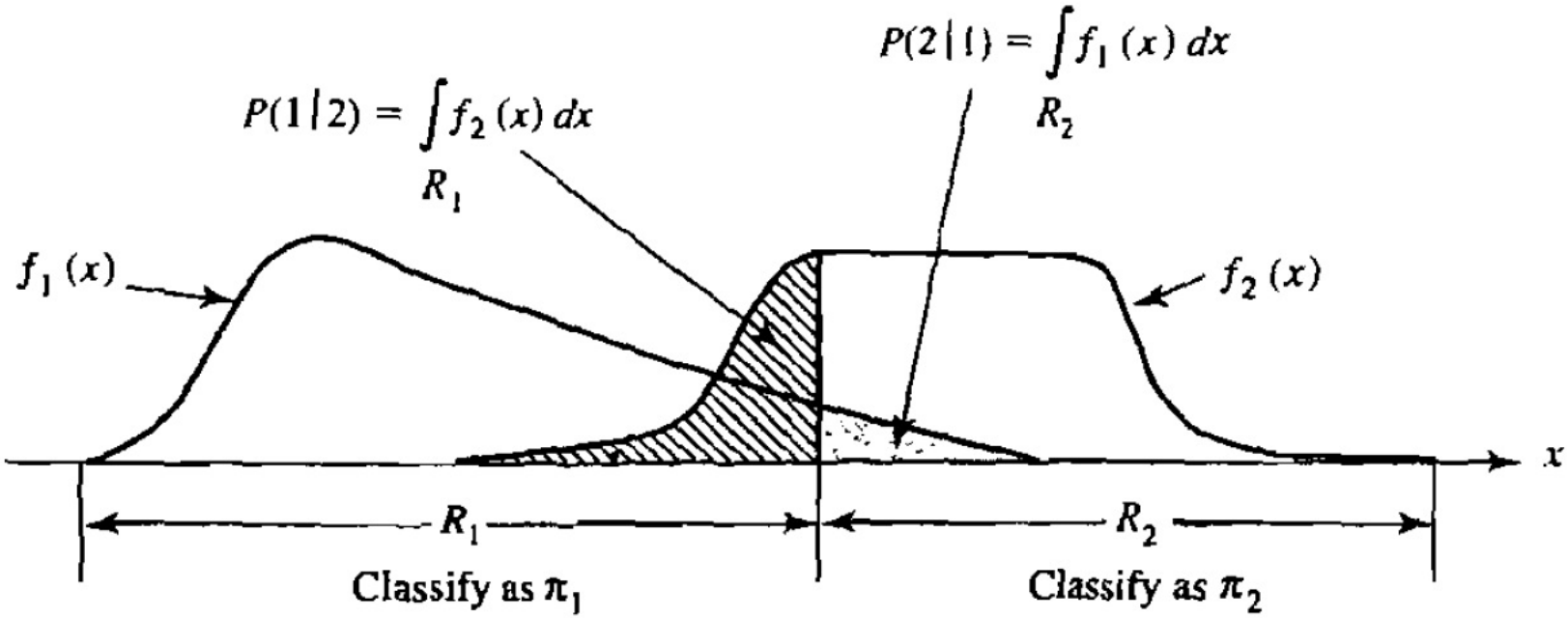
Example: linear discriminant analysis

The earlier results reduces to the linear test statistic

$$\psi(x) := \left\langle \mu_1 - \mu_2, \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_2}{2} \right) \right\rangle$$

where $< ., . >$ denotes the Euclidean inner product in $\mathbb{R}^d$.

- The optimal decision rule is based of thresholding $\psi(x)$, where the quality of the decision is obtained by computing the optimal probability of incorrect classification. For an example of incorrect classification, refer to the plot below that has been extracted from Johnson et al. (2002)


$P(2|1) = \int_{R_2} f_1(x)\,dx$
$P(1|2) = \int_{R_1} f_2(x)\,dx$
$f_1(x)$
$f_2(x)$
x
R_1
R_2
Classify as π_1
Classify as π_2

- If two classes are equally likely, the probability is given by

$$Err(\psi) = \frac{1}{2}\left\{P_1[\psi(X') \leq 0] + P_2[\psi(X'') > 0]\right\}$$

where X' and X'' are random vectors drawn from the distributions P_1 and P_2 respectively. Given the Gaussian assumption, the error probability can be expressed as

$$\begin{aligned} Err[\psi] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\gamma/2} \exp\left(\frac{-t^2}{2}\right) dt \\ &= \Phi\left(\frac{-\gamma}{2}\right) \end{aligned}$$

where Φ is the Gaussian CDF and where

$$\gamma = \sqrt{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}$$

- Assuming a Gaussian distribution and two labelled samples $\{x_i\}_{i=1}^{n_1}$ and $\{x_i\}_{i=n_1+1}^{n_2}$ which are drawn independently from P_1 and P_2, we may estimate the parameters of the log-likelihood ratio using their empirical counterparts, namely the sample mean $\hat{\mu}$ and the sample covariance matrix $\hat{\Sigma}$.
- Substituting these estimates in the log-likelihood ratio yields

$$\hat{\psi}(x) := \left\langle \hat{\mu}_1 - \hat{\mu}_2, \hat{\Sigma}^{-1} \left(x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right) \right\rangle$$

- Let us assume that the two classes are equally likely, the error probability by using a zero threshold is given by

$$Err[\hat{\psi}] = \frac{1}{2} \left\{ P_1[\hat{\psi}(X') \leq 0] + P_2[\hat{\psi}(X'') > 0] \right\}$$

- Let us now assume that the covariance matrix Σ is known a priori and is assumed to be the identity matrix, leading to the test statistic

$$\hat{\psi}(x) := \left\langle \hat{\mu}_1 - \hat{\mu}_2, x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right\rangle$$

- Kolmogorov showed that in high-dimensional asymptotic framework, in which $(n_1, n_2, d) \to \infty$ and with $d/n_i \to \alpha > 0$ for $i = 1, 2$ and the Euclidean distance $\|\mu_1 - \mu_2\|_2^2 \to \gamma > 0$, the error probability

$$Err[\hat{\psi}] \xrightarrow{p} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\right)$$

- **Question:** What happens when $d/n_i \to 0$? What about when $d/n_i \to \alpha > 0$?

Classical vs high-dimensional error rate

In the classical setting, the asymptotic error probability was given by

$$Err[\hat{\psi}] = \Phi\left(-\frac{\gamma}{2}\right)$$

with its high-dimensional analogue being

$$Err[\hat{\psi}] \xrightarrow{p} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\right)$$

Looking at the following plots, is the error probability better described by the classical or high-dimensional theory?

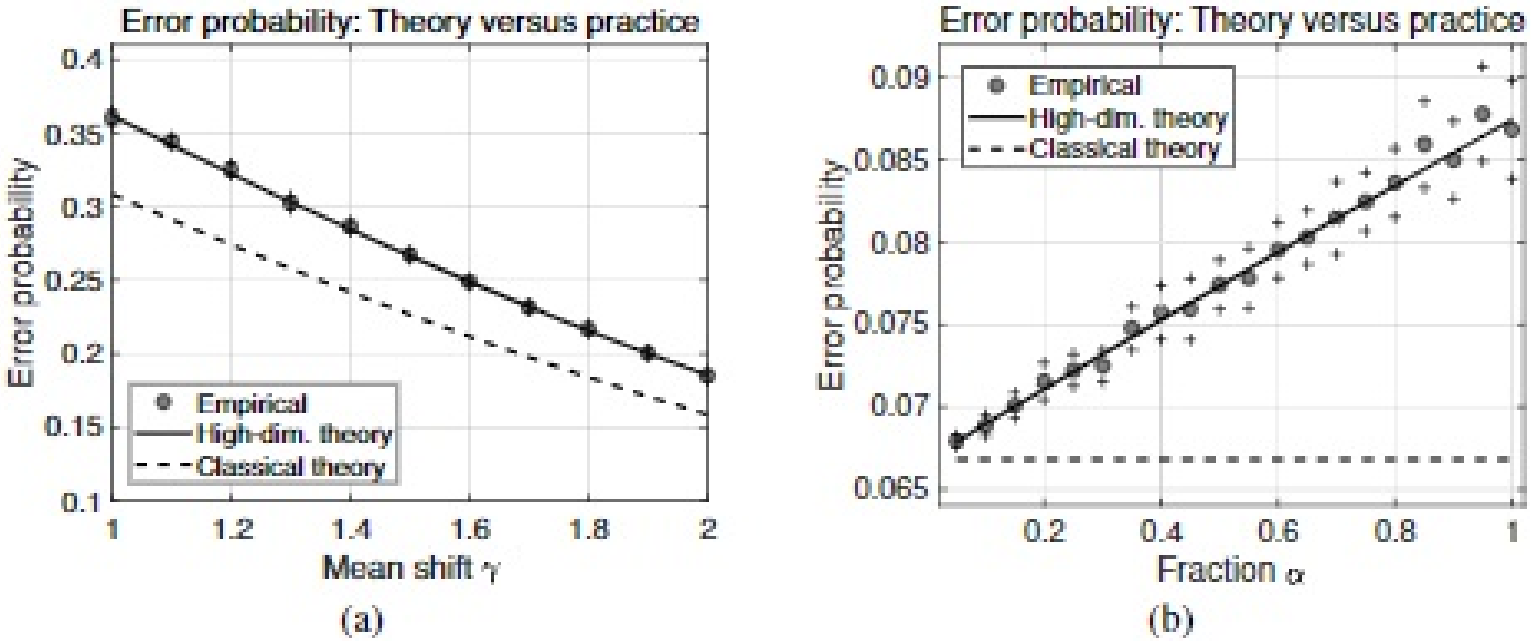


Figure 1.1 (a) Plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the mean shift parameter $\gamma \in [1, 2]$ for $d = 400$ and fraction $\alpha = 0.5$, so that $n_1 = n_2 = 800$. Gray circles correspond to the empirical error probabilities, averaged over 50 trials and confidence bands shown with plus signs, as defined by three times the standard error. The solid curve gives the high-dimensional prediction (1.6), whereas the dashed curve gives the classical prediction (1.2). (b) Plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the fraction $\alpha \in [0, 1]$ for $d = 400$ and $\gamma = 2$. In this case, the classical prediction $\Phi(-\gamma/2)$ plotted as a dashed line remains flat, since it is independent of α.

- High-dimensional phenomena outlined in many settings are unavoidable.
- In the classification problem earlier, if $d/n_i > 0$ it is impossible obtain optimal classification rate.
- Hope is that the data harbours some form of a low-dimensional structure.

- Classical asymptotics: $n \to \infty$ while other problem parameters remain fixed.
- High-dimensional asymptotics: $(n, d) \to \infty$, while enforcing that for some scaling function Ψ, the sequence $\Psi(n, d)$ remains fixed or $\Psi(n, d) \to \alpha \in [0, \infty)$. The scaling function Ψ may be as simple $\Psi(n, d) = d/n$, as shown earlier, or it may use other parameters, such as the sparsity parameter s [provide example].
- Non-asymptotic bounds: (n, d) as well as other parameters are viewed as fixed and high probability statements are made as a function of them.

Conclusion: The knowledge of tail bounds and concentration inequalities is crucial for assessing the performance of statistical estimators.

References

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

White, H. (2014). *Asymptotic theory for econometricians*. Academic press.