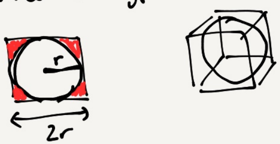
High Dimensional

The quality of BDR depends on the CCD estimates.
How does it work when x is high-dimensional?

"High dimensional spaces are <u>weird</u>!"
(do not trust your intuition.)

<u>Examples</u>:

(1) consider a hypercube & an inscribed hypersphere in $\mathbb{R}^d$.



volume of hypersphere: $V_d(r) = \dfrac{\pi^{d/2} r^d}{\underbrace{\Gamma(\frac{d}{2}+1)}}$

↖ Gamma function

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$$

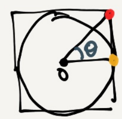
$$\Gamma(n+1) = n!$$

volume of hypercube: $(2r)^d$

let $f_d = \dfrac{\text{volume sphere}}{\text{volume cube}} = \dfrac{\pi^{d/2}}{2^d \Gamma(\frac{d}{2}+1)}$

d	1	2	3	$\longrightarrow \infty$
f_d	1	0.785	0.524	$\longrightarrow 0$

$\Rightarrow$ as d increases, the volume of corners (of cube) increases.



$c = [r, r, r, \ldots, r]$

$p = [r, 0, 0, \ldots, 0]$

$\|c\|^2 = d r^2$

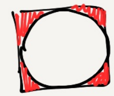
$\|p\|^2 = r^2$

$$\cos\theta = \frac{c^T p}{\|c\|\|p\|} = \frac{r^2}{\sqrt{d}\, r^2} = \frac{1}{\sqrt{d}}$$

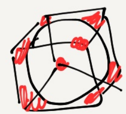
$\Rightarrow$ as d increases then $c \perp p$!
(corner is orthogonal to the axis)

$d=1$

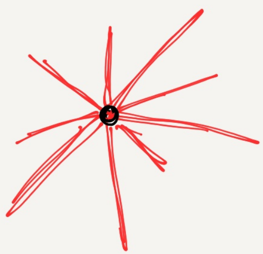
$d=2$



$d=3$



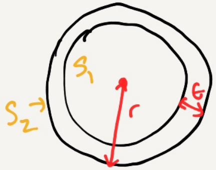
$d=$ large



CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Example 2

consider a hyperspere shell of thickness ϵ.



$$V_{shell} = V(S_2) - V(S_1) = \left(1 - \underbrace{\frac{V(S_1)}{V(S_2)}}\right) V(S_2)$$

$$\frac{V(S_1)}{V(S_2)} = \frac{\frac{(r-\epsilon)^d \cancel{\pi^{d/2}}}{\Gamma(\cancel{\frac{d}{2}+1})}}{\frac{r^d \cancel{\pi^{d/2}}}{\Gamma(\cancel{\frac{d}{2}+1})}} = \Big(\underbrace{1 - \frac{\epsilon}{r}}_{<1}\Big)^d$$

suppose $0 < \epsilon < r$,

as d increases, then $\frac{V(S_1)}{V(S_2)} \to 0$

$\Rightarrow V_{shell} \to V(S_2)$ as d increases

"All the volume is in the shell of the hypersphere"

Example 3 high-dim Gaussian

let $X \sim N(0, \sigma^2 I)$

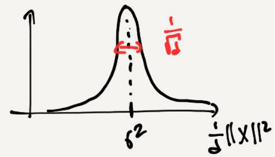
i.e. $x_i \sim N(0, \sigma^2)$ iid r.v.

Then, $E[\|X\|^2] = E[x_1^2 + x_2^2 + \dots + x_d^2] = d\sigma^2$

$$E\left[\tfrac{1}{d}\|X\|^2\right] = \sigma^2$$

Note: $\|X\|^2$ is a sum of iid r.v., thus by the central limit theorem it is concentrated around the mean as $d \to \infty$.

$$\tfrac{1}{d}\|X\|^2 \sim N\left(\sigma^2, \tfrac{1}{d}\right)$$

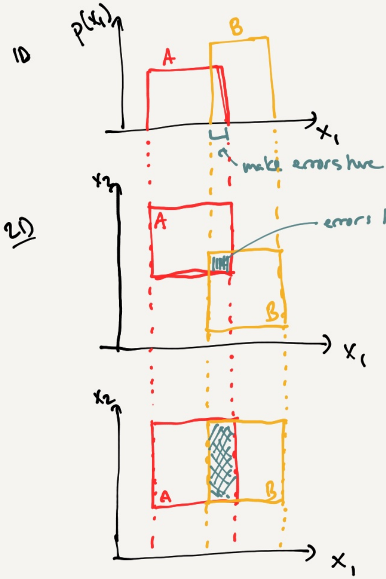


In high-dim, a Gaussian is essentially a shell of radius $\sigma\sqrt{d}$. Most of the density is in the shell.

(max density is still the mean)

Curse of Dimensionality

In theory, adding new features will not increase P(error)



Adding informative features
→ p(error) decreases.

Adding noninformative features
→ p(error) is the same as before.

In practice, for BDR, error increases as feat dim increases.

The problem: Quality of the CCD estimates.

Density estimates in high-dim require more training samples.

Roughly, desired training set size = $O(e^p)$, p = # of parameters

Solution:

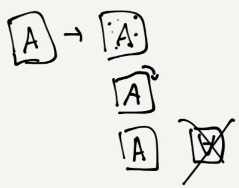
1) Reduce # of parameters (complexity of model)
   (e.g. full cov → diag cov)

→ 2) Reduce # of features (dimensionality reduction)
   → implicitly reduce # of parameters.

3) Create more data

   a) Bayesian estimation (virtual samples)

   b) data augmentation

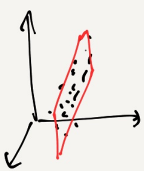


Linear Dimensionality Reduction

- Summarize correlated features w/ fewer features
- How to find these correlations?

linear projection of points onto line.

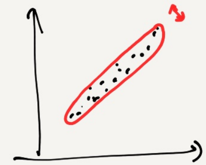
correlated data "lives" in a lower-dim subspace (w/ some noise)



$\mathbb{R}^3 \to \mathbb{R}^2$

Principal Component Analysis (PCA)

Idea: if the data lives in a subspace, then it will look flat in the full space.



If we fit a Gaussian, it will be "skinny" in some directions.

let (v_i, λ_i) be an eigenpair of covariance matrix Σ

$$\Sigma = V \Lambda V^T, \quad V = [v_1, \dots, v_d], \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

- each v_i defines an axis of ellipse
- each λ_i defines the width on that axis

Hence, the eigenvalues of Σ tell us which directions the data is flat.

$\Rightarrow$ select axis v_i w/ larger eigenvalues as "principal components".

PCA: Given the dataset $\{x_1, \dots, x_n\}$ & dim k

training

1) Calculate Gaussian

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$$

2) eigendecomp of Σ: $\Sigma = V \Lambda V^T$

3) order the eigenvalues: $\lambda_1 > \lambda_2 > \lambda_3 \dots > \lambda_d > 0$

4) select the top-k eigenvectors: $\Phi = [v_1, \dots, v_k]$

dim. reduction

5) project new point x onto Φ:

$$z = \Phi^T(x-\mu)$$

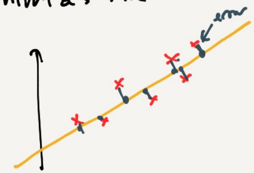
$\leftarrow$ PCA coefficients. (new feature vector) (use BDR or other classifiers)

Notes:

The selection of Φ w/ $\Phi^T\Phi = I$ also:

(1) maximizes the variance of the projected data into z. PS7-3

(2) minimizes the reconstruction error of training data. PS7-2



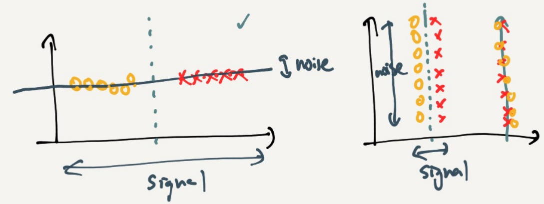
(3) can be implemented efficiently w/ SVD PS7-4 (tutorial)

(4) Select k? 1) pick k that works in the downstream task (classification)

2) pick k to preserve variance of data

$$p\% = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \quad (95\%)$$

(5)

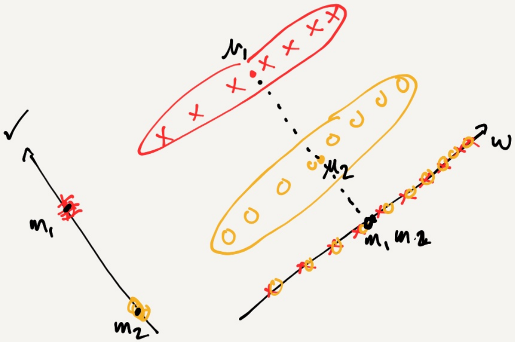


Assumption that signal variance is larger than the noise variance.

(6) PCA optimal for representation (but not necessarily for classification)

Fisher's Linear Discriminant (FLD)

(Linear Discriminant Analysis (LDA))



Find the projection that best separates the classes.

$z = w^T x$

Class statistics

	original space	1-d space
class mean	$\mu_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$	$m_j = w^T \mu_j$
class scatter	$S_j = \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j)^T$	$s_j = w^T S_j w$

IDEA: maximize the distance btwn projected means

$$(m_1 - m_2)^2 = (w^T(\mu_1 - \mu_2))^2$$

problem: w is unconstrained $\rightarrow$ need normalization

Fisher's Idea

$$w^* = \underset{w}{\operatorname{argmax}} \frac{\overbrace{(m_1 - m_2)^2}^{\text{btwn-class scatter}}}{\underbrace{s_1 + s_2}_{\text{within class scatter}}} = \underset{w}{\operatorname{argmax}} \frac{w^T S_B w}{w^T S_W w}$$

$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$

$S_W = S_1 + S_2$

generalized eigenvalue problem

$$\boxed{w^* = (S_1 + S_2)^{-1}(\mu_1 - \mu_2)}$$

"Fisher's Linear Discriminant"

Note: this hyperplane separate 2 Gaussian w/ cov $\frac{1}{n}(S_1 + S_2)$

$\Rightarrow$ FLD is optimal when 2 classes are Gaussians w/ equal covariance matrices.

$$(m_1 - m_2)^2 = [w^T(\mu_1 - \mu_2)]^2 = w^T \underbrace{(\mu_1 - \mu_2) \cdot (\mu_1 - \mu_2)^T}_{} w$$

$$= w^T S_B w$$