# High Dimensional

The quality of BDR depends on the CCD estimate

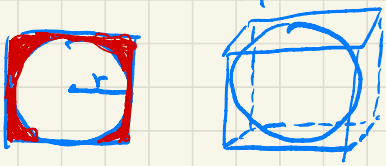
How does it work when $X$ is high-dimensional?

"High demensional spaces are weird"

(do not trust your intuition!)

## Examples

(1) Consider a hypercube and an inside hypersphere in $\mathbb{R}^d$




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

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

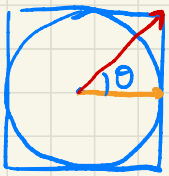
$\Gamma(n+1) = n!$

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

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

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

As $d$ increases, the volume of cube's corners increases




$C = (1, 1, \dots, 1)$

$p = (1, 0, \dots, 0)$

$\|C\|^2 = d \cdot 1^2 \qquad \|p\|^2 = 1^2$

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

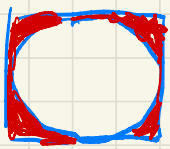
As $d$ increases, then $C \perp p$

corners are orthogonal to the axis

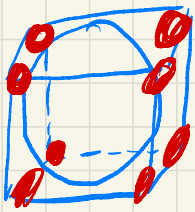
$d = 1$



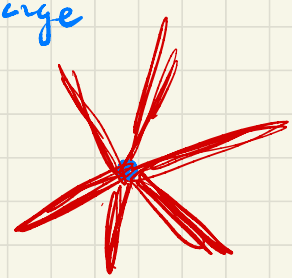
$d = 2$



$d = 3$

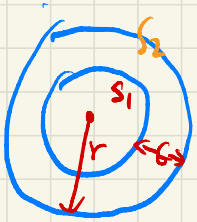


$d =$ large



Example 2

Consider a hypersphere shell of thickness $\varepsilon$




$$V_{shell} = V(S_2) - V(S_1)$$

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

$$\frac{V(S_1)}{V(S_2)} = \frac{(r-\varepsilon)^d \pi^{\frac{d}{2}} / \Gamma\left(\frac{d}{2}+1\right)}{r^d \pi^{\frac{d}{2}} / \Gamma\left(\frac{d}{2}+1\right)} = \left(1 - \frac{\varepsilon}{r}\right)^d$$

suppose $0 < \varepsilon < r$,

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

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

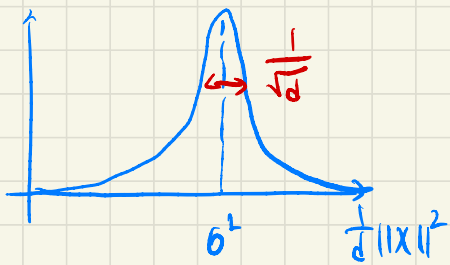
$= Ex_1^2 + Ex_2^2 + \dots + Ex_d^2$

$= E^2 x_1 + \text{var}(x_1) + E^2 x_2 + \text{var}(x_2) + \dots + E^2 x_d + \text{var}(x_d)$

$= d\sigma^2$

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

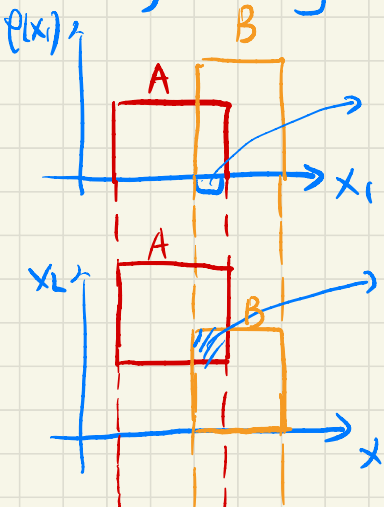
Note: $\|X\|^2$ is a sum of i.i.d. r.v., thus by the central limit theorem, it is concentrated around the mean as $d \to \infty$, $\frac{1}{d}\|X\|^2 \sim N\left(\sigma^2, \frac{1}{d}\right)$




In high-dimension, 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

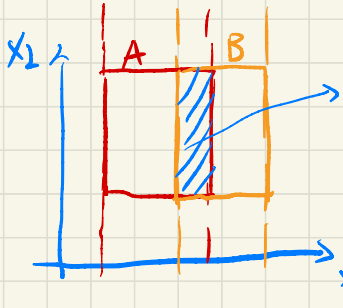
In theory, adding new features will not increase P(error)




make errors here

errors here:

Adding informative features → P(error) decreases




errors here:

Adding non informative features → P(error) is the same as before.

In practice, for BDR, error increases as feature dimension increases

Then problem: Quality of the (CI) estimates.

Density estimates in high-dim require more training samples.

Roughly, desired training size = O(c^p), p = # of parameters.

Solution:

1) Reduce # of parameters (complexity of model)
   (e.g. full covariance → diagonal covariance)

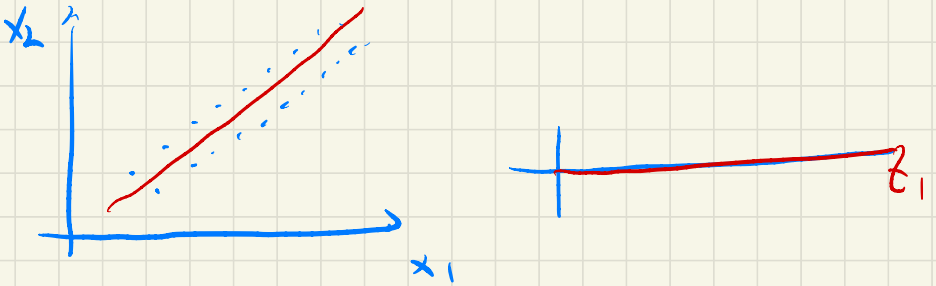
2) Reduce # of features (dimensionality reduction)
   → implicitly reduce # of parameters.

3) Create more data
   a) Bayesian estimate (virtual samples)

b) data augmentation

## Linear Dimensionality Reduction

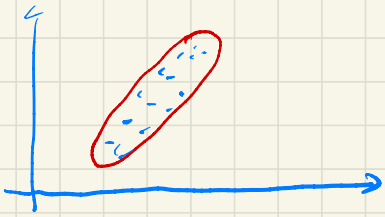
- Summarize correlated features with fewer features
- How to find these correlations?




correlated data "lives" in a lower-dim subspace with some noise

## 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.

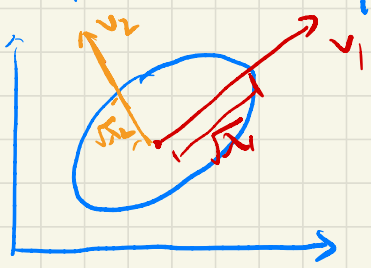
Let $(v_i, \lambda_i)$ be an eigenpair of covariance matrix $\Sigma = V \Lambda V^T$ $V = [v_1, v_2, \ldots v_d]$

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ 0 & \ddots & \\ & & \lambda_d \end{bmatrix}$$

- each $v_i$ defines an axis of ellipse
- each $\lambda_i$ defines the width on that axis

Hence, the eigenvalues of $\Sigma$ tell us which directions the data is flat.

⇒ Select axis with larger eigenvalues as "principle components."



PCA: Given the dataset $\{x_1, \cdots, x_n\}$ and dim $k$

Training

1) Calculate Gaussian:

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

2) eigen decomposition of $\Sigma$: $\Sigma = V\Lambda V^T$

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

4) select the top $k$ eigenvalues: $\Phi = [v_1, \cdots v_k]$

dim reduction

5) project new point $x$ onto $\Phi$

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

PCA coefficients

new feature vector, use BDR or other classifiers.

Notes:

The selection of $\Phi$ with $\Phi^T\Phi = I$ also:

(1) maximize the variance of the projected data

(2) minimize the reconstruction error of training data

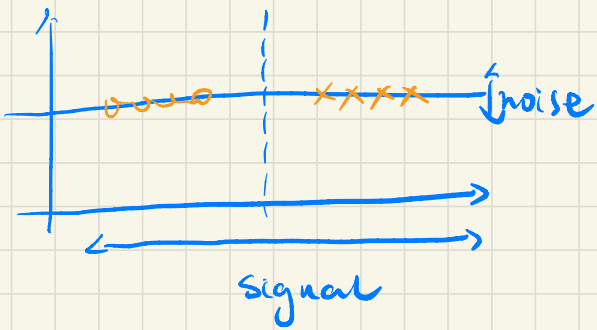
(3) can be implemented effectively with SVD.

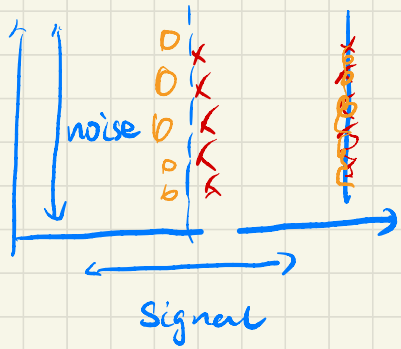
(4) Select k?

- pick k that works in the downstream task (classification)
- pick k to preserve variance of data.

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

(5)




Assumption that signal variance is larger than the noise variance.

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

# Linear Discriminant Analysis (LDA)

Find the projection that best seperates the classes

$$z = w^T x$$

Class statistics:

| | original space | 1-d space |
|---|---|---|
| class mean | $\mu_j = \frac{1}{n} \sum_{x_i \in c_j} x_i$ | $m_j = w^T \mu_j$ |
| 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 between projected means.

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

Problem: $w$ is unconstrained $\rightarrow$ need normalization.

Fisher's Idea:

$$w^* = \arg\max_w \frac{(m_1 - m_2)^2}{s_1 + s_2} = \arg\max_w \frac{w^T S_B w}{w^T S_w w}$$

($(m_1 - m_2)^2$: btwn-class scatter; $s_1 + s_2$: with-class scatter)

$$\Rightarrow w^* = S_w^{-1}(\mu_1 - \mu_2)$$

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

$S_w = S_1 + S_2$

Note this hyperplane seperate 2 Gaussian with cov $\frac{1}{n}(S_1 + S_2)$

$\Rightarrow$ FLD is optimal when 2 classes are Gaussian with equal covariance matrices.