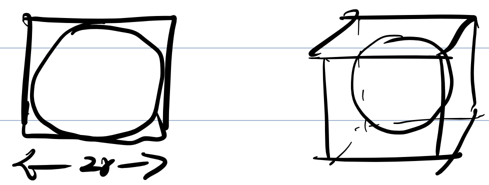
The quality of BDR depends on the CCD estimates
How does it work when x is high-dimensional?

"High dimensional spaces are wired!"
(do not trust your intuition)

Examples:

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



Volumn of hypersphere $V_d(r) = \dfrac{\pi^{\frac{d}{2}} \cdot r^d}{\underbrace{\Gamma(\frac{d}{2}+1)}_{\text{Gamma function}}}$

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

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

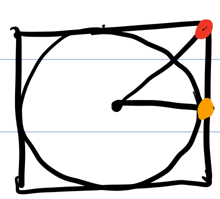
Volumn of hypercuble $(2r)^d$

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

d	1	2	3	∞
f_d	1	0.785	0.524	0

factorial overweights

As d increases, the volumn of the corner increases.



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

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

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

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

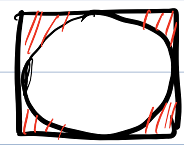
$$\cos\theta = \frac{c^T p}{\|c\|\|p\|} = \frac{r^2}{\sqrt{d}\, r^2} = \frac{1}{\sqrt{d}}$$

★ as $d \nearrow$ $\Rightarrow \cos\theta = 0 \Rightarrow c \perp p$!

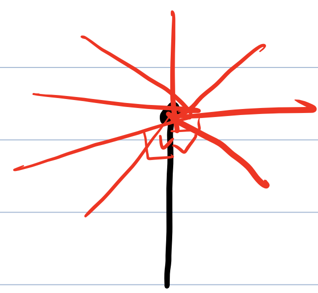
(corner is orthogonal to axis)

$d=1$

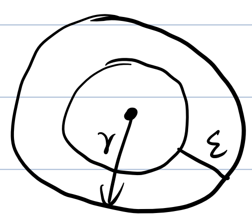
$d=2$



$d \Rightarrow \infty$



Example 2 a hypersphere shell of thickness ε



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

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

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

Suppose $0 < \varepsilon < r$

as $d \nearrow$, $\dfrac{V(S_1)}{V(S_2)} \to 0$

$$V_{shell} \to V(S_2)$$

All the volumn is in the shell of the hyperspheres

Example 3 high-dim Gaussian

let $x \sim \mathcal{N}(0, \alpha^2 I)$

i.e. $x_i \sim \mathcal{N}(0, \alpha^2)$ i.i.d. r.v.

Then $E[\|x\|^2] = E[\underbrace{x_1^2}_{0+\alpha^2} + x_2^2 + \cdots + x_d^2] = d\alpha^2$

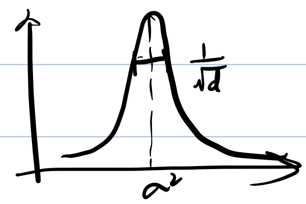
$E[\frac{1}{d}\|x\|^2] = \alpha^2$

Note $\|x\|^2$ is a sum of iid. r.v.

By central limit theorem. it is concentrated

around mean as $d \to \infty$

$\frac{1}{d}\|x\|^2 \sim \mathcal{N}(\alpha^2, \frac{1}{d})$



radius α^2

In high-dim, a Gaussian is essentially a shell of

radius $\alpha\sqrt{d}$ Most of the density is in the shell

$\Downarrow$

why $\|\frac{1}{\sqrt{d}} x\|^2$

However, the mean is still 0

though density is in the shell

Our intuition in 2d

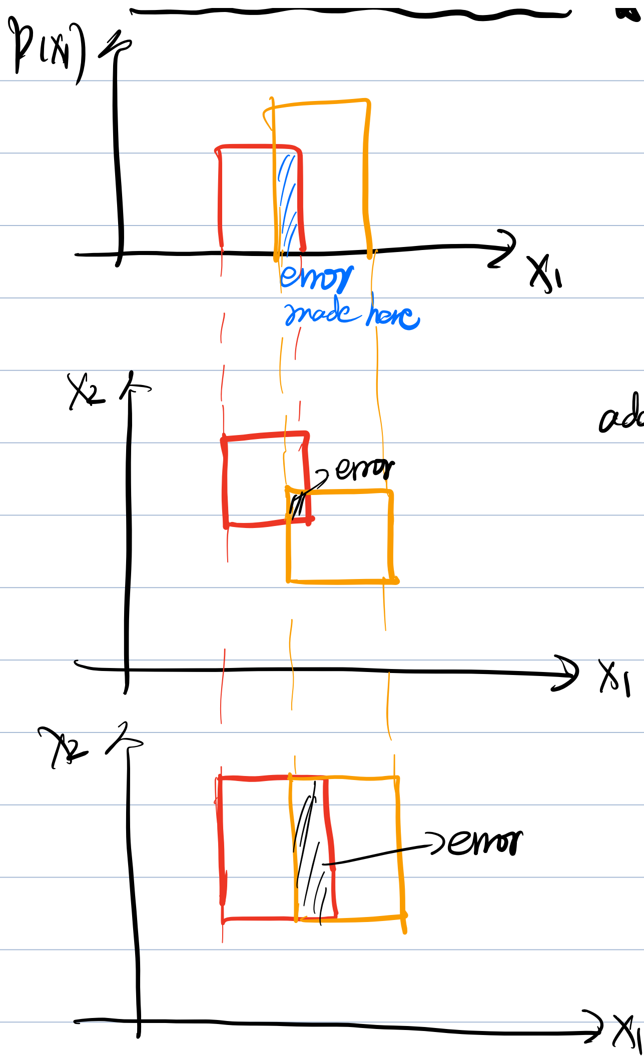
$\Rightarrow$ density should

be around the mean

Curse of dimensionality

In theory, add a feature will not

increase the error. ☆



adding informative features
→ p(error) decreases

Adding non-informative
features p(error) is
the same as
before

In practice, for BDR, error increases as the
feature dim increases

The problem is the quality of the CCD estimate.
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) → things we optimize

   eg. full cov ⇒ diag cov

2) Reduce # of features (dimensionality reduction)

   → implicitly reduce # of parameters

3) Create more data

   (a) Bayesian estimation (virtual samples)

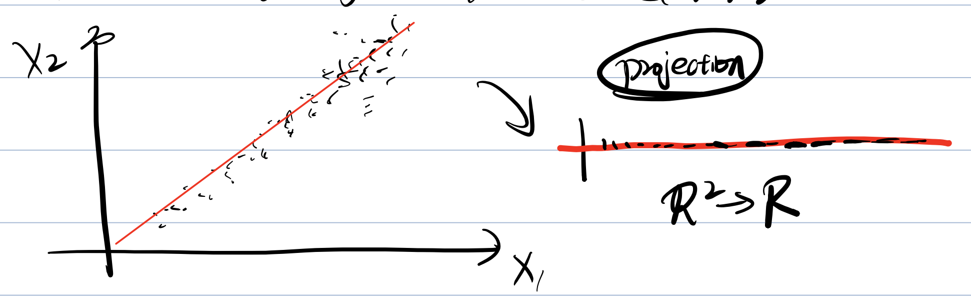
   (b) data augmentation.

   (eg. $\boxed{A} \Rightarrow \boxed{A}$ add noise

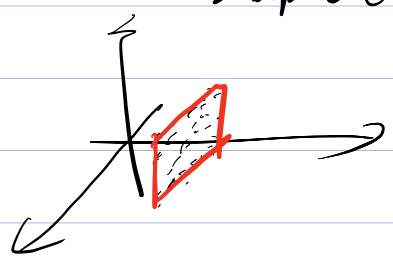
   $\boxed{A} \Rightarrow \boxed{A}$)

Linear Dimensionality Reduction

- Summarize correlated features w/ fewer features.
- How do we find these correlations?

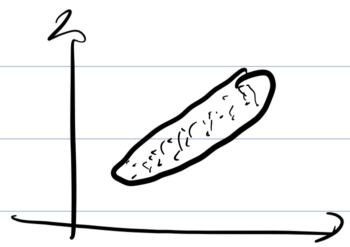


Correlated data "lives" in a lower-dim subspace (w/ some noise)



PCA (Principle Component Analysis)

Idea: If the data can fit into a subspace,
then it should be flat in full space.



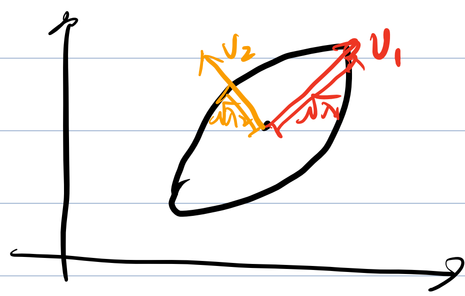
If we fit a Gaussian,
it will be "skinny" in
some directions.

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

$$\Sigma = U \Lambda U^T, \quad U = [v_1, v_2, \dots, v_d]$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \end{bmatrix}$$

- 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/ large eigenvalues
as "principle components".

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

1) Calculate Gaussian.

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

training

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

3) order the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$

4) select top k eigenvectors. $\Phi = [v_1, \ldots, v_k]$

dim reduction.

5) project new point x onto Φ

$$z = \Phi^T (x - \mu) \leftarrow \text{PCA coefficients}$$

mean (new feature vectors)

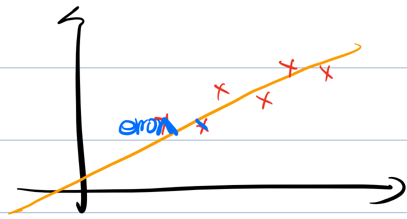
central the data

Notes:

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

ps (1) maximize the variance of the projected data into z

(2) minimize the construction error of training data



(3) Can be implemented efficiently w/ SVD P87-4

(4) How to select k?