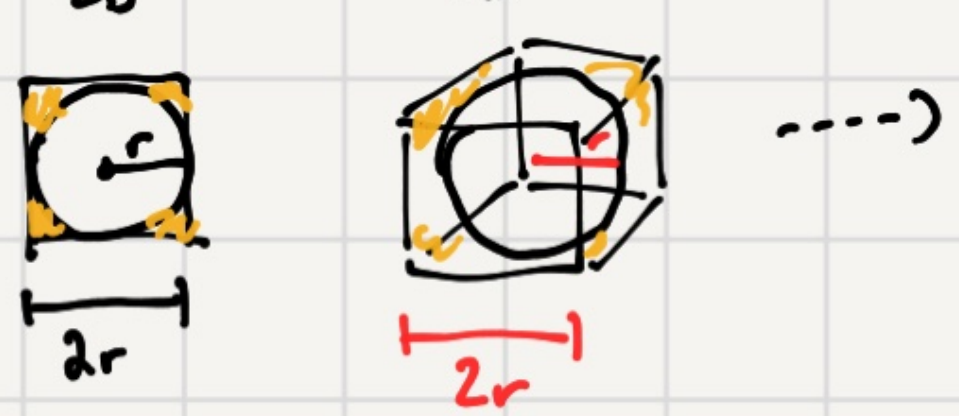
Lecture 7: Dimensionality

The quality of calculating BDR depends on the CCD estimates.
How does it work in high-dimensional space for x?

"high-dimensional spaces are weird."
Do not trust your intuition!

Examples

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



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

volume of hypersphere: $V_d(r) = \dfrac{\pi^{d/2} r^d}{\Gamma(\frac{d}{2}+1)}$, $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$

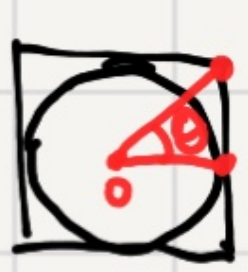
Gamma function
$\Gamma(n+1) = n!$
n is counting number

"factorial for real numbers"

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

$\lim_{d\to\infty} f_d = 0$ $\therefore$ as dim increases the volume of sphere decreases relative to the cube.

$\Rightarrow$ the volume of corners of cube increases.



$c = \begin{bmatrix} r \\ \vdots \\ r \end{bmatrix}$ corner

$p = \begin{bmatrix} r \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ axis

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

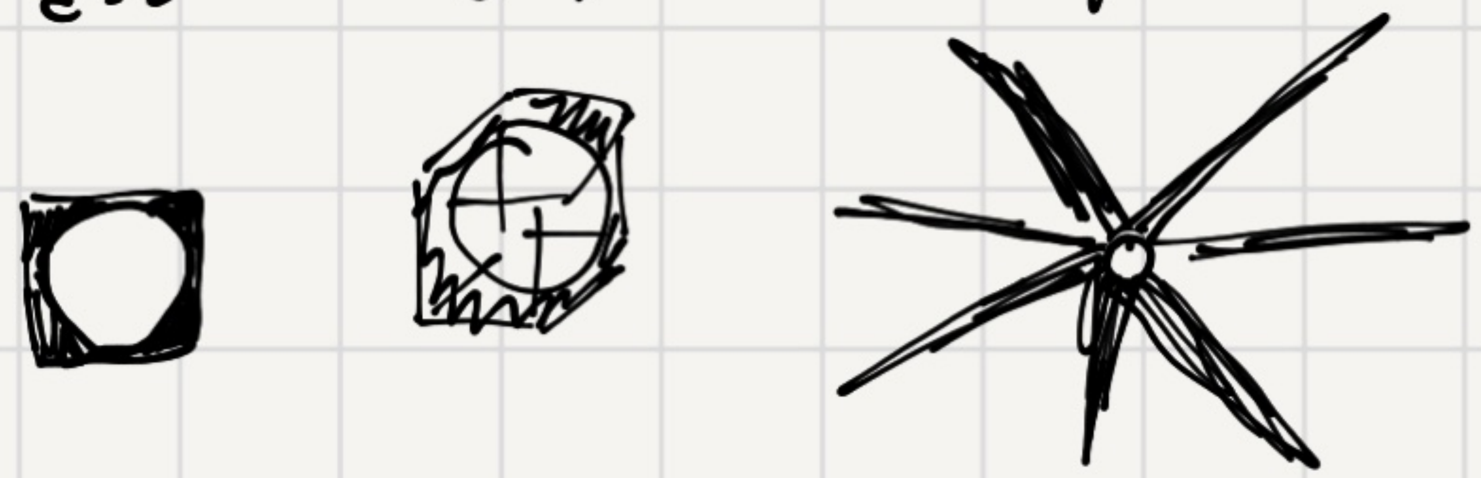
$\|c\|^2 = dr^2 \Rightarrow \|c\| = r\sqrt{d}$

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

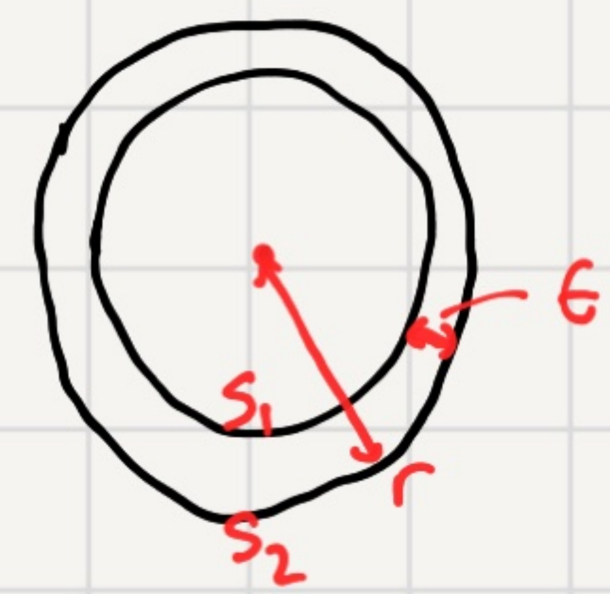
as $d \to \infty$, $\cos\theta \to 0 \Rightarrow c \perp p$

"The corner is orthogonal to the axis"

$d=2$ $d=3$ large d



Example 2: consider a hypersphere shell of thickness ϵ



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

$$\frac{V(S_1)}{V(S_2)} = \underbrace{\left(1 - \frac{\epsilon}{r}\right)^d}$$

for $0 < \epsilon < r$

as d increases $\dfrac{V(S_1)}{V(S_2)} \to 0$ (because $1 - \frac{\epsilon}{r} < 1$)

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

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

Example 3: high-dim Gaussian

let $X \sim N(0, \sigma^2 I_d)$, $X_i \sim N(0, \sigma^2)$ $\forall i$

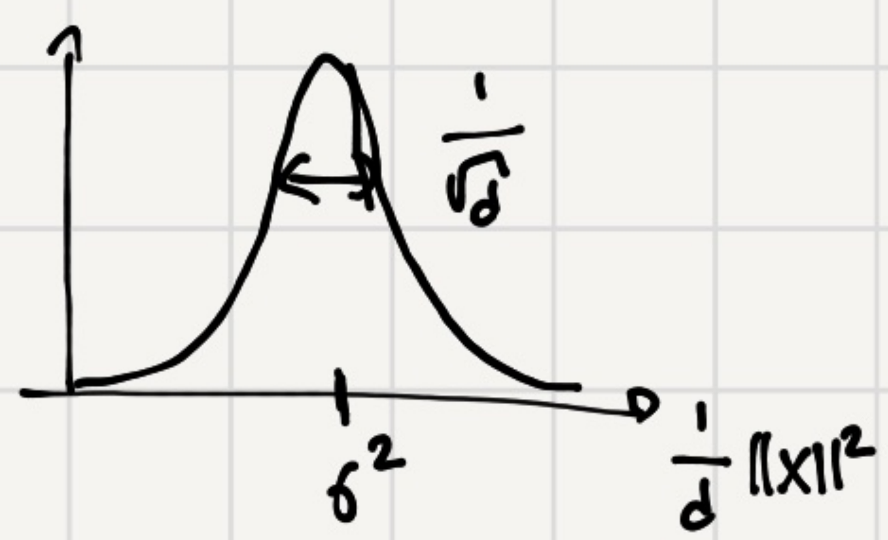
Then $E[\|X\|^2] = E[\underbrace{x_1^2} + \dots + \underbrace{x_d^2}] = d\sigma^2$

length-squared of X

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

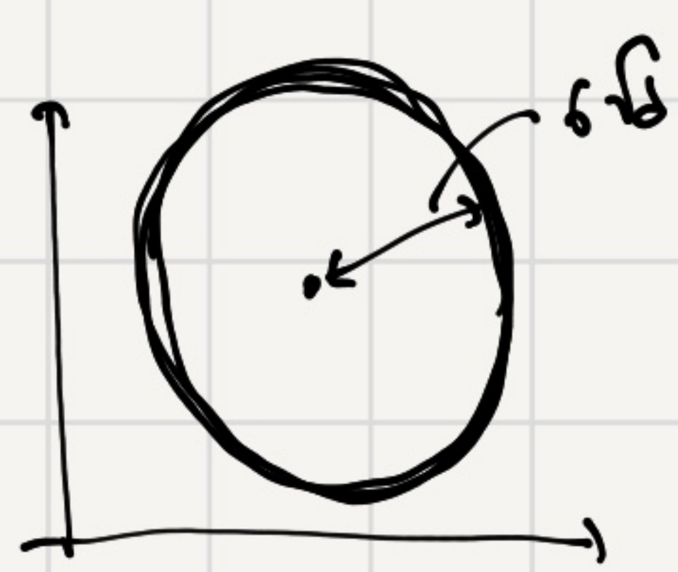
$\|X\|^2 = \sum_{i=1}^{d} \underbrace{x_i^2}_{r.v.}$ is a sum of iid r.v., thus by the central limit theorem it is concentrated around its mean.

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



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

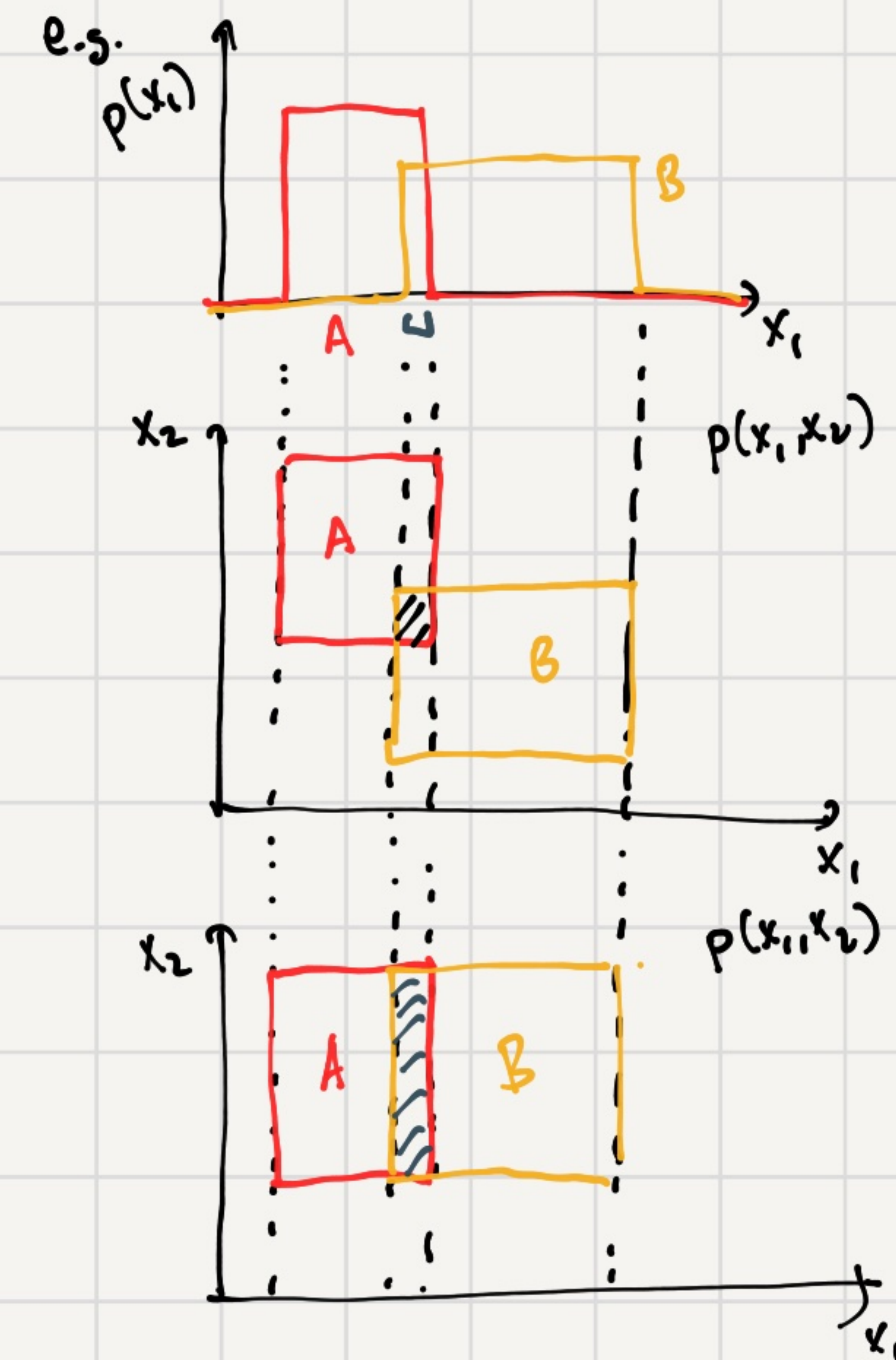
- the point of maximum density is still the mean.



Curse of dimensionality

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

e.g.



Add informative feature
→ overlap of A & B decreases
→ p(error) decrease

Add uninformative feature
→ overlap is same
→ p(error) is same

In practice, for BDR error can increase if we add more features!

Why? Quality of BDR depends on the quality of CCD estimates.
⇒ estimates in high-dim require more data.

e.g. histogram on unit cube $[0,1]^d$

10 bins / dimension:

to have one sample per bin on average:

$d=1 \Rightarrow 10$ samples
$d=2 \Rightarrow 100$ samples
$d=3 \Rightarrow 1000$ "

10^d samples

→ increases exponentially w/ the # of parameters.

Solution:

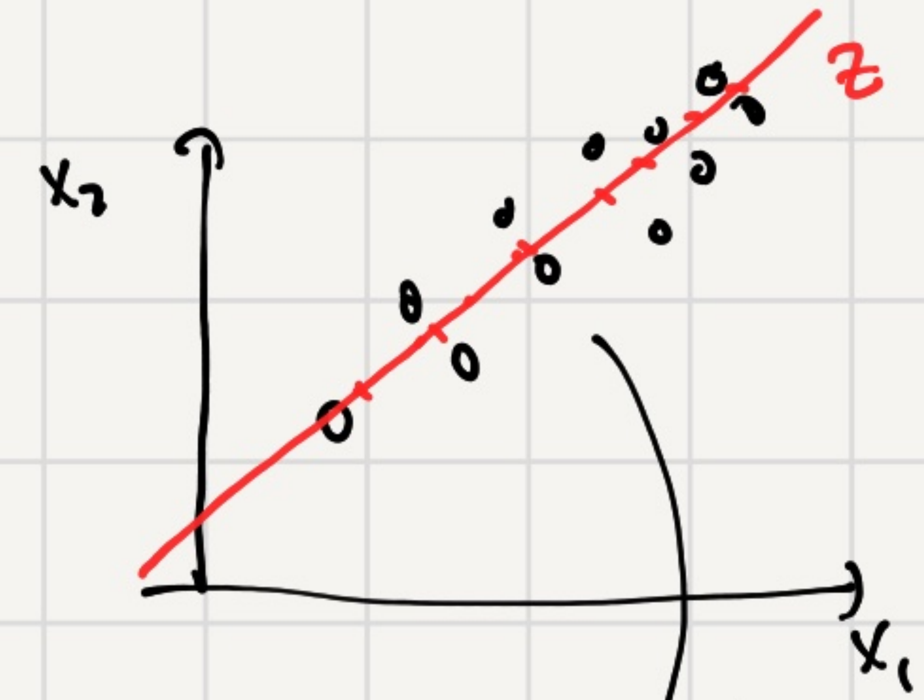
1) Reduce # of parameters (full cov → diag cov → isotropic cov)
   d^2 → d → 1

⇒ 2) Reduce # of features (dimensionality reduction)
   ⇒ implicitly reduces # of param.

3) Create more data
   a) Bayesian method (virtual samples)
   b) Data augmentation (perturb the inputs to make more data)

Linear dimensionality Reduction

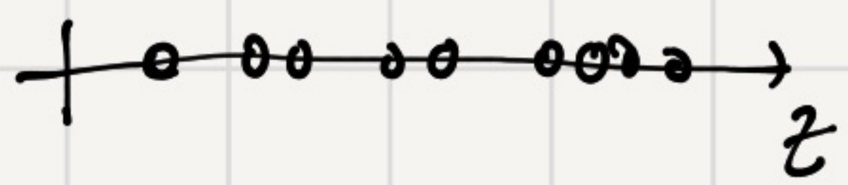
- summarize correlated features w/ fewer features.

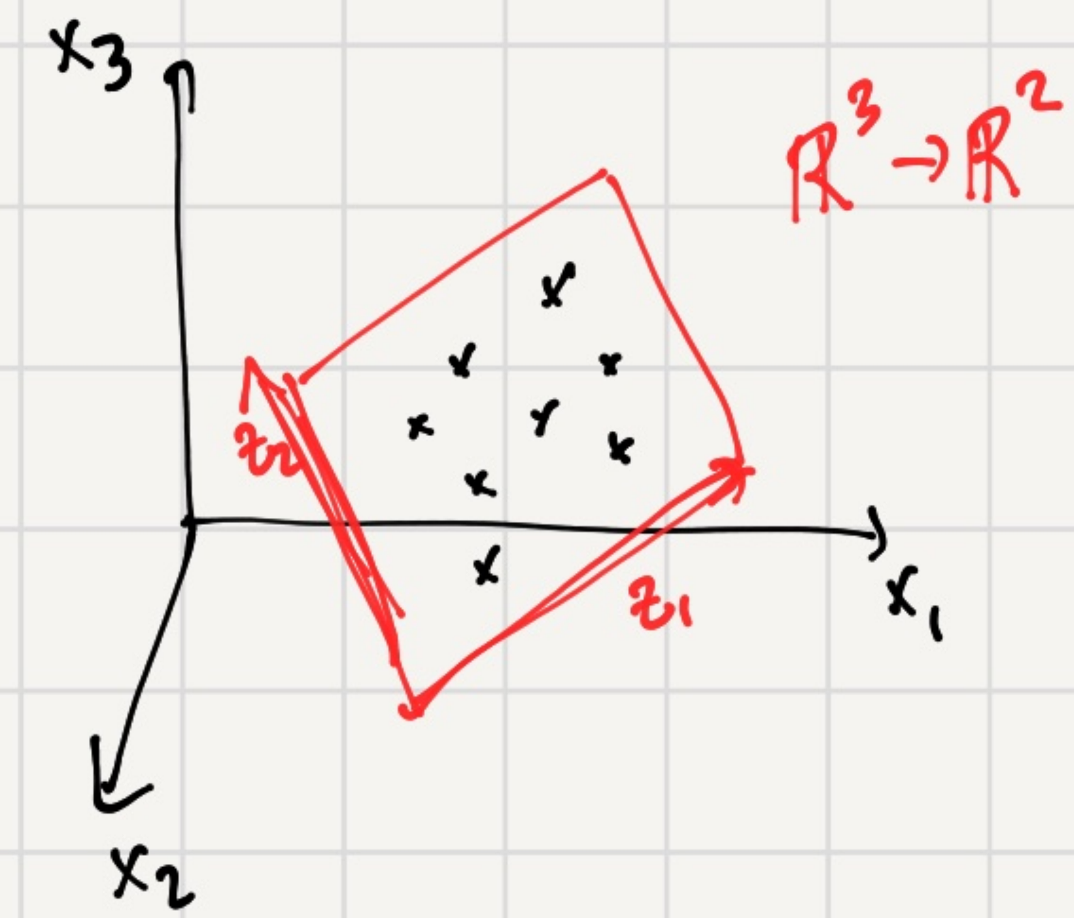


The correlated features "live" in a low-dimensional linear subspace of the original space

linear projection of samples onto a line.

$z = a^T x$

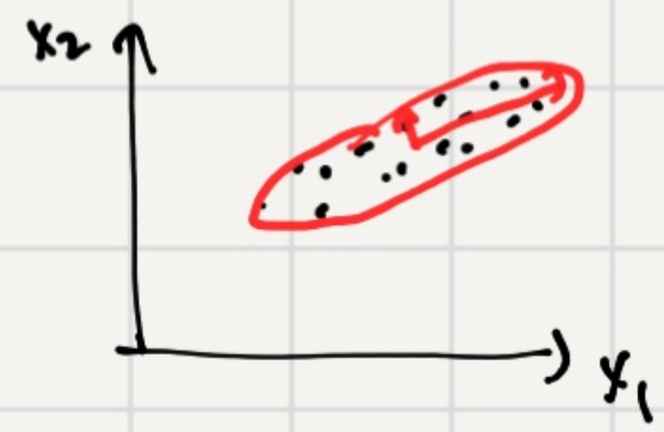




Principal Component Analysis (PCA)

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

$\Rightarrow$ if we fit a Gaussian, it will be skewed (skinny ellipses)



let (λ_i, v_i) be the eigenvalue/vector of cov Σ

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

- each v_i defines an axis of ellipse
- each λ_i defines its width.

$\Rightarrow$ the eigenvalues of Σ tell us which directions are flat

$\Rightarrow$ select axis v_i w/ largest eigenvalue as the <u>principal component</u>.

<u>PCA</u>: given $D = \{x_1, \dots, x_n\}$ & dim K

learning

1) Estimate Gaussian: $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$
2) eigendecomposition: $\Sigma = V \Lambda V^T$
3) order the eigenvalues: $\lambda_1 \geq \lambda_2 \cdots > 0$
4) Select the top-K eigenvectors: $\Phi = [v_1 \cdots v_K]$

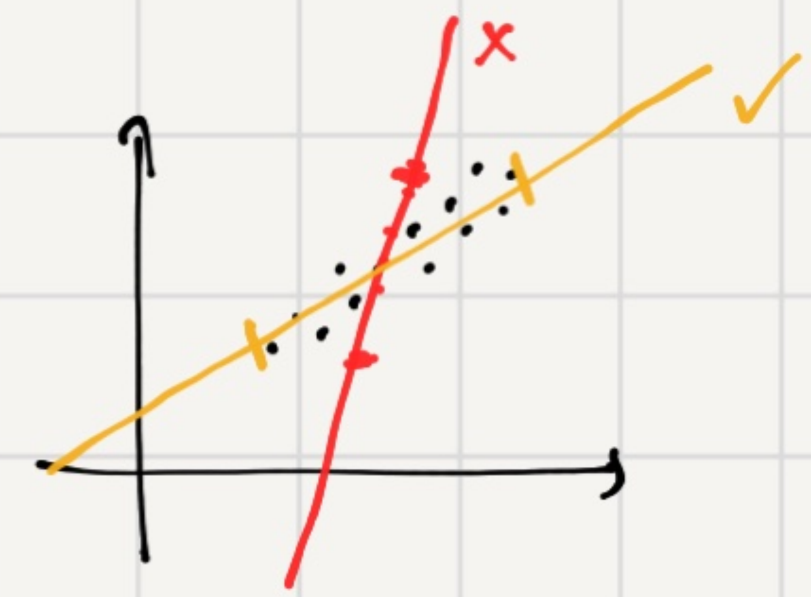
dim. reduction

5) project x onto Φ: $z = \Phi^T(x-\mu)$ ← "PCA coefficients"
6) use z as the new f.v. ⇝ BDR ....

<u>Note</u>: This selection of Φ w/ $\Phi^T\Phi = I$

1) maximizes the variance of the projected training data.

$z_i = \Phi(x_i - \mu)$ $\quad \|z_i\|^2$ $\quad$ (PS 7-3)



2) minimizes the reconstruction error of training data (PS 7-2)

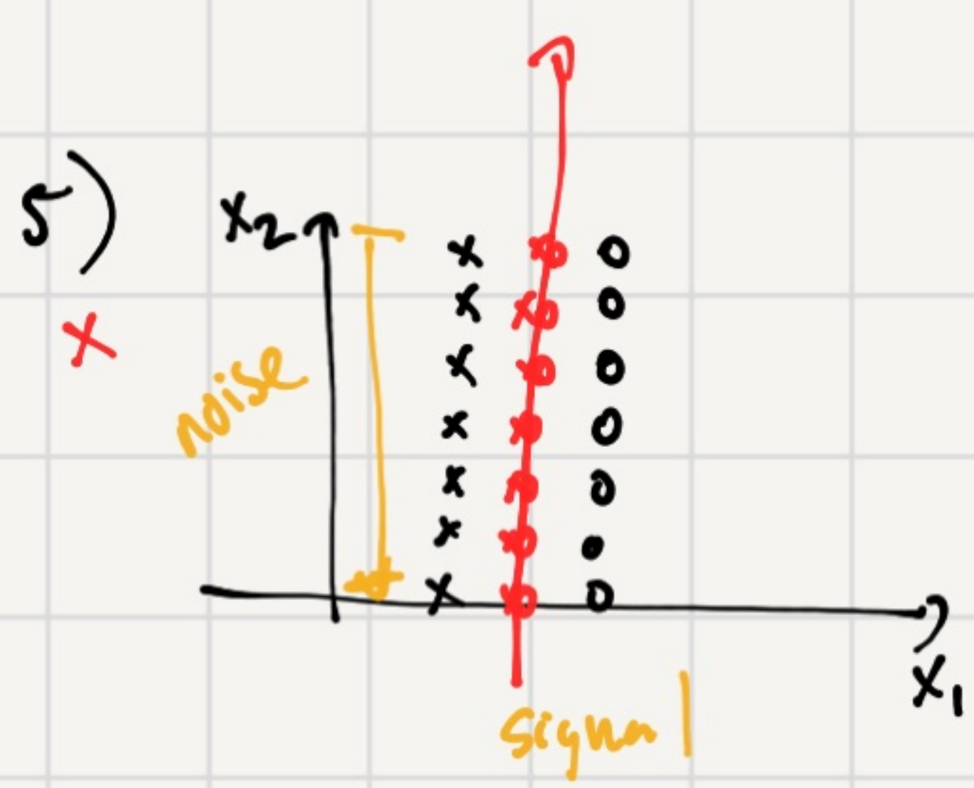
$$\sum_{i=1}^{n} \| \underbrace{x_i}_{\text{original}} - \underbrace{\Phi(z_i + \mu)}_{\text{reconstruction}} \|^2$$

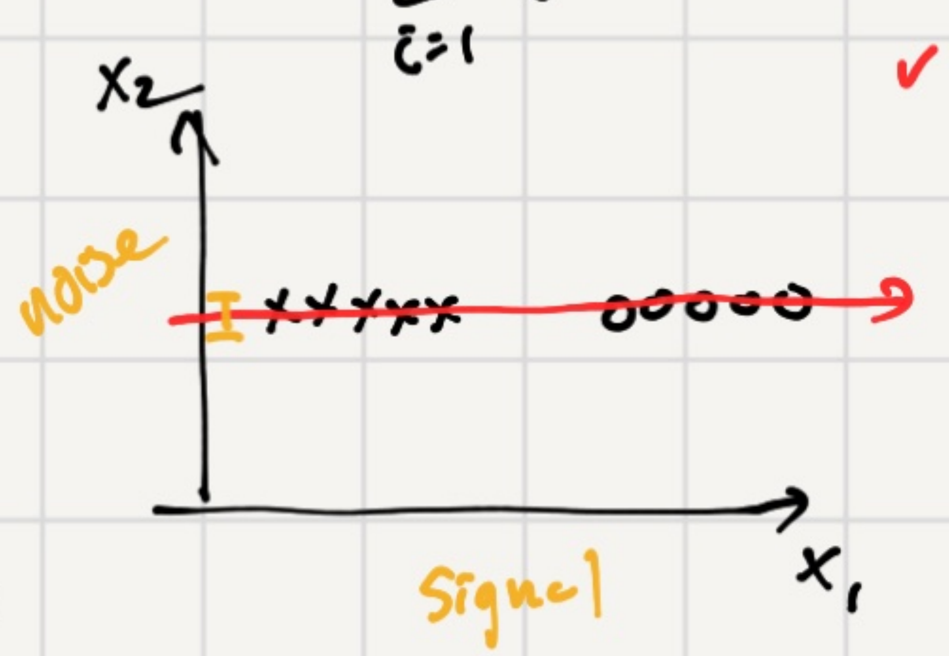
3) can implemented w/ SVD for high-dim data (PS 7-4)

4) How to select K?
   1) pick a K that works on downstream task
   2) pick K to preserve p% of variance

$$p = \frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$$

5)

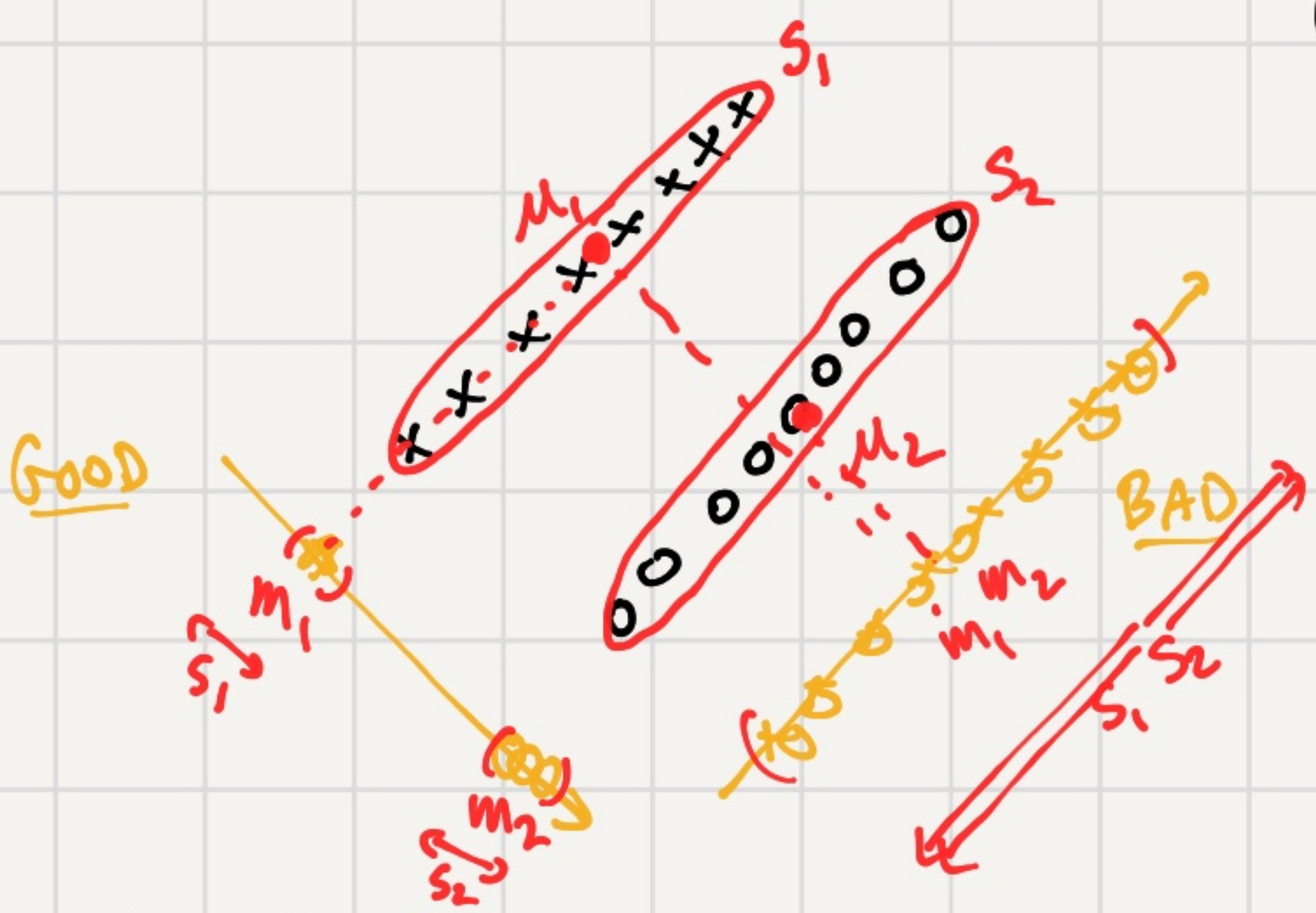




<u>Assumption</u>: the "noise" variance is smaller than the "signal" variance. → could cause problems for classification.

⚠ PCA is optimal for <u>representation of variance</u>, but suboptimal for classification.

Fisher's Linear Discriminant (FLD)



Goal: find the projection that maximally 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:

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

problem: w is unconstrained → need normalization

Fisher's Idea: $w^* = \arg\max_w \frac{(m_1 - m_2)^2}{s_1 + s_2}$ ← "between class scatter" ← "within class scatter"

$$= \arg\max_w \frac{w^T S_B w}{w^T S_W w}$$

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

← $S_W = S_1 + S_2$

(tutorial):

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

Note: hyperplane that separates 2 Gaussians w/ cov $\Sigma = \frac{1}{n}(S_1 + S_2)$

⇒ FLD is optimal when 2 classes are Gaussian w/ equal covariance.