Unsupervised Learning

Supervised: have both an input X and output Y

Want to predict Y from X $p(y|x)$

Unsupervised: only have an X

Want to summarize important patterns in X

$p(x)$

Ex. ① dimensionality reduction

rep data using fewer vars

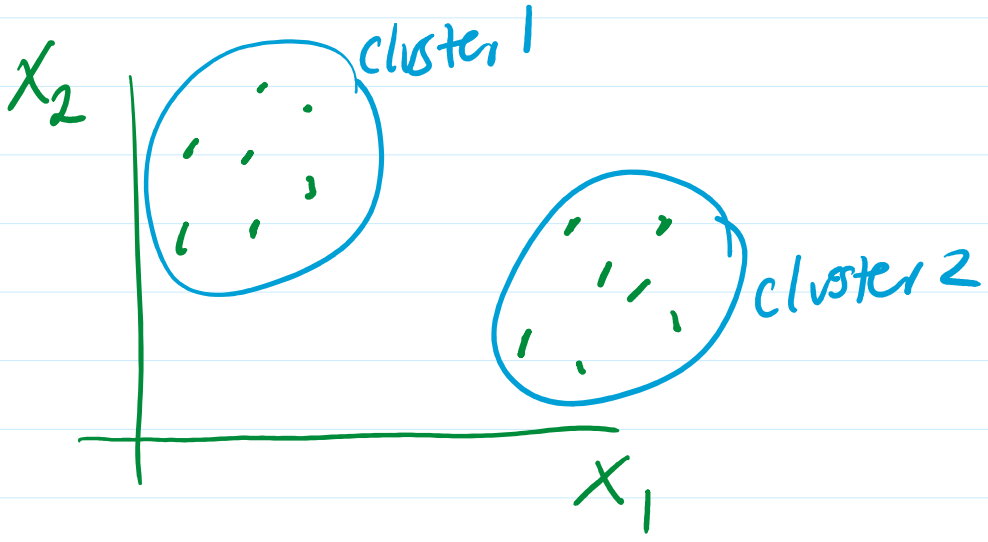
P covariates $\rightsquigarrow$ q covariates

$q \ll P$.

$q \ll p$.

finding high density subspaces

(2) Clustering: group data into similar "clusters"



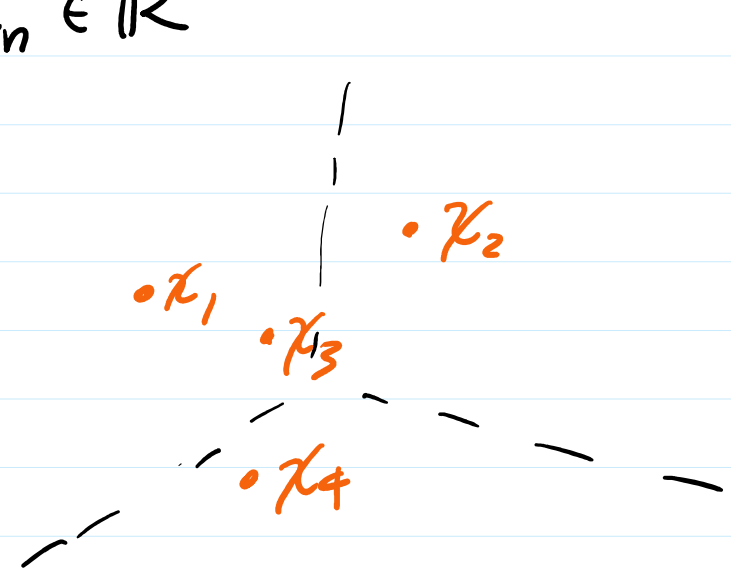
finding high density regions

Principal Components Analysis (PCA)

Dim'l reduction technique.

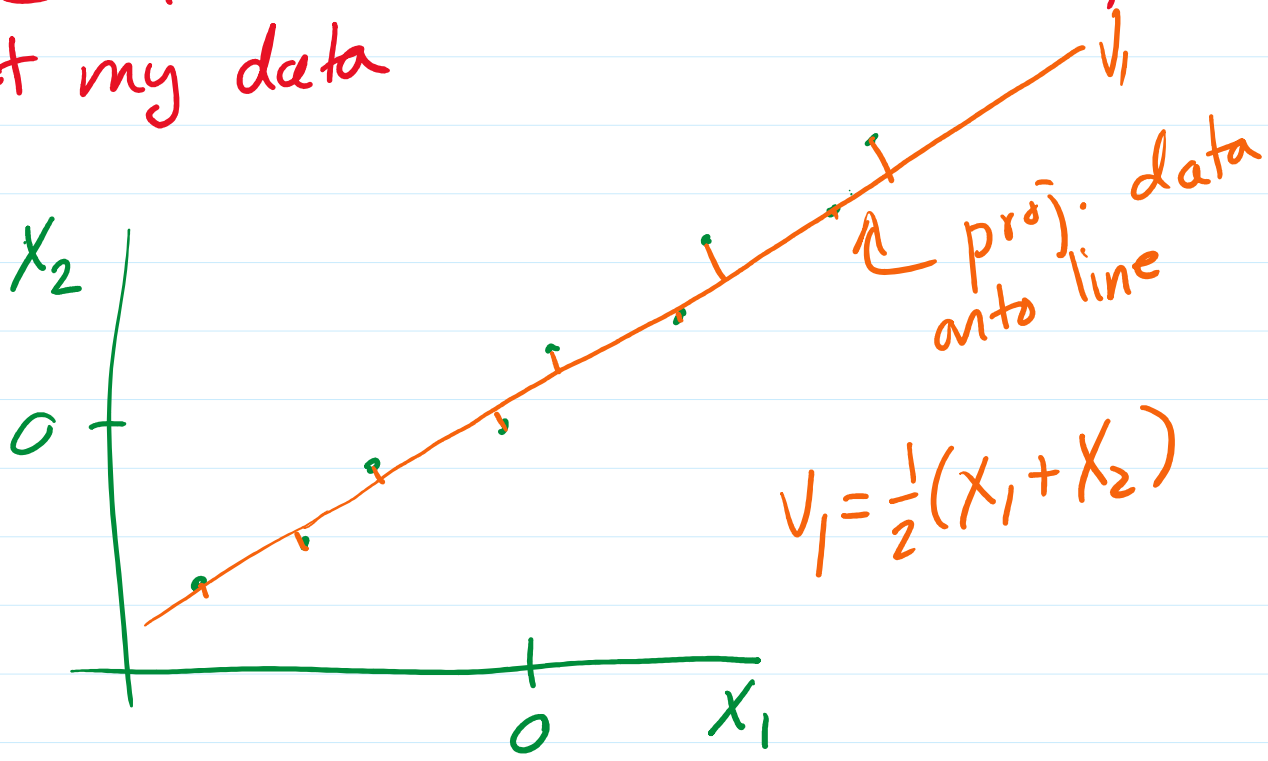
$$\underset{n \times p}{X} = \begin{bmatrix} — \text{ obs } 1 — \\ — \text{ obs } 2 — \\ \vdots \end{bmatrix} = \begin{bmatrix} | & | \\ \text{Var1} & \text{Var2} \cdots \\ | & | \end{bmatrix}$$

Visualize X row-wise so that n^{th} row $x_n \in \mathbb{R}^p$



Goal: find a lower dim'l subspace of $\mathbb{R}^p$ that doesn't lose too much info about my data

Ex.



Goals of PCA:

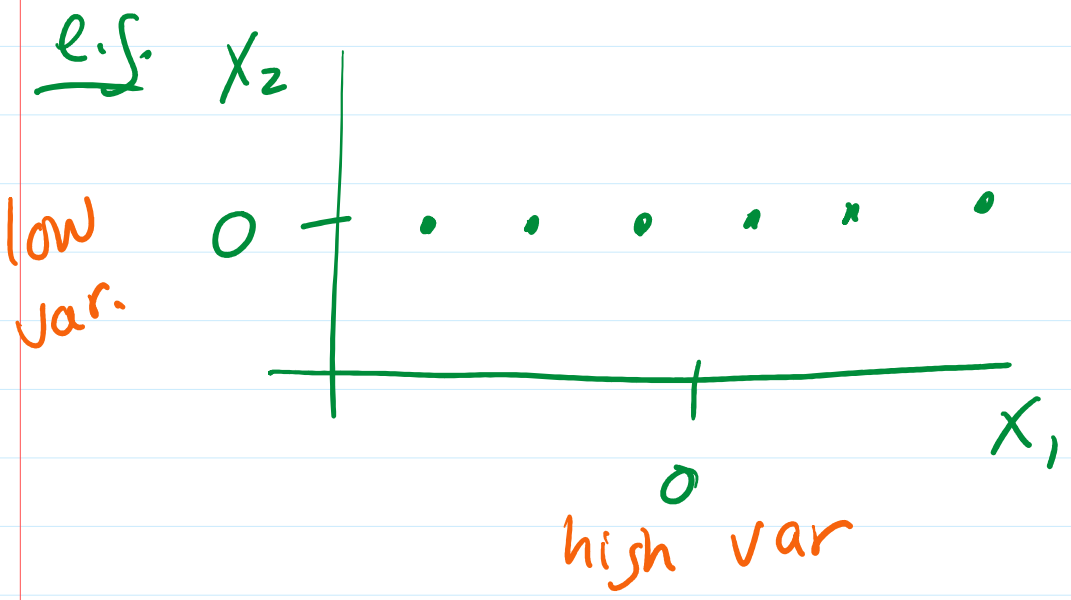
(1) reduce the number of vars

$X_1, \ldots, X_p \xrightarrow[\text{reduce}]{} Z_1, \ldots, Z_q$

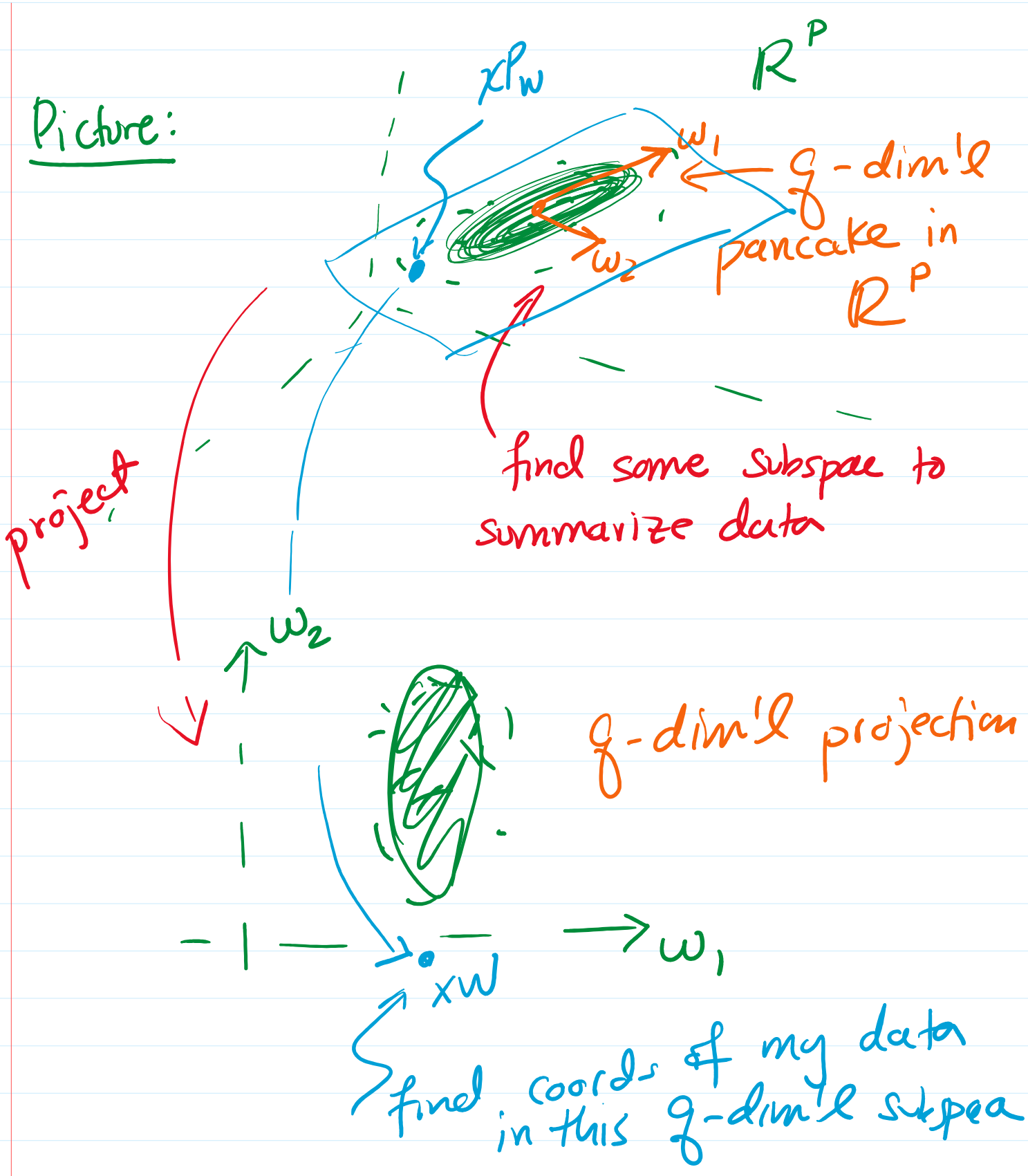
where $q \ll p$.

(2) don't lose too much info

Central dogma: variance = info



Picture:



Proj. matrices

Proj. matrices

Let W be the $p \times q$ mtx of basis elements.

The proj. mtx onto $\mathcal{C}(W)$ is

$$P_W = W(W^TW)^{-1}W^T$$

so that if $x \in \mathbb{R}^p$ (raw-vector) then

xP_W is the

coord (in $\mathbb{R}^p$) of the proj. pt.

If W is "orthogonal" (col orthonormal)

then $$P_W = WW^T$$

and so

$$xP_W = xWW^T$$

xP_W: coords of proj. pt in new basis

$1 \times p \quad p \times q \quad q \times p$

W, W^T: bases

coords of proj. in orig. space ($\mathbb{R}^p$)

$1 \times p$ $p \times q$ $q \times p$ — basis

PCA wants to find pos. of data pts in the lower dim'l space.

If $X_{N \times p}$ is my data mtx then

$$Z = \overbrace{XW}^{N \times q}$$

$N \times p$ $p \times q$

is the data mtx (expressed) embedded in this lower dim'l coords.

If Z_i is the i^{th} col of Z (i^{th} PC) (PCs)

X_j is the j^{th} col of X

W_i is i^{th} col W

then $Z_i = XW_i = X_1 W_{i1} + X_2 W_{i2} + \cdots + X_p W_{ip}$ (LC of cols of X)

for $i = 1, \ldots, q$

PCA: Find LCs of X_js to

(1) max var. of resulting Z_is

Subject to

(2) Z_is are uncorrelated (no redundancy)

(3) W_i are unit vectors

Aside: Let $x \in \mathbb{R}^N$

assume $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n = 0$.

Then $\hat{var}(x) = \frac{1}{N-1} \sum_n (x_n - \bar{x})^2$

$$= \frac{1}{N-1} \sum_n x_n^2$$

$$\propto x^T x$$

Similarly if $y \in \mathbb{R}^N$ is some other var then

$$\hat{cor}(x,y) \propto \hat{cov}(x,y) \propto x^T y$$

So uncorrelated $\simeq$ orthogonal

Similarly if X is an $N \times P$ data matrix (cols are mean-centred)

$\hat{Cov}(X)$ is a $P \times P$ mtx

where $\hat{Cov}(X)_{ij} = \hat{cov}(x_i, x_j)$

where $\text{Cov}(X)_{ij} = \text{cov}(x_i, x_j)$

$$\hat{\text{Cov}}(X)_{ii} = \hat{\text{Var}}(x_i)$$

then

$$\hat{\text{Cov}}(X) \propto \underbrace{X^T X}_{\substack{P \times N \; N \times P \\ P \times P}}.$$

PCA: $Z = XW$

Want: (1) max diag elements of $\hat{\text{Cov}}(Z)$
(max sum of diag)

(2) off diag elements of $\hat{\text{Cov}}(Z)$ are zero

(3) Cols of W to be unit vectors.

(3) lots of ... vectors.