Lecture 15

Can generally fit penalized methods

$$\hat{f} = \underset{f}{\operatorname{argmin}}\ L(f)$$

L: loss

penalize:

$$\hat{f} = \underset{f}{\operatorname{argmin}}\ L(f) + \lambda J(f)$$

λ: penalization strength

$J(f)$: measures complexity of f

E.g. $\beta \Leftrightarrow f$

ridge:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ L(\beta) + \lambda \|\beta\|_2^2$$

$\|\beta\|_2^2$: complexity

LASSO: " $+ \lambda \|\beta\|_1$

Can do this w/ other methods too.

E.G. Logistic Regression

$$y_n \mid \underset{\sim}{X}_n = \underset{\sim}{x}_n \overset{\text{indep}}{\sim} \text{Bern}\left(p_\beta(\underset{\sim}{x}_n)\right)$$

logistic fn

$$p_\beta(\underset{\sim}{x}_n) = 1 / \left(1 + \exp\left(-\underset{\sim}{x}_n^T \beta\right)\right)$$

Find $\hat{\beta}$ as MLE

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ P_\beta(\underset{\sim}{Y} \mid \underset{\sim}{X})$$

$$= \underset{\beta}{\operatorname{argmin}} - \log P_\beta(\underset{\sim}{Y} \mid \underset{\sim}{X})$$

negative log likelihood (NLL)

$$= \underset{\beta}{\operatorname{argmin}}\ L(\beta)$$

Can also penalize this fit,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ L(\beta) + \lambda \|\beta\|_2^2$$

$$\text{“} \qquad \text{“} + \lambda \|\beta\|_1$$

This has similar effects as in regression case.

Unsupervised Learning

Supervised Problem: have both input X and an output Y

Want to predict Y from X. $\mathbb{P}(Y|X)$

Unsupervised: only have X

Want to summarize important patterns in X. $\mathbb{P}(X)$

Ex.

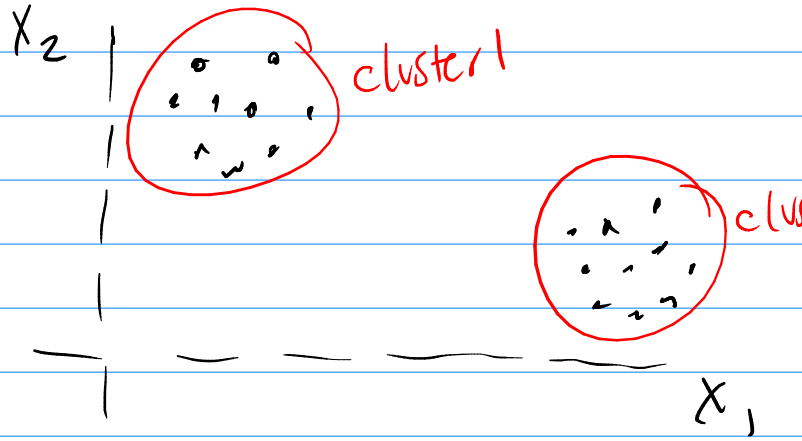
(1) **dimensionality reduction**:

represent data using fewer vars.

P covariates $\rightsquigarrow$ q covariates

$q \ll P$ high density subspaces of $\mathbb{P}(X)$

(2) **clustering**: group my data into similar "clusters"



high-density regions of $\mathbb{P}(X)$

Principal Components Analysis (PCA)

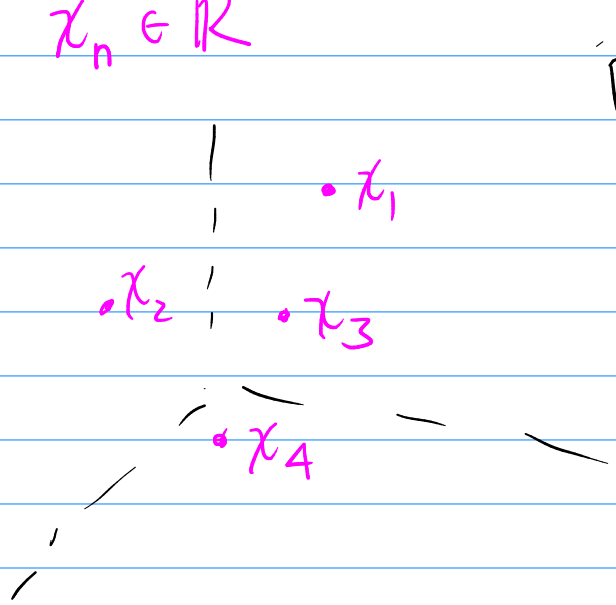
→ dim-reduction approach

Unsupervised!

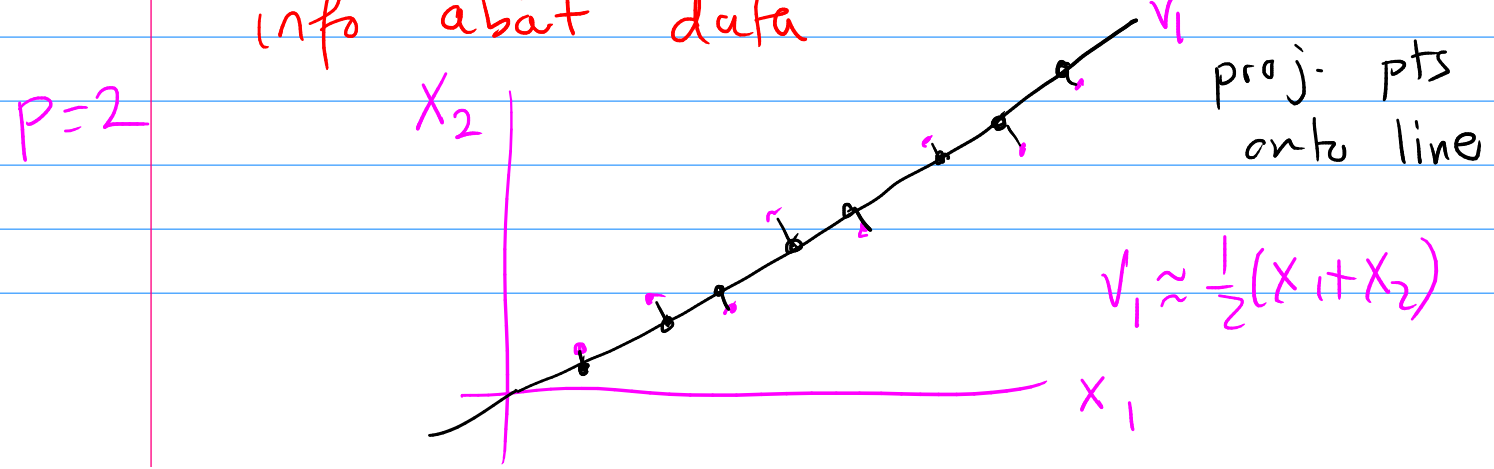
$$\underset{N \times P}{X} = \begin{bmatrix} — \text{ obs } 1 — \\ — \text{ obs } 2 — \\ \vdots \end{bmatrix} = \begin{bmatrix} | & | & \\ \text{Var1} & \text{Var2} & \cdots \\ | & | & \end{bmatrix}$$

visualize X row-wise

n^{th} row $x_n \in \mathbb{R}^p$



dim-reduction tries to find a lower dim'l subspace that doesn't lose too much info about data



Goals of PCA:

(1) reduce the number of vars

$X_1, X_2, \ldots, X_p \xrightarrow{\text{reduce}} Z_1, Z_2, \ldots, Z_q$ (principal components)

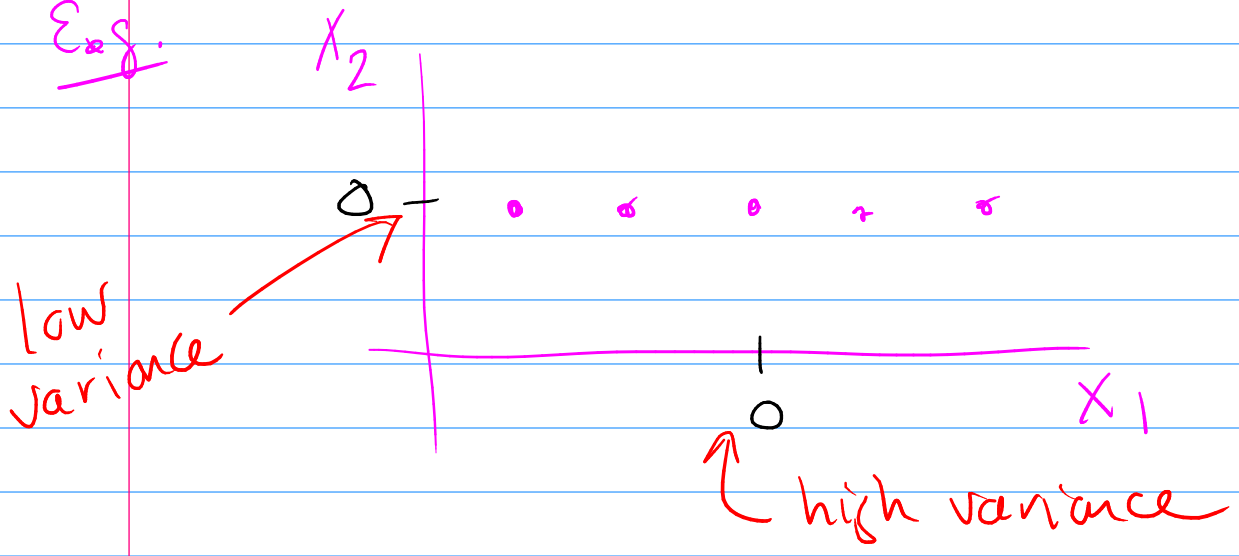
where $q \ll p$.

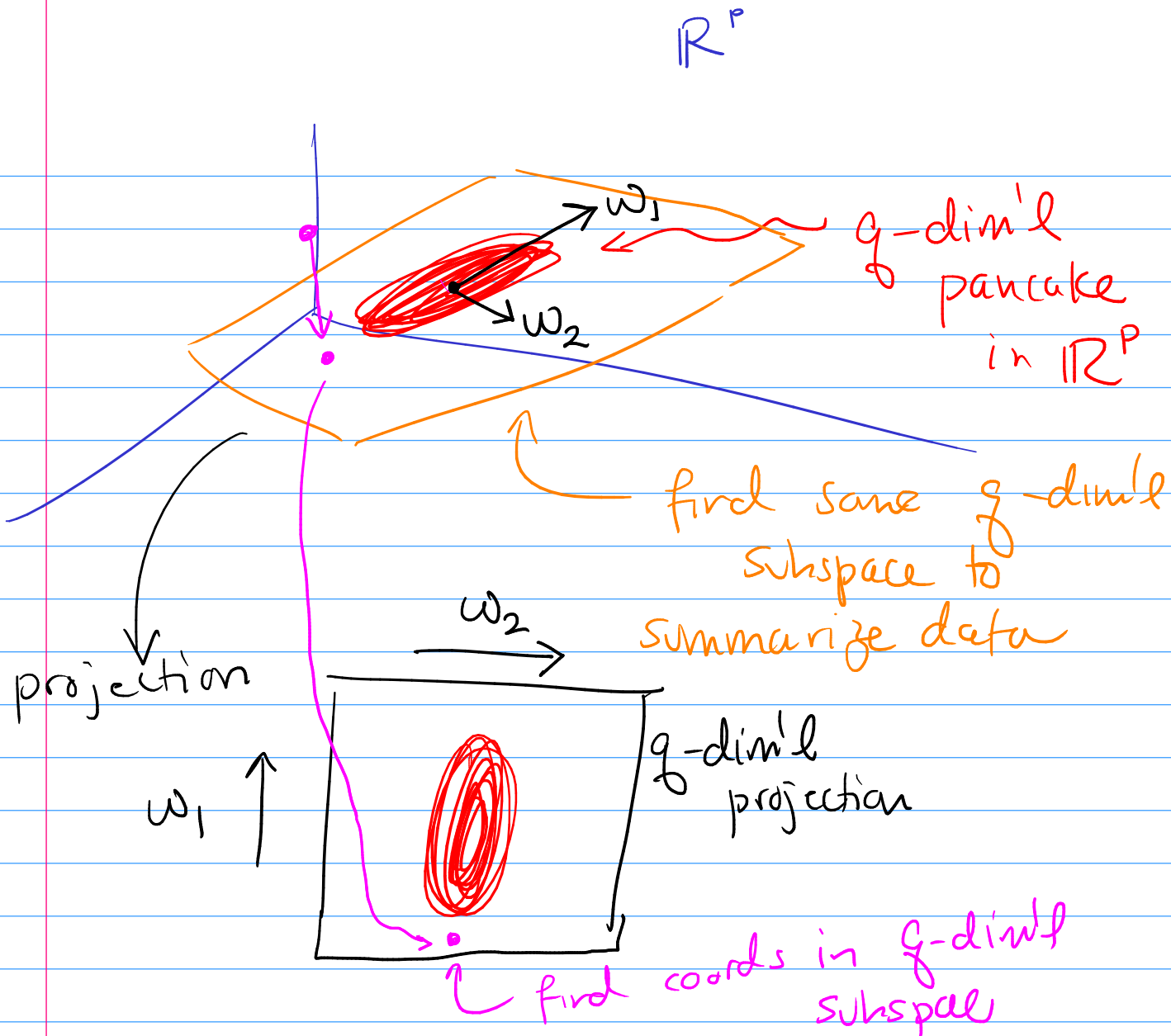
(2) don't want to lose too much info about my data

$\Updownarrow$

Central dogma: variance = information

Ex:





Proj. matrices

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

The proj. mtx onto $\text{col}(W)$ is

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

so that if $x \in \mathbb{R}^p$ then xP_W is the coord (in $\mathbb{R}^p$) of proj. pt.

If W has ortho-normal cols then

$$W^T W = I \quad \text{so}$$

$$P_W = W W^T \quad (P \times P)$$

so

$$\underbrace{\chi P_W}_{P} = \underbrace{\chi W}_{q} \underbrace{W^T}_{P \times P}$$

χP_W: coords in org. basis

χW: coordinates of proj. in W basis

PCA wants to calc positions of pts in the lower dim'l subspace.

If $X_{N \times P}$ is data mtx, W is $P \times q$ basis mtx then

$$Z = X W$$

$N \times P \quad P \times q$

Z: $N \times q$ mtx of data in lower-dim'l coords

If Z_i is the i^{th} col of Z

X_j " j^{th} col of X

$$Z_i = X W_i = X_1 W_{i1} + X_2 W_{i2} + \cdots + X_p W_{ip}$$

linear combn of X_js w/ weights specified by W

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 to be unit vectors