Principal component methods

Readings for today

- Chapter 6: Linear model selection and regularization. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer

Topics

1. Principal components analysis

2. Principal component regression

3. Partial least squares

Principal component analysis

Dimensionality

Dimensionality of a model: n x p

As n →p, dimensionality increases & model variance increases

$$\begin{pmatrix} x_{1,1} \\ \cdots \\ x_{n,1} \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,15} \\ \cdots & & & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,15} \end{pmatrix} \rightarrow \uparrow \text{model flexibility}$$

How do you reduce the dimensionality of your model?

(James et al. 2013)

How to deal with high model dimensionality

<u>So far we have covered:</u>

1. Feature selection by comparing lower dimensional variants of your model.
   - Best subset selection
   - Forward/Backward stepwise selection
2. Apply a sparsity constraint to your model during fitting.
   - Ridge regression
   - LASSO
   - Elastic Net

(James et al. 2013)

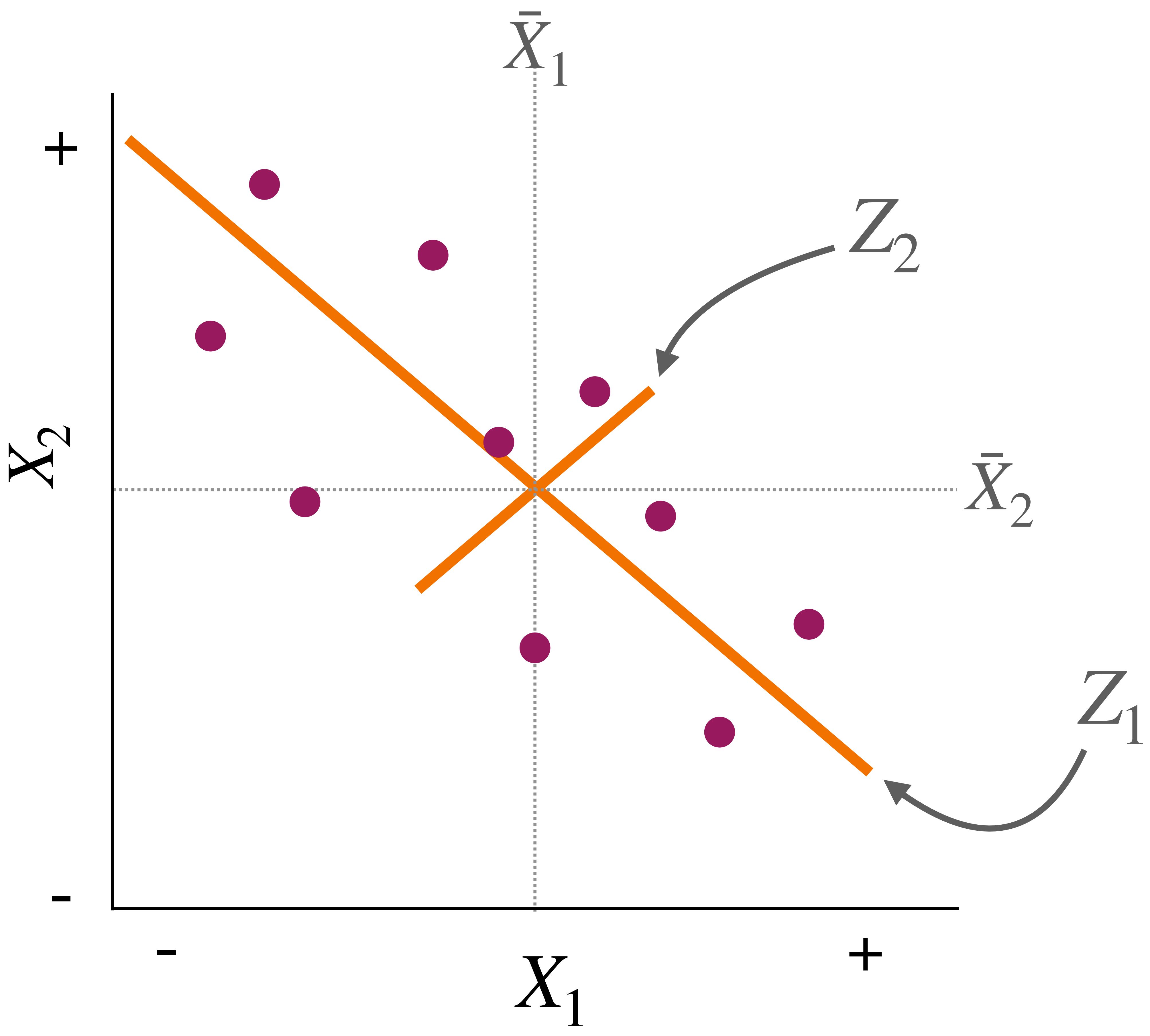
Principal component analysis (PCA)

What if you reduce the dimension of X itself?

<u>PCA:</u>

$$\underbrace{Z_m}_{\text{lower dimensional projection (component)}} = \sum_{j=1}^{p} \underbrace{\phi_{j,m}}_{\text{component loading}} \underbrace{X_j}_{\text{original data variable}}$$

Low dimensional components



The first principal component (Z_1) explains the most variance about the relationship between X_1 and X_2.

PCA algorithm

Step 1: Find the first component (Z_1) loading. $$\phi_1 = \arg\max(\frac{\phi_1' X' X \phi_1}{\phi_1' \phi_1})$$

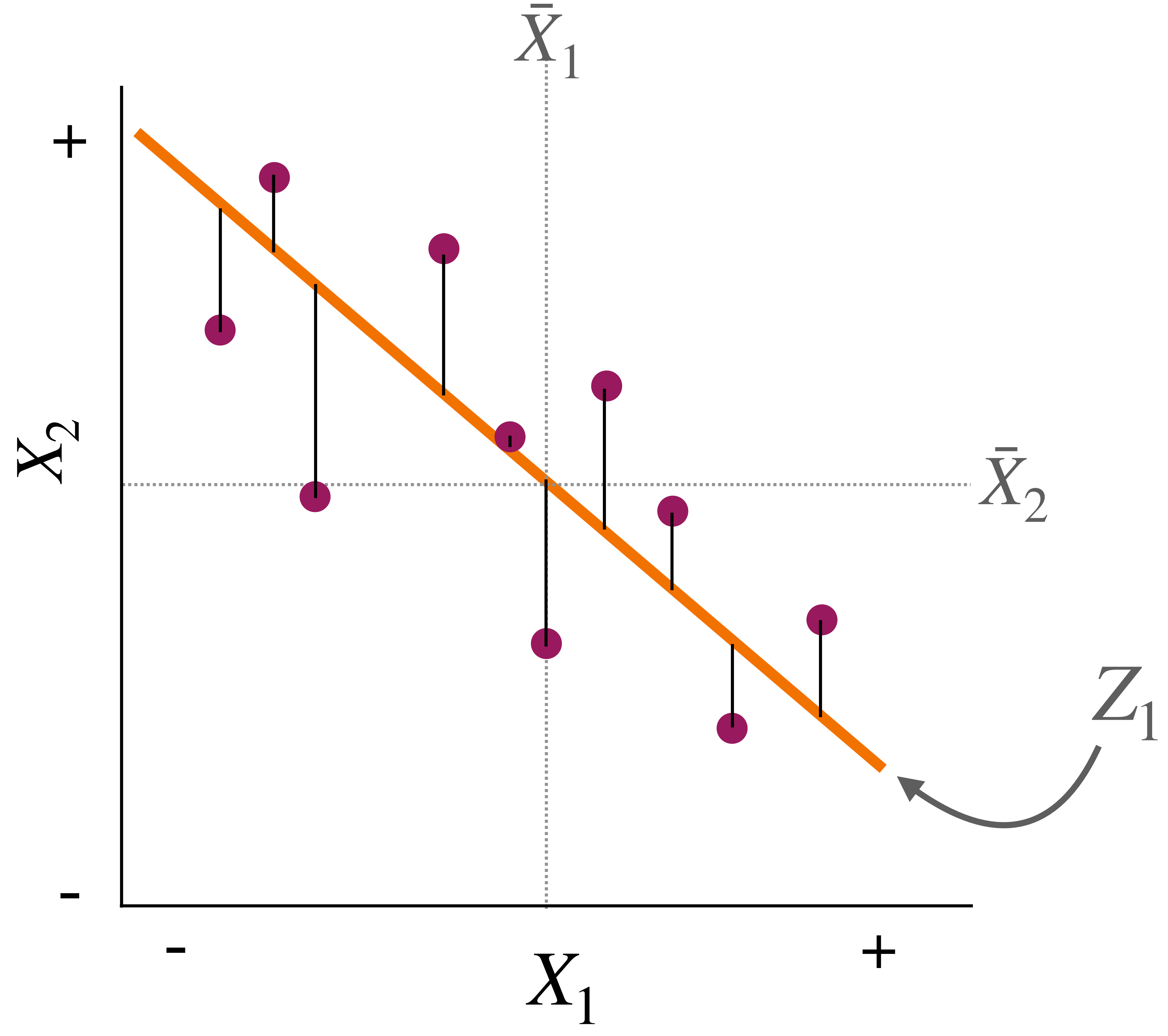
Step 2: Take the residuals after accounting for Z_1. $$\hat{X}_m = X - \sum_{s=1}^{m-1} X \phi_s \phi_s'$$

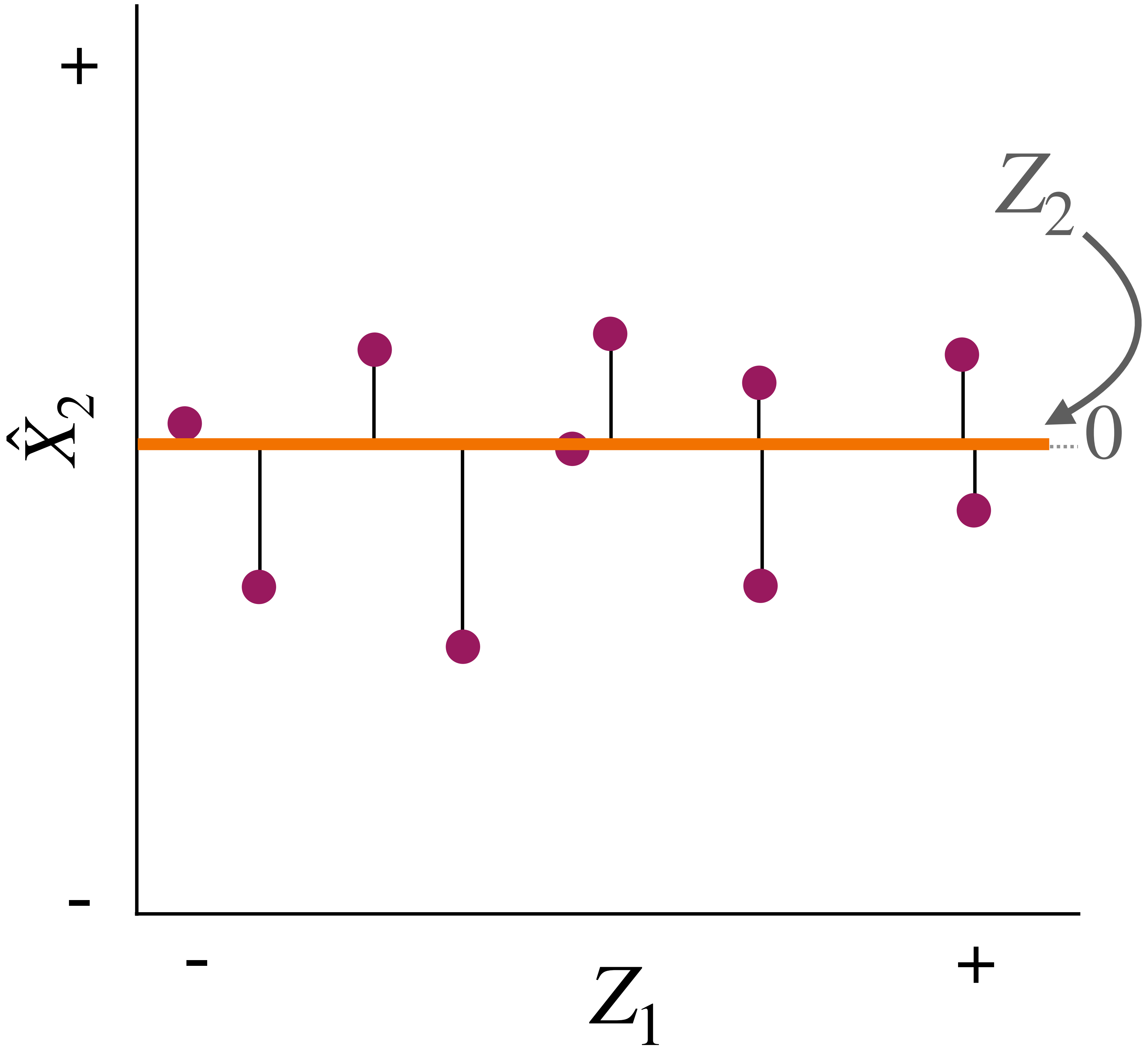
Step 3: Calculate the next component (Z_m). $$\phi_m = \arg\max(\frac{\phi_m' \hat{X}_m' \hat{X}_m \phi_m}{\phi_m' \phi_m})$$

Step 4: Repeat Steps 2-3 until $m = p$

(James et al. 2013)

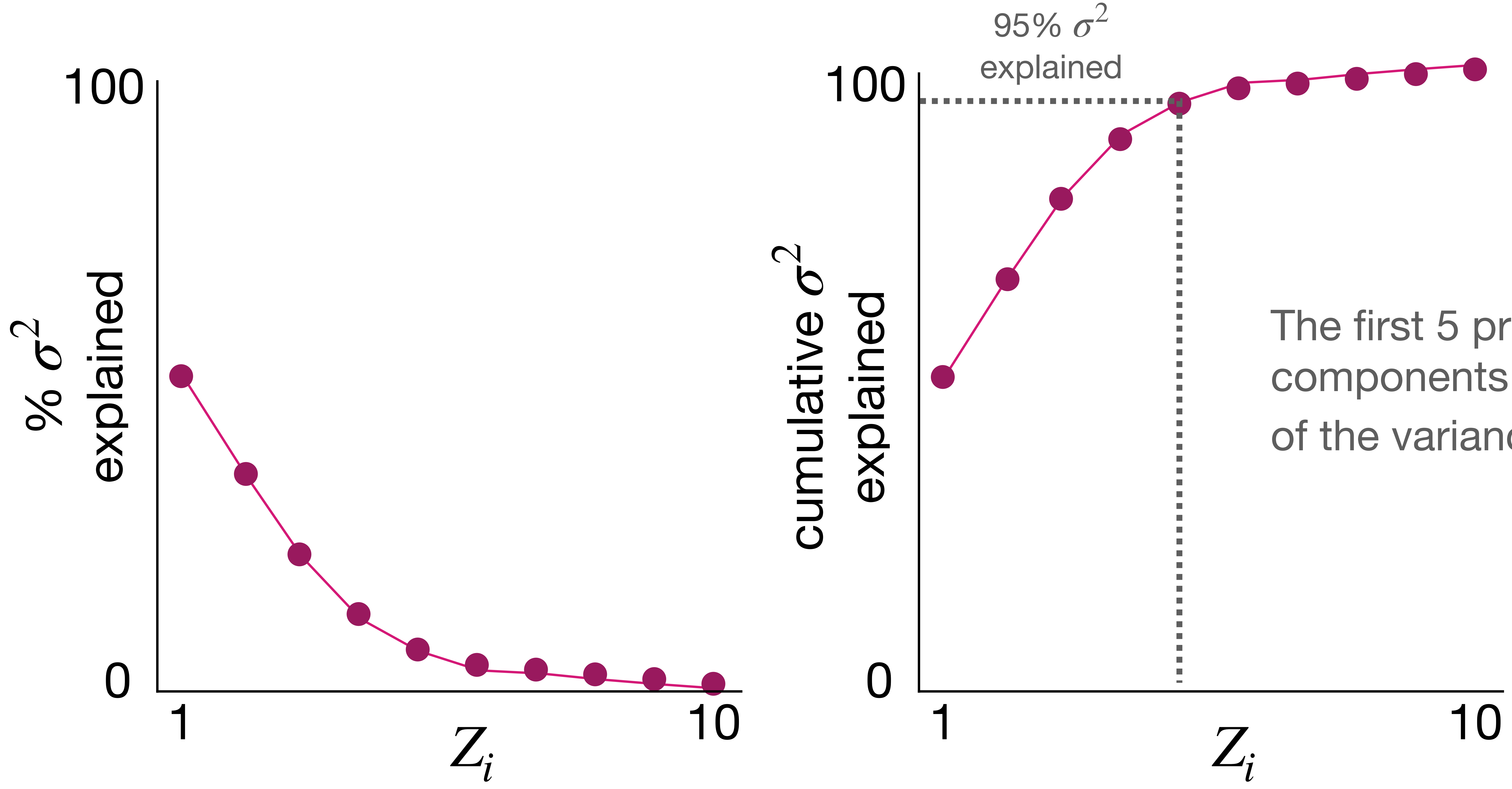
PCA algorithm





Example: 10 dimensional X

$$\begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & & \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \rightarrow n = 15,\ p = 10$$



The first 5 principal components explain 95% of the variance in X.

PCA vs. Factor Analysis

<u>PCA:</u> $$\underbrace{X}_{n \times p} = \underbrace{Z}_{n \times m} \underbrace{\phi^{-1}}_{m \times p}$$

<u>Factor Analysis (FA):</u> $$\underbrace{X^T}_{p \times n} = \underbrace{L}_{p \times m} \underbrace{F}_{m \times n}$$

Factor loading matrix with m specified up front.

- PCA is better for exploratory analysis.
- FA is better for hypothesis testing.
- PCA explains <u>all</u> variance in X.
- FA <u>assumes</u> lower dimensionality in X.

Principal component regression

Principal component regression (PCR)

Goal: Reduce the dimensions of X using PCA and use the strongest components in Z as your predictor variables.

PCA + linear regression:

$$k < p$$

# of components that explain $1 - \alpha$ percent variance in X

regression coefficients learned on $Z_1, \ldots, Z_k$

$$\hat{y}_i = \sum_{m=1}^{k} \hat{\theta}_m Z_{i,m}$$

principal component m

(James et al. 2013)

PCR algorithm

Step 1: Calculate $Z = \phi X$.

Step 2: Identify the k components that explain $1 - \alpha$ (e.g., $(1 - 0.05) = 95\,\%$) of the variance in X.

Step 3: Fit your regression model with the k components identified in Step 2 (i.e., learn $\hat{Y} = \hat{f}(Z_{1...k})$)

(James et al. 2013)

Projecting back to X

Can determine the regression weights in X (i.e., $\hat{\beta}_j$) that best resolve the bias-variance tradeoff via the coefficients learned in Z (i.e., $\hat{\theta}_j$).

PCR to linear regression:

$$\hat{y}_i = \sum_{m=1}^{k} \hat{\theta}_m Z_{i,m} \longrightarrow Z_m = \sum_{j=1}^{p} \phi_{j,m} X_j$$

$$= \sum_{m=1}^{k} \hat{\theta}_m \sum_{j=1}^{p} \phi_{j,m} X_j$$

$$= \sum_{j=1}^{p} \underbrace{\sum_{m=1}^{k} \hat{\theta}_m \phi_{j,m}}_{\hat{\beta}_j} X_j$$

Best regression model:

$$\hat{\beta}_j = \sum_{m=1}^{k} \hat{\theta}_m \phi_{j,m}$$

Even when n is close to p.

(James et al. 2013)

Partial least squares

Partial least squares (PLS)

Goal: Find the lower dimensions in X that *maximize* $Cov[X, Y]$

# of variables in Z ($k = p$)

# of variables in X

regression coefficients learned on $Z_1, \ldots, Z_k$

$$\hat{y}_i = \sum_{j=1}^{p} \sum_{m=1}^{k} \hat{\theta}_m \hat{\phi}_{j,m} x_{i,j}$$

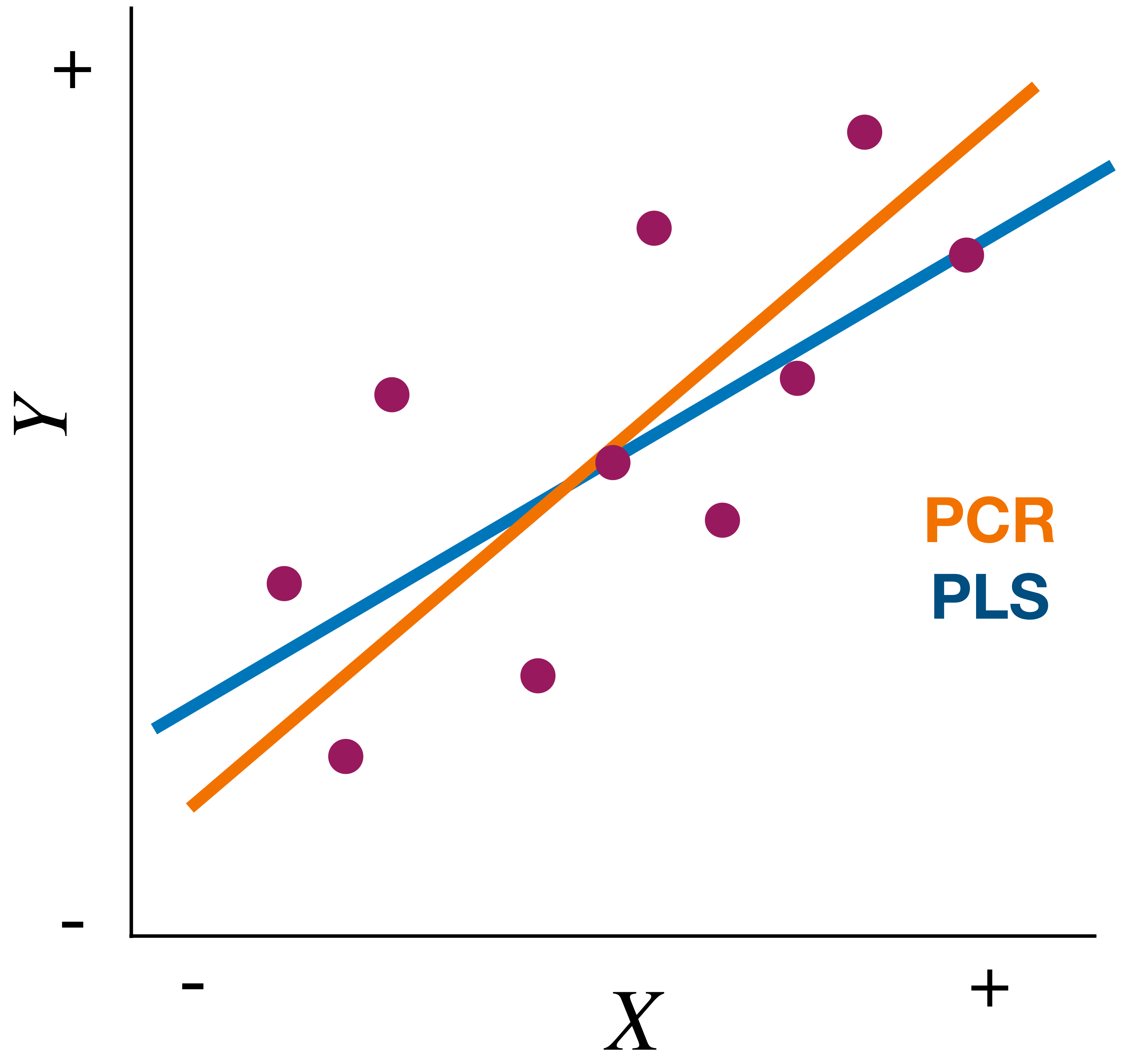
principal component loading learned by minimizing

$$\sum (y - \sum_{m=1}^{k} \hat{\theta}_m Z_{i,m})^2$$

$\hat{\phi}$ and $\hat{\theta}$ estimated in one step.

(James et al. 2013)

PLS vs. PCR



- PLS and PCR produce qualitatively different results depending on how the low dimensional components in X associate with Y.

(James et al. 2013)

Take home message

- Principal component methods offer an easy way of resolving the bias-variance tradeoff in high dimensionality (i.e., high variance) contexts by leveraging any correlational structure in your predictor variables.