Learning with Singular Vectors

CIS 520 Lecture
30 October 2015
Barry Slaff

Based on:
CIS 520 Wiki Materials
Slides by Jia Li (PSU)
Works cited throughout

Overview

Linear regression: Given **X, Y** find $\hat{w}$**:** $\hat{Y} = X\hat{w}$

Choose the best $\hat{w}$ for projecting **X** onto **Y.**

Goal: Build a model that predicts well with *new* data.

Motivating questions:

1. What are the best possible features to extract from **X**?
2. What if **X** has many more features than observations or many features of **X** are highly correlated?
3. What if each prediction is not a single number, but a vector **y**?

Overview

- **Ordinary Least Squares (OLS) Regression:** finds the projection direction for which the x's are maximally correlated with the y's
- **PCA:** new X features. Finds the directions of maximal covariance of the x's.
- **Principal Component Regression (PCR):** does PCA for dimensionality reduction on X, and then OLS using PC features.
- **Partial Least Squares (PLS) Regression:** new X and Y features. Finds the projection directions of X and Y that maximize their *covariance*.
- Regularization with Ridge Regression, PCR, and PLS.
- **Canonical Correlation Analysis:** new X and Y features. Finds the projection directions of X and Y that maximize their *correlation*.
- **PCA** and **CCA:** both using SVD to minimize reconstruction error or maximize variance/covariance

Linear Methods vs. Neural Nets

Linear methods: new features are linear combinations of original features

Neural nets are great!

Why use linear methods?

Tera Scale deep learning project:

10 million images (200 x 200 pixels), 1 billion parameters

Singular Value Decomposition

Singular value decomposition of matrix X: $\mathbf{X = UDV^T}$

X: the data matrix. (n x p).

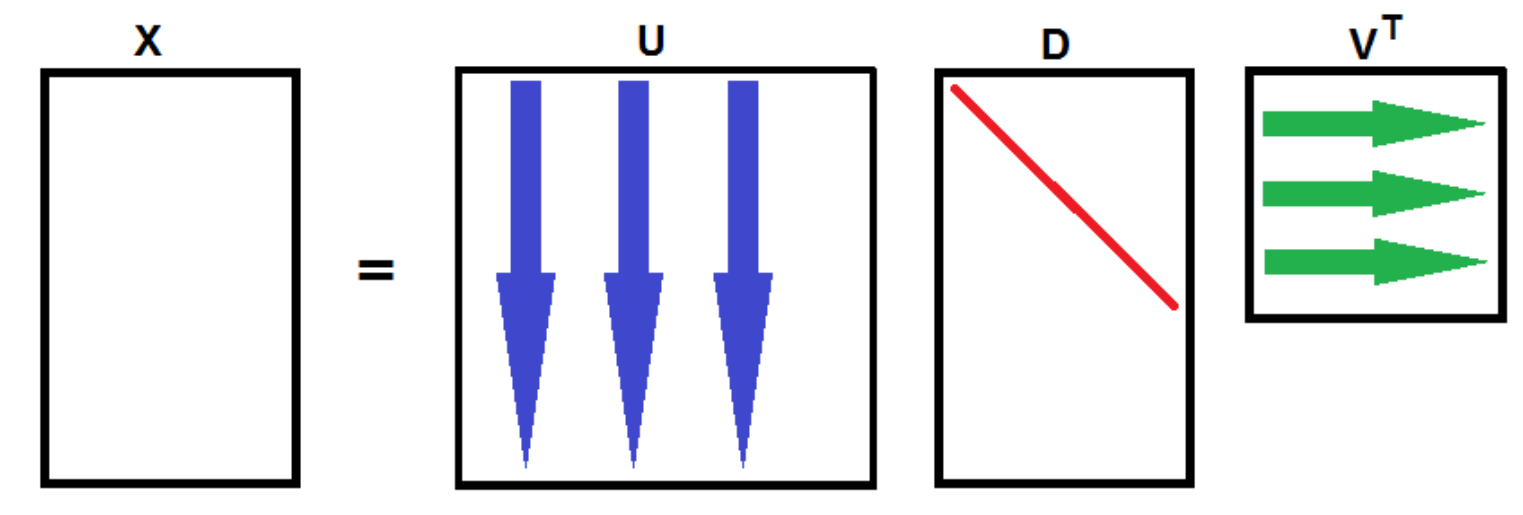
U: orthogonal, $\mathbf{U^TU=I}$. (n x n).
Columns of U are the *left singular vectors of X.*

D: diagonal. (n x p).
Diagonal elements of D are the *singular values of X*.
All non-negative; in *decreasing* order of magnitude down the diagonal.

V: orthogonal, $\mathbf{V^TV=I}$. (p x p).
Columns of V are the *right singular vectors of X*.

Singular Value Decomposition

Singular value decomposition of X: $\mathbf{X = UDV^T}$



Let k = min(n,p). Then: $\mathbf{X} = \sum_{i=1}^{k} D_{ii}\boldsymbol{u}_i\boldsymbol{v}_i^T$

Since all $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ are unit vectors, the importance of the i'th term in the sum is determined by the size of D_{ii}.

Singular Value Decomposition

$$\boldsymbol{X = UDV^T}, \qquad \boldsymbol{X^T X = V\left(D^T D\right)V^T}$$

The columns $\mathbf{v_1}, \ldots \mathbf{v_p}$ of **V** are the *eigenvectors* of the covariance matrix $\mathbf{X^T X}$. Hence we can write

$$\boldsymbol{X^T X} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} (D_{ii})^2 \, \boldsymbol{v_i v_i^T}$$

From before:

$$\boldsymbol{X} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} D_{ii} \boldsymbol{u_i v_i^T}$$

k = min(n,p).

D_{ii} are singular values of X, $(D_{ii})^2$ are eigenvalues of $\mathbf{X^T X}$

Principal Component Analysis

PCA finds the directions of max covariance of the X's:

If X is mean-centered, then PCA finds the directions

$$v_i = \underset{\substack{w_i \in \mathbb{R}^p \\ w_i^T w_i = 1}}{argmax} \ (Xw_i)^T (Xw_i)$$

such that v_i is uncorrelated with v_j for all j<i.

Principal Component Analysis

$\mathbf{X} \rightarrow \mathbf{X}_c = \mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{Z}\mathbf{V}^T$

$\mathbf{X}_c$ is (n x p), $\mathbf{Z}$ is (n x p), $\mathbf{V}$ is (p x p).

$\mathbf{Z}$ is the transformation of $\mathbf{X}$ into "PC space"
Column vector $\boldsymbol{z}_i$ is the i'th *PC score vector.*
Column vector $\boldsymbol{v}_i$ is the i'th *PC direction* or *loading.*

Since $\mathbf{V}$ is orthogonal, $\mathbf{X}_c\mathbf{V} = \mathbf{Z}\mathbf{V}^T\mathbf{V} = \mathbf{Z}$, and therefore:

$$\boldsymbol{z}_i = \boldsymbol{X}_c\boldsymbol{v}_i = \boldsymbol{u}_i D_{ii}$$

Hence $\boldsymbol{z}_i$ is the projection of the row vectors of $\mathbf{X}_c$ on the (unit) direction $\boldsymbol{v}_i$, scaled by D_{ii}.

Principal Component Analysis

$$X \rightarrow X_c = UDV^T = ZV^T$$

$$X_c^T X_c = \sum_{i=1}^{p} (D_{ii})^2 \, v_i v_i^T$$

"% Variance explained by the i'th principal component:"

$$= 100 \cdot \frac{(D_{ii})^2}{\sum_{j=1}^{p} (D_{jj})^2}$$

PCA

True or false: If $\mathbf{X}$ is any matrix, and $\mathbf{X}$ has singular value decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

then the principal component scores for $\mathbf{X}$ are the columns of

$$\mathbf{Z} = \mathbf{U}\mathbf{D}.$$

(a) True

(b) False

PCA

If X is mean-centered, then PCA finds…?

(a) Eigenvectors of $\mathbf{X}^T\mathbf{X}$

(b) Right singular vectors of $\mathbf{X}$

(c) Projection directions of max covariance of $\mathbf{X}$

(d) All of the above

PCA: Reconstruction Problem

PCA can be viewed as an L_2 optimization, minimizing distortion, the reconstruction error.

$$Z^*, V^* = \underset{\substack{Z \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{p \times k}, \\ v_i^T v_j = \delta_{ij}}}{argmin} |X_c - ZV^T|_F$$

Here we have constrained **Z**, **V** by dimension:

$\mathbf{X_c}$ is still (n x p).

Z is (n x k), with $k \leq p$.

V is (p x k).

If k=p then the reconstruction is perfect. k<p, not.

Sparse PCA

Apply L1-norm constraints to the PCA optimization problem to zero out loadings. (Another variation: L0-norm constraints.)

Similar to using an L1-norm penalty to zero out weights in penalized linear regression.

$$Z^*, V^* = \underset{\substack{Z \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{p \times k}, \\ v_i^T v_j = \delta_{ij}}}{argmin} |X_c - ZV^T|_F$$

Subject to:

$$|v_i|_1 < c_1 \text{ for } i = 1, \ldots, k$$
$$|z_i|_1 < c_2 \text{ for } i = 1, \ldots, k$$

Improves interpretability of PCA: "which PC scores really matter?"
See Zhou, Hastie, and Tibshirani, 2006.

Regularized PCA

- PCA, with feature selection
- Sparse PCA, possibly also with feature selection

Why regularize PCA?

PCR:
Principal Component Regression

Ordinary Least Squares (OLS) Regression finds the direction **w** for which the **x**'s are maximally correlated with (predictive of) the **y**'s.

PCR has two steps:

1. Do a PCA for dimensionality reduction
2. Do OLS regression using the PC features, usually with feature selection.

Toy Data

Suppose we generate toy data as follows:

- **X** is generated from normal random variables, all with the same mean and variance
- **Y** is generated as a linear combination of some of the **x**'s, plus noise

If n>p: compared to normal OLS, what performance would we expect for…?

(a) PCR using all the components

(b) PCR using a small number of components

PCR: Principal Component Regression

$$\mathbf{X} \rightarrow \mathbf{X_c} = \mathbf{ZV^T}$$

The columns $\mathbf{z_1}, \ldots \mathbf{z_k}$ can be used as features in supervised learning.

Ex: linear regression. Given training **X** and **Y**,

$$w^* = \underset{w \in \mathbb{R}^p}{arg\min} |Y - Zw|_2^2$$

If k=p: result is the *same* as linear regression with X, Y

If k<p: this is a form of *regularized* linear regression

So is ridge regression! How are PCR and Ridge fundamentally different?

PCR: Principal Component Regression

$\mathbf{X_c = UDV^T = ZV^T,}$ $\quad w^* = \underset{w \in \mathbb{R}^p}{arg\min} |Y - Zw|_2^{\,2}$

When the solution is unique, we can use the normal equation to write:

$$\hat{Y} = Zw^* = Z(Z^TZ)^{-1}Z^TY = UD(D^TU^TUD)^{-1}D^TU^TY$$

$$\hat{Y} = UU^TY = \sum_{i=1}^{n} \boldsymbol{u_i u_i^T Y}$$

UU^T is the (n x n) hat matrix.

Ridge Regression in terms of SVD

X = UDV^T, $w^* = \underset{w \in \mathbb{R}^p}{argmin} \left(|Y - Xw|_2^{\,2} + \gamma |w|_2^2 \right)$

Can solve: $\hat{Y} = Xw^* = X(X^TX + \gamma I)^{-1}X^TY$

$$\hat{Y} = UDV^T(V[D^TD + \gamma(V^TV)]V^T)^{-1}VD^TU^TY$$

$$\hat{Y} = UD(D^TD + \gamma I)^{-1}DU^T = U\tilde{D}U^T$$

$$\hat{Y} = \sum_{i=1}^{n} \frac{(D_{ii})^2}{(D_{ii})^2 + \gamma} \boldsymbol{u}_i \boldsymbol{u}_i^T = \sum_{i=1}^{n} \frac{\lambda_i^2}{\lambda_i^2 + \gamma} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

OLS vs. Ridge vs. PCR

OLS: $\mathbf{X = UDV^T}$ $\hat{Y} = \sum_{i=1}^{n} \boldsymbol{u_i u_i^T Y}$

Regularized methods:

PCR: $\mathbf{X_c = UDV^T}$ $\hat{Y} = \sum_{i=1}^{k} \boldsymbol{u_i u_i^T Y}$, $k \leq n$

Ridge: $\mathbf{X = UDV^T}$ $\hat{Y} = \sum_{i=1}^{n} \frac{D_{ii}^2}{D_{ii}^2+\gamma} \boldsymbol{u_i u_i^T Y}$

Ridge shrinks *all* the singular vectors, and keeps all.

PCR chooses the k "largest" singular vectors.

Ridge Shrinkage

Ridge: $\mathbf{X = UDV^T}$ $\hat{Y} = \sum_{i=1}^{n} \frac{D_{ii}^2}{D_{ii}^2+\gamma} \boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{Y}$

Which eigenvectors of $\mathbf{XX^T}$ does Ridge shrink the most (by % of original, for fixed gamma)?

(a) Largest eigenvalues

(b) Smallest eigenvalues

(c) All the same

Ridge Shrinkage Example

Suppose **X, Y** have a unique OLS solution.

Suppose $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and the nonzero singular values are 5, 4, 3, 2, and 1.

- What are the nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$?
- When constructing the hat matrix, how are these eigenvalues shrunk by PCR?
- When constructing the hat matrix, how are these eigenvalues shrunk by Ridge?

Partial Least Squares Regression

If **Y** is high-dimensional, we might want to do dimension reduction for *both* **Y** and **X**. Regress only the truly significant **Y** features against the truly significant **X** features.

PLS adjusts the PCA directions to a better job of predicting the y's.

PLS finds the projection directions of maximum covariance for **X** and **Y**. PLS is a kind of *canonical covariance analysis*.

(Comparison: PCA finds the projection directions of maximum covariance for **X** and **X**.)

Partial Least Squares Regression

PLS finds the projection directions of maximum covariance for **X** and **Y**. Basic idea*:

Project $\mathbf{X}_c$ down to **T**. Project $\mathbf{Y}_c$ down to **U.**

(**U**, **T** have the same dimension)

"Inner model": regress **U** on **T.**

One scalar regression weight per pair $\mathbf{u}_i$, $\mathbf{v}_i$.

Final model: regress **Y** on **X**

- Project each new **x** down into **T**-space
- Predict **u**'s based on **t**'s (inner model)
- Project each **u** up to each final **y-hat**.

*Historically, PLS could refer to one of many variant algorithms.

Partial Least Squares Regression

Find reduced-dimension representations **T** (of $\mathbf{X}_c$) and **U** (of $\mathbf{Y}_c$) such that each pair of corresponding columns $\mathbf{t}_i$, $\mathbf{u}_i$ are optimal in the following sense:

Let $$w_i^*,\ v_i^* = \underset{\substack{w_i \in \mathbb{R}^p,\ v_i \in \mathbb{R}^m \\ w_i^T w_i = v_i^T v_i = 1}}{argmax} (X_c w_i)^T (Y_c v_i)$$

Subject to: $(X_c w_i^*)^T (X_c w_j^*) = 0$ for all **j**<**i**.

Then: $t_i := X_c w_i^*$ and $u_i := Y_c v_i^*$

PLS Regression

The first singular value a_1 of $\mathbf{X}^T\mathbf{Y}$ has the interpretation

$$(a_1)^2 = \max_{|d|=|e|=1} d^T X^T Y e$$

For $\mathbf{w}_1$=**d** and $\mathbf{v}_1$=**e**, this is what we've computed above.

$\mathbf{w}_1$ is the first left singular vector of $\mathbf{X}^T\mathbf{Y}$.

$\mathbf{v}_1$ is the first right singular vector of $\mathbf{X}^T\mathbf{Y}$.

More on PLS:

Hoskuldsson A, "PLS Regression Methods," J. Chemometerics, 1988

Abdi H, Partial Least Squares (PLS) Regression:
https://www.utdallas.edu/~herve/Abdi-PLS-pretty.pdf

PCR and PLS Feature Scores

The process of initially computing the *feature scores* to be considered in principal component regression uses…?

The process of initially computing the *feature scores* to be considered in partial least squares regression uses…?

(a) The X matrix only

(b) The Y matrix only

(c) Both the X and Y matrices

OLS vs PCR vs PLS

Suppose I have a data set with

$$p = 400 \text{ features}, n = 100 \text{ observations}$$

If I want to learn a linear model, then what should I consider when using…

(a) Ordinary least squares regression

(b) Ridge regression

(c) Principal component regression (PCR)

(d) Partial least squares regression (PLS)

Canonical Correlation Analysis

Find the projection directions of maximum *correlation* for **X** and **Y**.

In PLS we compute (**X** and **Y** are mean-centered):

$$w_i^*, v_i^* = \underset{|w_i|=1,\ |v_i|=1}{argmax} (Xw_i)^T (Yv_i)$$

In *canonical correlation analysis (CCA)*, we compute:

$$w_i^*, v_i^* = \underset{|Xw_i|=1,\ |Yv_i|=1}{argmax} (Xw_i)^T (Yv_i)$$

Canonical Correlation Analysis

$$w_i^*, v_i^* = \underset{|Xw_i|=1,\, |Yv_i|=1}{argmax} (Xw_i)^T (Yv_i)$$

This is equivalent to finding

$$w^*, v^* = \underset{w,v \in \mathbb{R}^n}{argmax} \frac{w^T X^T Y v}{(w^T X^T X w)^{1/2} (v^T Y^T Y v)^{1/2}}$$

Let $X = UDV^T$. We define: $X^{1/2} = UD^{1/2}V^T$

Then the desired w_i^*, v_i^* are the singular vectors of:

$$(X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2}$$

Canonical Correlation Analysis

$$w_i^*, v_i^* = \underset{|Xw_i|=1,\ |Yv_i|=1}{argmax} (Xw_i)^T (Yv_i)$$

w_i^*, v_i^* are the singular vectors of:

$$(X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2}$$

$\mathbf{w_1}$ and $\mathbf{v_1}$ maximize the *correlation* between **Xw** and **Yv**.

$\mathbf{w_2}$ and $\mathbf{v_2}$ do the same and are orthogonal to (respectively) $\mathbf{w_1}$ and $\mathbf{v_1}$. Etc.

More:

http://www.cs.toronto.edu/~jepson/csc420/notes/introSVD.pdf,
http://www.ofai.at/~roman.rosipal/Papers/eig_book04.pdf

Canonical Correlation Analysis

Uses the singular vectors of: $(X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2}$

Correlation: re-scales the data, no units. Range -1 to 1.

Analog to auto-scaling: if $\mathbf{X^T X}$ is diagonal, then this divides each row of $\mathbf{X^T}$ by the corresponding diagonal element of $(\mathbf{X^T X})^{1/2}$.

In the general case where $\mathbf{X^T X}$ is not diagonal: this normalizes $\mathbf{X^T}$ by "removing" covariance.

"Whitens" the data.

PCA vs. CCA vs. PLS

Table 1. Cost functions optimized by the different methods

PCA	Maximize variance	$\frac{\mathbf{w}'\mathbf{S_{XX}}\mathbf{w}}{\mathbf{w}'\mathbf{w}}$
		$\mathbf{w}'\mathbf{S_{XX}}\mathbf{w}$ s.t. $\|\mathbf{w}\|^2 = 1$
	Minimize residuals	$\|(\mathbf{I} - \mathbf{w}\mathbf{w}')\mathbf{X}\|_F^2$
CCA	Maximize correlation	$\frac{\mathbf{w}_\mathbf{X}'\mathbf{S_{XY}}\mathbf{w_Y}}{\sqrt{\mathbf{w}_\mathbf{X}'\mathbf{S_{XX}}\mathbf{w_X}}\sqrt{\mathbf{w}_\mathbf{Y}'\mathbf{S_{YY}}\mathbf{w_Y}}}$
	Maximize fit	$\mathbf{w}_\mathbf{X}'\mathbf{S_{XY}}\mathbf{w_Y}$ s.t. $\|\mathbf{X}\mathbf{w_X}\|^2 = \|\mathbf{Y}\mathbf{w_Y}\|^2 = 1$
	Minimize misfit	$\|\mathbf{w}_\mathbf{X}'\mathbf{X} - \mathbf{w}_\mathbf{Y}'\mathbf{Y}\|^2$ s.t. $\|\mathbf{X}\mathbf{w_X}\|^2 = \|\mathbf{Y}\mathbf{w_Y}\|^2 = 1$
PLS	Maximize covariance	$\frac{\mathbf{w}_\mathbf{X}'\mathbf{S_{XY}}\mathbf{w_Y}}{\sqrt{\mathbf{w}_\mathbf{X}'\mathbf{w_X}}\sqrt{\mathbf{w}_\mathbf{Y}'\mathbf{w_Y}}}$
	Maximize fit	$\mathbf{w}_\mathbf{X}'\mathbf{S_{XY}}\mathbf{w_Y}$ s.t. $\|\mathbf{w_X}\|^2 = \|\mathbf{w_Y}\|^2 = 1$
	Minimize misfit	$\|\mathbf{w}_\mathbf{X}'\mathbf{X} - \mathbf{w}_\mathbf{Y}'\mathbf{Y}\|^2$ s.t. $\|\mathbf{w_X}\|^2 = \|\mathbf{w_Y}\|^2 = 1$

Bie et al: http://www.ofai.at/~roman.rosipal/Papers/eig_book04.pdf

PCA, PLS, CCA, MLR

5.3 Relation to other linear subspace methods

Instead of the two eigenvalue equations in 4 we can formulate the problem in one single eigenvalue equation:

$$\mathbf{B}^{-1}\mathbf{A}\hat{\mathbf{w}} = \rho\hat{\mathbf{w}} \tag{11}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{w}} = \begin{pmatrix} \mu_x\hat{\mathbf{w}}_x \\ \mu_y\hat{\mathbf{w}}_y \end{pmatrix}. \tag{12}$$

Solving the eigenproblem in equation 11 with slightly different matrices will give solutions to *principal component analysis* (PCA), *partial least squares (PLS) and* multivariate linear regression (MLR). The matrices are listed in table 1.

	A	B
PCA	$\mathbf{C}_{xx}$	$\mathbf{I}$
PLS	$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$	$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$
CCA	$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$	$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}$
MLR	$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$	$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$

Table 1: The matrices **A** and **B** for PCA, PLS, CCA and MLR.

From: Borga, M. 2001.

https://www.cs.cmu.edu/~tom/10701_sp11/slides/CCA_tutorial.pdf

Recap

OLS find direction of max correlation between x's and y's

PCA finds the directions of maximal covariance of the x's
(find the SVD of X or X'X)

PCR does a PCA for dimensionality reduction and then OLS
(usually with feature selection)

PLS adjusts the PCA directions to a better job of predicting the y's.
Finds the projection directions of X and Y which maximize their *covariance*.
Can be used when many features are correlated.

CCA finds the projection directions of X and Y that maximize their *correlation.*

SVD of the 'whitened' $X^T Y$: $(X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2}$

PCA and CCA are both using SVD to minimize reconstruction error or maximize variance/covariance