PCA, PCR, CCA and friends

Barry Slaff
modified by Lyle Ungar

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

Overview

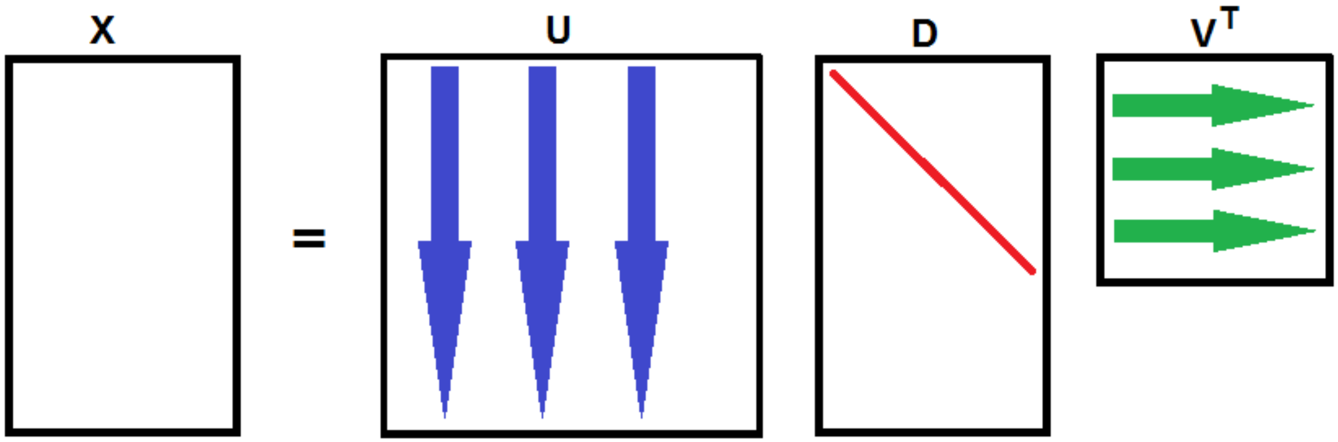
- **Ordinary Least Squares (OLS) Regression**: Finds the projection direction for which the x's are maximally correlated with the y's
- **PCA: Finds projection directions of the x's with maximal covariance**
- **Principal Component Regression (PCR):** Does PCA for dimensionality reduction on X, and then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

Singular Value Decomposition

- **Singular value decomposition of matrix X** (n x p)
  - $X = UDV^T$
- **U: orthogonal, $U^TU=I$** (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, $V^TV=I$** (p x p)
  - Columns of V are the *right singular vectors of X*.

Thin SVD

Singular value decomposition of X: **X = UDV^T**



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

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

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

The columns $\mathbf{v}_1, \ldots \mathbf{v}_p$ of **V** are the *eigenvectors* of the covariance matrix $\mathbf{X}^T\mathbf{X}$. Hence we can write

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

From before:

$$\boldsymbol{X} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} D_{ii} \boldsymbol{u}_i \boldsymbol{v}_i^T$$

k = min(n,p).

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

Principal Component Analysis

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

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

$\mathbf{X_c}$ is (n x p), **Z** is (n x p), **V** is (p x p).

Z is the transformation of **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 **V** is orthogonal, $\mathbf{X_cV} = \mathbf{ZV^TV} = \mathbf{Z}$, and therefore:

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

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

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

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

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

$$= 100 \cdot \frac{(D_{ii})^2}{\sum_{j=1}^{p} (D_{jj})^2} \quad = 100\, \lambda_i / \Sigma_i\, \lambda_i$$

PCA

True or false:

If X is any matrix, and X has singular value decomposition

$$X = UDV^T$$

then the principal component scores for X are the columns of

$$Z = UD.$$

(a) True

(b) False

PCA

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

(a) Eigenvectors of X^TX
(b) Right singular vectors of X
(c) Projection directions of maximum covariance of X
(d) All of the above

PCA: Reconstruction Problem

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

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

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

X_c is still (n x p).

Z is (n x k), with k≤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:

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

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

PCR: Principal Component Regression

PCR has two steps:

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

PCR: Principal Component Regression

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

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**,

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

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

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

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

$$\hat{Y} = Zw^* = Z(Z^T Z)^{-1} Z^T Y = UD(D^T U^T UD)^{-1} D^T U^T Y$$

$$\hat{Y} = UU^T Y = \sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{Y}$$

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: $\hat{Y} = Xw^* = X(X^T X + \gamma I)^{-1} X^T Y$

$$\hat{Y} = UDV^T (V[D^T D + \gamma(V^T V)]V^T)^{-1} V D^T U^T Y$$

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

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

λ_i here are singular values, not eigenvalues

OLS vs. Ridge vs. PCR

OLS: **X = UDV^T** $\hat{Y} = \sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{Y}$

Regularized methods:

PCR: **X$_c$ = UDV^T** $\hat{Y} = \sum_{i=1}^{k} \boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{Y}$, $k \leq n$

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

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

PCR chooses the k "largest" singular vectors.

Ridge Shrinkage

Ridge: **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 **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 $X = UDV^T$ and the nonzero singular values are 5, 4, 3, 2, and 1.

- What are the nonzero eigenvalues of XX^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?

Canonical Covariance Analysis

If Y is high-dimensional, we might want to do dimension reduction for *both* Y and X.

C*anonical covariance analysis* finds the projection directions for both X and Y to maximize their covariance.

or to best reconstruct X from Y and to reconstruct Y for X

(Comparison: PCA finds the projection directions of maximum covariance for X with itself.)

This is one type of Partial Least Squares (PLS), which find projections of x that explain all the y's.

PLS

Finds the projection directions of maximum covariance for **X** and **Y**.

Project $\mathbf{X}_c$ down to **T**. Project $\mathbf{Y}_c$ down to **U**
T and **U** are k-dimensional bases for **X** and **Y**, respectively

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

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

Final model: to predict **Y** from **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**.

***PLS can refer to many similar algorithms.**

Canonical Covariance Analysis (PLS)

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: $(X_c w_i^*)^T \left(X_c w_j^*\right) = 0$ for all **j**<**i**.

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

Canonical Covariance Analysis (PLS)

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 $\mathbf{w_1}=\mathbf{d}$ and $\mathbf{v_1}=\mathbf{e}$, 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

principal component regression uses…?

canonical covariance (PLS 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 features, n = 100 observations

Then use:

a) Ordinary least squares (OLS) 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 $\mathbf{X}^T\mathbf{X}$ is diagonal, then this divides each row of $\mathbf{X}^T$ by the corresponding diagonal element of $(\mathbf{X}^T\mathbf{X})^{1/2}$.

In the general case where $\mathbf{X}^T\mathbf{X}$ 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**: Finds the projection direction for which the x's are maximally correlated with the y's
- **PCA:** Finds projection directions of the x's with maximal covariance
  - SVD of **X'X**
- **Principal Component Regression (PCR):** Do PCA on X, and then OLS using PC features**.**
  - PCR zeros small eigenvectors; Ridge regression shrinks them all
- **Canonical Covariance Analysis:** Finds the projection directions of X and Y that maximize their *covariance*.
  - SVD of Y'X   a form of Partial Least Squares (PLS)
- **Canonical Correlation Analysis (CCA):** Finds the projection directions of X and Y that maximize their *correlation*.
  - SVD of $(X'X)^{-1/2}X'Y(Y'Y)^{-1/2}$
  - The whitening makes it scale invariant

All minimize reconstruction error and maximize variance/covariance