Machine Learning, STOR 565

The Sample Covariance Matrix and PCA

Andrew Nobel

January, 2021

Low-Dimensional Approximation of High-Dimensional Data

General Setting and Goals

Given: Data set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ centered so that $\sum_i \mathbf{x}_i = \mathbf{0}$

Goal: Find a subspace V of $\mathbb{R}^p$ such that

- $\dim(V)$ much less than p and n
- V captures most of the variability in the data points $\mathbf{x}_i$

Fitting criterion: Sum of squared distance between samples and projections

$$\text{Err}(\{\mathbf{x}_i\}, V) \;=\; \sum_{i=1}^{n} \|\mathbf{x}_i \;-\; \text{proj}_V(\mathbf{x}_i)\|^2$$

Overview of PCA

Step 1: Use samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to construct

- Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with rows $\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t$
- Sample covariance matrix $\mathbf{S} = n^{-1}\, \mathbf{X}^t \mathbf{X} \in \mathbb{R}^{p \times p}$

Step 2: Eigenanalysis of $\mathbf{S}$

- Principal component directions are eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ of S ordered by eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$
- $V_k = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ minimizes $\text{Err}(\{\mathbf{x}_i\}, V)$ over all k-dim subspaces
- $\text{Err}(\{\mathbf{x}_i\}, V_k) = \sum_{j=k+1}^{p} \lambda_j$

Data Matrix and Sample Covariance Matrix

Data Matrix

Given: Dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$

- Measurements of p numerical features on each of n samples
- Assume data centered so that $\sum_i \mathbf{x}_i = \mathbf{0}$

Form: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with n rows and p columns

- i'th row $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip}) = \mathbf{x}_i^t$ transpose of the ith sample
- j'th col $\mathbf{x}_{\cdot j} = (x_{1j}, \ldots, x_{nj})$ contains measurements of jth feature

Sample Covariance Matrix

Definition: The *sample covariance matrix* of $\mathbf{X}$ is given by

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^t\,\mathbf{X} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t$$

Note: $\mathbf{S} \in \mathbb{R}^{p \times p}$ and for each $1 \leq j, k \leq p$

$$\mathbf{S}_{j,k} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,x_{ik} \;=\; s(\mathbf{x}_{\cdot j}, \mathbf{x}_{\cdot k})$$

is the sample covariance of *features* j and k

Properties of the Sample Covariance

1. $\mathbf{S}$ is symmetric and non-negative definite
2. $\mathbf{S}$ has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
3. $\sum_{k=1}^{p} \lambda_k = n^{-1} ||\mathbf{X}||^2$
4. $\sum_{k=1}^{p} \lambda_k = \sum_{j=1}^{p} s^2(\mathbf{x}_{\cdot j})$, the aggregate variance of the features
5. $\text{rank}(\mathbf{S}) = \text{rank}(\mathbf{X}^t \mathbf{X}) = \text{rank}(\mathbf{X}) \leq \min(n, p)$
6. If $p > n$ then $\text{rank}(\mathbf{S}) < p$ and $\mathbf{S}$ is not invertible.

Principal Component Analysis (PCA)

One-dimensional case

Best One-Dimensional Subspace

Given: Data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ find 1-dim subspace V to minimize

$$\text{Err}(\{\mathbf{x}_i\}, V) \;=\; \sum_{i=1}^{n} \|\mathbf{x}_i \;-\; \text{proj}_V(\mathbf{x}_i)\|^2$$

- Any 1-dim $V = \{\alpha\, \mathbf{v} : \alpha \in \mathbb{R}\}$ for some $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\| = 1$
- In this case, $\text{proj}_V(\mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{v} \rangle\, \mathbf{v}$
- Easy calculation shows $\text{Err}(\{\mathbf{x}_i\}, V) = \sum_{i=1}^n ||\mathbf{x}_i||^2 - \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2$

Upshot: The following two optimization problems are equivalent

- Minimize $\text{Err}(\{\mathbf{x}_i\}, V)$ over 1-dim subspaces V
- Maximize $n^{-1} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2$ over $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\| = 1$

Best One-Dimensional Subspace

Fact: For each $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\| = 1$

1. $n^{-1} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = s^2(\langle \mathbf{x}_1, \mathbf{v} \rangle, \ldots, \langle \mathbf{x}_n, \mathbf{v} \rangle)$
2. $n^{-1} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = \mathbf{v}^t \mathbf{S} \mathbf{v}$

Solution (at last!)

Fischer-Courant theorem tells us that $\mathbf{v}^t \mathbf{S} \mathbf{v}$ is maximized when $\mathbf{v}$ is an eigenvector of $\mathbf{S}$ with maximum eigenvalue.

Principal Component Analysis (PCA)

General case

Principal Component Analysis

Recall setting

- Data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ with $\sum_i \mathbf{x}_i = \mathbf{0}$
- Data matrix $\mathbf{X}$ ($n \times p$) with rows $\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t$
- $\mathbf{S} = n^{-1} \mathbf{X}^t \mathbf{X}$ sample covariance of $\mathbf{X}$

Definition: Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be eigenvalues of $\mathbf{S}$, with corresponding *orthonormal* eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$

- $\mathbf{v}_j$ called the j'th *principal component direction* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- projection $\langle \mathbf{x}_i, \mathbf{v}_j \rangle \, \mathbf{v}_j$ is called the jth *principal component* of $\mathbf{x}_i$

Higher Order Principal Components

Definition: For $1 \leq k \leq p$ let $V_k = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ = span of k leading eigenvectors of $\mathbf{S}$. Easy to show that

$$\text{proj}_{V_k}(\mathbf{x}) \; = \; \sum_{j=1}^{k} \langle \mathbf{x}, \mathbf{v}_j \rangle \, \mathbf{v}_j$$

Fact: The subspace V_k minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i \; - \; \text{proj}_V(\mathbf{x}_i)\|^2$$

over k-dimensional subspaces V of $\mathbb{R}^p$. Moreover

$$\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i \; - \; \text{proj}_{V_k}(\mathbf{x}_i)\|^2 \; = \; \sum_{i=k+1}^{p} \lambda_i$$

Proportion of Variation Explained

Definition: The *proportion of variation explained* by the first k principal components, equivalently the subspace V_k, is given by

$$\gamma_k \;=\; \frac{\sum_{i=1}^n \|\text{proj}_{V_k}(\mathbf{x}_i)\|^2}{\sum_{i=1}^n \|\mathbf{x}_i\|^2} \;=\; \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}$$

In practice γ_k can be close to 1 for values of k as small as 4 or 5, meaning that first few PCs capture most of the variation in the data.

Variable Selection vs. Dimension Reduction

Variable selection methods remove selected features from consideration in downstream analyses. Underlying coordinates unchanged

$$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^t \quad \mapsto \quad \tilde{\mathbf{x}} = (x_2, x_5)^t$$

Dimension reduction methods like PCA replace observed features by smaller number of derived features, for example

$$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^t \ \mapsto \ \tilde{\mathbf{x}} = \langle \mathbf{x}, \mathbf{v}_1 \rangle \, \mathbf{v}_1 + \langle \mathbf{x}, \mathbf{v}_2 \rangle \, \mathbf{v}_2$$

- The derived features $\mathbf{v}_1, \mathbf{v}_2$ constitute new coordinate system
- Each derived feature may depend on all the observed features

Principal Component Analysis (PCA)

Examples

Women's Heptathlon Scores

Background: Seven-event competition over two days. Data from 25 athletes competing in the 1988 Olympics, in Seoul [1]

- Scores for each event
- Overall score

Questions

- What is a good way of combining individual scores to get overall score?
- If we use a linear combination, should each event be weighed the same?

Idea: Consider principle components

[1] From Everitt & Hothorn (2011). An introduction to applied multivariate analysis with R

1988 Women's Heptathlon Scores ($n = 25$, $p = 7$)

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

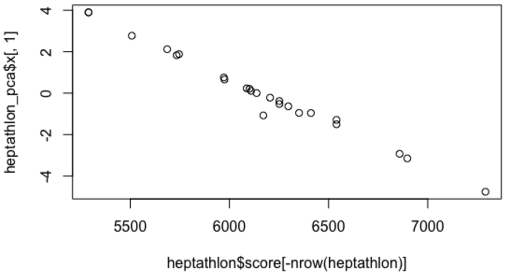
Principal Component Analysis

- Standardize the scores from each event, so each column of data matrix has mean 0 and variance 1
- Apply PCA to the resulting data matrix

```
R > heptathlon_pca <- prcomp(heptathlon[, -c("score")], scale = TRUE)
R > summary(heptathlon_pca)

                    PC1    PC2    PC3    PC4     PC5     PC6     PC7

St. Dev.         2.0793 0.9482 0.9109 0.68320 0.54619 0.33745 0.26204

Prop. of Var.    0.6177 0.1284 0.1185 0.06668 0.04262 0.01627 0.00981

Cum. Prop.       0.6177 0.7461 0.8646 0.93131 0.97392 0.99019 1.00000
```

Principal Component Analysis, cont.

- Approximately 75% of the variation is explained by the first two PCs.
- The overall score is *highly* correlated ($r = -.993$) with the first PC



Loadings of First Principal Components

Event	PC1	PC2	PC3
hurdles	0.4504	-0.0577	-0.1739
highjump	-0.3145	-0.6513	0.2088
shot	-0.4025	-0.0220	0.1535
run200m	0.4271	-0.1850	0.1301
longjump	-0.4510	-0.0249	0.2698
javelin	-0.2423	-0.3257	-0.8807
run800m	0.3029	-0.6565	0.1930

Note: Signs of loadings in PC1 coincide with ordering of scores

- Events where higher scores are better have negative coefficients
- Events where lower scores are better have positive coefficients

Text Analysis: The Federalist Papers

Federalist Papers

- 85 documents in all
- released between 1787 and 1788
- promoting the U.S. Constitution
- written by John Jay, James Madison, and Alexander Hamilton

Authorship

- authorship of 70 papers known
- 3 are collaborative
- authorship of remaining 12 disputed

From Documents to Data

Samples: Text of each document $n = 70$

- Ordered sequence of words

Variables: Counts of $p = 70$ function words

- Function words: common words used without much deliberation
- Examples: a, to, and, more, upon

Preprocessing: Standardize columns

- Center word counts to have mean zero
- Scale word counts to have variance one

PCA on Federalist Paper Data

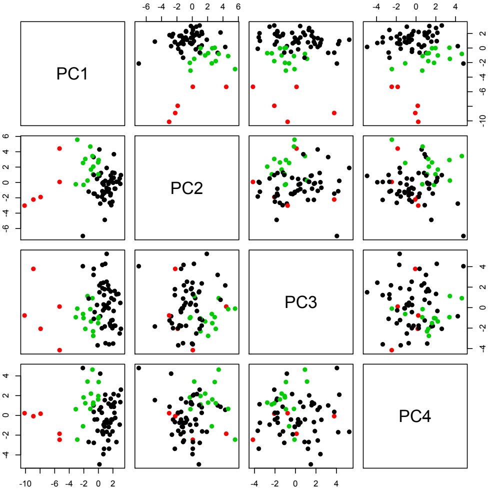


Figure: Projections of normalized word count data onto the first four principal components of the Federalist dataset. Colors represent known authorship: Madison = green, Jay = red, Hamilton = black

First PC Loadings: 8 words with largest +/- coefficients

"in",0.151791749764273
"there",0.157087053819256
"the",0.195915748087175
"a",0.198175928753355
"an",0.198737890289868
"this",0.233402747982087
"upon",0.241427130517209
"of",0.253236889522879

"and",-0.296914872485316
"one",-0.231054740057054
"more",-0.219232323121311
"their",-0.209819034770272
"also",-0.18953520090149
"into",-0.164657827937641
"than",-0.129280268238455
"our",-0.125302378571939