

"Introduction to Principal Components Analysis"
Francis and Wills (1999)

Krystian Hkiewicz

05.05.2022

Principal Components Analysis (PCA) - tool for simplifying datasets

Principal Components Analysis (PCA) - tool for simplifying datasets

= dimensionality-reduction method

Principal Components Analysis (PCA) - tool for simplifying datasets

= dimensionality-reduction method

= reduction of complexity

Why we want less dimensions?

Why we want less dimensions?

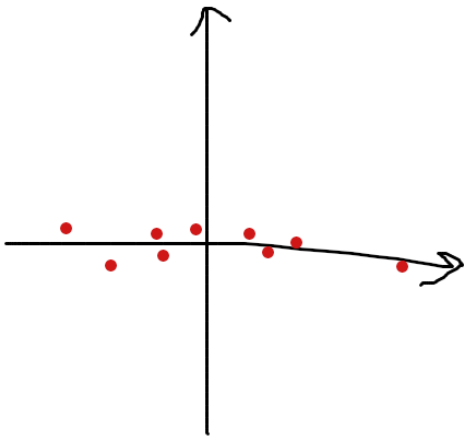
- To look for correlations (e.g. IQ)

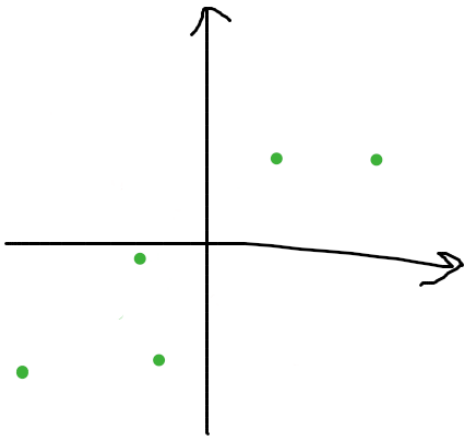
Why we want less dimensions?

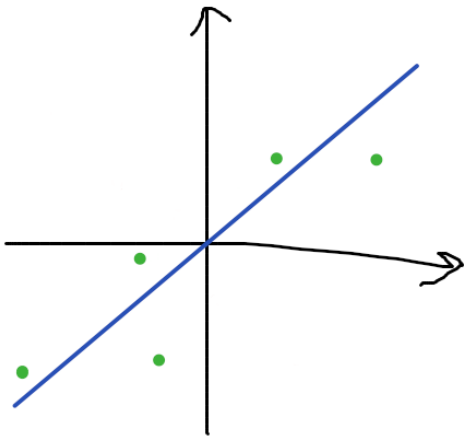
- To look for correlations (e.g. IQ)
- To visualize data

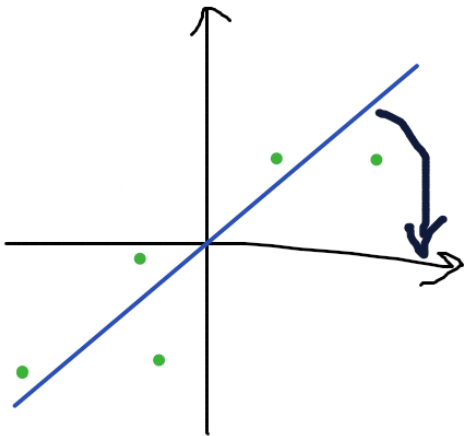
Why we want less dimensions?

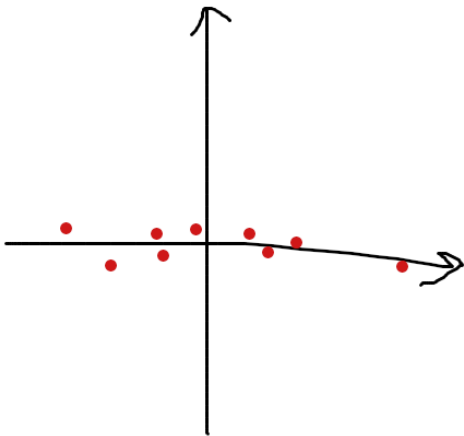
- To look for correlations (e.g. IQ)
- To visualize data
- For easier processing of the data

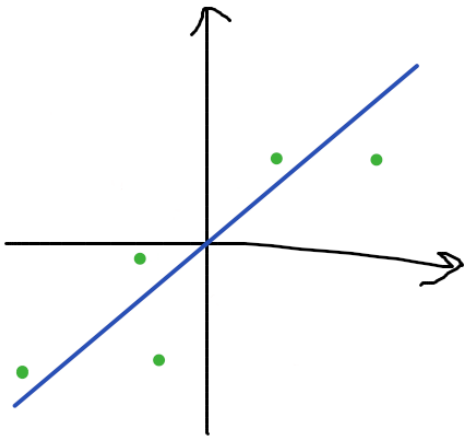


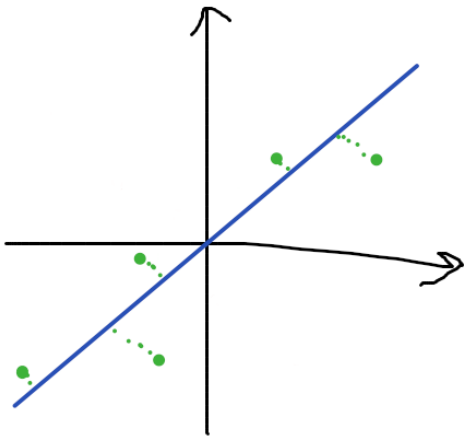


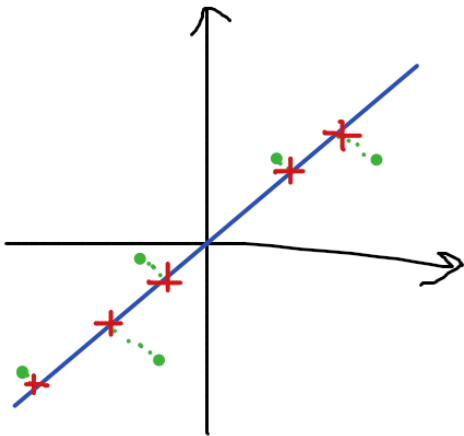


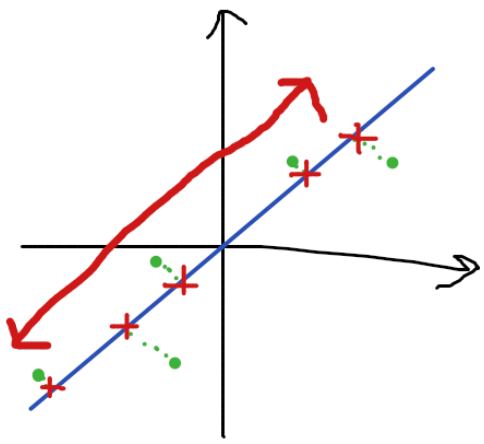










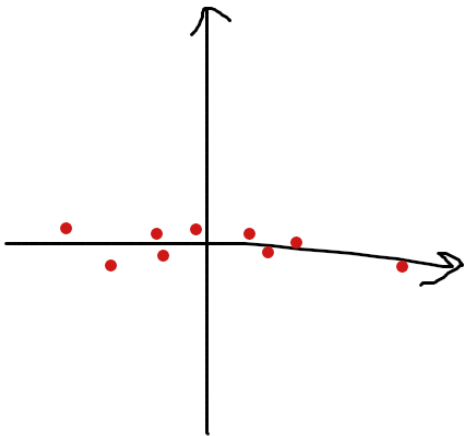


Principal Components Analysis

Click here for a GIF

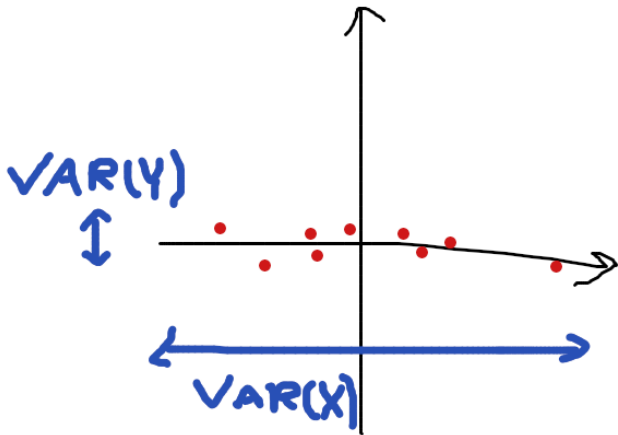
Source: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

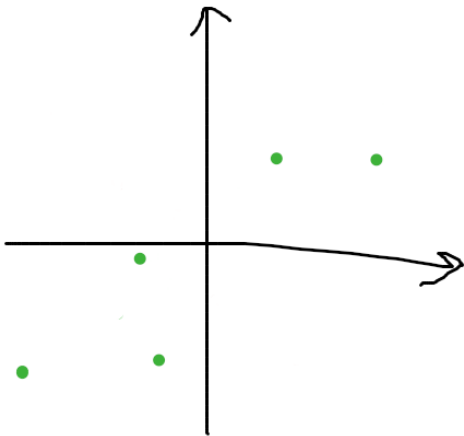
How to calculate the rotation?

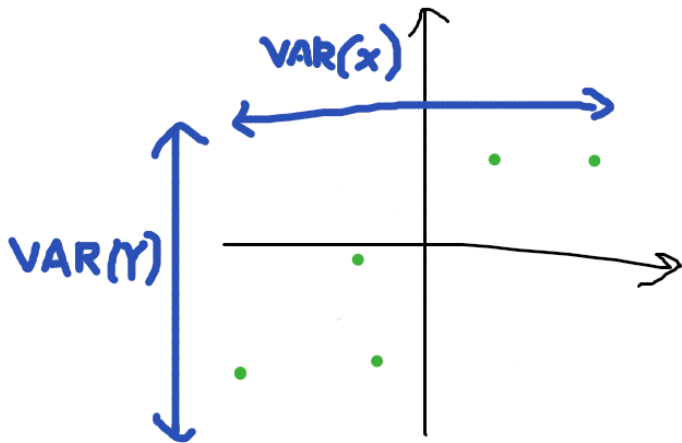


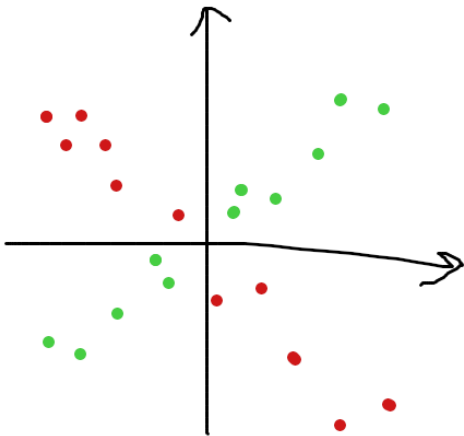
We can choose which axis to discard by measuring variance:

$$\text{var}(\mathbf{x}) \equiv \sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$$









Covariance matrix:

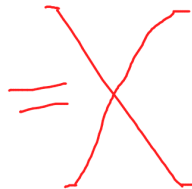
$$C_X = \frac{1}{N-1} X^T X$$



PG Name	$\log L_{1216}^a$	α_x	$\log \text{FWHM}$ $\text{H}\beta$	$\text{FeII}/$ $\text{H}\beta$	$\log \text{EW}$ [OIII]	$\log \text{FWHM}$ [CIII]
0947+396	45.66	1.51	3.684	0.23	1.18	3.520
0953+414	45.83	1.57	3.496	0.25	1.26	3.432
1001+054	44.93	...	3.241	0.82	0.85	3.424
1114+445	44.99	0.88	3.660	0.20	1.23	3.654
1115+407	45.41	1.89	3.236	0.54	0.78	3.403
1116+215	46.00	1.73	3.465	0.47	1.00	3.446
1202+281	44.77	1.22	3.703	0.29	1.56	3.434
1216+069	46.03	1.36	3.715	0.20	1.00	3.514
1226+023	46.74	0.94	3.547	0.57	0.70	3.477
1309+355	45.55	1.51	3.468	0.28	1.28	3.406
1322+659	45.42	1.69	3.446	0.59	0.90	3.351
1352+183	45.34	1.52	3.556	0.46	1.00	3.548
1402+261	45.74	1.93	3.281	1.23	0.30	3.229
1411+442	44.93	1.97	3.427	0.49	1.18	3.275
1415+451	45.08	1.74	3.418	1.25	0.30	3.434
1425+267	45.72	0.94	3.974	0.11	1.56	3.666
1427+480	45.54	1.41	3.405	0.36	1.76	3.300
1440+356	45.23	2.08	3.161	1.19	1.00	3.192
1444+407	45.92	1.91	3.394	1.45	0.30	3.479
1512+370	46.04	1.21	3.833	0.16	1.76	3.546
1543+489	46.02	2.11	3.193	0.85	0.00	...
1626+554	45.48	1.94	3.652	0.32	0.95	3.631
Number	22	21	22	22	22	21
Mean	45.56	1.57	3.498	0.56	0.99	3.446
Std dev'n	0.47	0.38	0.212	0.40	0.47	0.129

Data matrix

PG Name	$\log L_{1216}^a$	α_z	$\log \text{FWHM H}\beta$	$\text{FeII}/\text{H}\beta$	$\log \text{EW [OIII]}$	$\log \text{FWHM CIII]$
0947+396	45.66	1.51	3.684	0.23	1.18	3.520
0953+414	45.83	1.57	3.496	0.25	1.26	3.432
1001+054	44.93	...	3.241	0.82	0.85	3.424
1114+445	44.99	0.88	3.660	0.20	1.23	3.654
1115+407	45.41	1.89	3.236	0.54	0.78	3.403
1116+215	46.00	1.73	3.465	0.47	1.00	3.446
1202+281	44.77	1.22	3.703	0.29	1.56	3.434
1216+069	46.03	1.36	3.715	0.20	1.00	3.514
1226+023	46.74	0.94	3.547	0.57	0.70	3.477
1309+355	45.55	1.51	3.468	0.28	1.28	3.406
1322+659	45.42	1.69	3.446	0.59	0.90	3.351
1352+183	45.34	1.52	3.556	0.46	1.00	3.548
1402+261	45.74	1.93	3.281	1.23	0.30	3.229
1411+442	44.93	1.97	3.427	0.49	1.18	3.275
1415+451	45.08	1.74	3.418	1.25	0.30	3.434
1425+267	45.72	0.94	3.974	0.11	1.56	3.666
1427+480	45.54	1.41	3.405	0.36	1.76	3.300
1440+356	45.23	2.08	3.161	1.19	1.00	3.192
1444+407	45.92	1.91	3.394	1.45	0.30	3.479
1512+370	46.04	1.21	3.833	0.16	1.76	3.546
1543+489	46.02	2.11	3.193	0.85	0.00	...
1626+554	45.48	1.94	3.652	0.32	0.95	3.631
Number	22	21	22	22	22	21
Mean	45.56	1.57	3.498	0.56	0.99	3.446
Std dev'n	0.47	0.38	0.212	0.40	0.47	0.129



Francis and Willis (1999)



$$C_X = \frac{1}{N-1} X^T X$$

$\hat{\curvearrowright} K \times K$ MATRIX

$$C_X = \frac{1}{N-1}$$

$$X^T X$$

$\hat{\curvearrowright} K \times K$ MATRIX

Covariance is similar to correlation

We treat C_X as linear transformation and find eigenvalues and eigenvectors

We treat C_X as linear transformation and find eigenvalues and eigenvectors

Eigenvectors in PCA are **always** orthogonal

We treat C_X as linear transformation and find eigenvalues and eigenvectors

Eigenvectors in PCA are **always** orthogonal

Eigenvectors \rightarrow principal components

$$W = \begin{bmatrix} | & | & | & \dots \\ \text{PC1} & \text{PC2} & \text{PC3} & \dots \\ | & | & | & \dots \end{bmatrix}$$

← "LOADINGS"
← K x K MATRIX
↑ PRINCIPAL COMPONENTS

$$W = \begin{bmatrix} | & | & | & \dots \\ \text{PC1} & \text{PC2} & \text{PC3} & \dots \\ | & | & | & \dots \end{bmatrix}$$

← "LOADINGS"
← K x K MATRIX
↑ PRINCIPAL COMPONENTS

Principal components are sorted by their eigenvalues

$W = \begin{bmatrix} | & | & | & \dots \\ \text{PC1} & \text{PC2} & \text{PC3} & \dots \\ | & | & | & \dots \end{bmatrix}$

← "LOADINGS"

← $K \times K$ MATRIX

↑ PRINCIPAL COMPONENTS

Principal components are sorted by their eigenvalues

Eigenvalues \sim fraction of variance explained

ORIGINAL DATA

$$T = XW$$

LOADINGS

SCORES

(OUR DATA AFTER
THE TRANSFORMATION)

$$T_{N \times K} = X_{N \times K} W_{K \times K}$$

$$W_p = \begin{bmatrix} | & | & & | \\ PC_1 & PC_2 & \dots & PC_m \\ | & | & & | \end{bmatrix}$$

$\mathbb{K}^{n \times p}$ MATRIX

$$T_r$$

$$N \times r$$

$$=$$

$$X$$

$$N \times k$$

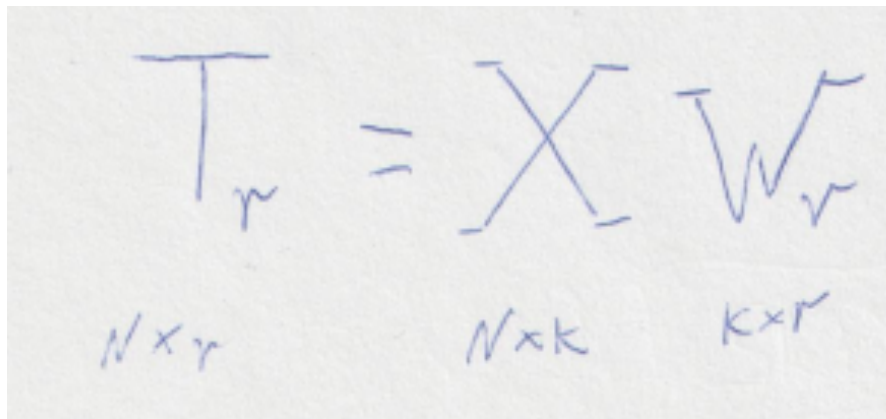
$$W_r$$

$$k \times r$$

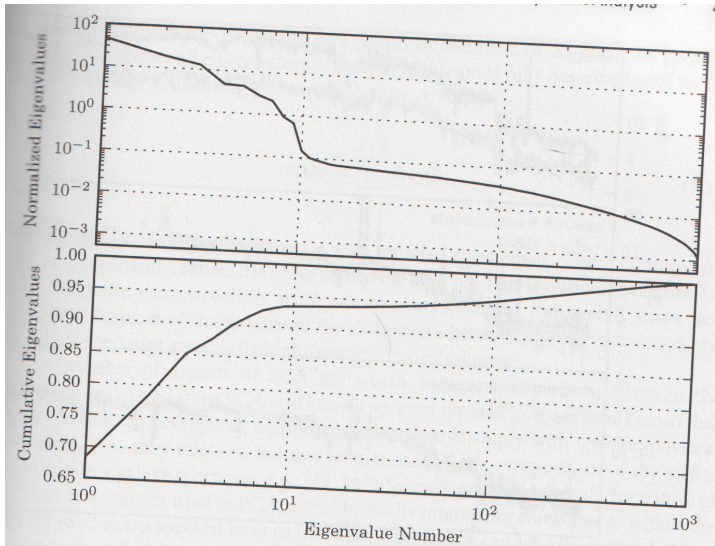
PCA is a **linear transformation** where the new axes are in direction of with the maximum variance

PCA is a hierarchical coordinate system

Choosing the level of truncation (r)

$$\begin{matrix} T_r \\ N \times r \end{matrix} = \begin{matrix} X \\ N \times k \end{matrix} \begin{matrix} W_r \\ k \times r \end{matrix}$$


The scree plot



Ivezić et al. "Statistics, Data Mining, and Machine Learning in Astronomy"

Choosing the level of truncation

- Finding a "knee" in the scree plot (Cattell 1966)

Choosing the level of truncation

- Finding a "knee" in the scree plot (Cattell 1966)
- Percentage of the explained variance (75%-95%)

Choosing the level of truncation

- Finding a "knee" in the scree plot (Cattell 1966)
- Percentage of the explained variance (75%-95%)
- Eigenvalues higher than one (Gutteman-Kaiser criterion)

M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 514

Automatic choice of dimensionality for PCA

Thomas P. Minka

MIT Media Laboratory, Vision and Modeling Group

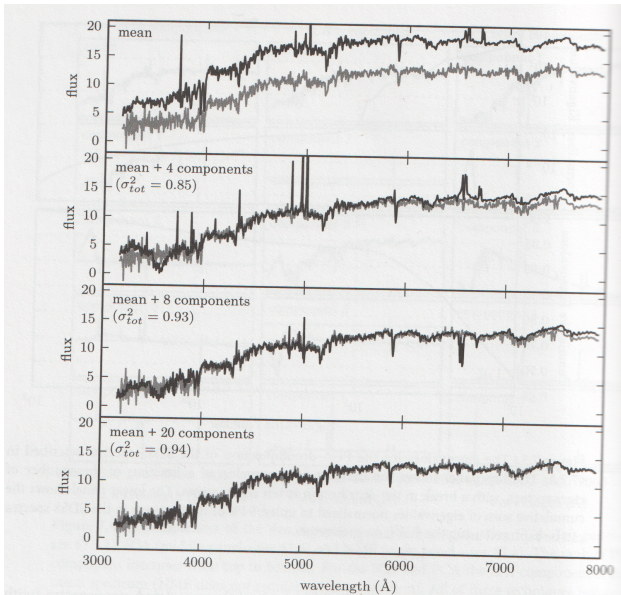
20 Ames Street; Cambridge, MA 02139

tpminka@media.mit.edu

December 29, 2000 (revised September 2, 2008)

Abstract

A central issue in principal component analysis (PCA) is choosing the number of principal components to be retained. By interpreting PCA as density estimation, this paper shows how to use Bayesian model selection to determine the true dimensionality of the data. The resulting estimate is simple to compute yet guaranteed to pick the correct dimensionality, given enough data. The estimate involves an integral over the Steifel manifold of k -frames, which is difficult to compute exactly. But after choosing an appropriate parameterization and applying Laplace's method, an accurate and practical estimator is obtained. In simulations, it is more accurate than cross-validation and other proposed algorithms, plus it runs much faster.



Ivezić et al. "Statistics, Data Mining, and Machine Learning in Astronomy"

Scaling of the data

First we need to subtract the mean

Scaling of the data

First we need to subtract the mean

Then divide the data by variance

Scaling of the data

First we need to subtract the mean

Then divide the data by variance

Alternatively, we can normalize the initial data (e.g. spectra)

PG Name	$\log L_{1216}^a$	α_z	$\log \text{FWHM } H\beta$	$\text{FeII}/H\beta$	$\log \text{EW } [\text{OIII}]$	$\log \text{FWHM } \text{CIII}]$
0947+396	45.66	1.51	3.684	0.23	1.18	3.520
0953+414	45.83	1.57	3.496	0.25	1.26	3.432
1001+054	44.93	...	3.241	0.82	0.85	3.424
1114+445	44.99	0.88	3.660	0.20	1.23	3.654
1115+407	45.41	1.89	3.236	0.54	0.78	3.403
1116+215	46.00	1.73	3.465	0.47	1.00	3.446
1202+281	44.77	1.22	3.703	0.29	1.56	3.434
1216+069	46.03	1.36	3.715	0.20	1.00	3.514
1226+023	46.74	0.94	3.547	0.57	0.70	3.477
1309+355	45.55	1.51	3.468	0.28	1.28	3.406
1322+659	45.42	1.69	3.446	0.59	0.90	3.351
1352+183	45.34	1.52	3.556	0.46	1.00	3.548
1402+261	45.74	1.93	3.281	1.23	0.30	3.229
1411+442	44.93	1.97	3.427	0.49	1.18	3.275
1415+451	45.08	1.74	3.418	1.25	0.30	3.434
1425+267	45.72	0.94	3.974	0.11	1.56	3.666
1427+480	45.54	1.41	3.405	0.36	1.76	3.300
1440+356	45.23	2.08	3.161	1.19	1.00	3.192
1444+407	45.92	1.91	3.394	1.45	0.30	3.479
1512+370	46.04	1.21	3.833	0.16	1.76	3.546
1543+489	46.02	2.11	3.193	0.85	0.00	...
1626+554	45.48	1.94	3.652	0.32	0.95	3.631
Number	22	21	22	22	22	21
Mean	45.56	1.57	3.498	0.56	0.99	3.446
Std dev'n	0.47	0.38	0.212	0.40	0.47	0.129

Francis and Wills (1999)

PG Name	log EW Ly α	log EW CIV	CIV/ Ly α	log EW CIII]	SiIII/ CIII]	NV/ Ly α	λ 1400/ Ly α
0947+396	2.08	1.78	0.45	1.24	0.306	0.179	0.143
0953+414	2.19	1.78	0.40	1.24	0.164	0.189	0.093
1001+054	2.25	1.76	0.40	1.43	0.443	0.462	0.174
1114+445	2.27	1.85	0.42	1.48	0.222	0.175	0.092
1115+407	1.90	1.51	0.33	1.14	0.385	0.228	0.134
1116+215	2.14	1.71	0.34	1.20	0.440	0.254	0.126
1202+281	2.72	2.41	0.69	1.87	0.164	0.154	0.098
1216+069	2.12	1.95	0.54	1.20	0.037	0.121	0.056
1226+023	1.64	1.44	0.45	1.00	0.280	0.174	0.018
1309+355	2.01	1.68	0.41	1.15	0.303	0.131	0.064
1322+659	2.19	1.85	0.41	1.30	0.291	0.135	0.097
1352+183	2.14	1.80	0.41	1.29	0.357	0.203	0.116
1402+261	1.91	1.59	0.39	1.09	0.568	0.227	0.161
1411+442	...	1.88	...	1.42	0.314	...	0.093
1415+451	2.32	1.78	0.29	1.40	0.688	0.210	0.142
1425+267	...	2.17	...	1.43	0.398	...	0.055
1427+480	2.03	1.82	0.49	1.21	0.265	0.126	0.117
1440+356	2.14	1.54	0.21	1.05	0.747	0.141	0.092
1444+407	1.99	1.34	0.21	1.06	0.809	0.335	0.164
1512+370	2.02	2.05	0.75	1.28	0.228	0.182	0.050
1543+489	1.93	1.60	0.44	0.398	0.174
1626+554	2.14	1.80	0.39	1.36	0.197	0.217	0.118
Number	20	22	20	21	21	20	22
Mean	2.11	1.78	0.421	1.279	0.362	0.212	0.108
Std dev'n	0.21	0.24	0.131	0.194	0.199	0.091	0.043

Francis and Wills (1999)

Table 3. Results of Eigenanalysis – The Principal Components^a

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	6.4505	2.8157	1.5879	0.6257	0.5698
Proportion	0.496	0.217	0.122	0.048	0.044
Cumulative	0.496	0.713	0.835	0.883	0.927
Variable	PC1	PC2	PC3	PC4	PC5
$\log L_{1216}$	0.053	0.535	-0.123	-0.029	-0.405
α_x	0.295	-0.198	0.079	0.485	-0.155
FWHM $H\beta$	-0.330	0.077	-0.357	-0.082	-0.141
Fe II/ $H\beta$	0.341	-0.140	0.003	-0.487	-0.212
$\log EW [O III]$	-0.310	0.016	0.255	0.394	-0.095
$\log FWHM C III]$	-0.198	0.077	-0.623	0.054	0.402
$\log EW Ly\alpha$	-0.177	-0.502	-0.006	-0.143	0.033
$\log EW CIV$	-0.336	-0.262	0.048	-0.050	-0.303
$CIV/Ly\alpha$	-0.342	0.062	0.025	-0.074	-0.584
$\log EW C III]$	-0.262	-0.413	-0.124	-0.176	-0.008
$Si III]/C III]$	0.342	-0.149	-0.018	-0.311	-0.116
$N V/Ly\alpha$	0.231	-0.050	-0.573	0.107	-0.288
$\lambda 1400/Ly\alpha$	0.223	-0.351	-0.225	0.441	-0.216

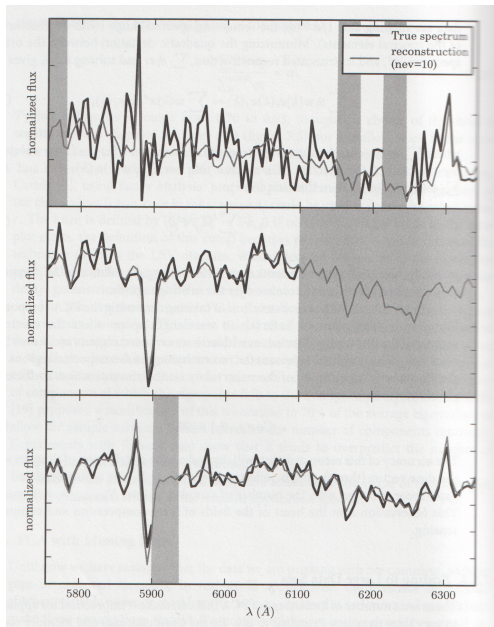
Francis and Wills (1999)

Significance of variables can be checked by performing the PCA without these variables

PCA with missing data

Solved by weighted calculation of PCA (Everson & Sirovich 1995)

gap \rightarrow weight=0



Ivezić et al. "Statistics, Data Mining, and Machine Learning in Astronomy"

A principal component analysis of quasar UV spectra

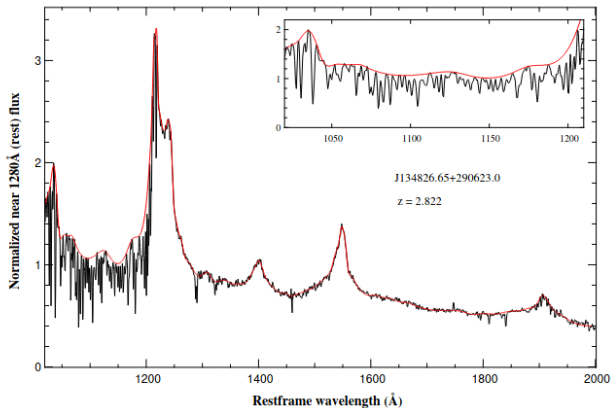


Fig. 3. Spectrum and continuum of the quasar SDSS J134826.65+290623.0. This quasar belongs to the sample of SDSS $z \sim 3$ quasars that is used to derive the Principal Component Analysis eigenvectors (Section 3). Our estimate of the continuum is shown with the thick red line and a zoom in the Lyman- α forest region is shown in the inset.

- Large amount of memory required for calculation

- Large amount of memory required for calculation -> Incremental PCA

Problems with PCA

- Large amount of memory required for calculation -> Incremental PCA
- Problems with data interpretation

- Large amount of memory required for calculation -> Incremental PCA
- Problems with data interpretation -> simulations

Light curve analysis of variable stars using Fourier decomposition and principal component analysis

S. Deb¹ and H. P. Singh^{1,2}



Received: 8 July 2009 | Accepted: 1 September 2009

Abstract

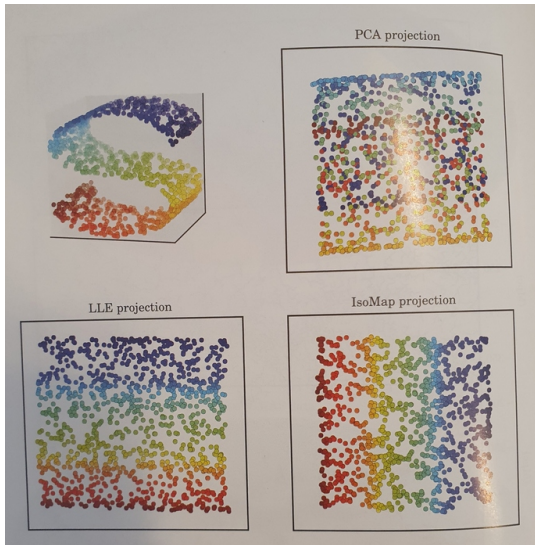
Context. Ongoing and future surveys of variable stars will require new techniques to analyse their light curves as well as to tag objects according to their variability class in an automated way.

Aims. We show the use of principal component analysis (PCA) and Fourier decomposition (FD) method as tools for variable star light curve analysis and compare their relative performance in studying the changes in the light curve structures of pulsating Cepheids and in the classification of variable stars.

Methods. We have calculated the Fourier parameters of 17 606 light curves of a variety of variables, e.g., RR Lyraes, Cepheids, Mira Variables and extrinsic variables for our analysis. We have also performed PCA on the same database of light curves. The inputs to the PCA are the 100 values of the magnitudes for each of these 17 606 light curves in the database interpolated between phase 0 to 1. Unlike some previous studies, Fourier coefficients are not used as input to the PCA.

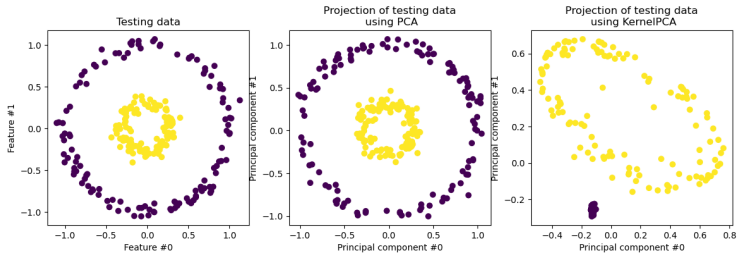
Results. We show that in general, the first few principal components (PCs) are enough to reconstruct the original light curves compared to the FD method where 2 to 3 times more parameters are required to satisfactorily reconstruct the light curves. The computation of the required number of Fourier parameters on average needs 20 times more CPU time than the computation of the required number of PCs. Therefore, PCA does have some advantages over the FD method in analysing the variable stars in a larger database. However, in some cases, particularly in finding the resonances in fundamental mode (FU) Cepheids, the PCA results show no distinct advantages over the FD method. We also demonstrate that the PCA technique can be used to classify variables into different variability classes in an automated, unsupervised way, a feature that has immense potential for larger databases in the future.

Non-linear PCA



Ivezić et al. "Statistics, Data Mining, and Machine Learning in Astronomy"

Non-linear PCA



Source: scikit-learn manual examples

Principal component analysis tomography in near-infrared integral field spectroscopy of young stellar objects. I. Revisiting the high-mass protostar W33A

F. Navarete^{1*}, A. Daminieli¹, J. E. Steiner^{1†}, R. D. Blum²

¹ Universidade de São Paulo, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Rua do Matão 1226, Cidade Universitária São Paulo-SP, 05508-090, Brasil

² NSF's Optical-Infrared Astronomy Research Laboratory P.O. Box 26732, Tucson, AZ 85719, USA

Accepted 2021 February 4. Received 2021 February 4; in original form 2020 November 29

ABSTRACT

W33A is a well-known example of a high-mass young stellar object showing evidence of a circumstellar disc. We revisited the *K*-band NIFS/Gemini North observations of the W33A protostar using principal components analysis tomography and additional post-processing routines. Our results indicate the presence of a compact rotating disc based on the kinematics of the CO absorption features. The position-velocity diagram shows that the disc exhibits a rotation curve with velocities that rapidly decrease for radii larger than $0''.1$ (~ 250 AU) from the central source, suggesting a structure about four times more compact than previously reported. We derived a dynamical mass of $10.0^{+4.1}_{-2.2} M_{\odot}$ for the “disc+protostar” system, about $\sim 33\%$ smaller than previously reported, but still compatible with high-mass protostar status. A relatively compact H_2 wind was identified at the base of the large-scale outflow of W33A, with a mean visual extinction of ~ 63 mag. By taking advantage of supplementary near-infrared maps, we identified at least two other point-like objects driving extended structures in the vicinity of W33A, suggesting that multiple active protostars are located within the cloud. The closest object (Source B) was also identified in the NIFS field of view as a faint point-like object at a projected distance of $\sim 7,000$ AU from W33A, powering extended *K*-band continuum emission detected in the same field. Another source (Source C) is driving a bipolar H_2 jet aligned perpendicular to the rotation axis of W33A.

0
0
0
0

1
1
1
3

2
2
2
5

3
3
3
5

4
4
4
6

5
5
5
5

6
6
6
0

7
7
7
5

8
8
8
8

9
9
9
5

VanderPlas "Python Data Science Handbook"



VanderPlas "Python Data Science Handbook"

0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9

VanderPlas "Python Data Science Handbook"

The Impact of Measurement Error on Principal Component Analysis

KRISTOFFER HERLAND HELLTON and MAGNE THORESEN

Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo

ABSTRACT. We investigate the effect of measurement error on principal component analysis in the high-dimensional setting. The effects of random, additive errors are characterized by the expectation and variance of the changes in the eigenvalues and eigenvectors. **The results show that the impact of uncorrelated measurement error on the principal component scores is mainly in terms of increased variability and not bias.** In practice, the error-induced increase in variability is small compared with the original variability for the components corresponding to the largest eigenvalues. This suggests that the impact will be negligible when these component scores are used in classification and regression or for visualizing data. **However, the measurement error will contribute to a large variability in component loadings, relative to the loading values, such that interpretation based on the loadings can be difficult.** The results are illustrated by simulating additive Gaussian measurement error in microarray expression data from cancer tumours and control tissues.