Exploratory data analysis

Up to now, we've dealt with simple statistical problems.

Primary goals were to
- **detect and quantify a single analyte.**
- **develop relationships between an analyte and a response.**
- **optimize an experiment design and methods used to measure a response.**
- **confirmatory data analysis.**

Confirmatory data analysis

When we obtain a set of samples and make one type of measurement.

Many analytical methods are developed to quantify a single analyte or a limited number of analytes.

All other factors are held constant or eliminated.

Exploratory data analysis

- When we obtain many measurements from a number of samples and attempt to learn something about our sample beyond simple numbers.
- 'Real world' problems are typically much more complex. A true understanding of a system may only be possible if many factors are considered.

Complex samples

Complex sample consists of many components.

- Each may contribute to the overall properties of the sample.
- A measurement of any single component or property is unlikely to tell you much about what the sample is.
- Any type of sample can be either simple or complex based on the type of information desired regarding the sample.

Complex samples

Examples

Gasoline

Its overall performance as a fuel is not based on the amount of any single component.

Coffee

This material contains hundreds of components. The flavor can't be attributed to any single component.

Complex samples

With current analytical tools, its possible to detect and quantify most materials in a complex sample.

Knowing that information, its still impossible to state what the original sample was or be able to precisely reproduce it.

Example - perfume reproductions.

When more is better

Exploratory data analysis attempts to detect and evaluate underlying trends a data set.

This is accomplished by collecting as much information about a problem as possible and multivariate data analysis.

The introduction of the personal computer made it possible for routine evaluation of complex data sets (many variables and samples.)

When more is better

Example.

Assume you are doing QA/QC for a fertilizer company.

You are provided with representative samples at 30 minute intervals. If there is a problem, you must stop production. If you are wrong - you are fired!

Let's see what happens to you level of knowledge as we increase the amount of data.

When more is better

Time	% N
7:00 am	15.1
7:30	14.9
8:00	14.6
8:30	14.8
9:00	1.4

The 9:00 value appears low.

What should you do?

When more is better

A simple statistical calculation for the first four samples shows:

mean = 14.9, s_x = 0.21

Your 9:00 sample is -9.6 s.

So you know that the value is significantly lower (different) than the first four.

You don't know why!

Your analysis could be bad or something could be truly wrong in the plant.

When more is better

Time	% N	%P
7:00 am	15.1	6.2
7:30	14.9	6.4
8:00	14.6	5.9
8:30	14.8	6.0
9:00	1.4	0.6

By evaluating two components in our sample, we now know more.

When more is better

Another statistical evaluation shows that for the first four samples:

	% N	%P
mean	14.9	6.1
s_x	0.2	0.2

The 9:00 sample is low by 9.6 s for both nitrogen and phosphorous.

You can be pretty confident that something is wrong with the sample. But what?

When more is better

Time	% N	%P	%K
7:00 am	15.1	6.2	20.1
7:30	14.9	6.4	21.4
8:00	14.6	5.9	19.2
8:30	14.8	6.0	19.0
9:00	1.4	0.6	1.9

You decide to look at all of the 'active components' in the sample.

When more is better

All of the components are low by about the same amount.

You immediately call the operator in charge of blending the chemical additives with the 'inert' filler - fixing the problem.

You boss give you a promotion!

Multivariate leverage

As the amount if data is increased:

- The amount of information also increased
- Your potential for understanding a problem can improve.

We can also work with any type of information.

- Quantitative and qualitative data
- Data from any sort of analysis.

Multivariate leverage

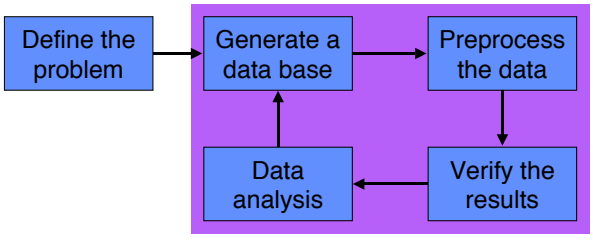
Sample	%N	%P	%K	%S	%O	%C	%Fe
1	15.1	6.2	20.1	0.23	30.1	2.5	0.02
2	15.3	6.1	19.3	0.12	29.2	1.6	0.01
3	14.8	5.9	21.4	0.22	28.8	3.1	0.03
4	16.3	6.9	20.2	0.15	31.5	2.0	0.10
5	12.7	6.1	20.1	0.23	33.5	2.2	0.02
6	15.9	5.8	20.2	0.19	20.9	2.6	0.05
7	15.9	4.3	20.3	0.28	27.5	1.8	0.04
8	10.3	7.1	22.1	0.23	27.9	2.5	0.01
9	20.1	6.6	20.1	0.22	30.3	2.5	0.03
10	15.9	6.6	20.4	0.22	33.1	2.9	0.02

While more data/information is good, we reach a point where we can no longer simply look at it to gain understanding.

Data exploration

Exploratory data analysis presents us with a set of tools to evaluate complex data sets.

The basic steps include:

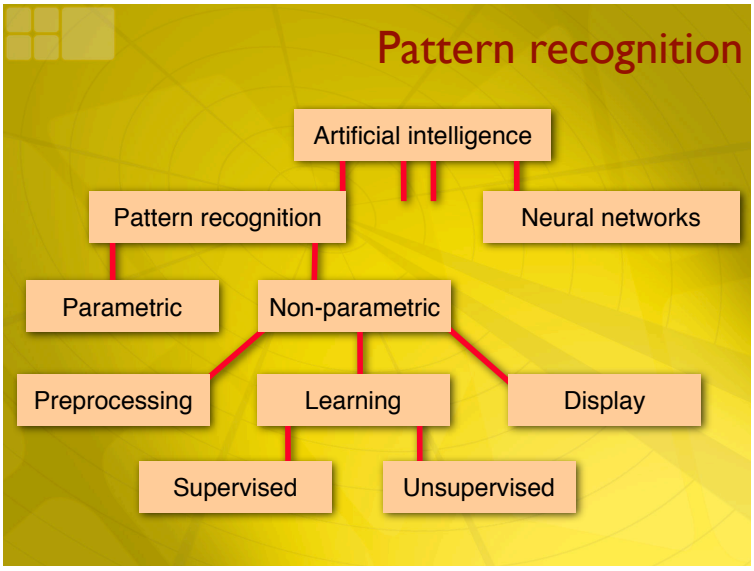


Pattern recognition

The goal is to be able to extract useful information for complex data sets. One way to do this is to detect and evaluate patterns in our data set.

We have several general types of tools available to use.

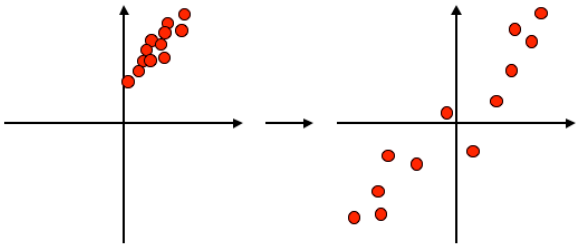
Pattern recognition



Pattern recognition

Preprocessing

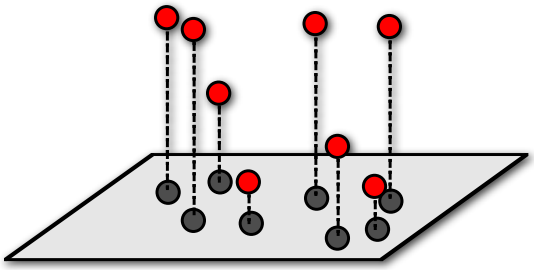
Data transformations such as scaling.



Pattern recognition

Display

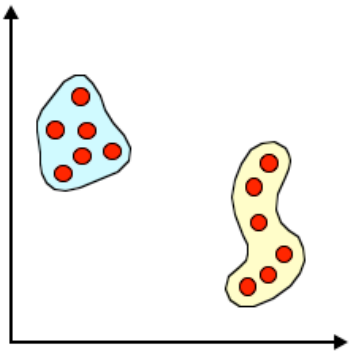
Projection of our data into a limited number of dimensions.



Pattern recognition

Unsupervised learning

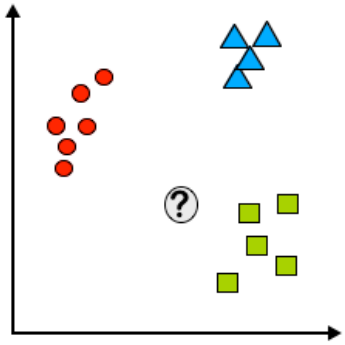
Methods that require no initial assumptions.
Examples - cluster analysis and PCA.



Pattern recognition

Supervised learning

Methods that require initial assumptions or a model. SIMCA and KNN are examples.



For most systems, we want an overdetermined dataset with at least three samples for each measured variable.

This is not always possible but the ratio of samples to variables should always be greater than one.

Data

Data

Methods assume that nearness in n-dimensional space reflects similarities in measured properties.

Each variable is treated as a dimension so a data set with 10 measured properties would be considered as existing in 10-dimensional space.

Since we typically have a large number of dimensions, we need a 'standard' way of working with our data.

The data matrix

The first step is to convert our data into a matrix where:

	1	2	3	...	NV
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	...	$X_{1,NV}$
2	$X_{2,1}$	.	.	...	.
3	$X_{3,1}$	.	.	...	.
...	...	.	.	...	.
NP	$X_{NP,1}$	.	.	...	$X_{NP,}$

The data matrix

Cases

A row of data where each value corresponds to measured properties of a specific sample

Variables or features

A column of data which corresponds to one measured property for all samples.

While many of our methods would still work if the definitions were reversed, its useful if we have a 'standard' matrix.

Pre-processing methods

We typically must initially convert our data so that all measurements can be compared.

It would be difficult to directly relate pH of a solution to the peak area resulting from its chromatographic analysis.

Qualitative data must also be converted to a form that we can process.

Initial data evaluation

Category data.

Convert to a numerical form.

Examples

hot/cold, day/night, gender

- convert to 1 and 0

color - convert to RGB index

Your goal is to convert descriptive information into a representative numerical format.

Initial data evaluation

Missing data

- Some samples may be missing one or more variables.
- Its best to avoid this by only using cases that are complete.
- If you must use incomplete data then you have several filling options.

Initial data evaluation

Filling options

Mean fill. Use the average for the other cases.

Random fill. Generate random values in the appropriate range.

PCA fill. Use an estimate based on other features.

All are bad as they change the nature of your data.

Initial data evaluation

Constant variables.
If a given measurement always gives the same value then eliminate it. It will only contain noise.

Redundant variables.
If two or more variables are strongly correlated (cc > 0.97) then remove all but one. Also, don't include two measurements of the same thing.
Example - Na via ISE and AA

Translation and scaling of data

The goal is to make all variables directly comparable.

	ppm		
	Cl	Fe	I
1	245	1.1	0.0001
2	233	1.4	0.0002
3	290	4.5	0.0001
4	300	7.2	0.0003

In this example, while the units are the same, the range and average values differ dramatically.

Translation and scaling of data

An evaluation of the ranges shows that for Cl and Fe:

$Range_{Cl}$ = 300 - 233 = 67

$Range_{Fe}$ = 7.2 - 1.1 = 6.1

On a percentage basis though,

$Range_{Cl}$ = 25.1

$Range_{Fe}$ = 150

So Fe actually has a larger variance range.

Scaling

One common approach would be to mean-center our values.

$$x' = x_{ik} - \overline{x}_k$$

Our data becomes

Cl	Fe
-22	-2.5
-34	-2.2
+57	+0.9
+33	+3.6

While all data is now centered around 0, Cl still swamps out Fe.

Scaling

Another approach is range-scaling

$$x' = \frac{(x_{ik} - x_{ik\,\min})}{(x_{\max} - x_{\min})}$$

Our data becomes

Cl	Fe
0.18	0.00
0.00	0.049
0.85	0.056
1.0	1.0

The problem with this approach is that while it is very sensitive to outliers, it falls apart if you have several points clustered at a high or low value.

One of the best approaches is autoscaling.

- Use mean-centering and units of standard deviation.
- In essence, you are converting the data into the 'reduced variable.' Actual units are 'standard deviation.'
- All variables will have the same units and occur over the same range.
- Total variance of each variable = 1.

Autoscaling

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

$$s_k = \left[\frac{\sum \left(x_{ik} - \bar{x}_k \right)^2}{N - 1} \right]^{1/2}$$

Total variance of your autoscaled matrix will be = NV

Autoscaling

If your variables are already correlated - in the same units - you can use:

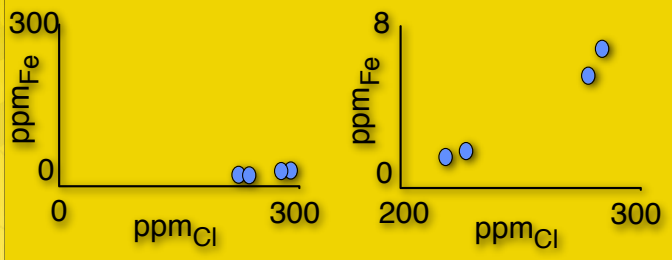
$$x'_{ik} = \frac{\left(x_{ik} - \bar{x}_k \right)}{\left[\sum \left(x_{ik} - \bar{x}_k \right)^2 \right]^{1/2}}$$

This results in a variance of 1/(NP-1) for each feature and NV/(NP-1) overall.

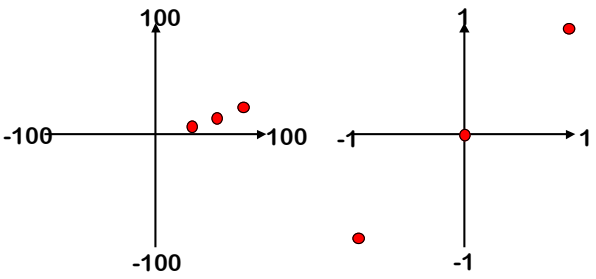
Autoscaling

We commonly do a type of autoscaling when we produce a graph.

Example - ppm_{Cl} vs. ppm_{Fe}



Autoscaling

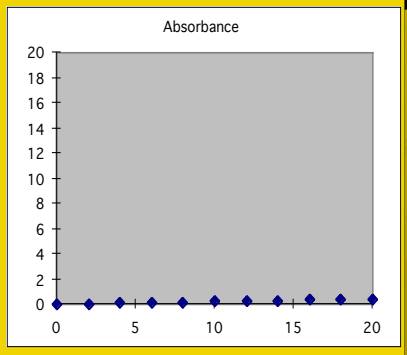


Autoscaling insures that all features are expressed with the same units and weight.

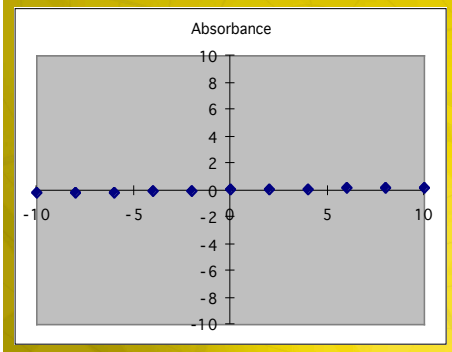
Scaling example

Original data

ppm As	Absorbance
0	0.0004
2	0.0492
4	0.0905
6	0.1325
8	0.1706
10	0.2296
12	0.2604
14	0.3051
16	0.3422
18	0.4018
20	0.4366

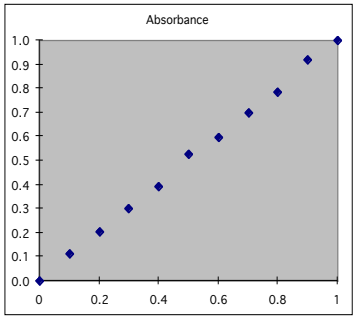


Mean Centered Scaling example



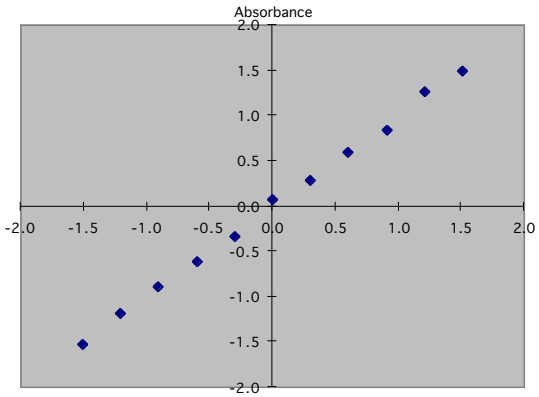
ppm As	Absorbance
-10	-0.2195
-8	-0.1707
-6	-0.1294
-4	-0.0874
-2	-0.0493
0	0.0097
2	0.0405
4	0.0852
6	0.1223
8	0.1819
10	0.2167

Range Scaling example

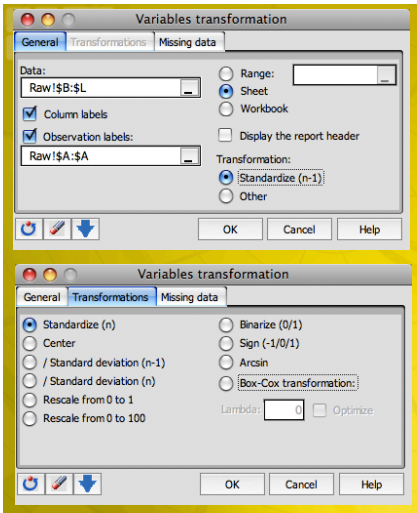


ppm As	Absorbance
0	0.0000
0.1	0.1118
0.2	0.2066
0.3	0.3027
0.4	0.3901
0.5	0.5254
0.6	0.5960
0.7	0.6985
0.8	0.7836
0.9	0.9203
1.0	1.0000

Autoscaling example



ppm As	Absorbance
-1.0508	-1.521
-1.206	-1.183
-0.905	-0.896
-0.603	-0.606
-0.302	-0.342
0.000	0.067
0.302	0.281
0.603	0.590
0.905	0.848
1.206	1.261
1.508	1.502



You have several scaling options that can be accessed via 'Variables Transformation.'

Standardize (n-1) is the same as autoscaling.

Results are best saved to a new sheet or workbook.

Scaling in XLStat

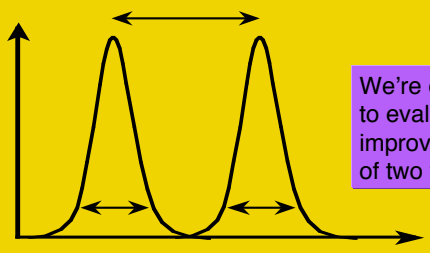
XLStat results

	Original data:		Transformed data:	
	ppm As	Absorbance	ppm As	Absorbance
	0	0.0004	-1.508	-1.521
	2	0.0492	-1.206	-1.183
	4	0.0905	-0.905	-0.897
	6	0.1325	-0.603	-0.606
	8	0.1706	-0.302	-0.342
	10	0.2296	0.000	0.067
	12	0.2604	0.302	0.281
	14	0.3051	0.603	0.591
	16	0.3422	0.905	0.848
	18	0.4018	1.206	1.261
	20	0.4366	1.508	1.502
Mean	10	0.2199	2.0186E-17	1.2112E-16
Stdev	6.633	0.144	1.000	1.000
Variance	44	0.021	1	1

Feature weighting

Weighting can be used to:

- ✓ Measure the discriminating ability of a variable in category separation.
- ✓ Improve your classification results.



We're essentially trying to evaluate and/or improve the resolution of two features.

Variance weighting

Weighting for 2 categories (I and II) based on the ratio of the intercategory variance to the sum of the intracategory variances.

$$w_{k(I,II)} = 2\frac{\frac{1}{N_I}\sum x_I^2 + \frac{1}{N_{II}}\sum x_{II}^2 - \frac{2}{N_I N_{II}}\sum x_I^2 \frac{1}{N_{II}}\sum x_{II}^2}{\frac{1}{N_I}\sum\left(x_I - \bar{x}_I\right)^2 + \frac{1}{N_{II}}\sum\left(x_{II} - \bar{x}_{II}\right)^2}$$

Intracategory - within group variance.
Intercategory - between group variance.
So, we're weighted based on F values.

Approach can be used to calculate feature weights, giving a measure of their ability to discriminate.

Fisher weights

An alternative to variance weighting.

$$w_{k(I,II)} = \frac{\left|\bar{x}_I - \bar{x}_{II}\right|}{\frac{1}{N_I}\sum\left(x_I - \bar{x}_I\right)^2 + \frac{1}{N_{II}}\sum\left(x_{II} - \bar{x}_{II}\right)^2}$$

Simply replaced the numerator with the difference of the category means.

A Fisher weight may actually go to zero for a nondiscriminating feature so the overall weight can be calculated as:

$$w_k = \frac{1}{N_J}\sum_{J=1}^{N_J} w_{k(J)}$$

Once the weight of each variable has been calculated for each category pair, you can use it for scaling:

$$x'_{ik} = w_k x_{ik}$$

This can be done before, after or in place of autoscaling.

Either Fisher or variance weights can be used.

Feature weighting

Example

- In a study, 119 paper samples were assayed for 13 trace elements by neutron activation analysis.
- Each paper could be classified based on paper grade (40 types) and manufacturer (9 companies).
- Goal - can we identify the paper grade and manufacturer based on trace element composition.

- First, the weights were calculated for each category.
- This is done by taking the average of each element based on paper grade.
- Weights are then calculated based on paper source.

Example

Example

	Paper Grade		Source	
	Variance	Fisher	Variance	Fisher
Na	7.39	6.62	1.49	0.048
Al	66.95	22240.00	3.03	0.650
Cl	9.96	137.10	1.67	0.085
Ca	13.24	11.08	1.98	0.141
Ti	17.94	12.78	1.66	0.092
Cr	8.67	41.53	1.75	0.106
Mn	13.01	15.53	2.37	0.182
Zn	4.87	18.24	1.99	0.163
Sb	10.19	31.68	1.92	0.138
Ta	2.06	1.71	1.25	0.013

$N_{grade} = 40$ $N_{source} = 9$

Example

What do the weights show?

Paper grade

All weights are large.

All can provide a way to classify grade.

You might want to consider only using 4-6 variables with the largest weights to save time and money.

	Paper Grade		Source	
	Variance	Fisher	Variance	Fisher
Na	7.39	6.62	1.49	0.048
Al	66.95	22240.00	3.03	0.650
Cl	9.96	137.10	1.67	0.085
Ca	13.24	11.08	1.98	0.141
Ti	17.94	12.78	1.66	0.092
Cr	8.67	41.53	1.75	0.106
Mn	13.01	15.53	2.37	0.182
Zn	4.87	18.24	1.99	0.163
Sb	10.19	31.68	1.92	0.138
Ta	2.06	1.71	1.25	0.013

$N_{grade} = 40$ $N_{source} = 9$

Example

Paper Source

This would be harder to do since the weights are smaller.

However, they are still > 1 for variance weighting, so it can be done.

Again, it would be best to pick the 4-6 variables with the largest weights.

	Paper Grade		Source	
	Variance	Fisher	Variance	Fisher
Na	7.39	6.62	1.49	0.048
Al	66.95	22240.0	3.03	0.650
Cl	9.96	137.10	1.67	0.085
Ca	13.24	11.08	1.98	0.141
Ti	17.94	12.78	1.66	0.092
Cr	8.67	41.53	1.75	0.106
Mn	13.01	15.53	2.37	0.182
Zn	4.87	18.24	1.99	0.163
Sb	10.19	31.68	1.92	0.138
Ta	2.06	1.71	1.25	0.013

$N_{grade} = 40$ $N_{source} = 9$

Eigenvector rotations

In general, if we treat our data set as a matrix, we are free to translate it.

This does not alter the significance of any of the information.

This translation can be some form of scaling or weighting. $X' = X \cdot a$

We can also rotate the matrix by multiplying by a transform matrix.

$$X' = XA^T$$

Eigenvector rotations

This rotation changes the coordinates of our matrix but not its variance.

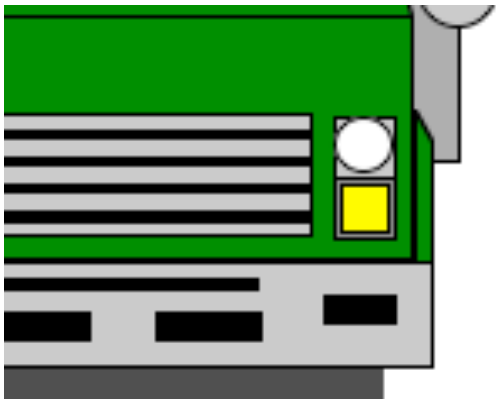
Autoscaling and eigenvector rotations work together to give us the best possible viewpoint for our dataset.

As an example, lets say that you are going to purchase your first truck.

Example

From this vantage, its difficult to make any sort of choice.

It might not even be a truck.

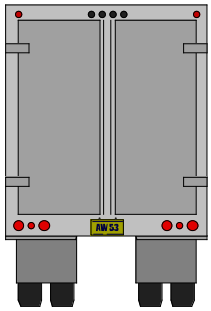


Here, we are too close.

Example

This is an 'autoscaled' view.

Its centered and full scale.



Unfortunately, from this angle, we only get a limited amount of information

Example

A scaled, rotated view of our example

We get as much information from a single view as possible. Some information still can't be seen.

Eigenvector rotations

The goal is to rotate our matrix so that we have the maximum amount of variation present in the minimum number of axes.

Eigenvector rotation

Create a new set of orthogonal axis.

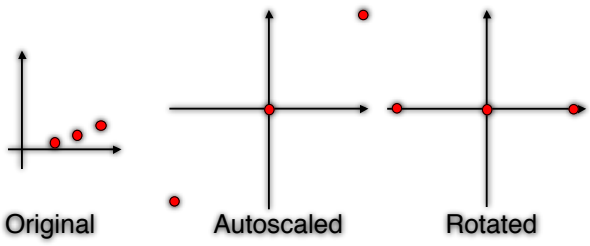
$\sigma^2_{EV1} > \sigma^2_{EV2} > \sigma^2_{EV3} > \ldots > \sigma^2_{EVNV}$

Data structure is not changed.

Eigenvector rotations

- These rotations are accomplished by diagonalization of either the correlation or covariance matrix.
- Which matrix you use will be based on the actual pattern recognition method is being evaluated.
- We'll discuss the differences as we introduce the various methods.

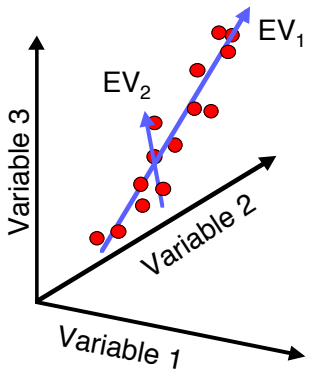
Eigenvector rotations



In this example, our original data is reduced to one variable after it is scaled and rotated.

Why?

Eigenvector rotations



This example shows both the original variables and the resulting eigenvectors

Information obtained

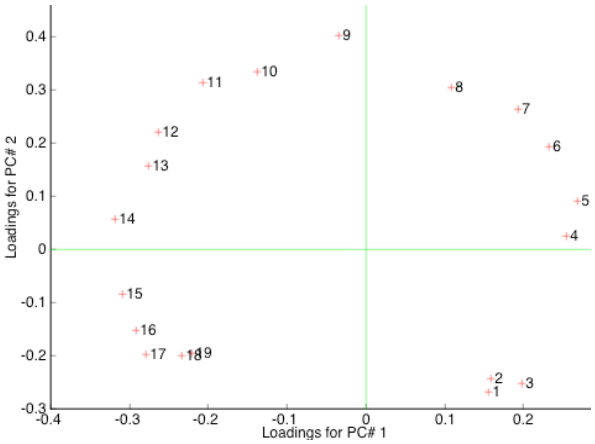
An eigenvector rotation results in a series of loadings and scores along with a residual.

Loading - 1-D array

Contains the eigenvector coefficients required for the rotation to a specific score.

Loading coefficients show the relative significance or contribution of each of our original variables.

Loading example



Autoscaled arson related samples, 19 variables

Information obtained

Score

A linear combination of the original variables where:

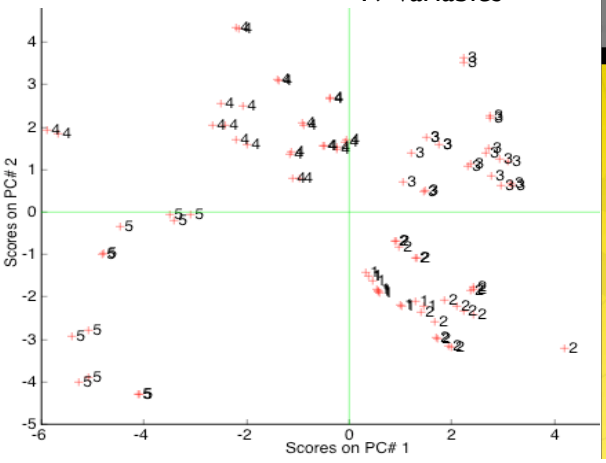
$$score_{1,1} = EV_{1,1}\, var_1 + EV_{1,2}\, var_2 \ldots EV_{N,NV}\, var_{NV}$$

Each score reflects the contribution of all variables for a specific case.

This results in related variables being combined into a single variable and a significant data reduction

Score example

Autoscaled arson related samples. 19 variables



Information obtained

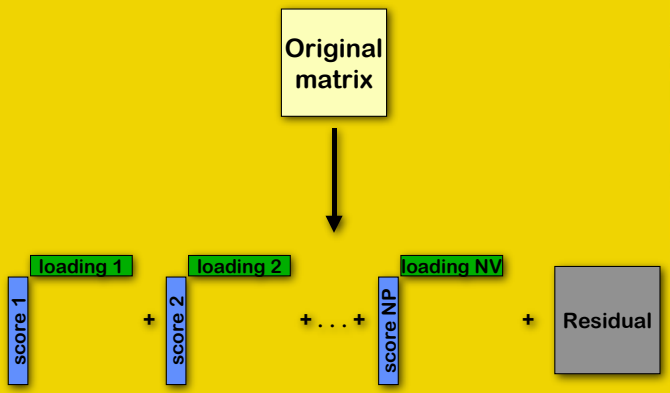
Residual

The portion of the original array that could not be correlated.

This could be random noise.

Many methods do not require a complete eigenvector solution. So, the residual could also be any remaining information that had yet to be used when the method terminated.

Eigenvector rotation components



Eigenvalues

- Another term that we typically obtain is the eigenvalue.
- One eigenvalue for each eigenvector.
- It indicates how much of the original information is contained in each eigenvector.

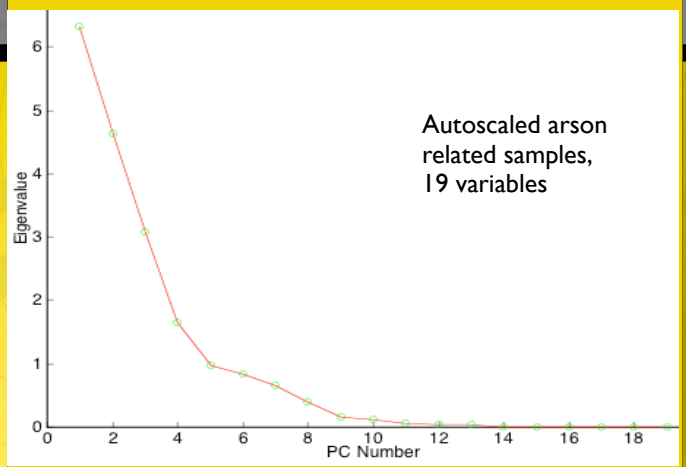
Eigenvalues

Assume that our data had been scaled such that the total variance was NV.

You can then determine how much of the original information is contained in each eigenvector by

$$\%\,variance = \frac{eigenvalue_i}{NV}$$

Eigenvalue example



Autoscaled arson related samples, 19 variables

Advantages of eigenvector rotation

Inherent data reduction

It is often possible to reduce complex data sets to 2-5 eigenvector / score sets and still express the majority of the information.

Display

Reduction of the number of variables makes it easier to evaluate our data.

Noise reduction

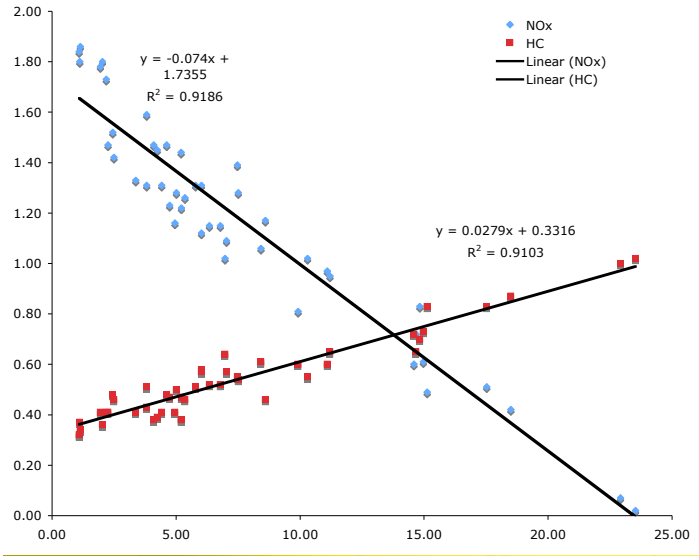
Truly random noise never correlates so it remains in the residual matrix.

Example

- The car exhaust problem from the first exam.
- CO, NOx and HC levels from a set of 'tailpipe' tests.
- We already know that the three measurements are correlated.
- Now, let's look that the effect of autoscaling and an eigenvector rotation.

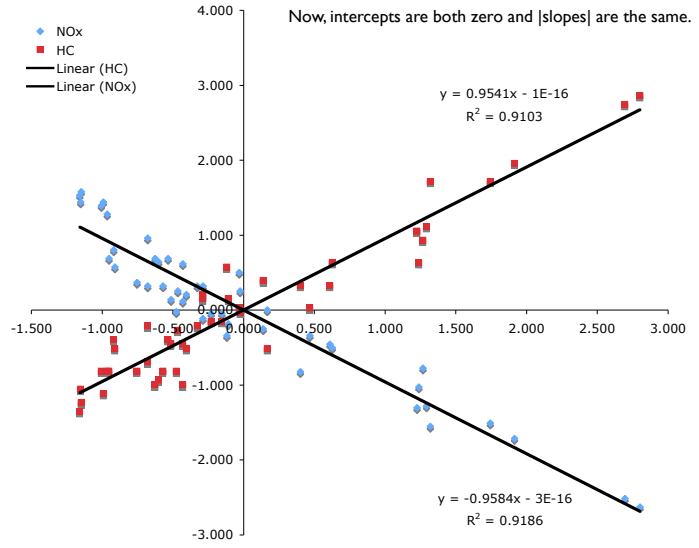
Original data

Car	CO	NOx	HC
1	5.01	1.28	0.50
2	14.67	0.72	0.65
3	8.60	1.17	0.46
4	4.42	1.31	0.41
5	4.95	1.16	0.41
6	4.24	1.45	0.39
7	7.51	1.28	0.54
8	10.30	1.02	0.55
9	14.59	0.60	0.72
10	6.98	1.02	0.64
11	17.53	0.51	0.83
12	4.10	1.47	0.38
13	5.21	1.44	0.38
14	11.10	0.97	0.60
15	9.92	0.81	0.60
16	14.97	0.61	0.73
17	15.13	0.49	0.83
18	7.04	1.09	0.57
19	1.14	1.86	0.34
20	3.38	1.33	0.41
21	1.12	1.80	0.37
22	23.53	0.02	1.02
23	18.50	0.42	0.87
24	22.92	0.07	1.00



Autoscaled

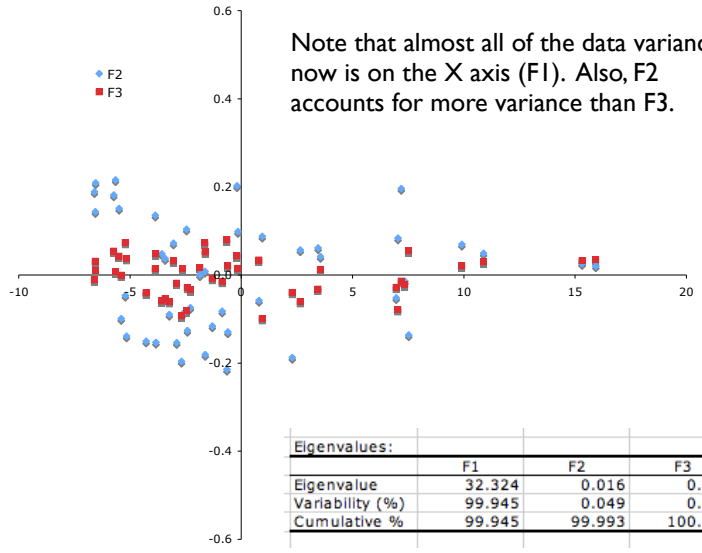
Car	CO	NOx	HC
1	-0.465	0.252	-0.272
2	1.239	-1.028	0.631
3	0.168	0.000	-0.514
4	-0.569	0.321	-0.815
5	-0.476	-0.022	-0.815
6	-0.601	0.641	-0.935
7	-0.024	0.252	-0.031
8	0.468	-0.342	0.029
9	1.225	-1.302	1.053
10	-0.118	-0.342	0.571
11	1.744	-1.508	1.716
12	-0.626	0.686	-0.996
13	-0.430	0.618	-0.996
14	0.609	-0.457	0.330
15	0.401	-0.822	0.330
16	1.292	-1.280	1.114
17	1.320	-1.554	1.716
18	-0.107	-0.182	0.149
19	-1.148	1.578	-1.237



Rotated data

Cars	F1	F2	F3
1	-2.638	-0.088	0.014
2	7.037	0.084	-0.079
3	0.947	0.088	-0.099
4	-3.231	-0.090	-0.060
5	-2.692	-0.196	-0.093
6	-3.422	0.037	-0.055
7	-0.145	0.099	0.015
8	2.656	0.057	-0.061
9	6.968	-0.052	-0.028
10	-0.651	-0.214	0.080
11	9.909	0.069	0.020
12	-3.563	0.048	-0.059
13	-2.454	0.104	-0.081
14	3.458	0.062	-0.032
15	2.294	-0.187	-0.040
16	7.346	-0.014	-0.022
17	7.518	-0.136	0.054
18	-0.599	-0.129	0.021
19	-6.544	0.209	0.011
20	-4.269	-0.151	-0.041

Note that variable labels have been replace to reflect the fact that these are no longer our original ones.



Note that almost all of the data variance now is on the X axis (F1). Also, F2 accounts for more variance than F3.

Eigenvalues:			
	F1	F2	F3
Eigenvalue	32.324	0.016	0.002
Variability (%)	99.945	0.049	0.007
Cumulative %	99.945	99.993	100.000