UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DEPARTMENT OF
INDUSTRIAL ENGINEERING

Machine Learning Laboratory #1

Academic year 2024-2025

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: `pierantonio.facco@unipd.it`
URL: `https://research.dii.unipd.it/capelab/`

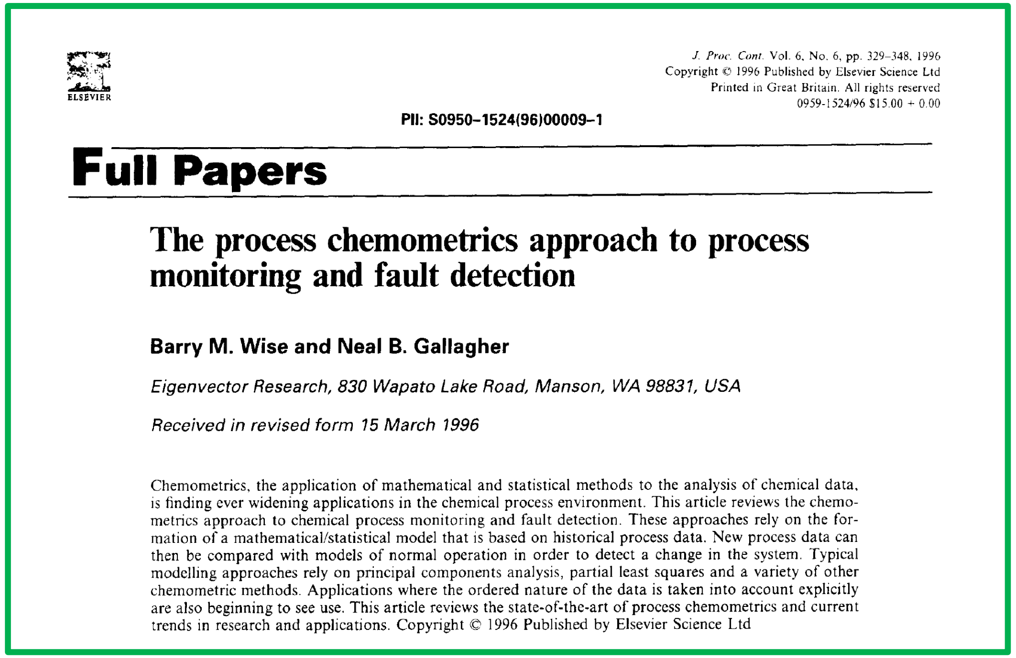
Summary

- Software:
  - PLS_Toolbox® for Matlab®

- **Hands-on-pc laboratory**
  - Example #1: health and alcohol consumption

Multivariate statistics software

- Matlab® (by MathWorks)
  - Statistics toolbox
  - PLS_Toolbox ® (by Eigenvector research Inc.)*
    - http://www.eigenvector.com/software/pls_toolbox.htm
- R
  - free
- Python
  - free
- Other software…
  - SPSS
  - Stata
  - SAS
  - JMP



J. Proc. Cont. Vol. 6, No. 6, pp. 329–348, 1996
Copyright © 1996 Published by Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0959-1524/96 $15.00 + 0.00

PII: S0950–1524(96)00009–1

Full Papers

The process chemometrics approach to process monitoring and fault detection

Barry M. Wise and Neal B. Gallagher

Eigenvector Research, 830 Wapato Lake Road, Manson, WA 98831, USA

Received in revised form 15 March 1996

Chemometrics, the application of mathematical and statistical methods to the analysis of chemical data, is finding ever widening applications in the chemical process environment. This article reviews the chemometrics approach to chemical process monitoring and fault detection. These approaches rely on the formation of a mathematical/statistical model that is based on historical process data. New process data can then be compared with models of normal operation in order to detect a change in the system. Typical modelling approaches rely on principal components analysis, partial least squares and a variety of other chemometric methods. Applications where the ordered nature of the data is taken into account explicitly are also beginning to see use. This article reviews the state-of-the-art of process chemometrics and current trends in research and applications. Copyright © 1996 Published by Elsevier Science Ltd

*Wise, B. M., and N. B. Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control* **6**, 329–348.

Matlab® codes

- Principal component analysis PCA* on **X** = singular value decomposition on the correlation of **X**

  - `[U,S,V] = svd(cov(auto(X)));`
  - `P=V;` *loadings* **P**
  - `T=X*P;` *scores* **T**

autoscaled **X**

X *covariance* $\operatorname{cov}(\mathbf{X}) = \dfrac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{N-1}$

$$\boxed{\mathbf{X} = \mathbf{t}_1\mathbf{p}^{\mathrm{T}}{}_1 + \mathbf{t}_2\mathbf{p}^{\mathrm{T}}{}_2 + \ldots + \mathbf{t}_k\mathbf{p}^{\mathrm{T}}{}_k + \mathbf{E}} \Leftrightarrow \boxed{\operatorname{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\,\mathbf{p}_i}$$

*Jackson, J. E. (2003). *A User's Guide to Principal Components*. Wiley-Interscience, Hoboken, N.J.

Geladi, P., and B.R., Kowalski 1986. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185, 1–17.

PCA Matlab® code: NIPALS algorithm

```
function [T,P,E]=pca_profFacco(X)
T=[];
P=[];
for n=1:min(min(size(X)),20)
    n
    te=1;
    tn=1;
    if n==1
        X=X;
    else
        X=E;
    end
    m=1;
    while (m<=1000)&&((norm(te)/norm(tn))>10^-12)
        m
        if m==1000
            display('no convergence');
            norm(te)/norm(tn)
        else
            if m==1
                xn=X(:,1);
                tn=xn;
            else
                tn=t1;
            end
            p1=((tn'*X)/(tn'*tn))';
            p1=p1/norm(p1);
            t1=(X*p1)/(p1'*p1);
            te=t1-tn;
            m=m+1;
        end
    end
    T=horzcat(T,t1);
    P=horzcat(P,p1);
    E=X-t1*p1';
end
```

Example #1
Health and alcohol consumption

Health and alcohol consumption

- Available data:
  - 10 observations (on the rows)
    - nations
  - 5 variables (on the columns)
    - wine, beer and liquor consumption, heart disease rate life expectation

	liquor consumption [L/year]	wine consumption [L/year]	beer consumption [L/year]	life expectancy [years]	heart disease rate [10^5 cases/year]
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58.0	25.1	78	94.1
Switzerland	1.7	46.0	65.0	78	106.4
Australia	1.2	15.7	102.1	78	173.0
Grait britain	1.5	12.2	100.0	77	199.7
USA	2.0	8.9	87.8	76	176.0
Russia	3.8	2.7	17.1	69	373.6
Czech Republic	1.0	1.7	140.0	73	283.7
Japan	2.1	1.0	55.0	79	34.7
Mexico	0.8	0.2	50.4	73	36.4

- Objective: **exploratory analysis** of the data

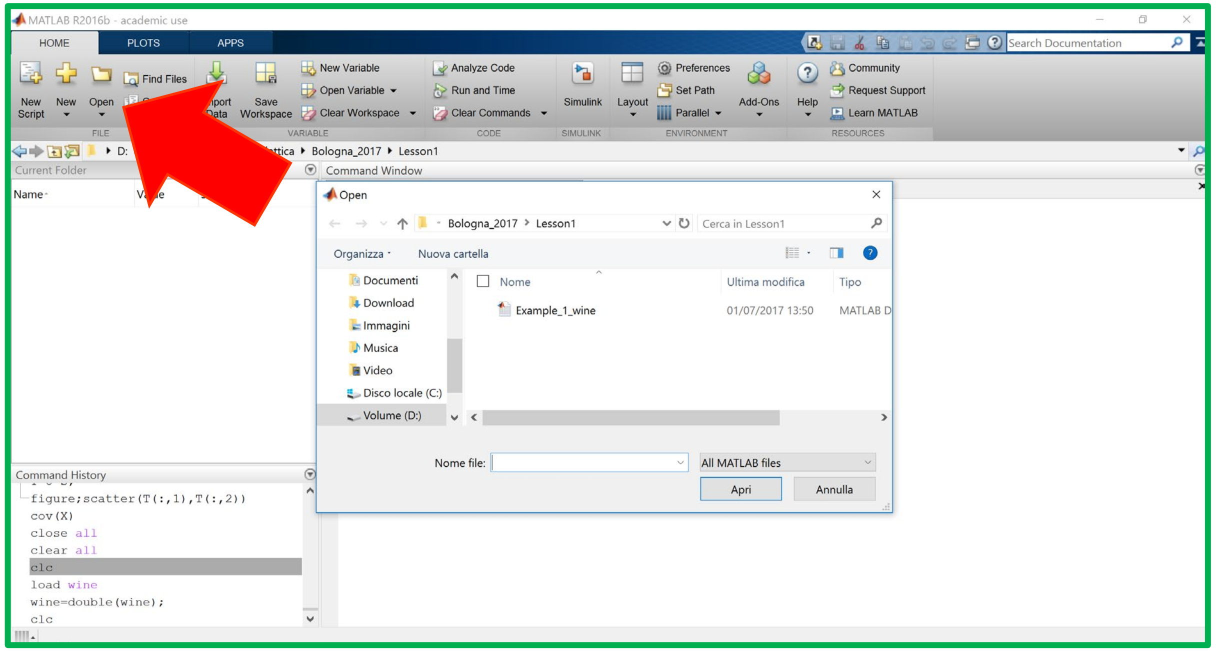
Loading data in Matlab

- Command:
  - **`load wine`**

*this is a **table** dataset*

- Alternative:
  - Matlab® user interface
  - select «Open»
  - the path for the dataset **`wine.mat`**
    - ○ shared in Moodle

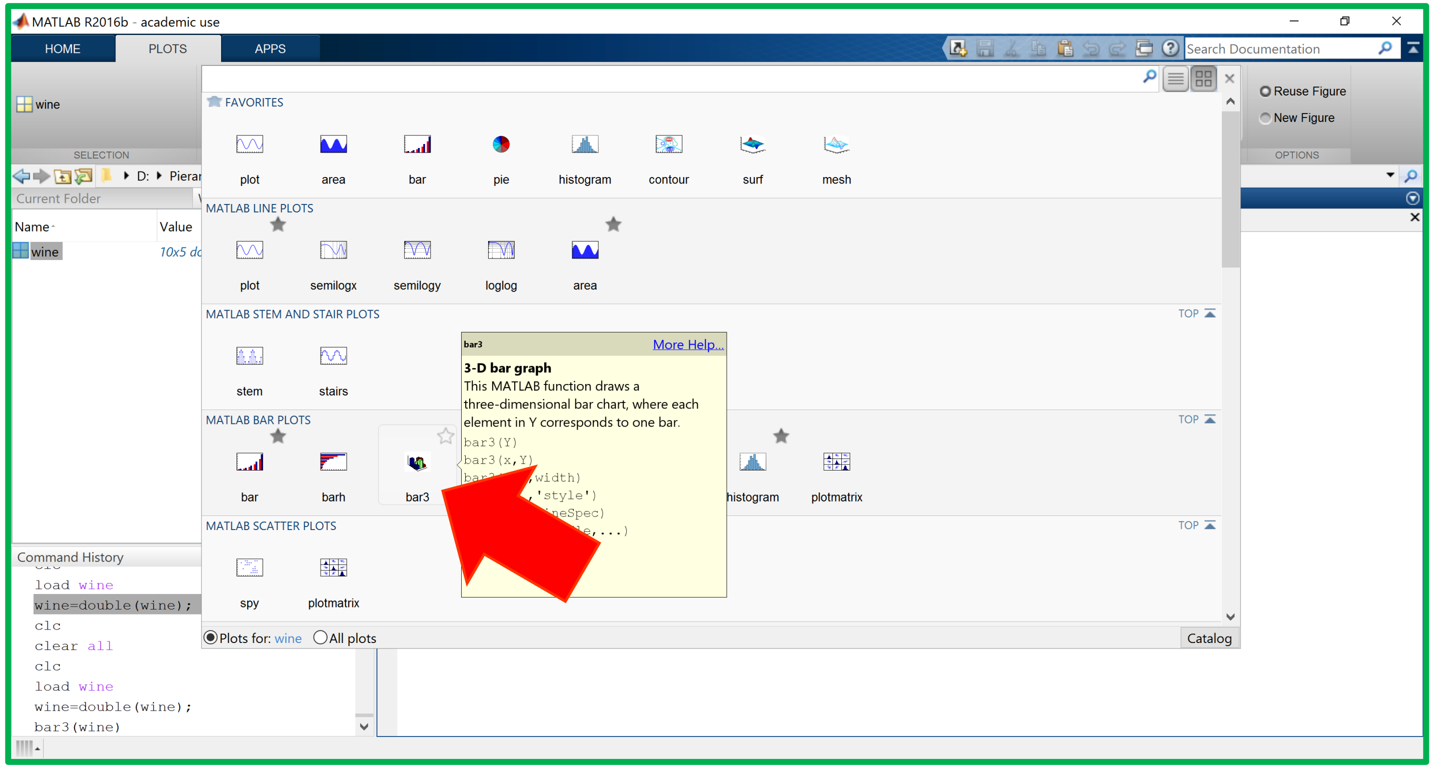
*this is a **double precision** dataset*



```
figure;scatter(T(:,1),T(:,2))
cov(X)
close all
clear all
clc
load wine
wine=double(wine);
clc
```

Initial data visualization

- Commands:
  - `bar3(wine)` or `bar3(double(wine))`
- User interface:
  - select the dataset
  - «Plot» ➡ «bar3»



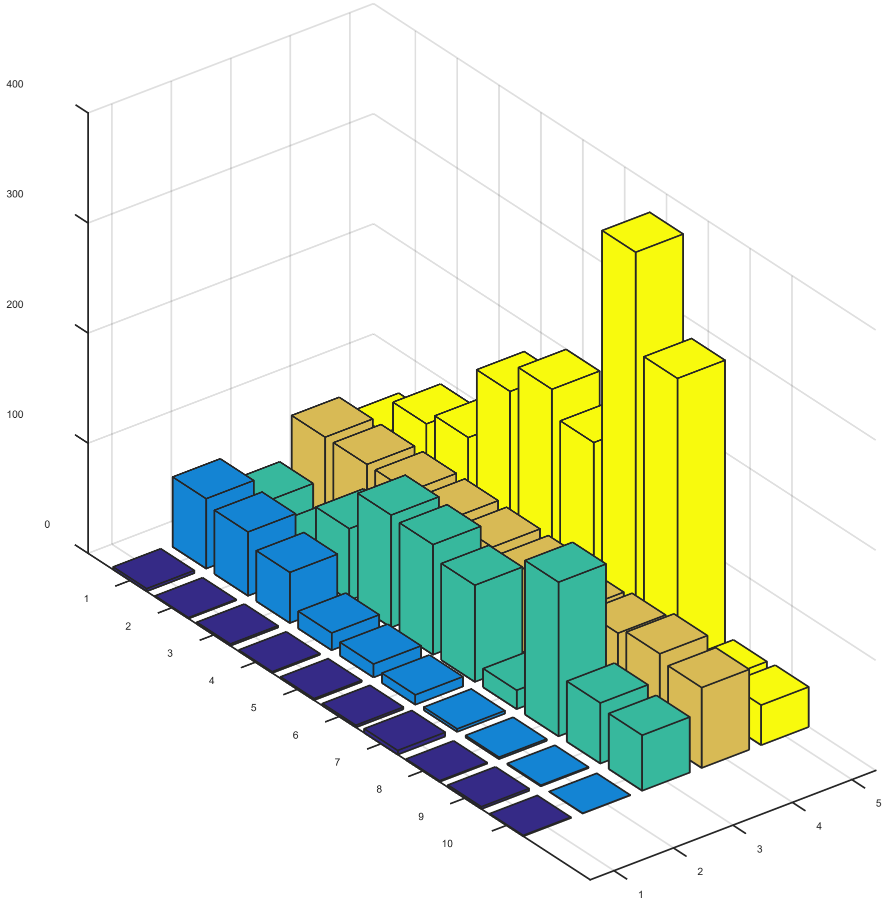
bar3 More Help...

3-D bar graph
This MATLAB function draws a three-dimensional bar chart, where each element in Y corresponds to one bar.

```
bar3(Y)
bar3(x,Y)
bar    ,width)
       ,'style')
        neSpec)
           le,...)
```

```
load wine
wine=double(wine);
clc
clear all
clc
load wine
wine=double(wine);
bar3(wine)
```

Visual data exploration

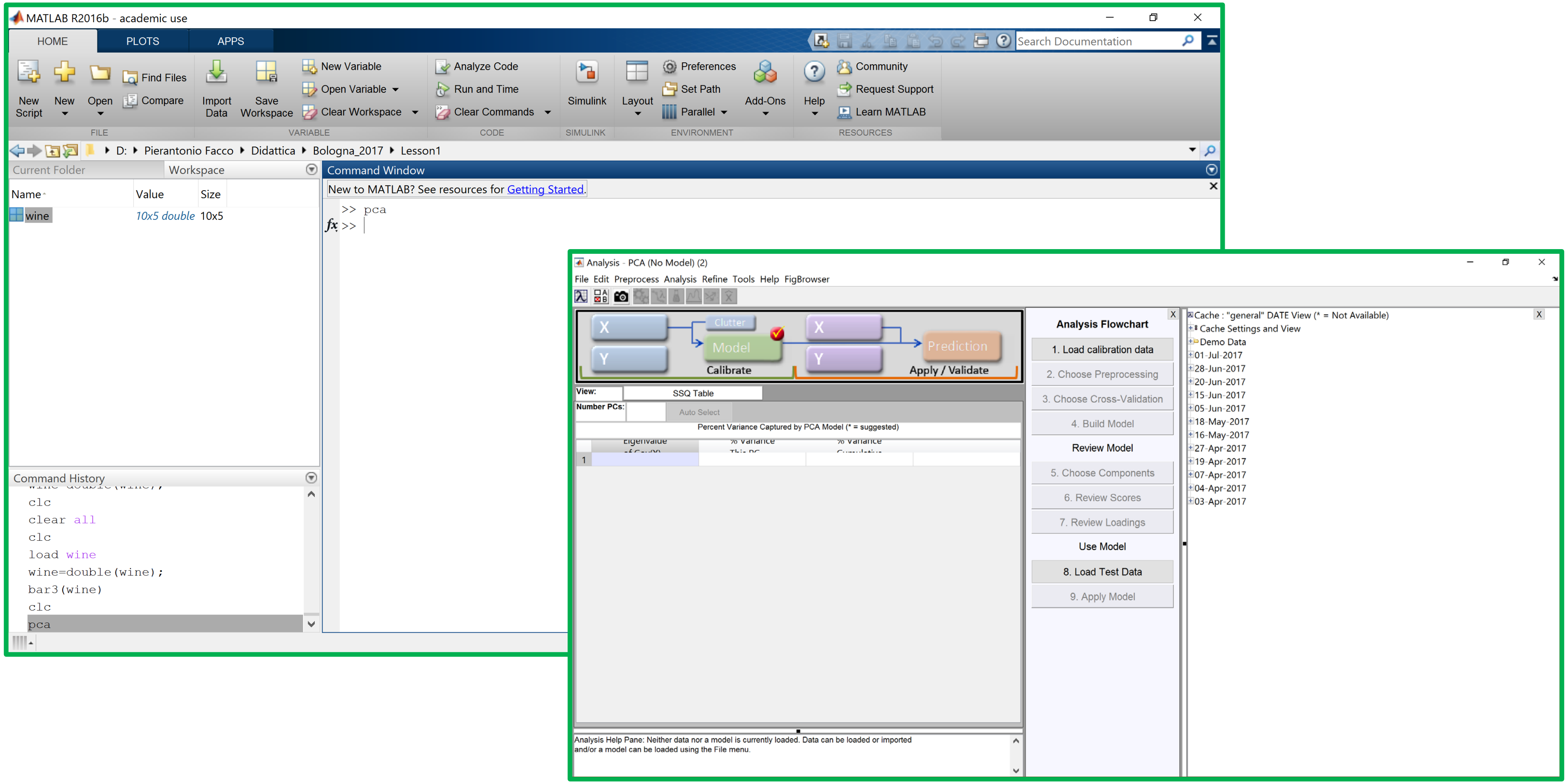


data have <u>different scales</u>

<u>AUTOSCALING</u> *to give the same weight to all the data (same importance to all the variables)*

PCA in the PLS_Toolbox

- Command:
  - **`pca`**



Name	Value	Size
wine	10x5 double	10x5

New to MATLAB? See resources for Getting Started.

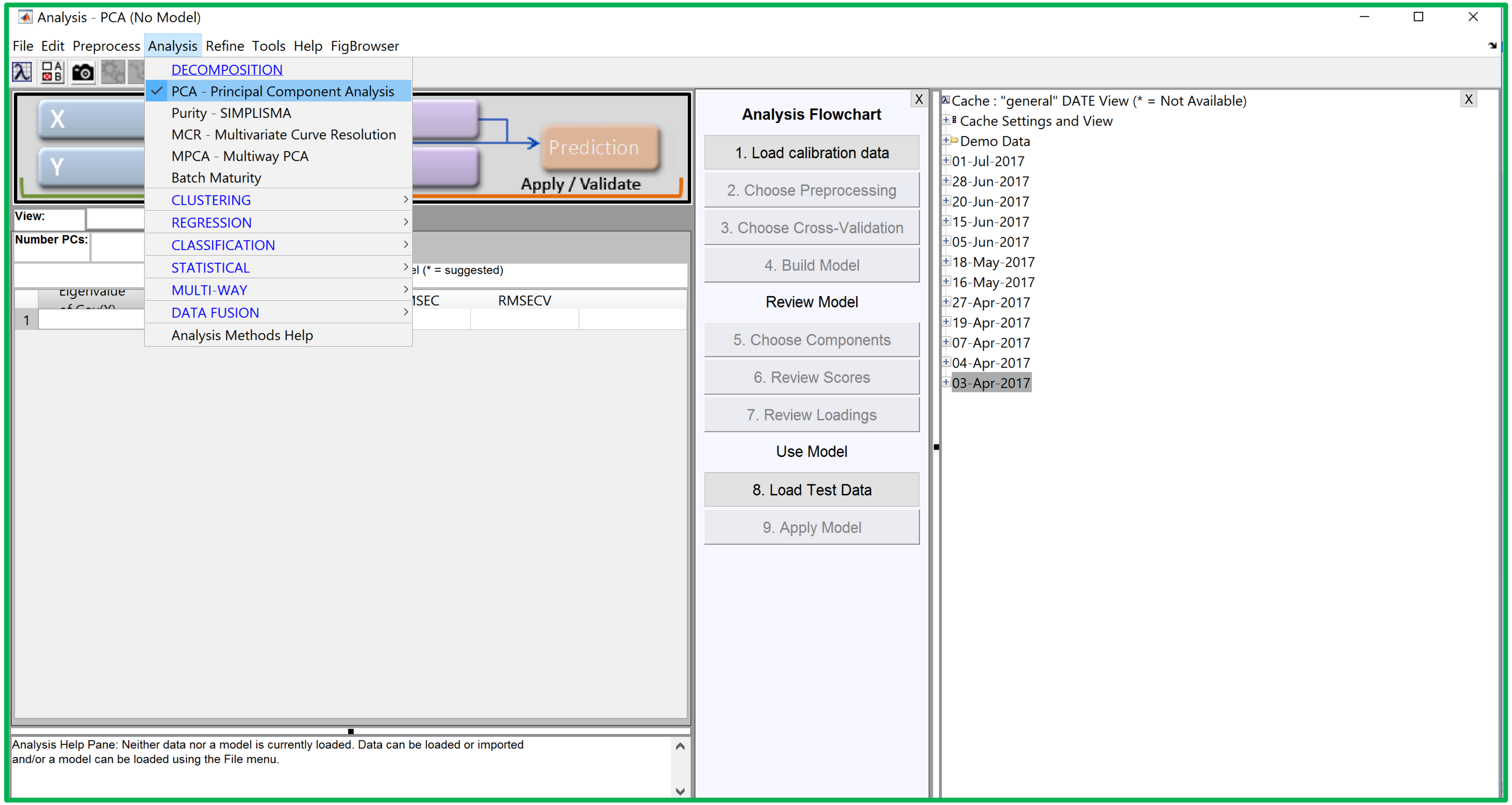
```
>> pca
fx >>
```

```
clc
clear all
clc
load wine
wine=double(wine);
bar3(wine)
clc
pca
```

Percent Variance Captured by PCA Model (* = suggested)

Analysis Help Pane: Neither data nor a model is currently loaded. Data can be loaded or imported and/or a model can be loaded using the File menu.

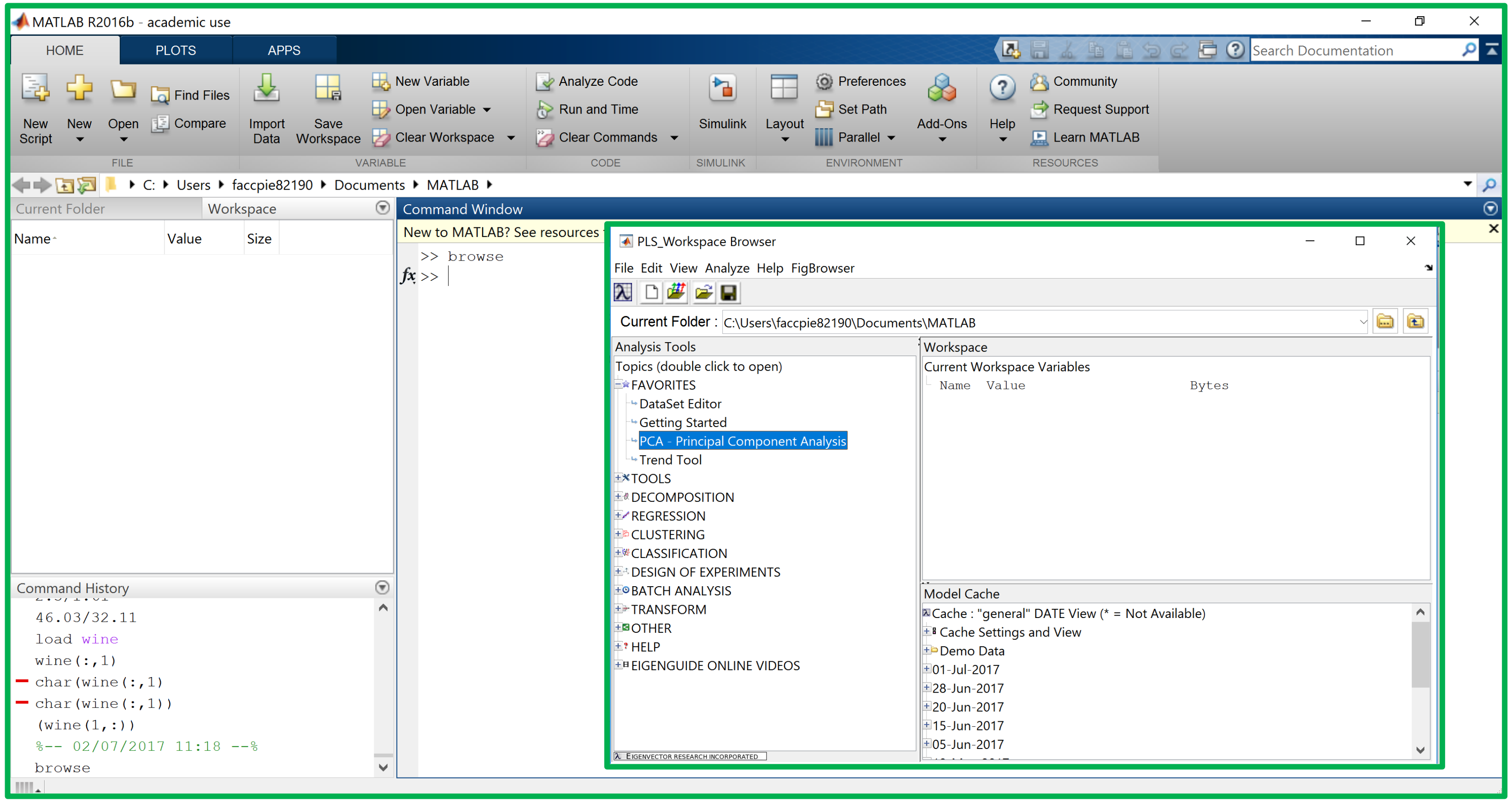
Select PCA



Analysis Help Pane: Neither data nor a model is currently loaded. Data can be loaded or imported and/or a model can be loaded using the File menu.

Alternative method

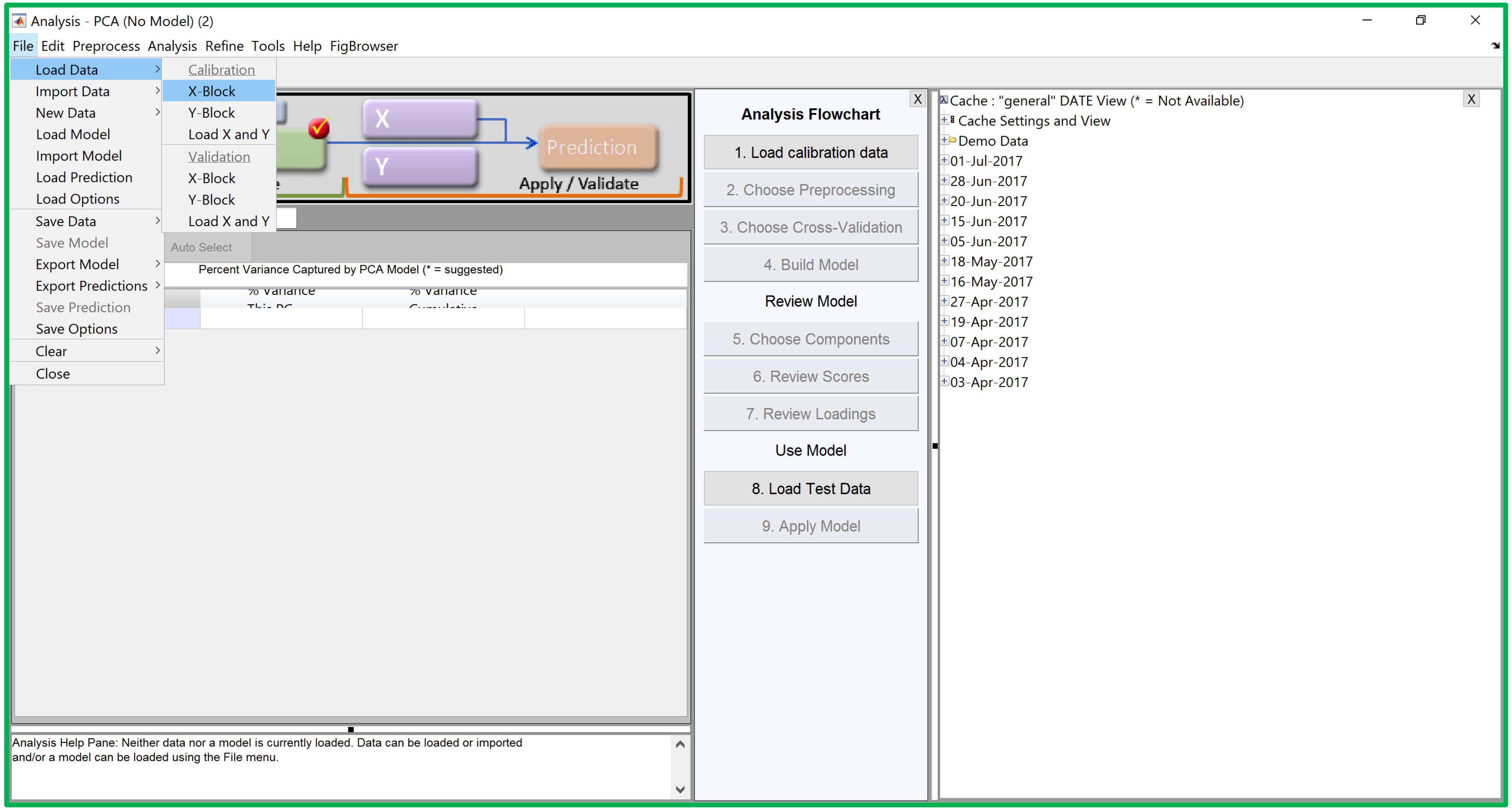
- Command:
  - **`browse`** … and select PCA



```
>> browse
>>
```

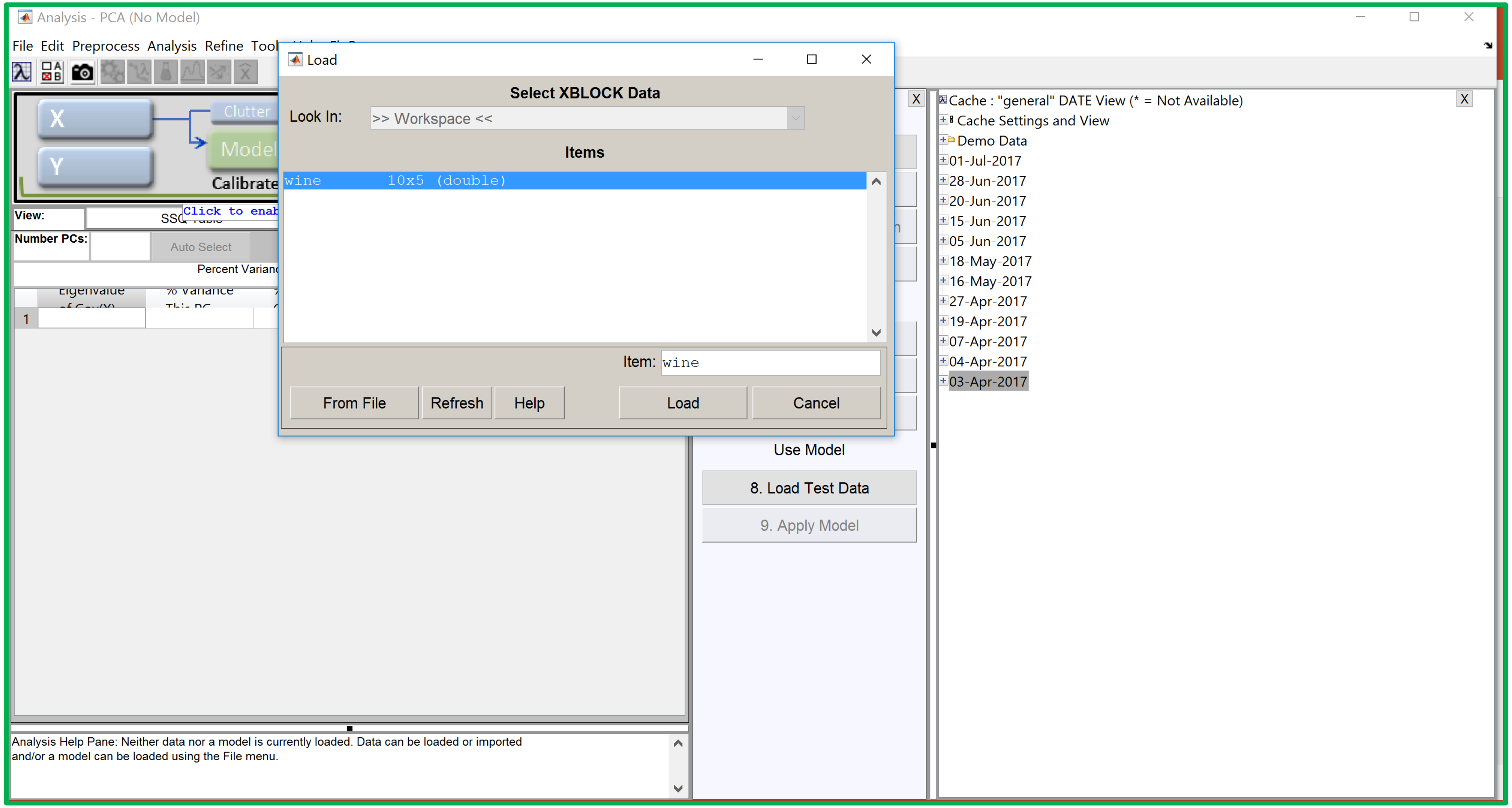
```
46.03/32.11
load wine
wine(:,1)
char(wine(:,1)
char(wine(:,1))
(wine(1,:))
%-- 02/07/2017 11:18 --%
browse
```

Data upload in the toolbox (1/2)



Analysis Help Pane: Neither data nor a model is currently loaded. Data can be loaded or imported and/or a model can be loaded using the File menu.

Data upload in the toolbox (2/2)



Select XBLOCK Data

Look In: >> Workspace <<

Items

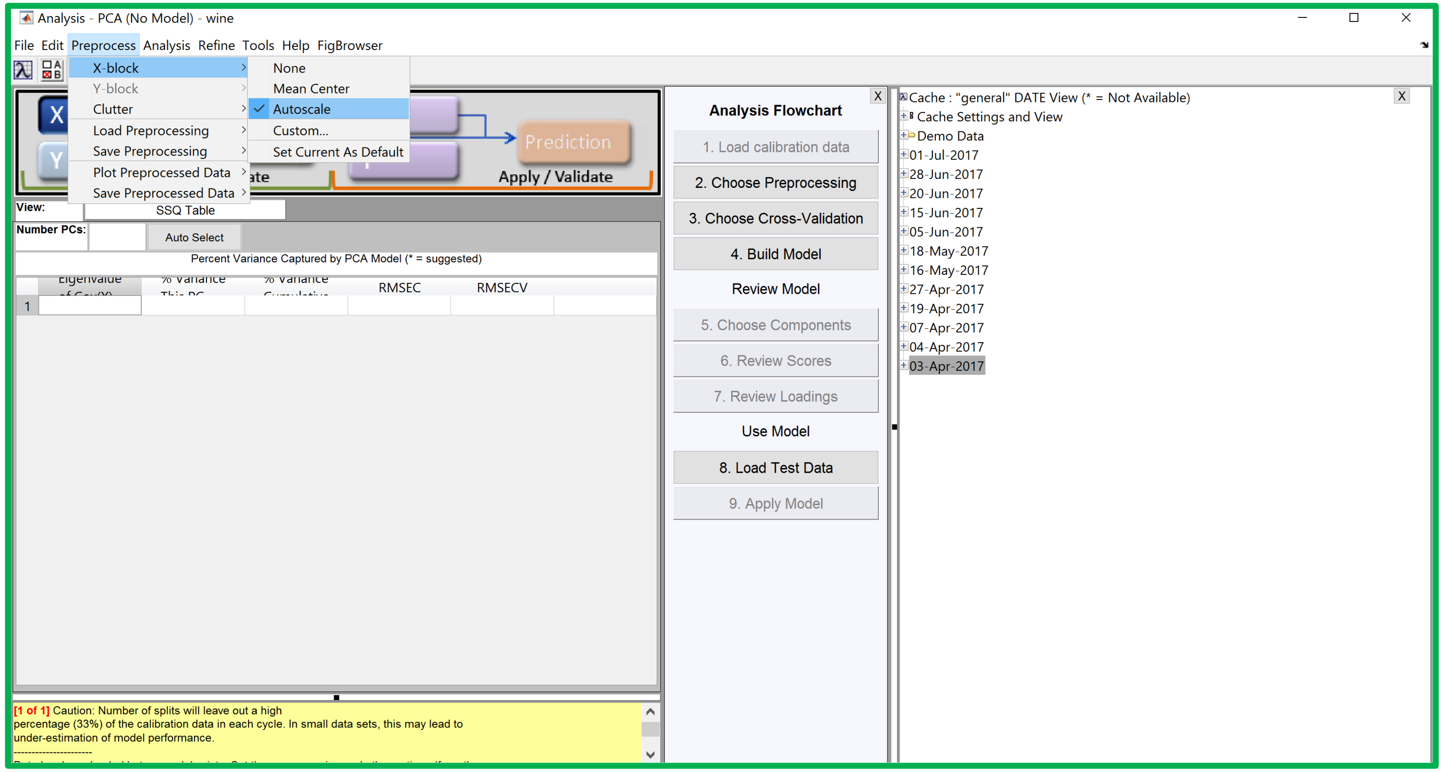
```
wine        10x5  (double)
```

Item: wine

Analysis Help Pane: Neither data nor a model is currently loaded. Data can be loaded or imported and/or a model can be loaded using the File menu.

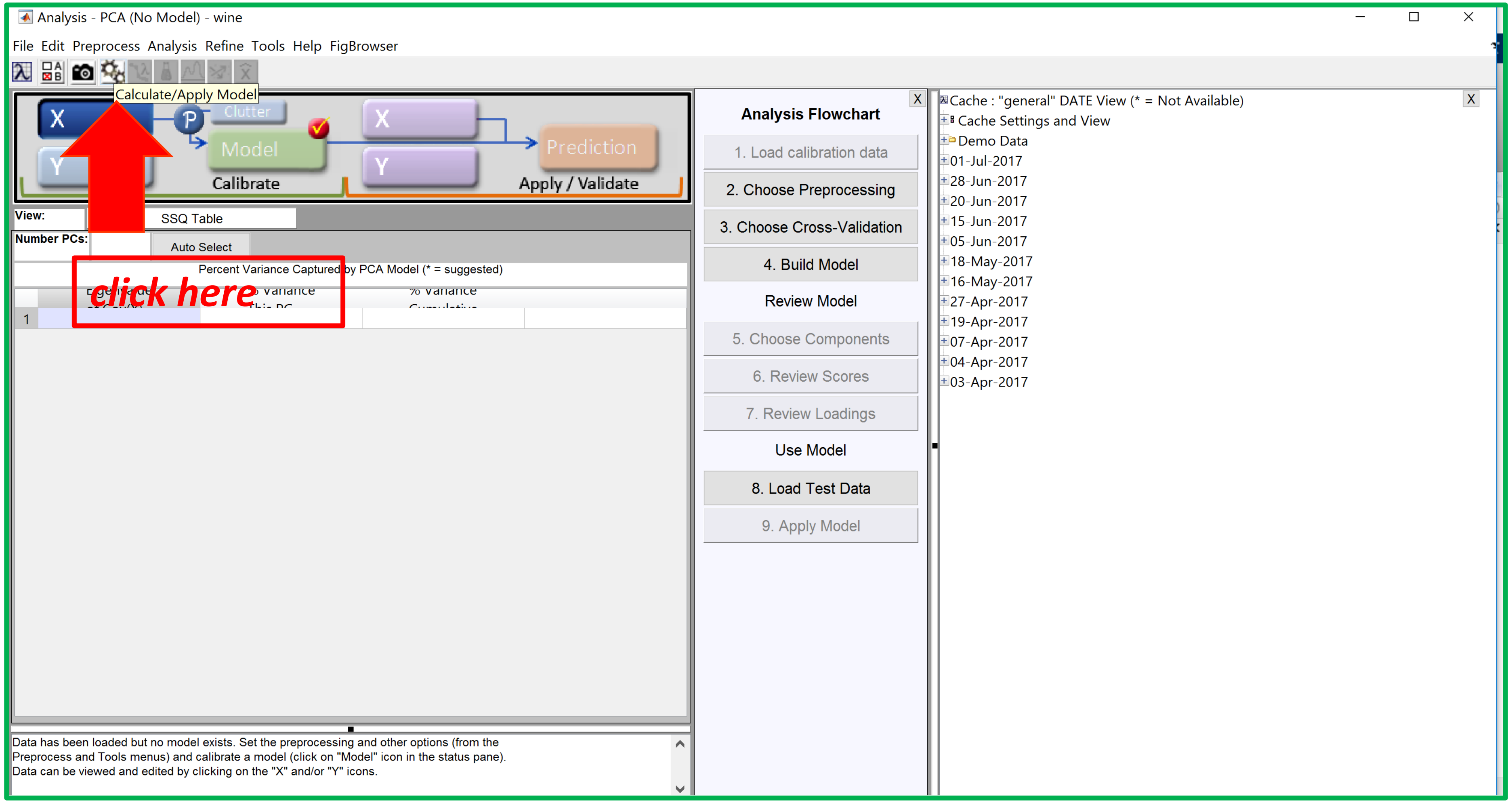
Select pretreatment

- Variable autoscaling
  - mean centering
  - scaling to unit variance



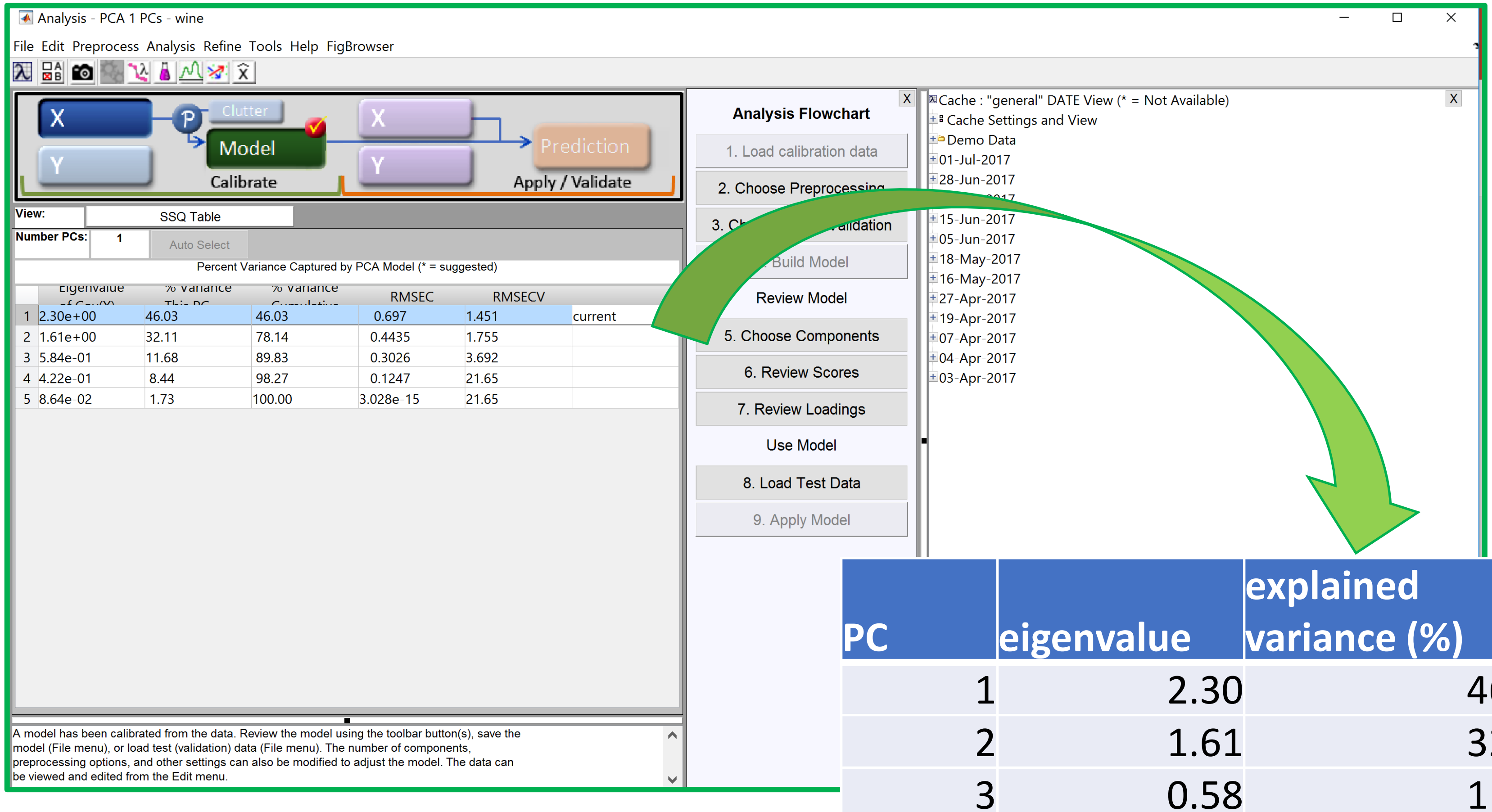
[1 of 1] Caution: Number of splits will leave out a high
percentage (33%) of the calibration data in each cycle. In small data sets, this may lead to
under-estimation of model performance.

PCA model building



Data has been loaded but no model exists. Set the preprocessing and other options (from the Preprocess and Tools menus) and calibrate a model (click on "Model" icon in the status pane).
Data can be viewed and edited by clicking on the "X" and/or "Y" icons.

Eigenvalues and explained variance



View: SSQ Table

Number PCs: 1

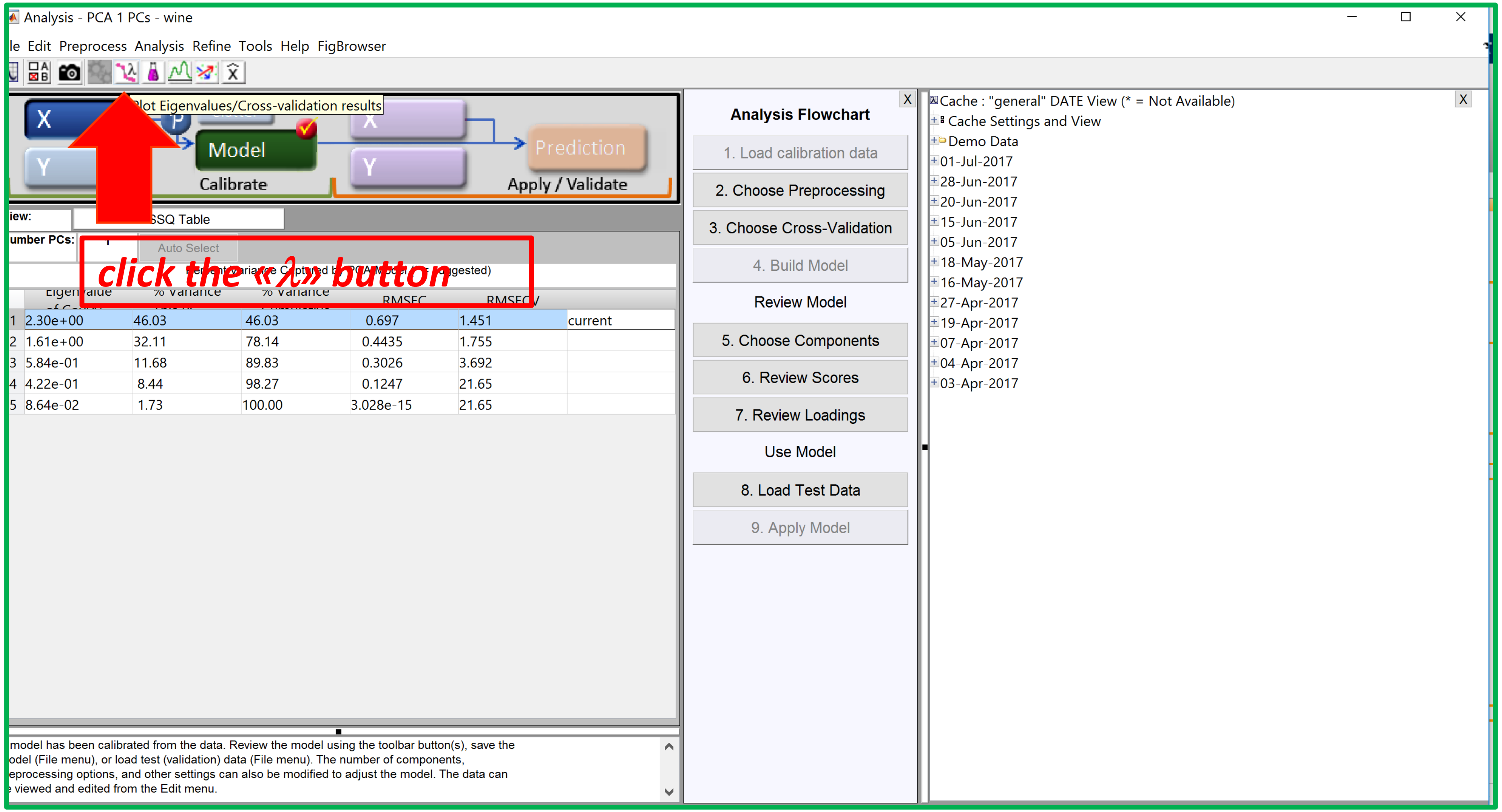
Percent Variance Captured by PCA Model (* = suggested)

	Eigenvalue	% Variance	% Variance	RMSEC	RMSECV	
1	2.30e+00	46.03	46.03	0.697	1.451	current
2	1.61e+00	32.11	78.14	0.4435	1.755	
3	5.84e-01	11.68	89.83	0.3026	3.692	
4	4.22e-01	8.44	98.27	0.1247	21.65	
5	8.64e-02	1.73	100.00	3.028e-15	21.65	

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be modified to adjust the model. The data can be viewed and edited from the Edit menu.

PC	eigenvalue	explained variance (%)	cumulative explained variance (%)
1	2.30	46.03	46.03
2	1.61	32.11	78.14
3	0.58	11.68	89.82
4	0.42	8.44	98.27
5	0.09	1.73	100.00

Select the right number of PCs



iew: SSQ Table

umber PCs: 1

click the «λ» button

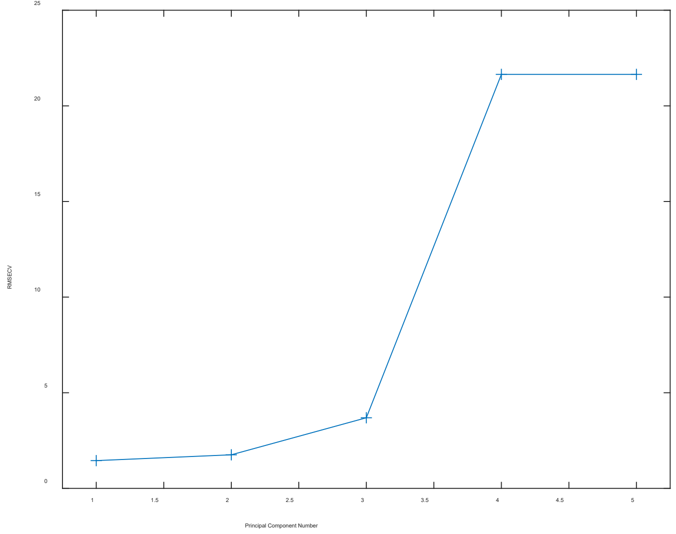
	Eigenvalue	% Variance	% Variance	RMSEC	RMSECV	
1	2.30e+00	46.03	46.03	0.697	1.451	current
2	1.61e+00	32.11	78.14	0.4435	1.755	
3	5.84e-01	11.68	89.83	0.3026	3.692	
4	4.22e-01	8.44	98.27	0.1247	21.65	
5	8.64e-02	1.73	100.00	3.028e-15	21.65	

model has been calibrated from the data. Review the model using the toolbar button(s), save the odel (File menu), or load test (validation) data (File menu). The number of components, eprocessing options, and other settings can also be modified to adjust the model. The data can viewed and edited from the Edit menu.

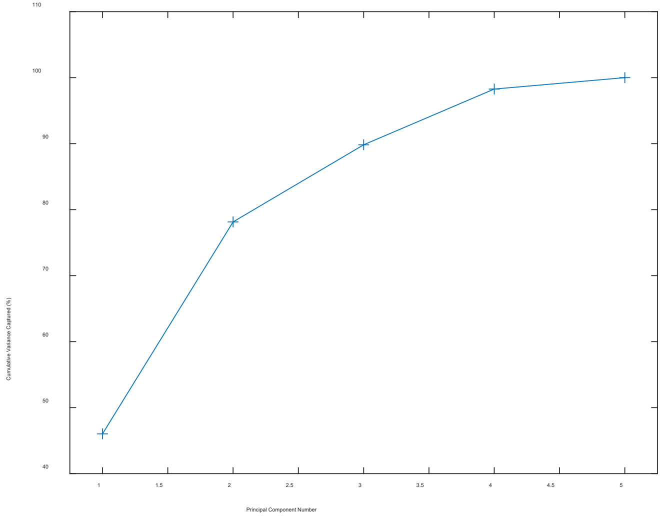
How to select the correct number of PCs

- Eigenvalue >1
- Scree test
  - identify the “elbow” on the figure of the explained variance vs. number of PCs
- Cross validation
  - identify the minimum of the RMSECV

RMSECV

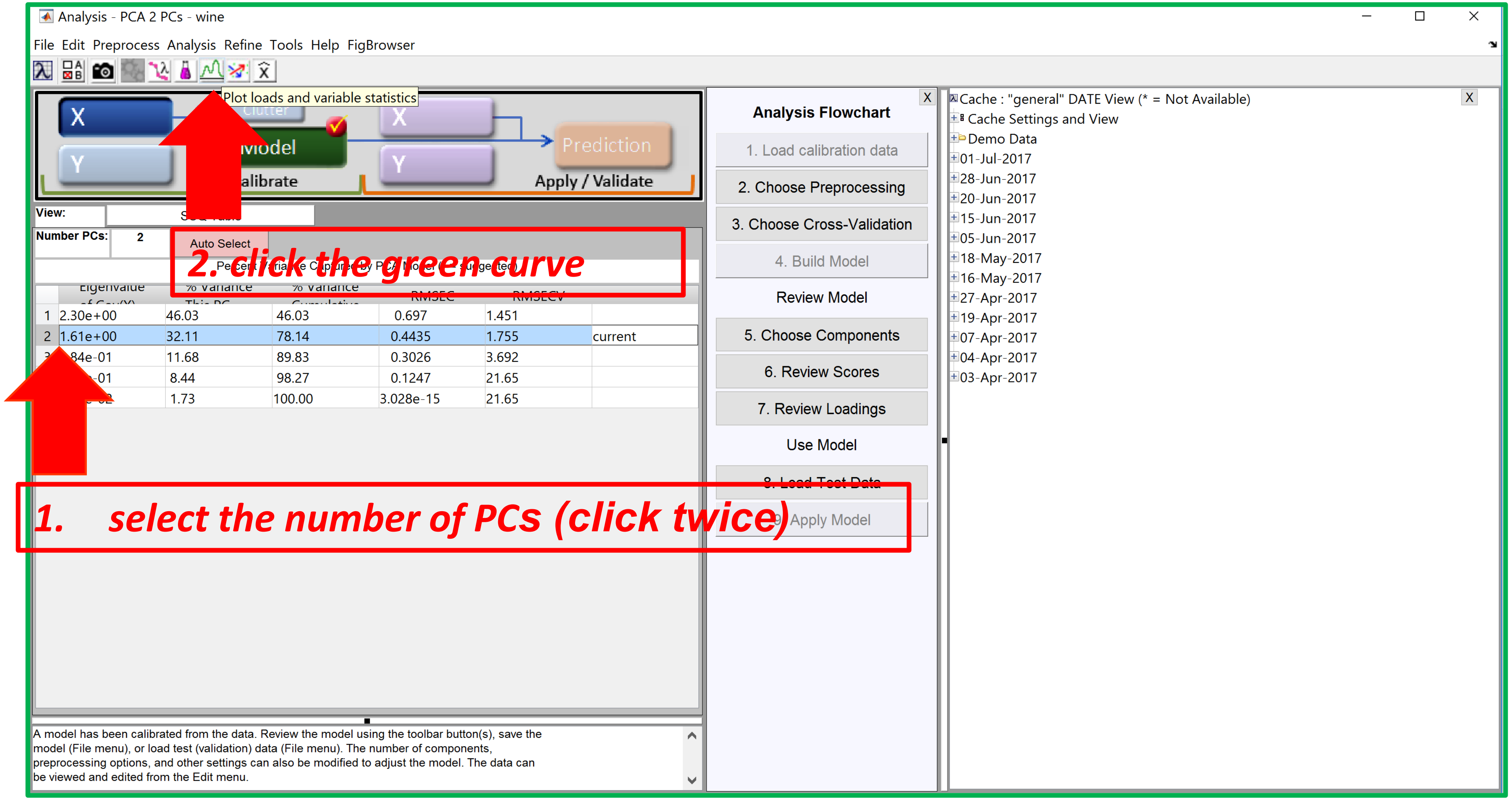


explained variance



Valle, S., W. Li, W., S. J. Qin (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res.* **38**, 4389-4401

Loading plot



Number PCs: 2

Percent Variance Captured by PCA Model (* = suggested)

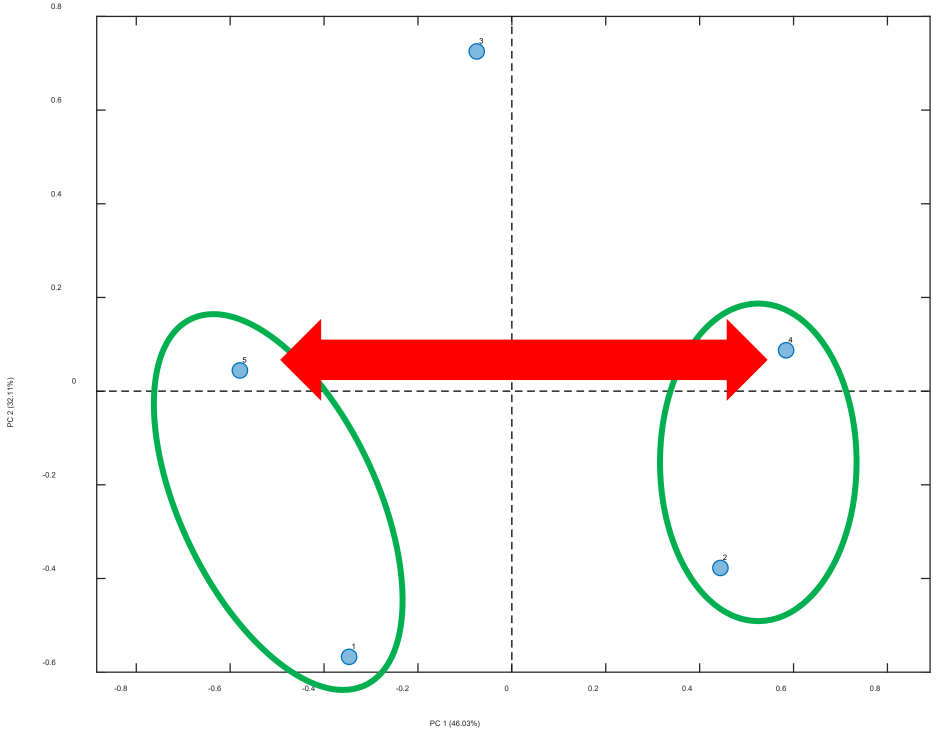
	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	2.30e+00	46.03	46.03	0.697	1.451	
2	1.61e+00	32.11	78.14	0.4435	1.755	current
[illegible]	[illegible]84e-01	11.68	89.83	0.3026	3.692	
[illegible]	[illegible]-01	8.44	98.27	0.1247	21.65	
[illegible]	[illegible]	1.73	100.00	3.028e-15	21.65	

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be modified to adjust the model. The data can be viewed and edited from the Edit menu.

2. click the green curve

1. select the number of PCs (click twice)

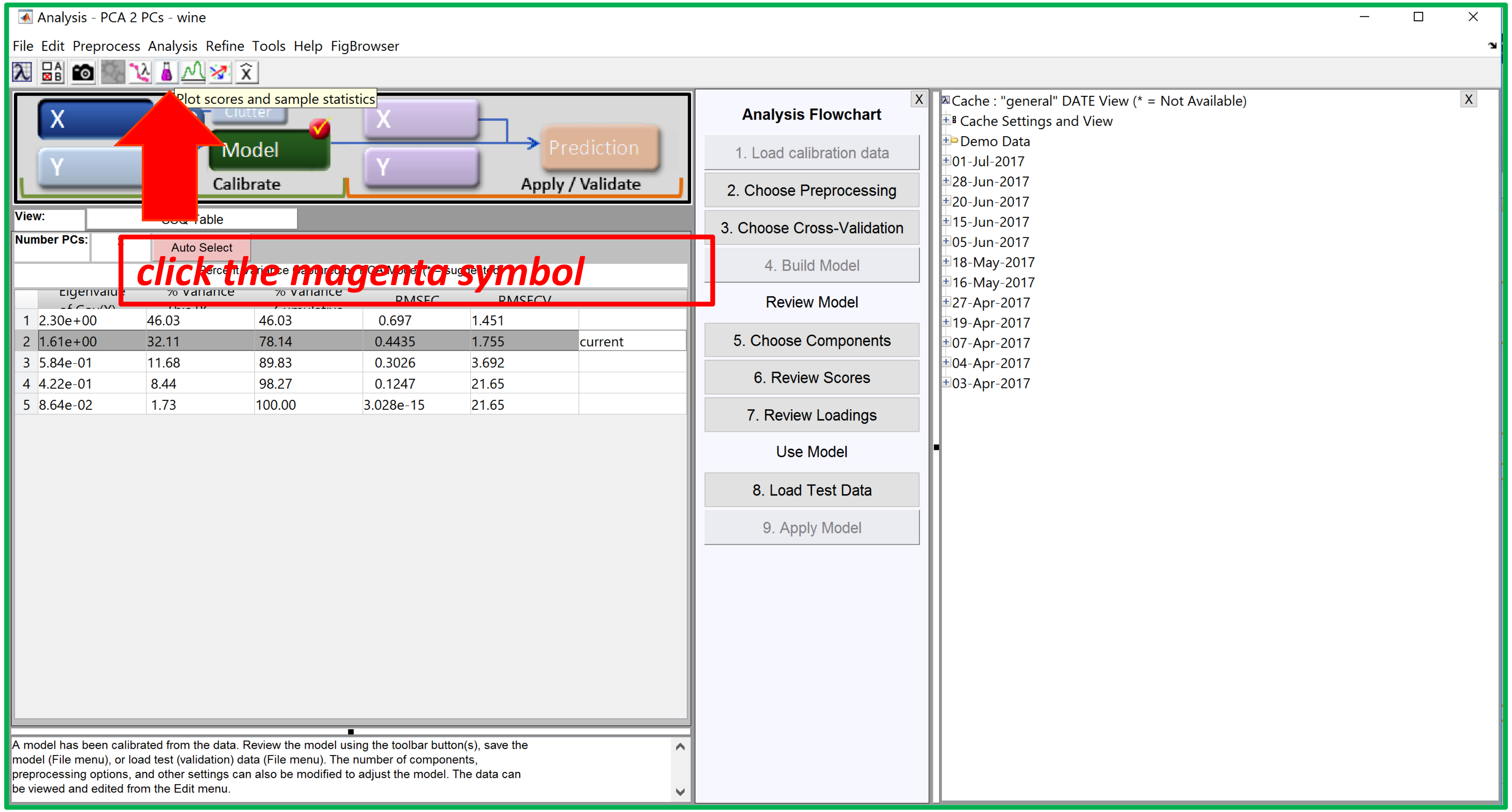
Variables correlation



	variable	units
1	liquor consumption	L/year
2	wine consumption	L/year
3	beer consumption	L/year
4	life expectation	years
5	heart disease frequency	1/(100000 year)

- PC1 (46% explained variance)
  - **positive correlation**
    - variables 2-4: wine consumption – life expectation
    - variables 1-5: liquor consumption – heart disease frequency
  - **anti-correlation**
  - 2-4/1-5: wine consumption(+ life expectation) and heart disease frequency (and liquor consumption)
- PC2 (32%)
  - beer consumption seems to be independent life expectation and heart disease frequency and partially anti-correlated to wine and liquor consumption

Score plot



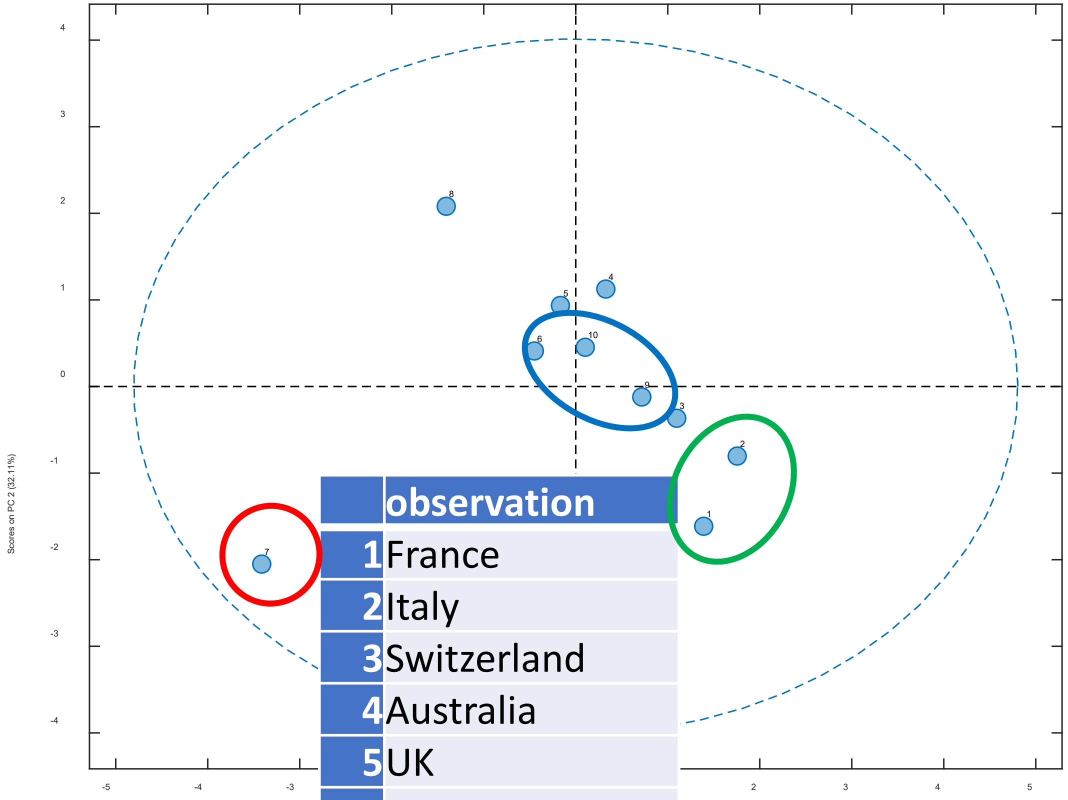
click the magenta symbol

Percent Variance Captured by PCA Model (* = suggested)

	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	2.30e+00	46.03	46.03	0.697	1.451	
2	1.61e+00	32.11	78.14	0.4435	1.755	current
3	5.84e-01	11.68	89.83	0.3026	3.692	
4	4.22e-01	8.44	98.27	0.1247	21.65	
5	8.64e-02	1.73	100.00	3.028e-15	21.65	

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be modified to adjust the model. The data can be viewed and edited from the Edit menu.

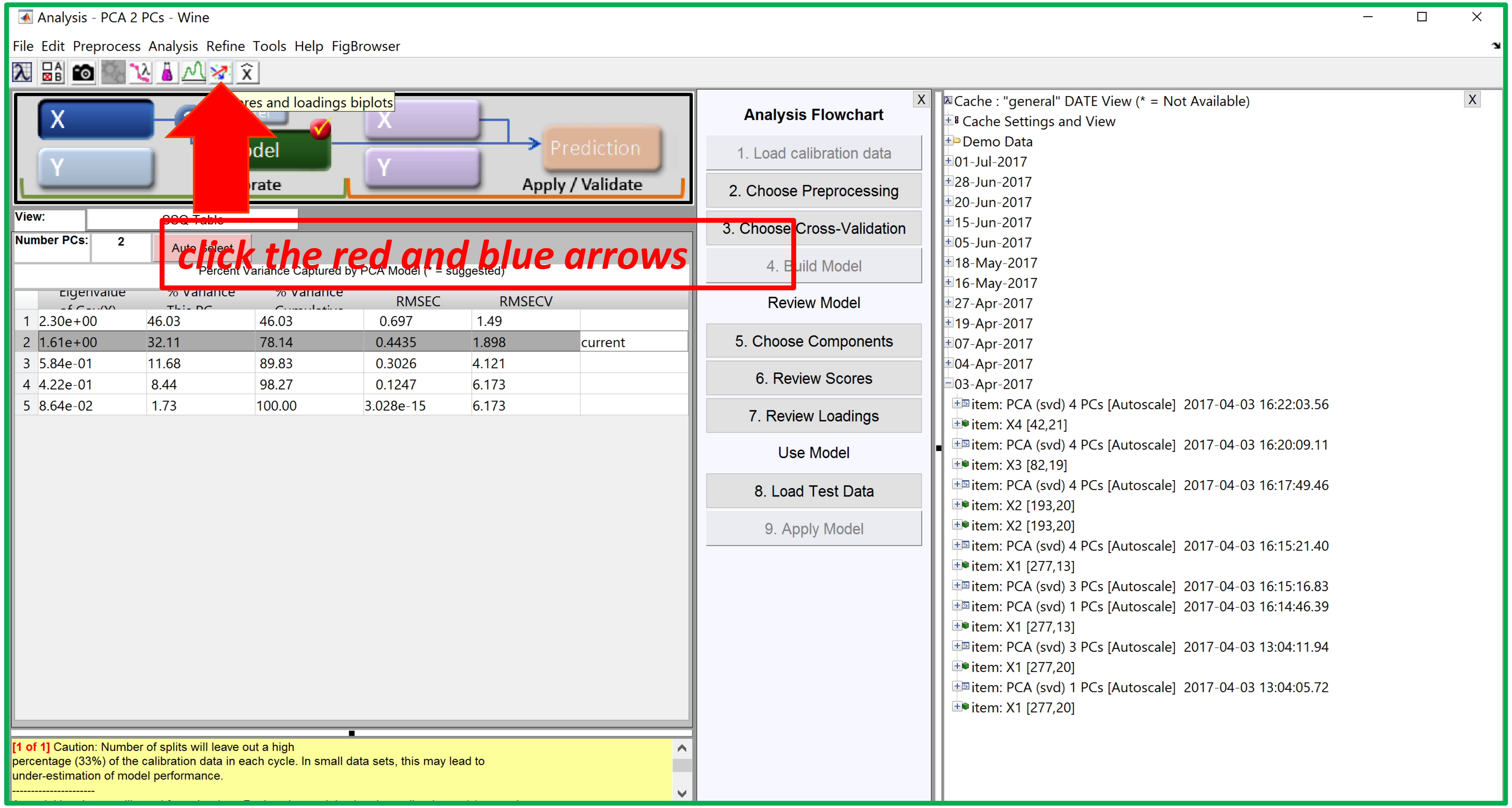
Relations among observations



	observation
1	France
2	Italy
3	Switzerland
4	Australia
5	UK
6	USA
7	Russia
8	Czech Rep.
9	Japan
10	Mexico

- France (1) and Italy (2), are **very similar**: high PC1
- Russia (7) is **very dissimilar** from other nations on both PC1 and PC2
- Czech Rep. (8) display the highest PC2
- USA (6), Japan and Mexico (10) are **“average observations”** because they are close to the origin

Biplot



click the red and blue arrows

Percent Variance Captured by PCA Model (* = suggested)

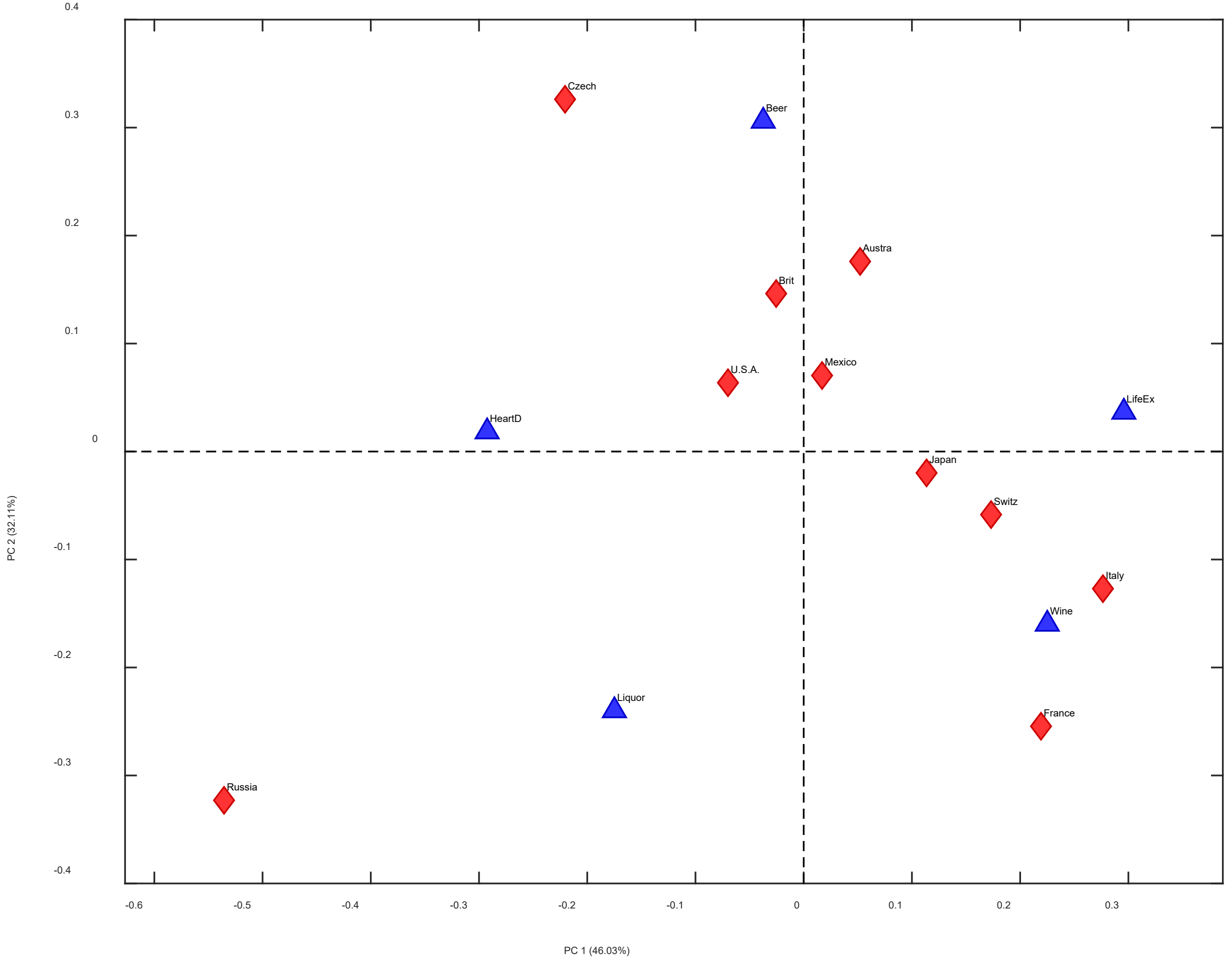
	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	2.30e+00	46.03	46.03	0.697	1.49	
2	1.61e+00	32.11	78.14	0.4435	1.898	current
3	5.84e-01	11.68	89.83	0.3026	4.121	
4	4.22e-01	8.44	98.27	0.1247	6.173	
5	8.64e-02	1.73	100.00	3.028e-15	6.173	

[1 of 1] Caution: Number of splits will leave out a high percentage (33%) of the calibration data in each cycle. In small data sets, this may lead to under-estimation of model performance.

Cache : "general" DATE View (* = Not Available)
- Cache Settings and View
- Demo Data
- 01-Jul-2017
- 28-Jun-2017
- 20-Jun-2017
- 15-Jun-2017
- 05-Jun-2017
- 18-May-2017
- 16-May-2017
- 27-Apr-2017
- 19-Apr-2017
- 07-Apr-2017
- 04-Apr-2017
- 03-Apr-2017
  - item: PCA (svd) 4 PCs [Autoscale] 2017-04-03 16:22:03.56
  - item: X4 [42,21]
  - item: PCA (svd) 4 PCs [Autoscale] 2017-04-03 16:20:09.11
  - item: X3 [82,19]
  - item: PCA (svd) 4 PCs [Autoscale] 2017-04-03 16:17:49.46
  - item: X2 [193,20]
  - item: X2 [193,20]
  - item: PCA (svd) 4 PCs [Autoscale] 2017-04-03 16:15:21.40
  - item: X1 [277,13]
  - item: PCA (svd) 3 PCs [Autoscale] 2017-04-03 16:15:16.83
  - item: PCA (svd) 1 PCs [Autoscale] 2017-04-03 16:14:46.39
  - item: X1 [277,13]
  - item: PCA (svd) 3 PCs [Autoscale] 2017-04-03 13:04:11.94
  - item: X1 [277,20]
  - item: PCA (svd) 1 PCs [Autoscale] 2017-04-03 13:04:05.72
  - item: X1 [277,20]

Biplot: joint score and loading reading

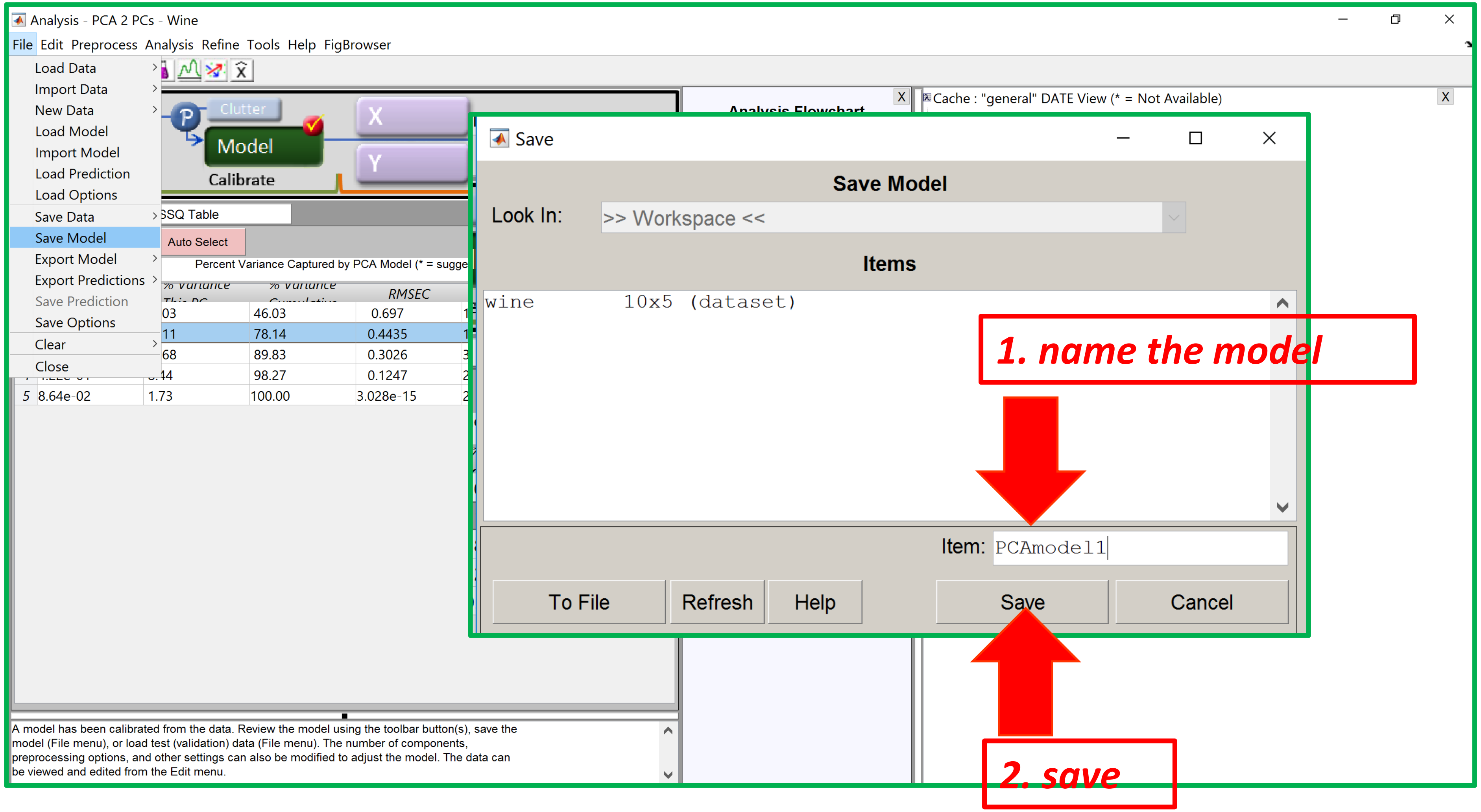
bi-plot = score plot + loading plot



- Japan, Italy, Switzerland and France are the nations with higher life expectation
- France and Italy have the highest wine consumption
- Russia has the lowest life expectation and the highest rate of heart disease and liquor consumption (but low wine and beer consumption)
- Czech Rep. has the highest beer consumption

	liquor consumption	wine consumption	beer consumption	life expectancy	heart disease rate
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58.0	25.1	78	94.1
Switzerland	1.7	46.0	65.0	78	106.4
Australia	1.2	15.7	102.1	78	173.0
Grait britain	1.5	12.2	100.0	77	199.7
USA	2.0	8.9	87.8	76	176.0
Russia	3.8	2.7	17.1	69	373.6
Czech Republic	1.0	1.7	140.0	73	283.7
Japan	2.1	1.0	55.0	79	34.7
Mexico	0.8	0.2	50.4	73	36.4

Save PCA model



Look In: >> Workspace <<

wine 10x5 (dataset)

Item: PCAmodel1

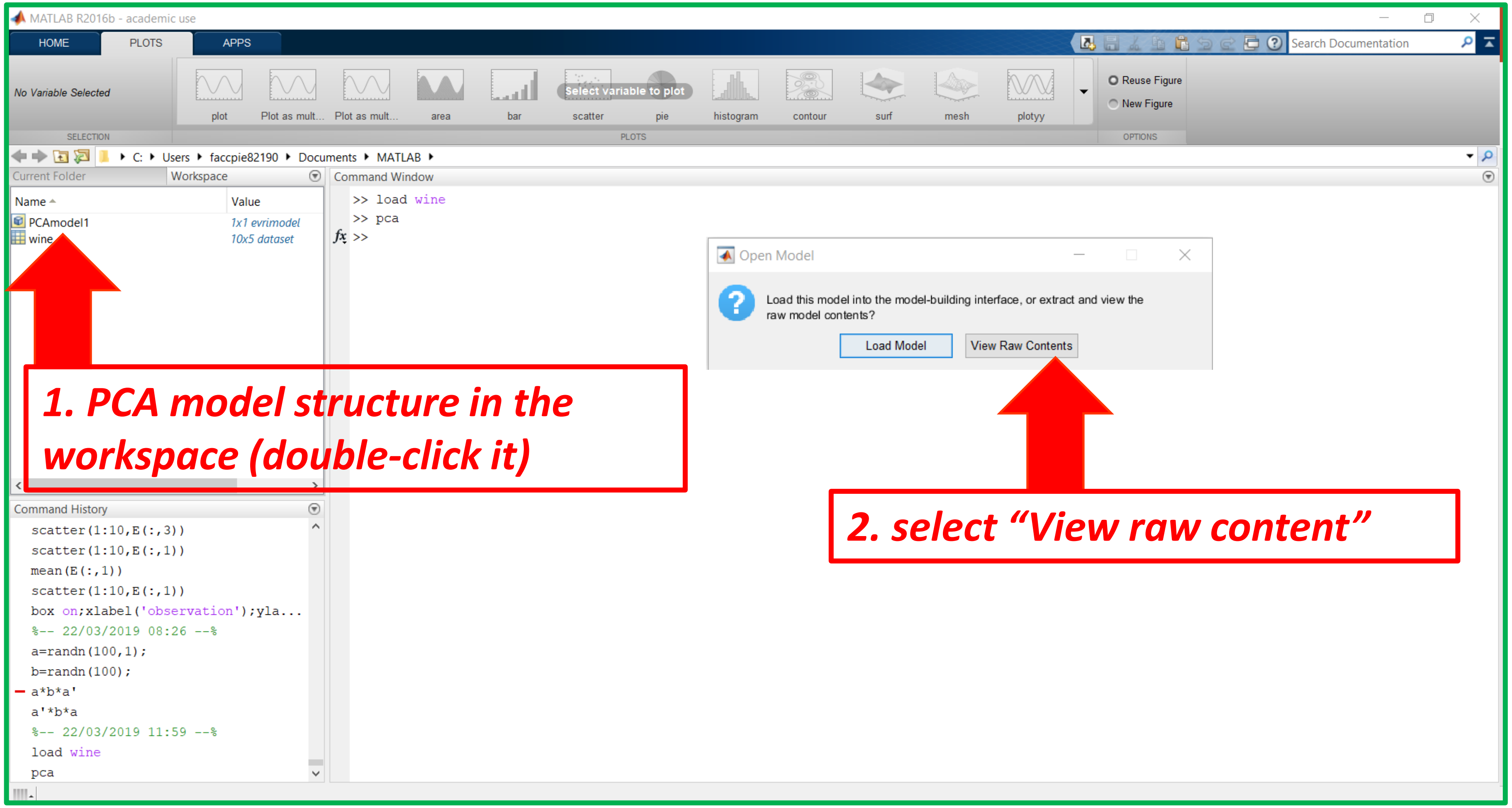
		% Variance This PC	% Variance Cumulative	RMSEC
		03	46.03	0.697
		11	78.14	0.4435
		68	89.83	0.3026
		.14	98.27	0.1247
5	8.64e-02	1.73	100.00	3.028e-15

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be modified to adjust the model. The data can be viewed and edited from the Edit menu.

1. name the model

2. save

PCA model is saved in a data structure



Name	Value
PCAmodel1	1x1 evrimodel
wine	10x5 dataset

```
>> load wine
>> pca
fx >>
```

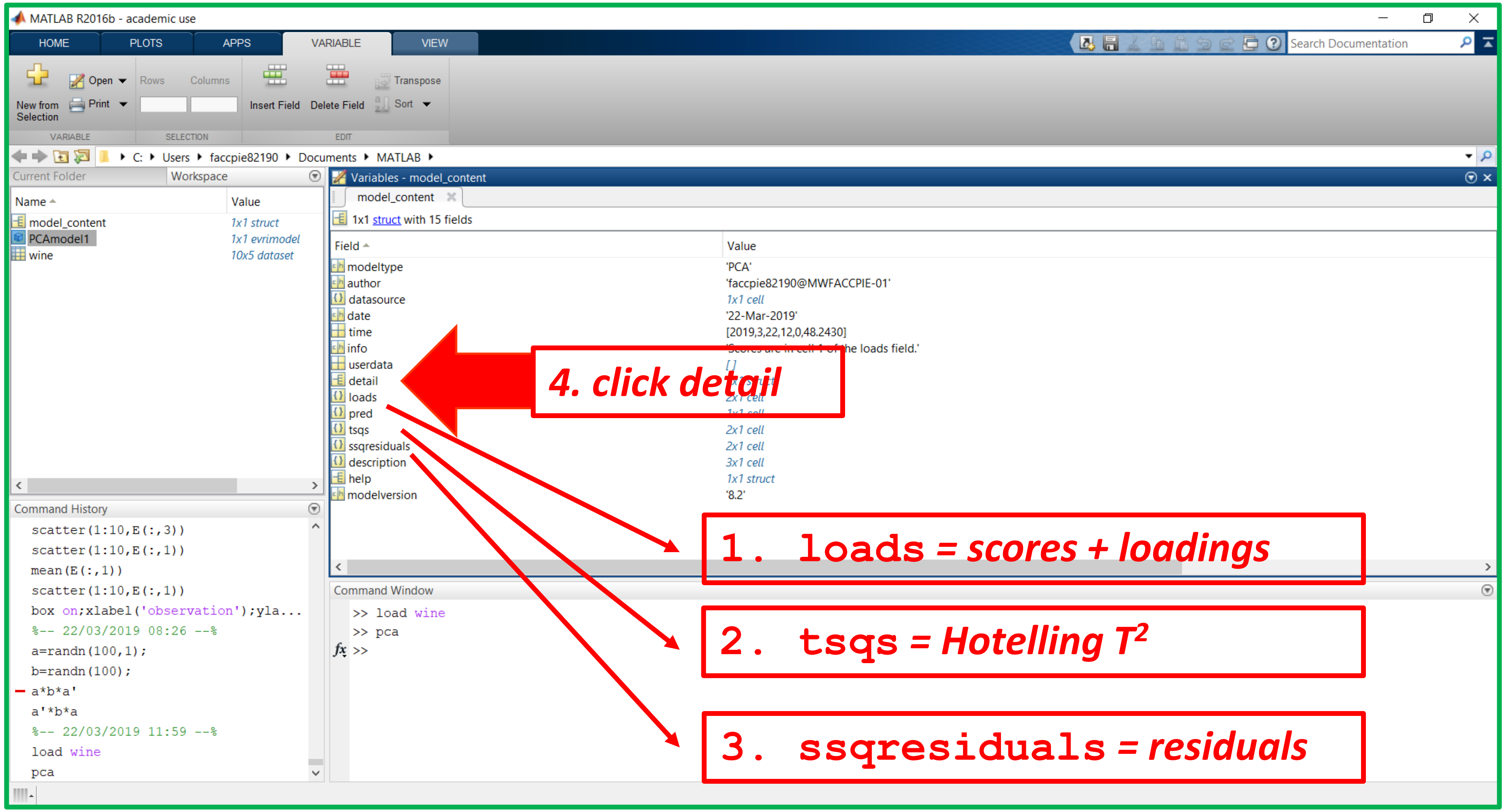
Load this model into the model-building interface, or extract and view the raw model contents?

```
scatter(1:10,E(:,3))
scatter(1:10,E(:,1))
mean(E(:,1))
scatter(1:10,E(:,1))
box on;xlabel('observation');yla...
%-- 22/03/2019 08:26 --%
a=randn(100,1);
b=randn(100);
a*b*a'
a'*b*a
%-- 22/03/2019 11:59 --%
load wine
pca
```

1. PCA model structure in the workspace (double-click it)

2. select "View raw content"

Access details on PCA model



Name	Value
model_content	*1x1 struct*
PCAmodel1	*1x1 evrimodel*
wine	*10x5 dataset*

1x1 struct with 15 fields

Field	Value
modeltype	'PCA'
author	'faccpie82190@MWFACCPIE-01'
datasource	*1x1 cell*
date	'22-Mar-2019'
time	[2019,3,22,12,0,48.2430]
info	'Scores are in cell 1 of the loads field.'
userdata	*[]*
detail	*1x1 struct*
loads	*2x1 cell*
pred	*1x1 cell*
tsqs	*2x1 cell*
ssqresiduals	*2x1 cell*
description	*3x1 cell*
help	*1x1 struct*
modelversion	'8.2'

Command History

```
scatter(1:10,E(:,3))
scatter(1:10,E(:,1))
mean(E(:,1))
scatter(1:10,E(:,1))
box on;xlabel('observation');yla...
%-- 22/03/2019 08:26 --%
a=randn(100,1);
b=randn(100);
a*b*a'
a'*b*a
%-- 22/03/2019 11:59 --%
load wine
pca
```

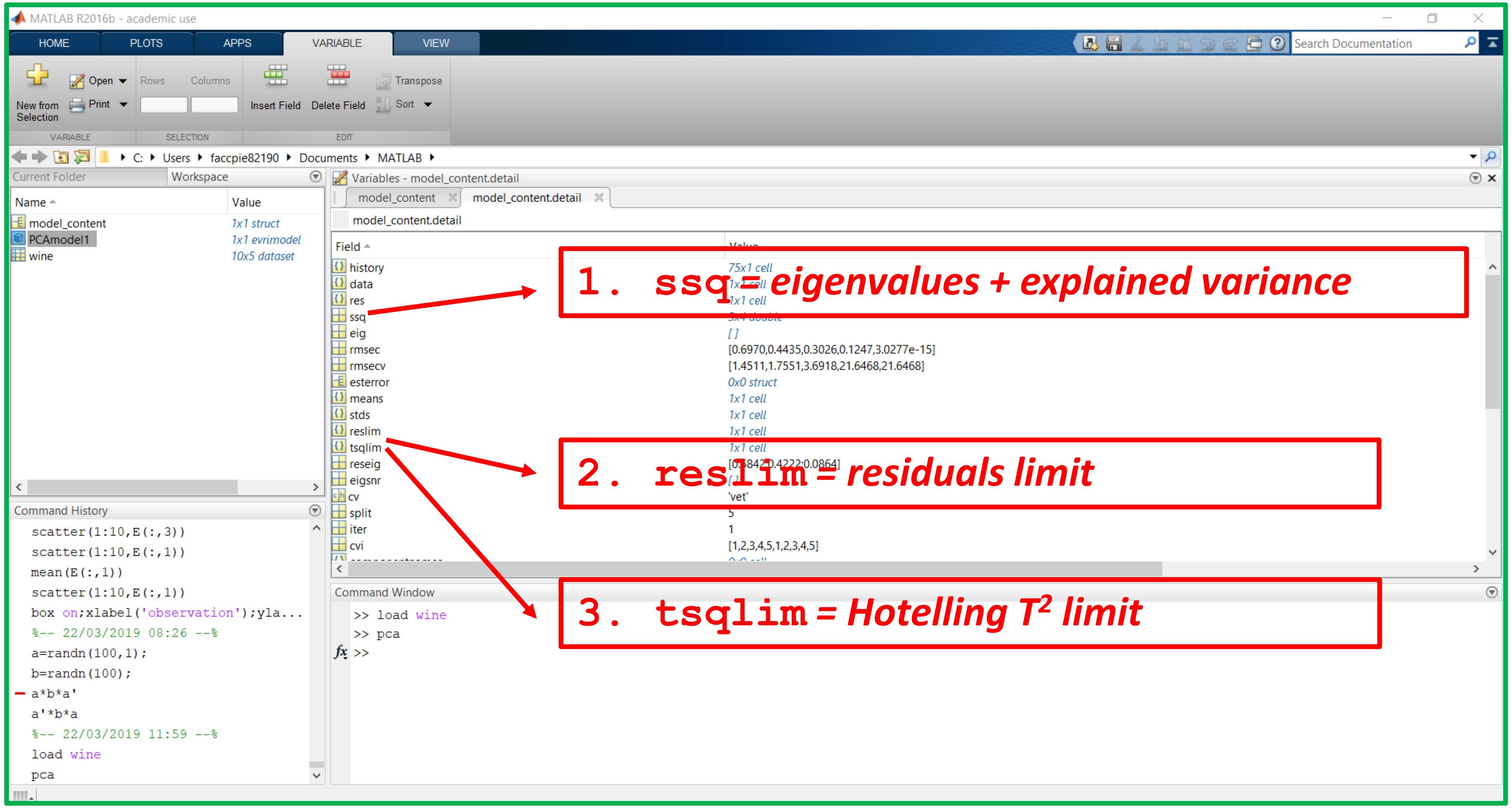
Command Window

```
>> load wine
>> pca
fx >>
```

4. click detail

1. `loads` = *scores + loadings*
2. `tsqs` = *Hotelling T^2*
3. `ssqresiduals` = *residuals*

Other details about PCA



Name	Value
model_content	1x1 struct
PCAmodel1	1x1 evrimodel
wine	10x5 dataset

model_content.detail

Field	Value
history	75x1 cell
data	1x1 cell
res	1x1 cell
ssq	5x4 double
eig	[ ]
rmsec	[0.6970,0.4435,0.3026,0.1247,3.0277e-15]
rmsecv	[1.4511,1.7551,3.6918,21.6468,21.6468]
esterror	0x0 struct
means	1x1 cell
stds	1x1 cell
reslim	1x1 cell
tsqlim	1x1 cell
reseig	[0.5842;0.4222;0.0864]
eigsnr	[ ]
cv	'vet'
split	5
iter	1
cvi	[1,2,3,4,5,1,2,3,4,5]

```
scatter(1:10,E(:,3))
scatter(1:10,E(:,1))
mean(E(:,1))
scatter(1:10,E(:,1))
box on;xlabel('observation');yla...
%-- 22/03/2019 08:26 --%
a=randn(100,1);
b=randn(100);
a*b*a'
a'*b*a
%-- 22/03/2019 11:59 --%
load wine
pca
```

```
>> load wine
>> pca
fx >>
```

1. **ssq** = ***eigenvalues + explained variance***

2. **reslim** = ***residuals limit***

3. **tsqlim** = ***Hotelling T^2 limit***

Model details and flexibility

- The model is available in the Matlab® *Workspace* in a *structure* with the assigned name
- A lot of useful information are stored in the structure:
  - PC explained variance
  - scores and loadings
  - model and sample diagnostics
  - confidence limits
  - etc…
- The model building and processing is **much faster writing and running Matlab® code**

PCA with PLS_Toolbox® and Matlab® commands

```
clear all
close all
clc

% load alcohol consumption dataset
load wine

% data visualization
X=double(wine);
bar3(wine);xlabel('variable');ylabel('country');

% PCA with the PLS Toolbox
opca.display='off';
opca.plots='none';
opca.preprocessing='autoscale';
pca1=pca(X,2,opca);

% scores
T=pca1.loads{1,1};
figure;scatter(T(:,1),T(:,2));box on;xlabel('PC1 score');ylabel('PC2 score')

% loadings
P=pca1.loads{2,1};
```

```
% PCA matrix reconstruction
Xhat=T*P';
Xa=auto(X);

% residuals
E=Xa-Xhat;

% eigenvalues and explained variance
lambda=pca1.detail.ssq(:,2);
2.3014/sum(lambda)
lambda./sum(lambda)
lambda./sum(lambda)*100
cumsum(lambda./sum(lambda)*100)

% Q residuals
Q=sum(E.^2,2);

% Hotelling T^2
T2=diag(T*inv(diag(lambda(1:2)))*T');

% singular value decomposition
[U,Lsvd,Psvd]=svd(cov(auto(X)));
Tsvd=Xa*Psvd;
```

Today homework

- Replicate through commands what we did with the PLS_Toolbox graphic user interface
  - `[PCAmodel1]=pca(X,2);`
  - `[predict]=pca(Xnew,PCAmodel1);`
- Practice with command
  - `svd`
- Extract all the mathematical entities:
  - scores
  - loadings
  - residuals
  - eigenvalues
  - etc…

… per sempre a fianco a me!