Topic 7. Cluster Analysis

Case 4: Value at Risk

using Daily Observations on Asset Prices

Reading Assignment

Berry and Linoff (2000)

- Pages 104–110. K-means.

Topic Objective

- Describe the problem
- Describe a visualization method
- Describe cluster analysis tools
- Apply to the value at risk problem
- Evaluate results

Value at Risk

The number of assets in a value at risk computation usually exceeds 100, which limits the methods that can be employed.

The classical approach is to estimate the mean vector and the variance-covariance matrix of daily asset returns and use an assumption of multivariate normality to compute the 5% quantile of a portfolio comprised of a weighted sum of the assets.

The mean and variance can be estimated unconditionally. Or a more sophisticated conditional analysis using a GARCH specification of the variance-covariance matrix can be used.

Value at Risk

Some approaches robustify the analysis by using conditional means and quantiles or by using fat-tailed distributions instead of the normal.

The most sophisticated approaches use extreme value theory. But the extreme value theory approach is severely hampered by the lack of an adequate multivariate distribution theory.

All of these approaches are model based and therefore at risk of serious error if model assumptions are violated.

Being model based, they suffer from the curse of dimensionality and all have limits on the number of assets that can be analyzed.

Conditional Behavior

One interesting fact that has been noted in the literature on value at risk is that asset correlations are different in expansions and contractions of the economy.

Similarly for bull and bear markets.

Value at Risk

We shall use data mining techniques – clustering methods in particular – to gain insights on value at risk issues.

We shall take the view of a leveraged investor who does not want to risk either bankruptcy or an involuntarily liquidation of securities in a market crash.

We shall also see if asset groupings change in rallies.

Most importantly, our methods are not model based and there is no limit on the number of assets that can be considered.

The Data

The data are one day holding period returns from 1971 through 2006 on the following assets:

- Treasury bonds at 1, 3, 5, 7, and 10 year maturities
- Euro to Dollar and Yen to Dollar exchange rates
- The Fama-French value weighted portfolios
- Nasdaq, NYSE, and S&P500 index portfolios
- Value weighted two-digit SIC Code portfolios

The Data

All data, data sources, code, graphics, etc. are on the website in directory (folder)

datamine/cases/assets

There are 79 portfolios and 8442 days of returns. The data contain a few missing values.

Severe missing value problems account for the lack of short term Treasury debt, commercial paper, and corporate debt.

Data Structure

1. The features are the returns for each trading day; $p = 8442$.

2. The cases are the portfolios; $n = 79$.

Labeling Conventions

1. Days in a yyyymmdd numerical format label the features.

2. Previously, the targets labeled the cases, e.g. "default", "paid in full". Here each case has an individual label. They are ...

Table 25. Case Labels, 1 of 4

sp500	S&P500 index portfolio
nyse	NYSE index portfolio
nasdaq	Nasdaq index portfolio
tcm1y	One year US Treasury bill
tcm3y	Three year US Treasury bond
tcm5y	Five year US Treasury bond
tcm7y	Seven year US Treasury bond
tcm10y	Ten year US Treasury bond
yen	Japanese Yen to US Dollar exchange rate
euro	First DM then Euro to US Dollar exchange rate
SL	Fama-French small cap, low book to market
SM	Fama-French small cap, medium book to market
SH	Fama-French small cap, high book to market
BL	Fama-French large cap, low book to market
BM	Fama-French large cap, medium book to market
BH	Fama-French large cap, high book to market
S01	Agricultural production-crops
S02	Agricultural production-livestock
S07	Agricultural services
S08	Forestry
S09	Fishing, hunting, and trapping
S10	Metal mining
S12	Coal mining
S13	Oil and gas extraction
S14	Nonmetallic minerals, except fuels

Table 25 (continued). Case Labels, 2 of 4

S15	General building contractors
S16	Heavy construction contractors
S17	Special trade contractors
S20	Food and kindred products
S21	Tobacco manufactures
S22	Textile mill products
S23	Apparel and other textile products
S24	Lumber and wood products
S25	Furniture and fixtures
S26	Paper and allied products
S27	Printing and publishing
S28	Chemicals and allied products
S29	Petroleum and coal products
S30	Rubber and miscellaneous plastics products
S31	Leather and leather products
S32	Stone, clay, glass, and concrete products
S33	Primary metal industries
S34	Fabricated metal products
S35	Industrial machinery and equipment
S36	Electrical and electronic equipment
S37	Transportation equipment
S38	Instruments and related products
S39	Miscellaneous manufacturing industries
S41	Local and interurban passenger transit
S42	Motor freight transportation and warehousing
S43	U.S. Postal Service

Table 25 (continued). Case Labels, 3 of 4

S44	Water transportation
S45	Transportation by air
S46	Pipelines, except natural gas
S47	Transportation services
S48	Communications
S49	Electric, gas, and sanitary services
S50	Wholesale trade--durable goods
S51	Wholesale trade--nondurable goods
S52	Building materials, hardware, garden supply, & mobile
S53	General merchandise stores
S54	Food stores
S55	Automotive dealers and gasoline service stations
S56	Apparel and accessory stores
S57	Furniture, home furnishings and equipment stores
S58	Eating and drinking places
S59	Miscellaneous retail
S60	Depository institutions
S61	Nondepository credit institutions
S62	Security, commodity brokers, and services
S63	Insurance carriers
S64	Insurance agents, brokers, and service
S65	Real estate
S67	Holding and other investment offices
S70	Hotels, rooming houses, camps, and other lodging place
S72	Personal services
S73	Business services

Table 25 (continued). Case Labels, 4 of 4

S75	Automotive repair, services, and parking
S76	Miscellaneous repair services
S78	Motion pictures
S79	Amusement and recreational services
S80	Health services
S81	Legal services
S82	Educational services
S83	Social services
S84	Museums, art galleries, botanical & zoological garden
S86	Membership organizations
S87	Engineering and management services
S88	Private households
S89	Miscellaneous services
S91	Executive, legislative, and general government
S92	Justice, public order, and safety
S93	Finance, taxation, and monetary policy
S94	Administration of human resources
S95	Environmental quality and housing
S96	Administration of economic programs
S97	National security and international affairs

Derived Features

We are interested in portfolio performance in crashes and rallies.

For rallies, these will be the returns on the S&P 500 that exceed 4.5%.

For crashes, these will be the returns on the S&P 500 that are less than -4.5%.

These features are in the following data sets:

datamine/cases/assets/rally.csv
datamine/cases/assets/crash.csv

They have the comma separated value format, which makes them easy to read in Excel. Here are some examples ...

Table 26. Some Crash Features and Cases

yyyy	1986	1987	1987	1987	1988	1989	1997	1998	2000	2001
mmdd	0911	1016	1019	1026	0108	1013	1027	0831	0414	0917
sp500	-4.82	-5.18	-19.57	-8.34	-6.75	-6.12	-6.77	-6.83	-5.97	-4.90
nyse	-4.47	-4.91	-18.35	-8.10	-6.00	-5.72	-6.39	-6.15	-5.26	-4.68
BL	-4.71	-5.35	-17.91	-9.05	-6.84	-6.12	-6.80	-7.69	-6.21	-4.64
BM	-4.49	-4.94	-18.80	-7.63	-5.86	-5.40	-6.48	-5.10	-5.07	-5.43
BH	-4.37	-4.26	-17.54	-7.00	-4.62	-5.26	-4.96	-4.07	-4.81	-6.30
nasdaq	-3.73	-3.85	-11.32	-8.96	-3.17	-3.10	-7.17	-8.66	-9.81	-6.82
SL	-4.21	-4.50	-13.03	-10.36	-3.48	-3.96	-7.40	-7.1	-8.57	-5.43
SM	-2.81	-3.90	-11.11	-7.21	-2.60	-3.45	-5.13	-5.35	-5.52	-4.56
SH	-3.04	-4.09	-10.87	-7.51	-2.55	-3.05	-4.80	-5.05	-4.76	-5.61
tcm10y	-2.05	-0.45	0.73	1.65	-1.29	1.57	0.94	0.38	0.85	-0.57
tcm5y	-1.03	-0.09	0.68	1.01	-0.60	0.88	0.71	0.14	0.61	-0.34
tcm1y	-0.09	0.04	0.37	0.19	-0.10	0.08	0.16	0.01	0.06	0.01
yen	1.03	0.42	-0.65	-0.46	0.44	-0.70	0.11	-1.16	-0.61	0.18
euro	1.36	0.03	-1.33	-0.70	0.70	-0.88	-0.66	-0.76	-0.42	0.01

Table 27. Some Rally Features and Cases

yyyy	1974	1982	1987	1987	1997	1998	2000	2001	2002	2002	2002
mmdd	1009	0817	1021	1029	1028	0908	0316	0103	0724	0729	1015
sp500	4.64	4.79	8.86	4.89	4.96	5.08	4.73	5.06	5.75	5.43	4.73
nyse	4.49	4.45	8.79	4.49	3.94	4.59	4.78	2.73	5.28	5.23	4.40
BM	3.49	4.34	8.28	4.16	3.81	4.42	4.77	1.13	4.88	5.83	4.78
BH	3.27	4.76	7.61	3.36	2.13	3.41	5.26	-0.06	4.38	4.45	4.91
SL	2.44	2.23	8.58	5.49	3.05	4.74	1.06	5.47	4.11	4.85	4.27
SM	1.86	1.92	6.28	3.27	1.36	3.19	1.93	2.79	3.55	4.64	3.87
SH	1.50	2.20	6.54	3.01	1.34	2.33	1.40	2.14	2.59	4.12	3.53
nasdaq	2.52	1.72	7.36	5.57	4.60	6.03	2.90	14.27	5.00	5.81	5.12
BL	5.18	4.50	9.65	6.06	5.23	5.50	4.80	6.45	6.09	5.41	4.57
tcm10y	0.00	3.90	0.91	1.10	-0.66	-0.19	0.28	-2.09	-0.19	-1.82	-2.31
tcm1y	-0.19	0.50	0.11	0.06	-0.11	0.00	0.01	0.07	0.02	-0.13	-0.10
tcm5y	-0.23	1.73	0.60	0.46	-0.48	0.05	0.14	-0.86	-0.05	-1.06	-1.36
yen	-0.36	0.00	0.21	-0.67	-1.80	-1.42	0.05	-0.41	-0.88	0.91	0.50
euro	-1.81	-0.52	0.30	-1.41	-1.91	-0.53	-0.15	-0.09	-0.12	1.05	0.45

Moving On

This concludes our discussion of the case and data per se.

We shall now consider the principal components visualization method.

Visualization Methods

It is hard to make progress if one cannot see what one is doing.

Therefore, most cluster analysis tools have visualization methods associated with them, as we shall see.

Principal components is a visualization method that does not rely on a tool. Consequently, it is good place to start.

Principal Components

The goal of principal components is to find a new coordinate system that reveals the structure of the data:

$$(x_1, x_2) \to (c_1, c_2) \quad \text{two features}$$

$$(x_1, x_2, x_3) \to (c_1, c_2, c_3) \quad \text{three features}$$

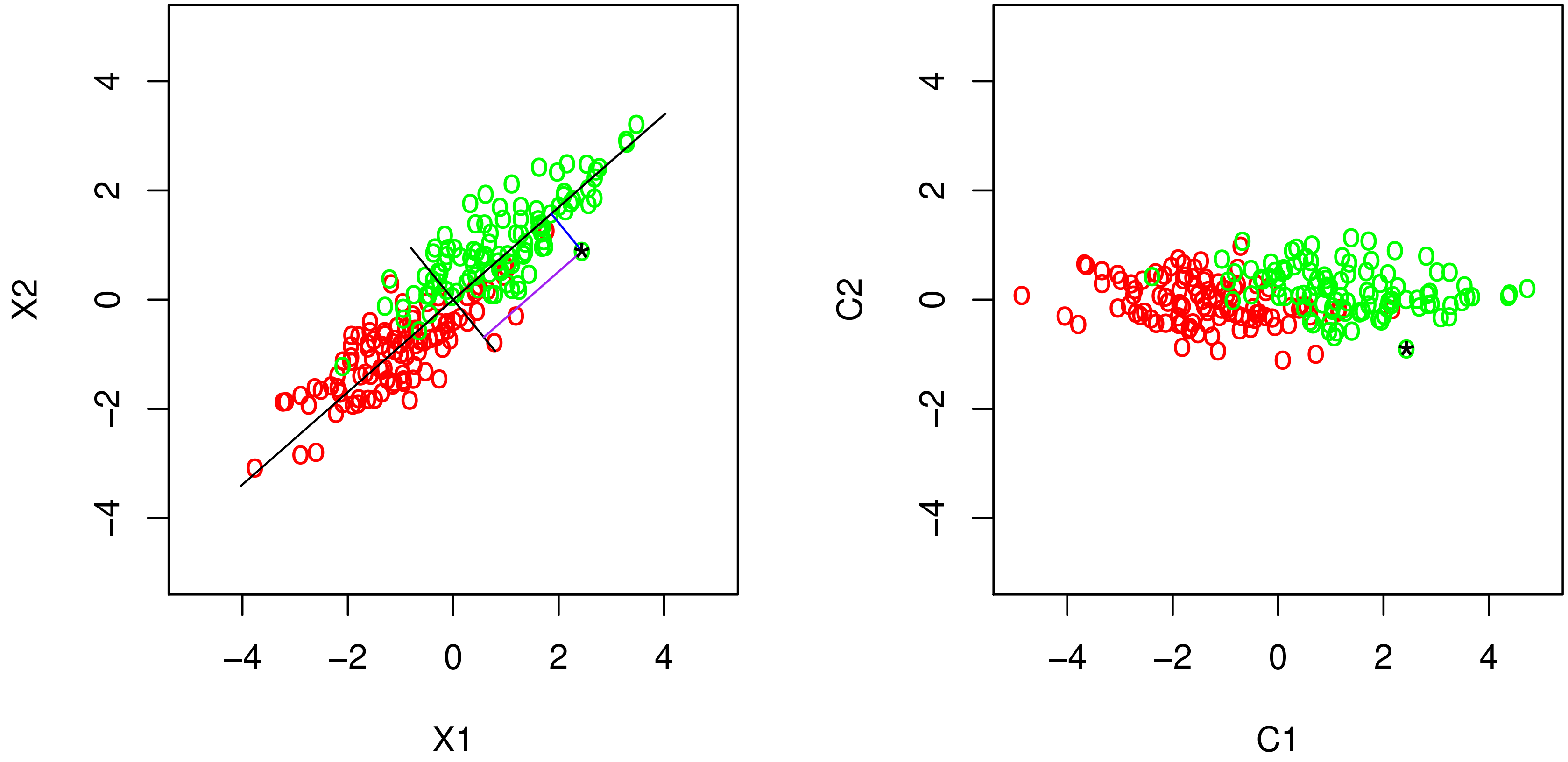
The c_i are the projections onto the new coordinate system and are called the principal componant values or principal components for short. We can think of them as derived features.

The new coordinate system has these propertiess

1. The axes of the new coordiate system are at right angles to each other in the old coordinate system.

2. $\text{Var}(c_1) > \text{Var}(c_2) > \text{Var}(c_3)$.

Fig 75. Principal Components: Two Dimensions



The values (x_1, x_2) of two features for 200 cases from two clusters, red and green, are plotted in the left panel. The new axes are the black lines. The value c_1 for a point with values (x_1, x_2) is computed by dropping a perpendicular from (x_1, x_2), black star, to the longer axis, blue line. Similarly, c_2 for the shorter axis, purple line. The principal component values (c_1, c_2) are plotted in the right panel.

Principal Component Loadings

The principal component values (c_1, c_2) in Figure 75 are computed using a formula of the form

$$\begin{aligned} c_1 &= w_{11}x_1 + w_{12}x_2 \\ c_2 &= w_{21}x_1 + w_{22}x_2 \end{aligned}$$

The weights w_{ij} are called principal component loadings.

They are also called principal component directions because they define the axes in the left panel of Figure 75; i.e., the right endpoint of the long axis is $(w_{11},\, w_{12})$.

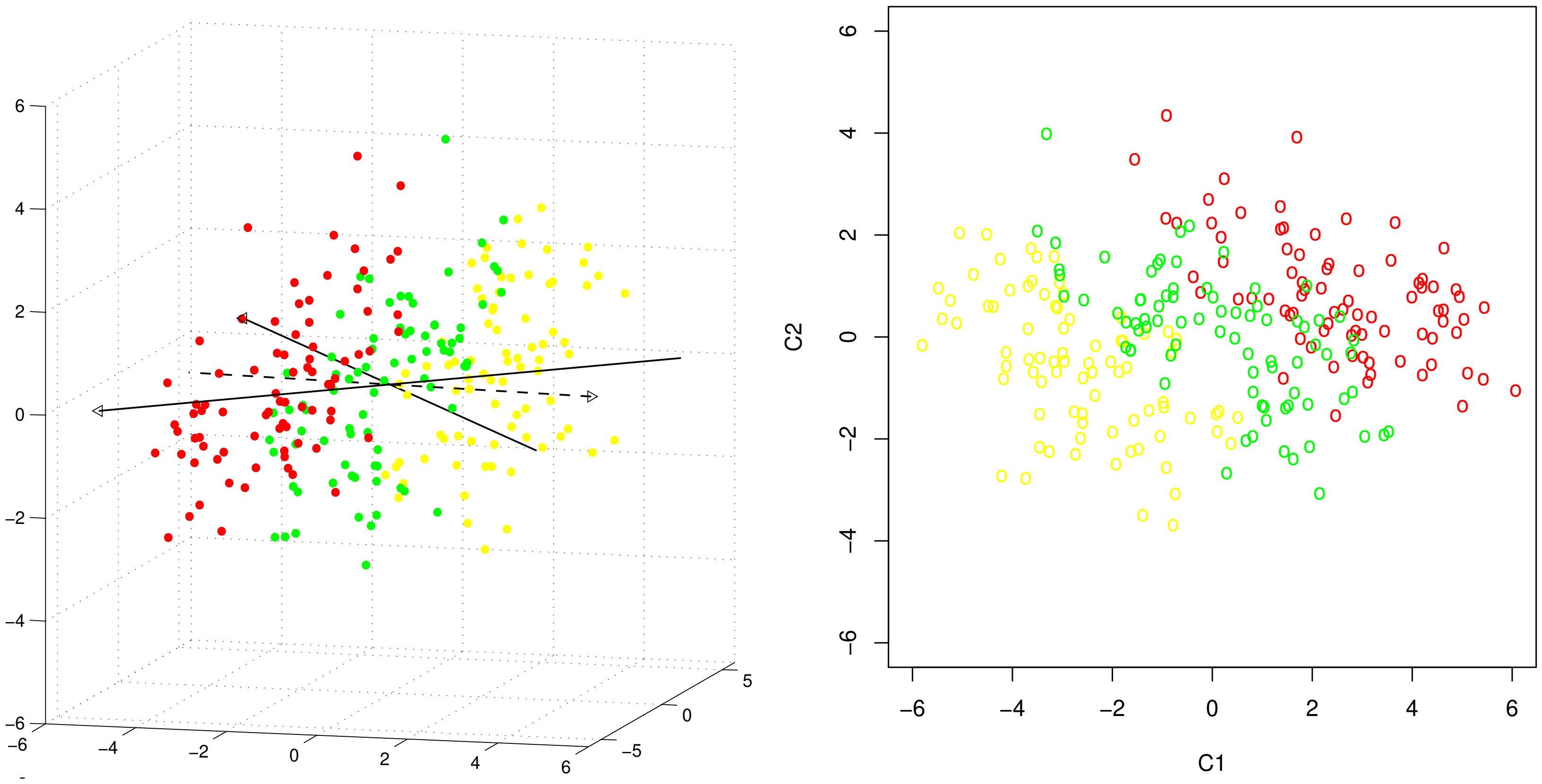
Data Simplification

The property

$$\text{Var}(c_1) > \text{Var}(c_2) > \text{Var}(c_3)$$

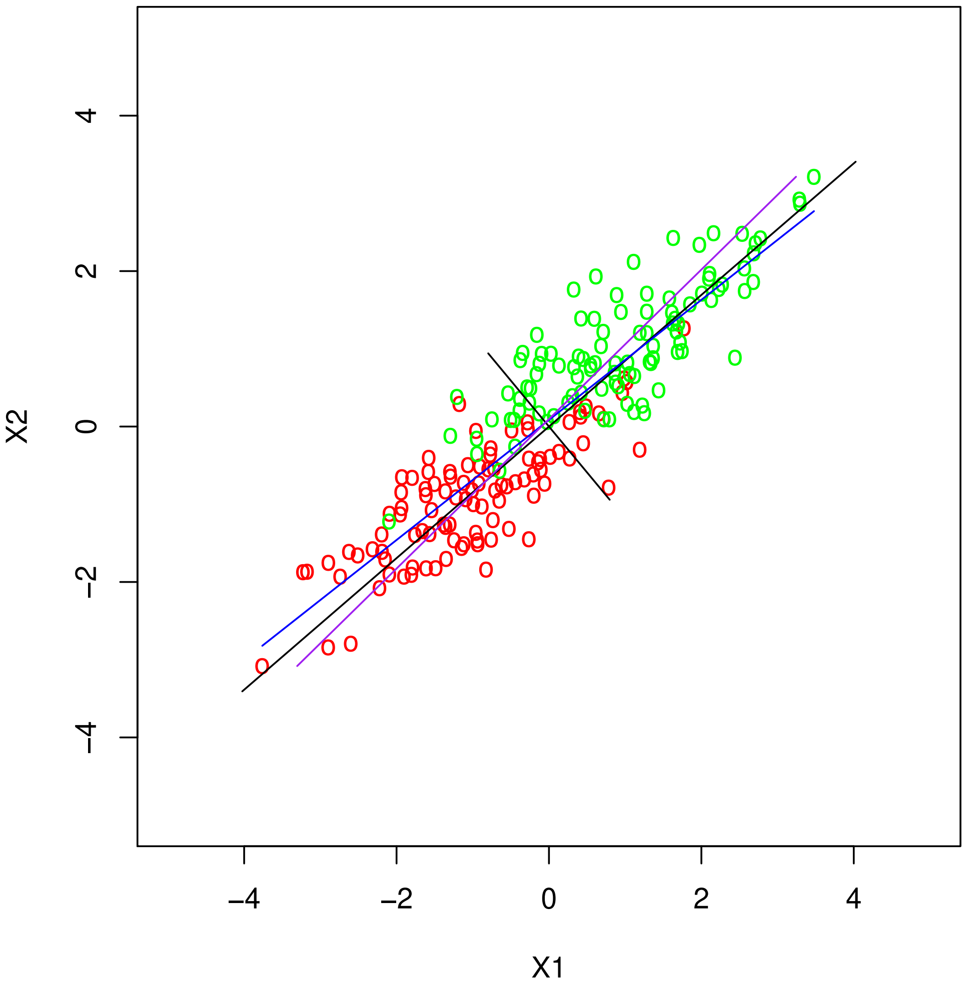
usually allows a dimensionality reduction.

Fig 76. Principal Components: Three Dimensions



The values (x_1, x_2, x_3) of three features for 240 cases from three clusters, red, and green, and yellow, are plotted in the left panel. The first two principal component directions are the black lines; the third is the dashed line. Plotting only the first two principal component values (c_1, c_2), right panel, achieves a dimensionality reduction with no loss of information.

Fig 77. The Relation of Regression to Principal Components



This is the left panel of Figure 75 with regression lines added. The blue line is the regression of x_2 on x_1. The purple is x_1 on x_2.

Principal Components Details

The next six slides discuss the math and computations for principal components.

They are included for completeness. They will not be discussed in class. They may be skipped if they do not interest you.

Hereafter I will denote feaures by r_i rather than x_i because the case deals with returns and r is the customary notation for returns.

Principal Components - The Math

Attention is restricted to features C_j of the form

$$c_{ij} = \sum_{k=1}^{p} r_{ik} w_{kj}$$

where $(r_{i1}, r_{i2}, \ldots., r_{ip})$ are the features for portfolio i, which can be thought of as a row from either Tables 26 or 27. The weights are required to satisfy $\sum_{k=1}^{p} w_{kj}^2 = 1$.

The first principal component has weights w_{k1}, called the component directions or the factor loading, that maximize

$$\text{Var}(C_j) = \frac{1}{n} \sum_{i=1}^{n} \left(c_{ij} - \bar{c}_j\right)^2$$

where n is the number of portfolios and $\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} c_{ij}$.

Principal Components - The Math

The second principal component has weights w_{k2} that satisfy

$$\text{Cov}(C_1, C_j) = \frac{1}{n} \sum_{i=1}^{n} (c_{i1} - \bar{c}_1) \left(c_{ij} - \bar{c}_j \right) = 0$$

and maximize

$$\text{Var}(x_i) = \frac{1}{n} \sum_{i=1}^{n} \left(c_{ij} - \bar{c}_j \right)^2$$

The third is uncorrelated with both C_1 and C_2 and maximizes variance.

And so on.

Principal Components - The Computations

Let X be a matrix with rows $(r_{i1}, r_{i2}, \ldots., r_{ip})$; X has $n =$ 79 rows and $p =$ 10 columns for the crash data. Let $\bar{x}$ be the average of the rows, which is a 1 by $p =$ 10 vector for the crash data.

Compute the following eigen-value eigen-vector decomposition

$$X'X - n\bar{x}'\bar{x} = WDW',$$

where D is a diagonal matrix containing the sorted eigen values, largest first.

The columns of W are the principal component directions; the columns C_j of $C = XW$ contain the principal component values c_{ij} .

In some applications, ours included, $\bar{x}$ is put to zero. In ours, the results with and without the mean correction are similar.

Principal Component Loadings

Principal components are derived features that are linear combinations of the features; that is, features whose value c_{ij} for case i and feature j has the form

$$c_{ij} = \sum_{k=1}^{p} r_{ik} w_{kj}$$

The weights w_{kj} are called loadings or component directions.

The next two slides display the loadings for crashes and rallies.

Crash Loadings

```
Loadings:
   Comp.1 Comp.2
1   0.170      0
2   0.199      0
3   0.664  0.348
4   0.368      0
5   0.214      0
6   0.222      0
7   0.252      0
8   0.255 -0.118
9   0.226      0
10  0.283 -0.919
```

The S&P500's crash features are

-4.82 -5.18 -19.57 -8.34 -6.75 -6.12 -6.77 -6.83 -5.97 -4.90

so the value of its first principal component is

(0.170)(-4.82) + (0.199)(-5.18) + ... + (0.283)(-4.90) = -26.90

and of its second is

0 + 0 + (0.348)(-19.57) + ... + (-0.919)(-4.90) = -3.13

Rally Loadings

	Comp.1	Comp.2
1	0.197	0.142
2	0.214	0
3	0.538	0.301
4	0.296	0
5	0.179	0
6	0.267	-0.131
7	0.288	0.237
8	0.295	-0.889
9	0.298	0
10	0.317	0
11	0.275	0

The S&P500's rally features are

4.64 4.79 8.86 4.89 4.96 5.08 4.73 5.06 5.75 5.43 4.73

so the value of its first principal component is

(0.197)(4.64) + (0.214)(4.79) + ... + (0.275)(4.73) = 17.99

and of its second is

(0.142)(4.64) + 0 + ... + 0 = 0.07

Carrying On

Let's see how principal components does with our problem.

For crashes, the next figure shows how much of the total variance each of the principal components accounts for.

The slide after that plots the first two principal component values.

Fig 78. Crash Principal Component Variances

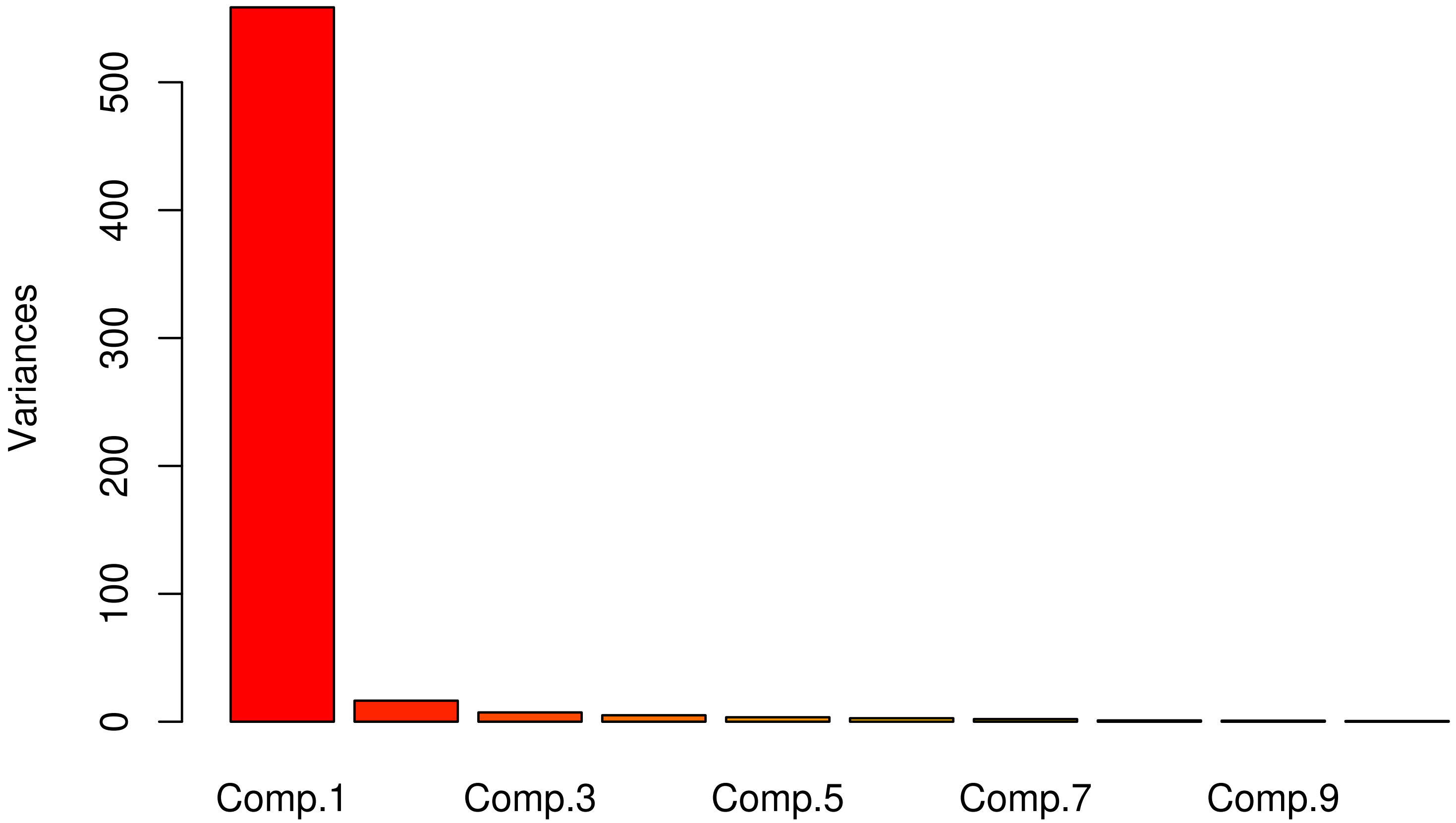
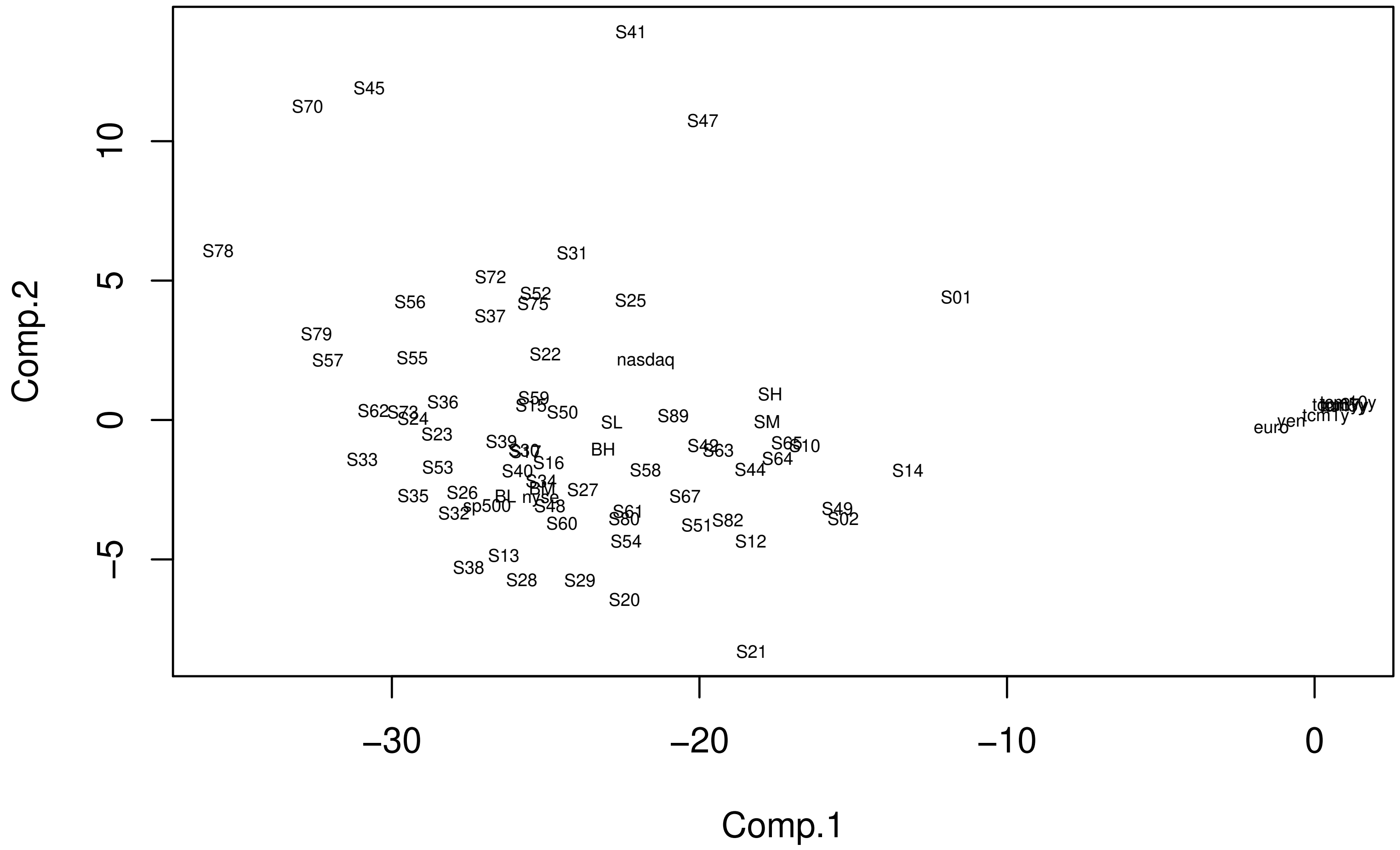


Fig 79. Performance in Crashes



Carrying On

The next two slides are the same thing for rallies.

Fig 80. Rally Principal Component Variances

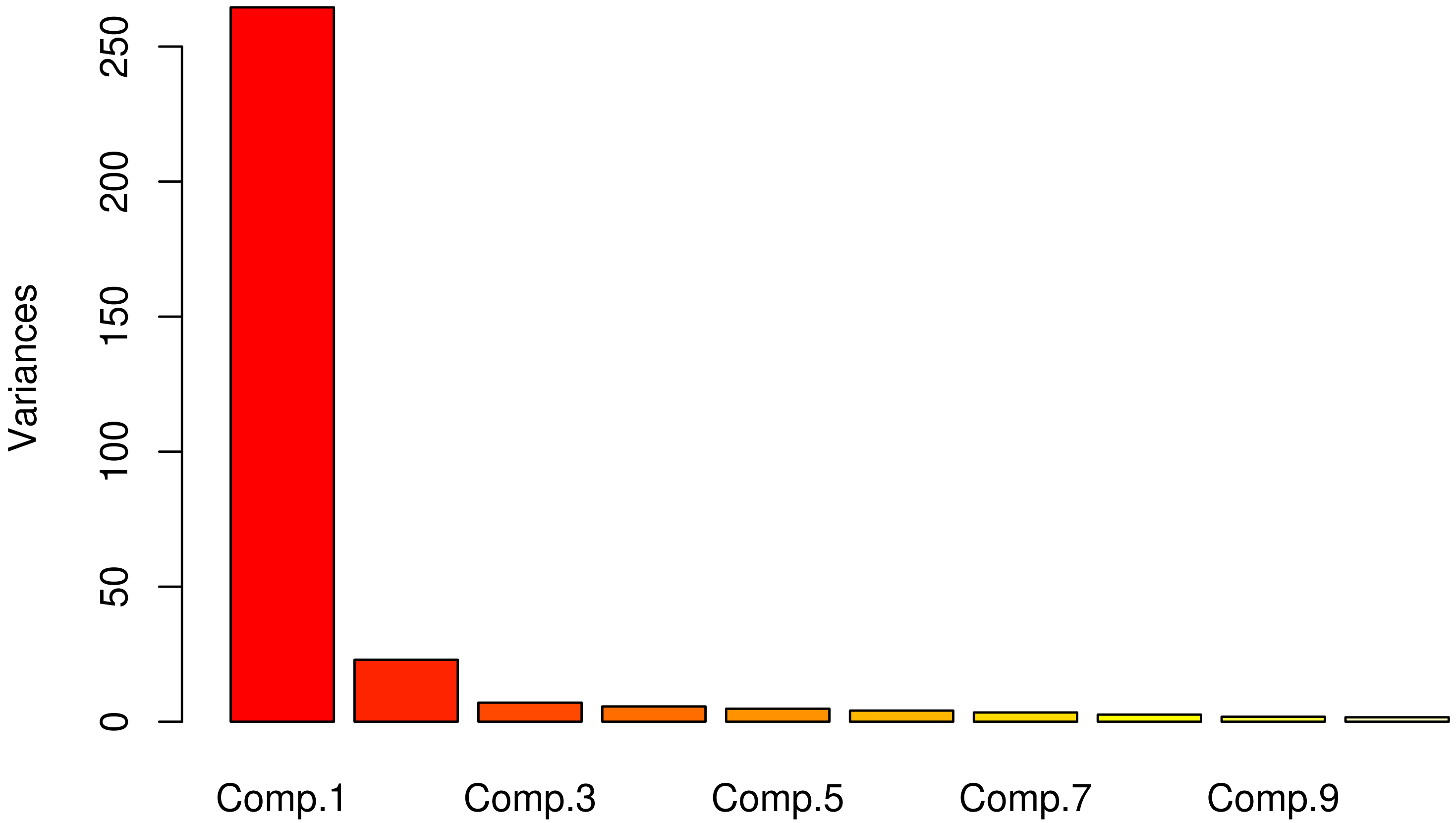
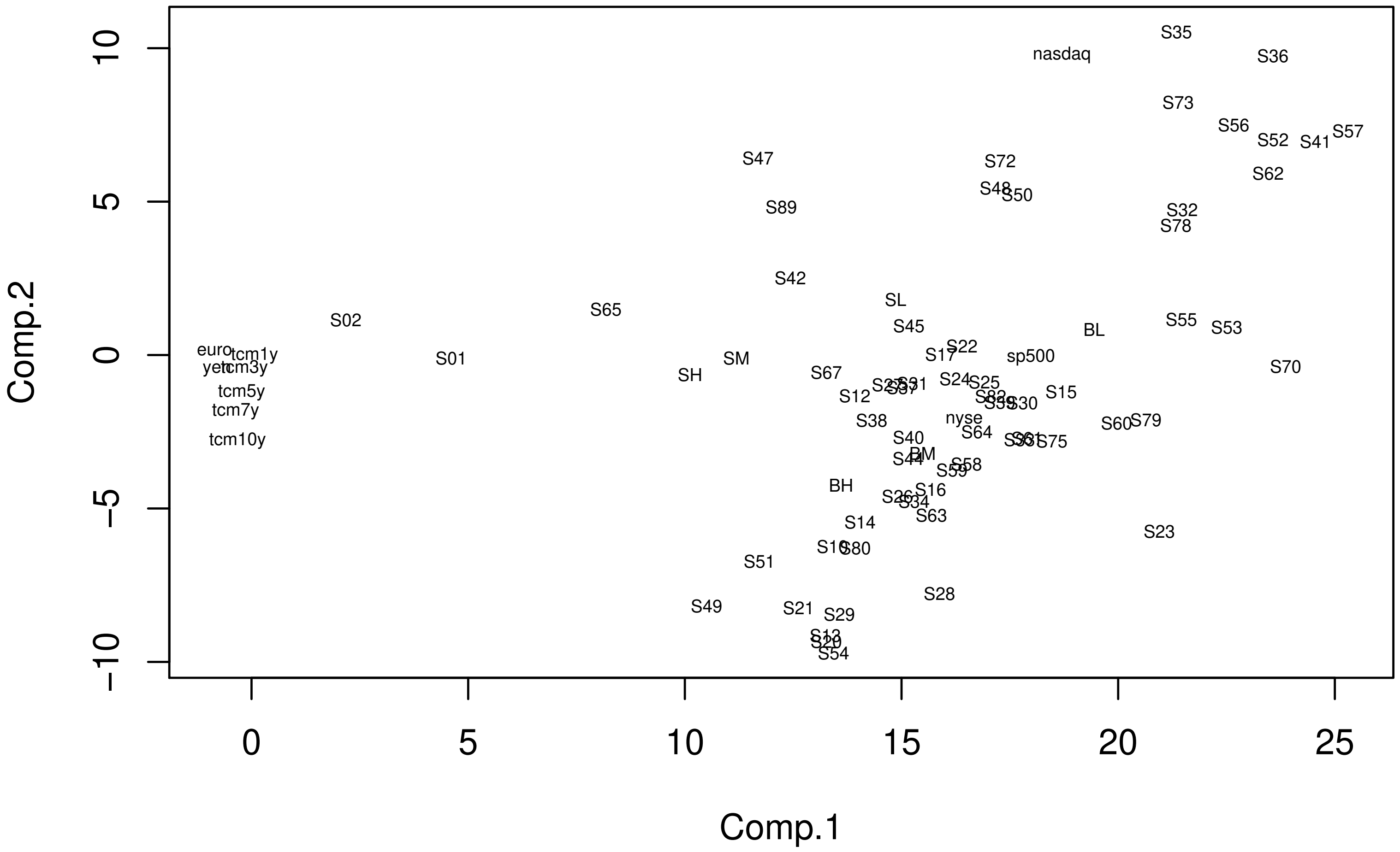


Fig 81. Performance in Rallies



Appraisal

The first principal component accounts for the overwhelming amount of the variation.

There are certainly two groups of portfolios, those that perform like bonds and those like stocks.

There is a hint that stocks can be split into two groups, those like the Nasdaq index and those like the S&P 500 index.

Unsupervised Learning

In our application, an operational definition of unsupervised learning is that we want to associate these 79 labels to a much smaller number of labels that could replace them for value at risk purposes.

Stated differently, the goal is to replace meaningless labels with targets that have meaning and could even be used subsequently to train a classifier.

Hastie, Tibshirani, and Friedman, p. 438

With supervised learning there is a clear measure of success, or lack thereof, that can be used to judge adequacy in particular situations and to compare the effectiveness of different methods over various situations.

It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to heuristic arguments not only for motivating the algorithms [...] but also for judgments as to the quality of the results.

This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly.

Strategy

In this situation it seems best to try several different methods and try to synthesize the results.

That is what we shall do.

Moving On

Our next topic is hierarchical clustering.

Most clustering methods, including hierarchical clustering, require a measure of dissimilarity between cases ...

Dissimilarities

Following standard convention, let X by a matrix where features are columns and cases are rows; X has n rows and p columns and would look like Table 26 but with more rows. The features for case i would be in row i and will be denoted here by x_i.

The most commonly used measure of dissimilarity between two cases x_i and $x_{i'}$ is Euclidean distance

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

To attach less importance to large differences, one can use

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|$$

Dissimilarities

Another common choice is based on the correlation between case x_i and case $x_{i'}$

$$D(x_i, x_{i'}) = 1 - \mathrm{Corr}(x_i, x_{i'}).$$

Dissimilarities can be mixed or weighted, e.g.

$$D(x_i, x_{i'}) = \frac{1}{2}(x_{i1} - x_{i'1})^2 + |x_{i2} - x_{i'2}| + \cdots$$

But whatever is done, it must be that $D(x_i, x_i) = 0$ for the same case and that $D(x_i, x_{i'}) > 0$ for different cases.

Dissimilarities

For ordinal variables like Small, Medium, Large, one can code them as -1, 0, 1 and use a numeric dissimilarity measure.

For categorical variables or when no mathematical expression seems right, one can assign values to the $D(x_i, x_{i'})$ by hand, making sure that for identical cases $D(x_i, x_{i'}) = 0$ and that $D(x_i, x_{i'})$ is positive otherwise.

Value at Risk Dissimilarities

Let X be all assets and returns under consideration, which would have $n = 79$ rows and $p = 8442$ columns. Our crash dissimilarity measure is

$$D(x_i, x_{i'}) = \sum_{SP500_j < -4.5} (x_{ij} - x_{i'j})^2$$

and our rally dissimilarity measure is

$$D(x_i, x_{i'}) = \sum_{SP500_j > 4.5} (x_{ij} - x_{i'j})^2$$

There are other dissimilarity measures one might consider such as taking the sum over all dates in a recession or expansion.

Same Song, Second Verse

As just seen, choosing a measure of dissimilarity involves two choices:

1. What (derived) features dissimilarity will depend upon.

2. How differences in these (derived) features are to be measured.

The first is far more important: The better the features the better the tool's performance.

As we have seen more than once, even simple tools work well with good features and often sophisticated tools cannot overcome bad features.

Moving On

Let's now consider the hierarchical clustering toolkit ...

Hierarchical Clustering

Hierarchical clustering builds what is called a dendogram and looks like a decision tree.

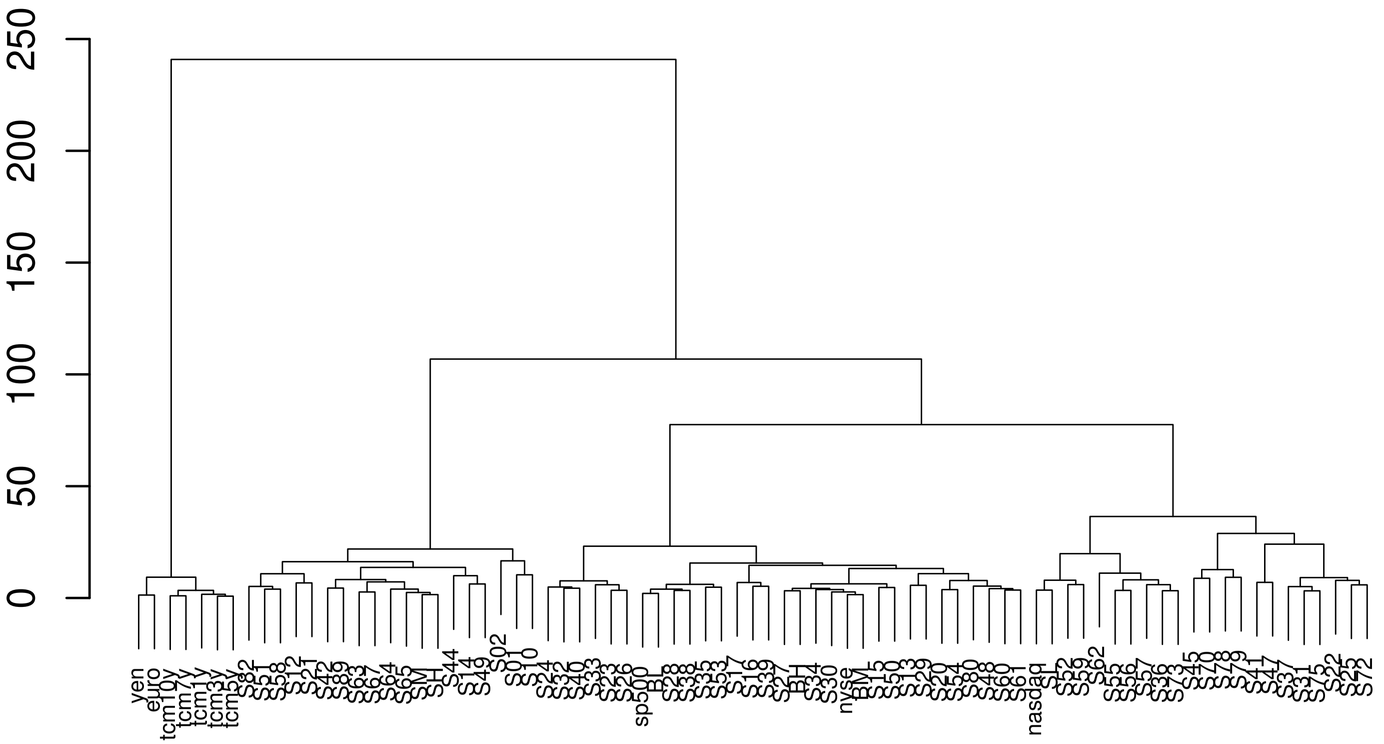
The clusters at each level of the dendogram are created by merging clusters at the next lower level.

At the lowest level, each cluster contains a single case.

At the highest level there is only one cluster containing all the data.

Here is an example ...

Fig 82. Ward Dendogram for Crashes



Interpretation

If we cluster at the level of the dendogram marked 150, we would have two groups:

1. Bonds and foreign exchange.

2. Stock portfolios.

If we cluster at the level of the dendogram marked 50, we would have four groups:

1. Bonds and foreign exchange.

2. Stock portfolios like small cap value stocks.

3. Stock portfolios like the S&P500 Index.

4. Stock portfolios like the Nasdaq Index.

How to Make a Dendogram

Start at the bottom and pair each case with the case that has the smallest dissimilarity.

Then work up and recursively merge the pair of clusters that has the smallest group dissimilarity.

Draw the join at a vertical height that is the value of the group dissimilarity between the two joined groups.

This is aglomerative or bottom-up clustering; divisive or top-down hierarchical clustering is little used.

What has not yet been defined is "group dissimilarity" and this is where methods differ. We define it next ...

Group Dissimilarity

Single Linkage takes group dissimilarity to be the dissimilarity between the two closest pair of cases, one from each group.

Complete Linkage takes group dissimilarity to be the dissimilarity between the two farthest cases pair of cases, one from each group.

Average Linkage takes group dissimilarity to be the average of the dissimilarities between all pairs of cases, one from each group.

Ward's Linkage aims at finding compact spherical clusters. Tries to find the grouping that minimizes within SS&CP. Mimics K-means.

Which to Use?

Presenting the arguments for and against each method would take twenty slides and not leave us with a definitive conclusion.

If the data actually falls into a few small tight groups, all methods will find them.

If not, since evaluation is subjective anyway, it seems best to examine them all to see if any of them suggest something interesting.

Here we go ...

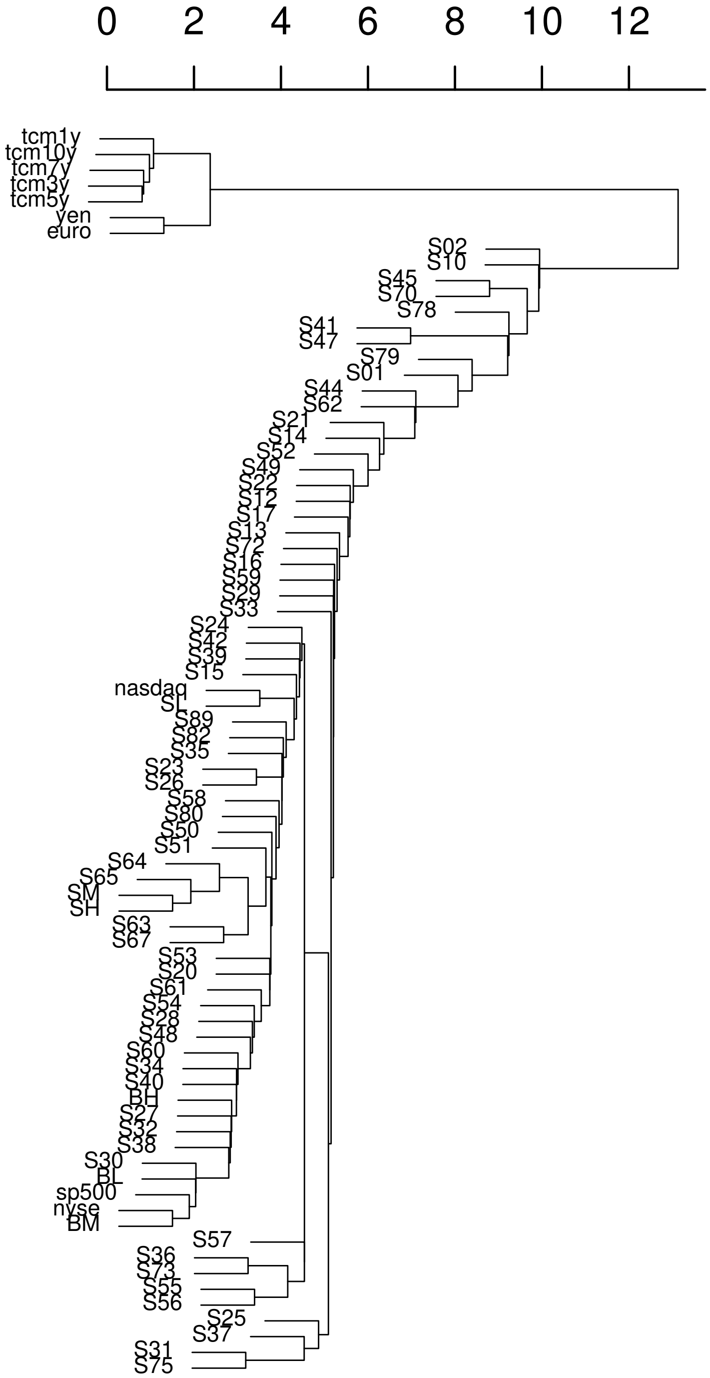


Fig 83. Single Linkage Dendogram for Crashes

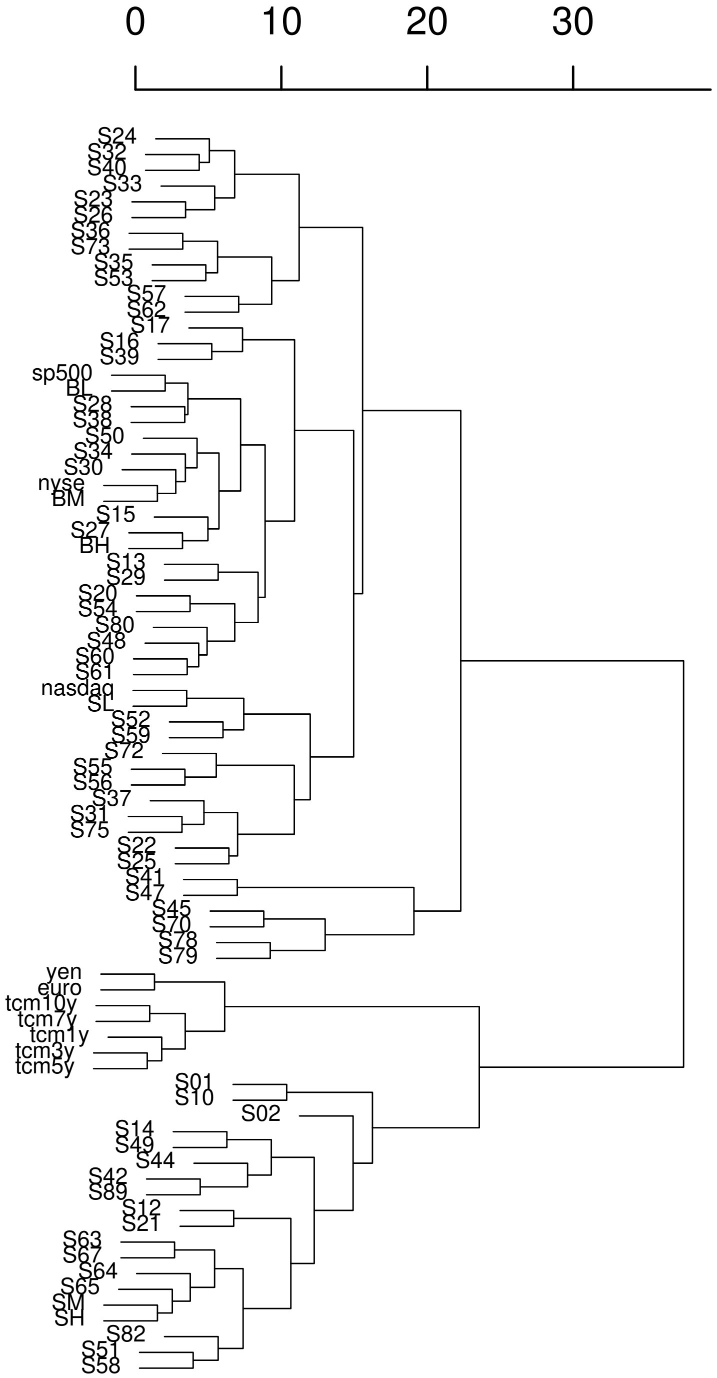


Fig 84. Complete Linkage Dendogram for Crashes

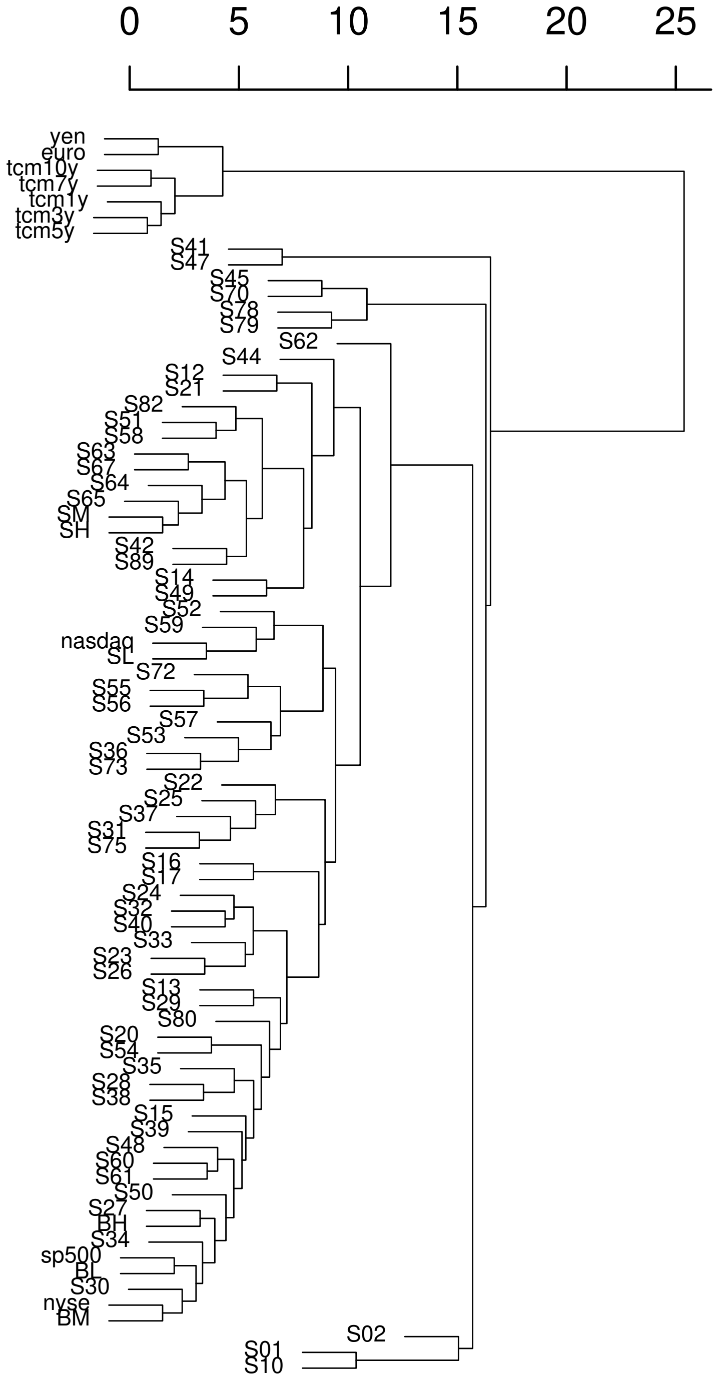


Fig 85. Average Linkage Dendogram for Crashes

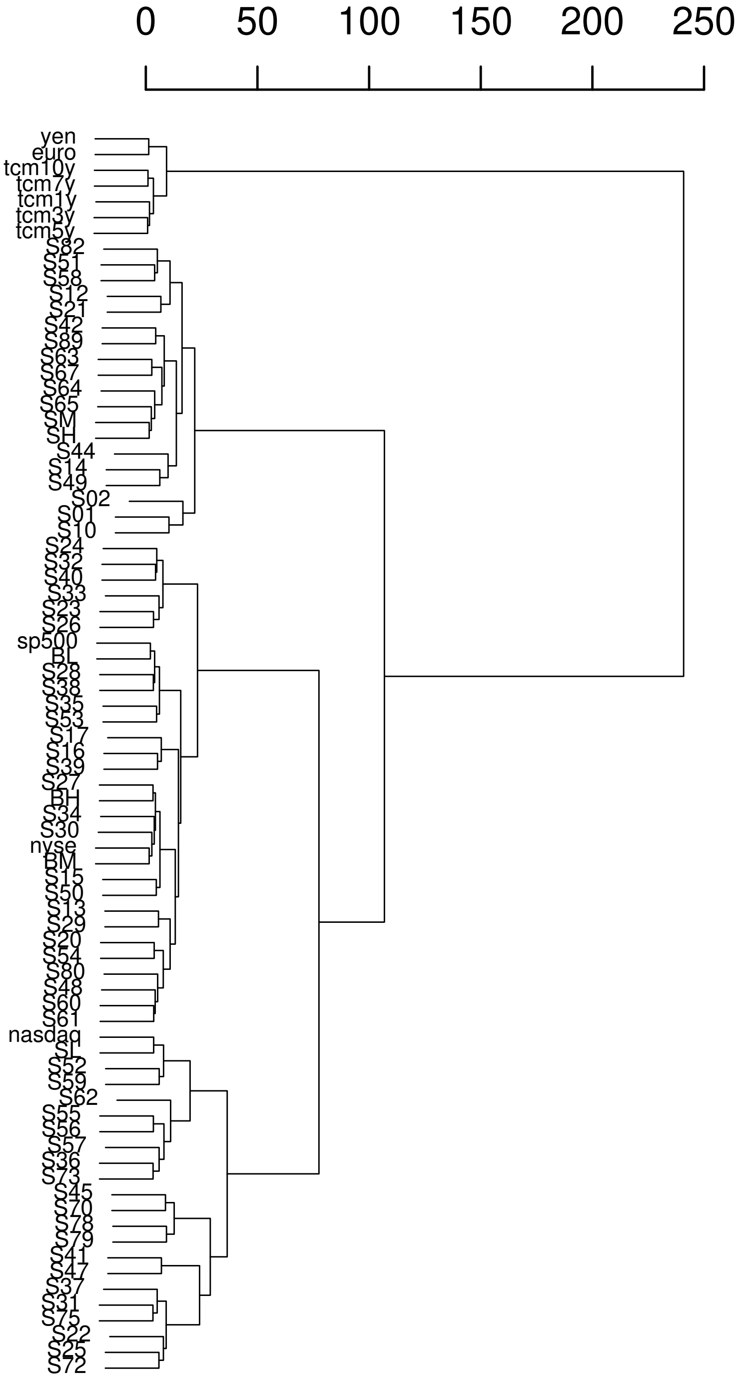


Fig 86. Ward Linkage Dendogram for Crashes

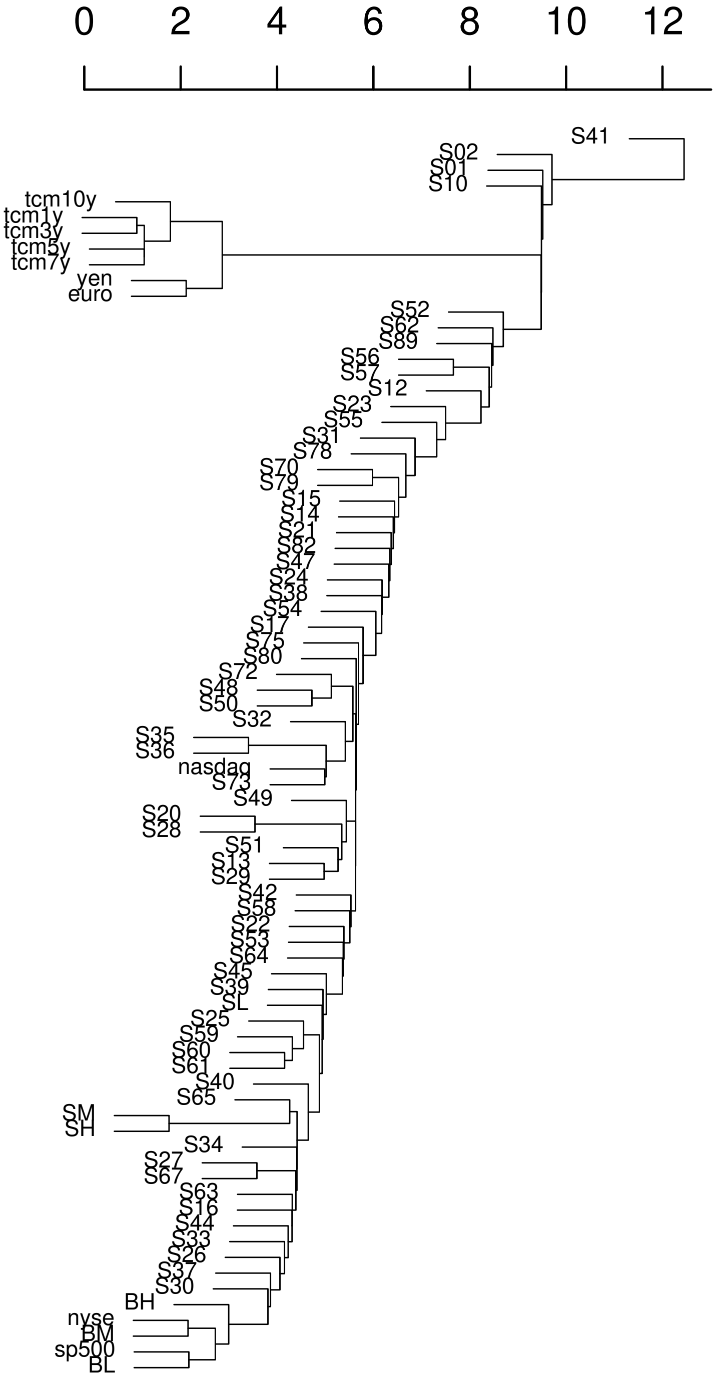


Fig 87. Single Linkage Dendogram for Rallies

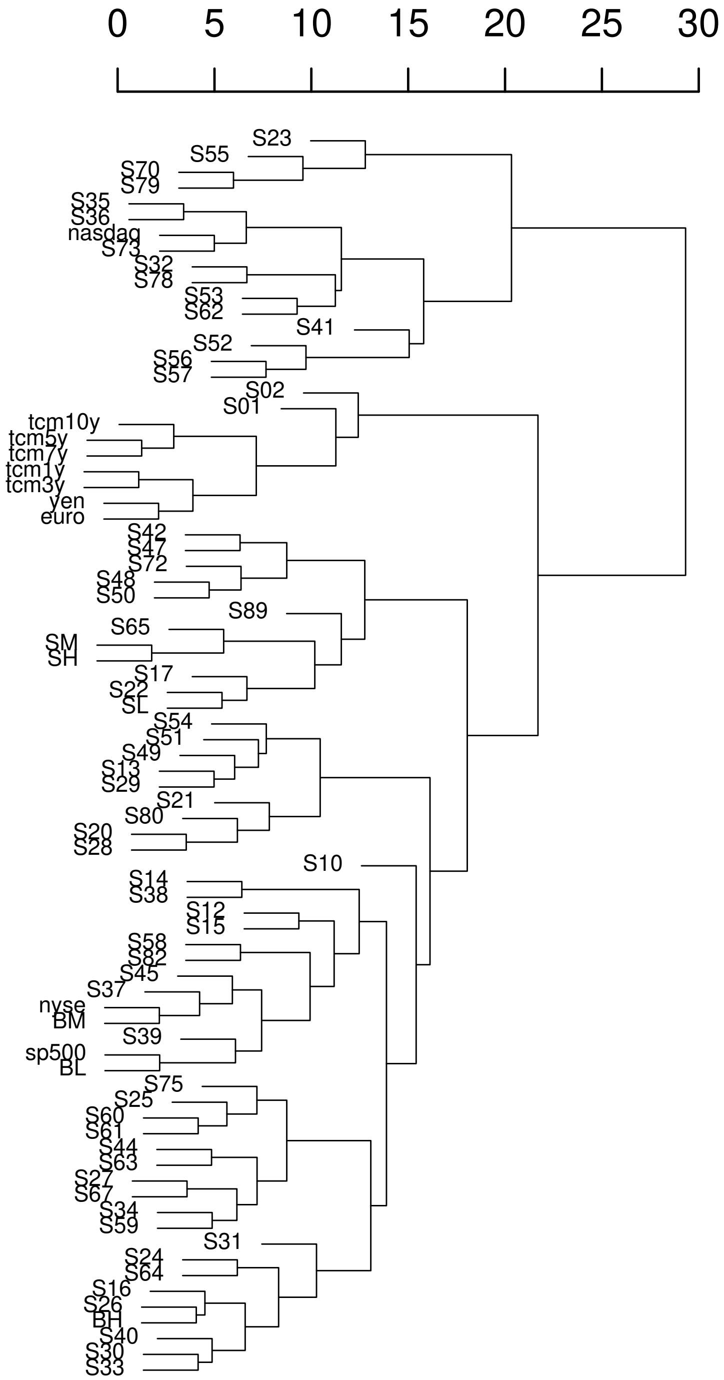


Fig 88. Complete Linkage Dendrogram for Rallies

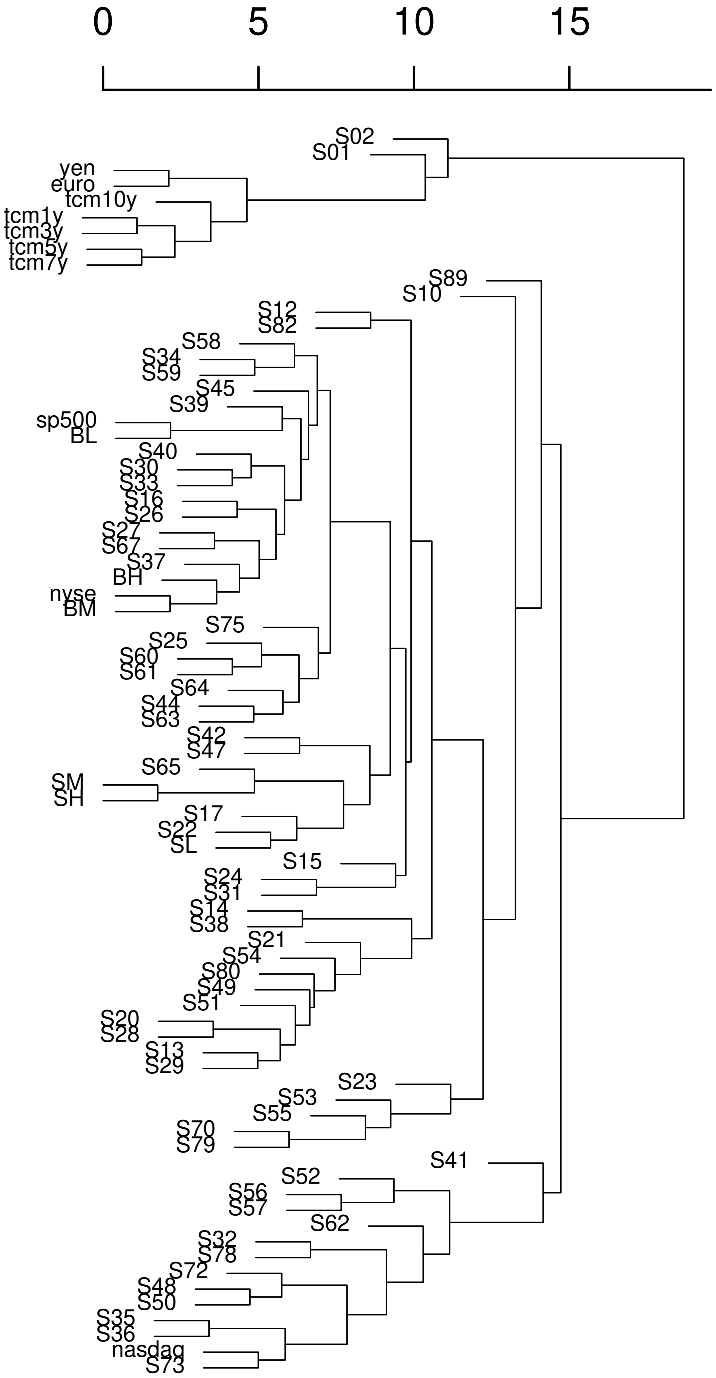
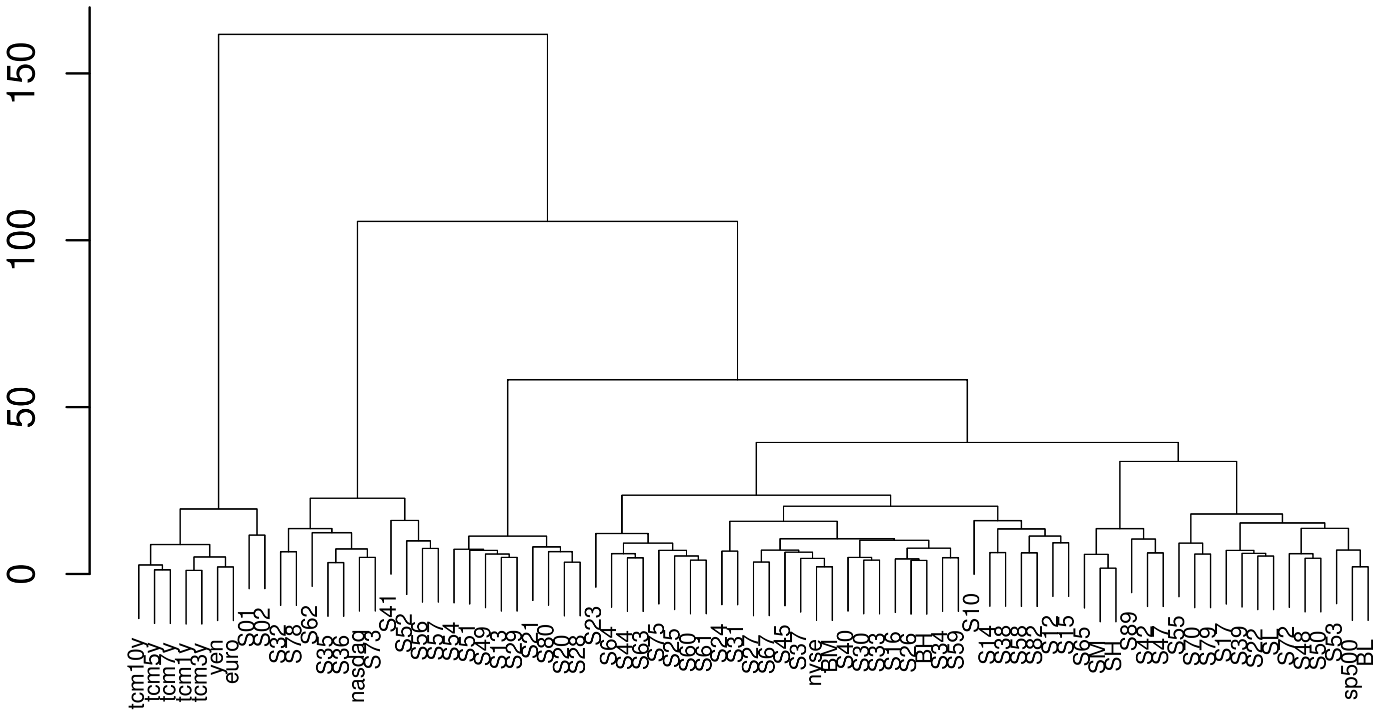


Fig 89. Average Linkage Dendogram for Rallies

Fig 90. Ward Linkage Dendogram for Rallies



Interpretation

Interpretation is subjective and we are all entitled to our own opinion. Mine is that we have three groups:

1. Bonds and foreign exchange.

2. Stock portfolios like the S&P500 Index.

3. Stock portfolios like the Nasdaq Index.

Moving On

Lastly, we consider K-means, the most popular clustering method.

It's main disadvantage is that the user has to tell it how many clusters there are, which is the K of K-means, and give it a center for each cluster.

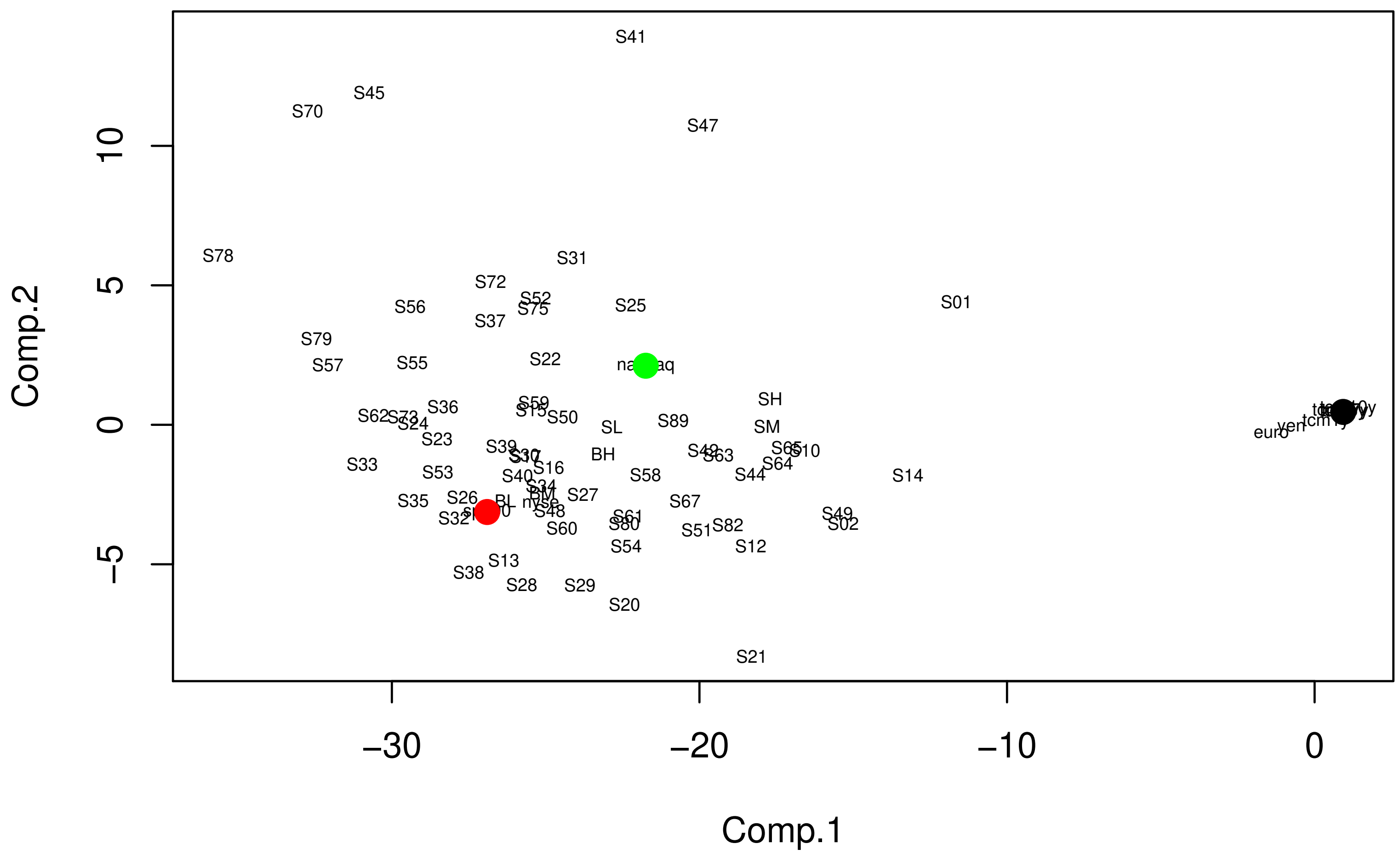
Since we now have a good idea how many clusters there are and know a member from each — tcm5y, sp500, and nasdaq — this will not be a problem for us.

We will use these three members as our initial cluster centers.

K-Means, Iter 0

The next figure plots our three initial cluster centers together with all cases.

Fig 91. K-Means Initialization, Crashes



K-Means, Assign Cases to Centers

The next step is to assign cases to the cluster centers.

One computes the Euclidean distance from the case's features to the cluster center's features. The case is assigned to the cluster for which the distance to the cluster center is smallest.

To say the same thing with math, if c_k denotes a center cluster with feature values c_{kj} and x_i denotes a case with feature values x_{ij}, then one computes

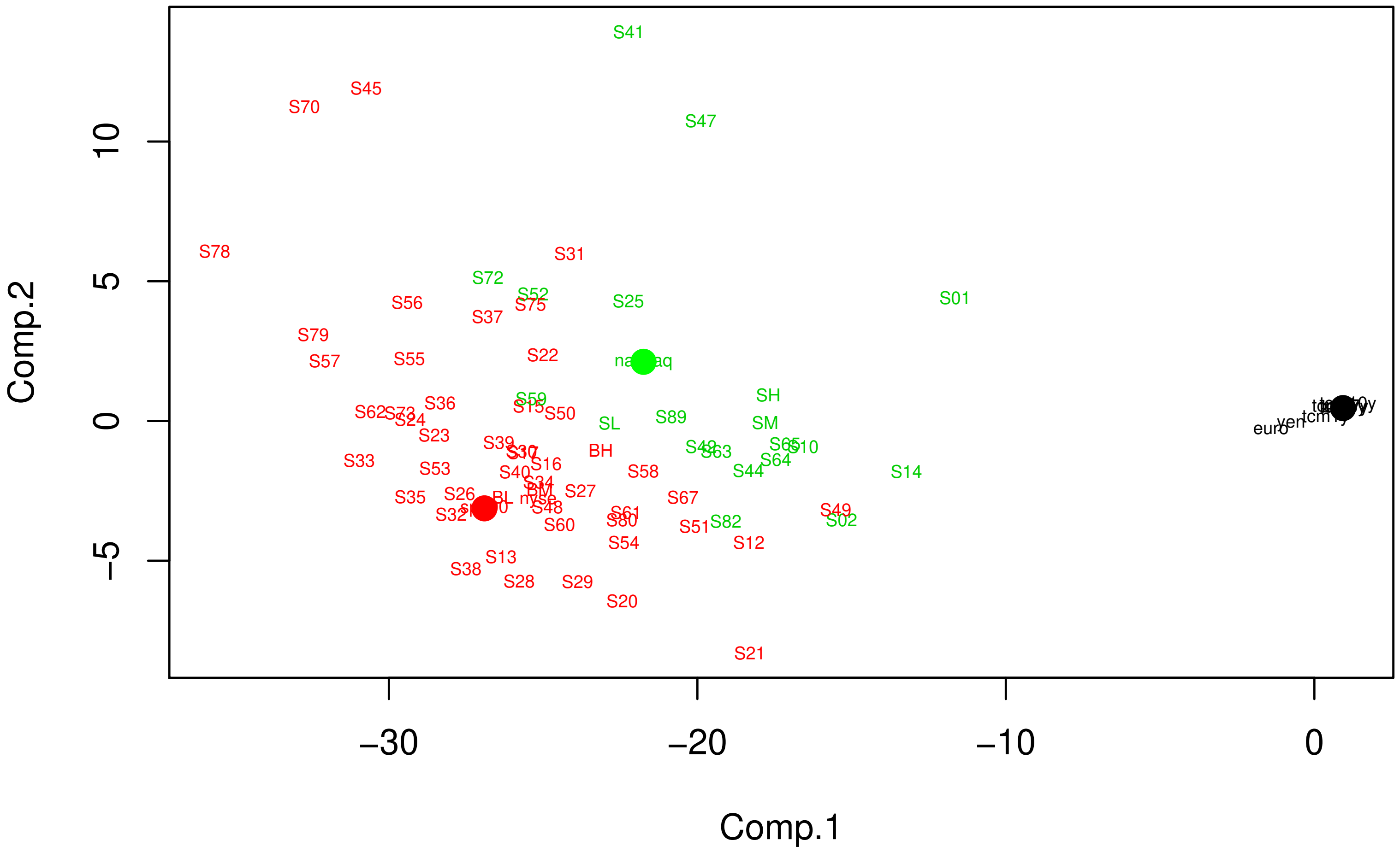
$$D(x_i, c_k) = \sum_{j=1}^{p} (x_{ij} - c_{kj})^2$$

for $k = 1, \ldots, K$. Case x_i is assigned to that cluster k for which $D(x_i, c_k)$ is smallest.

K-Means, Iter 1

The next figure shows the assignment cases to our initial cluster centers.

Fig 92. K-Means Iter 1, Crashes



K-Means, Recompute Cluster Centers

The next step is to recompute cluster centers.

For each feature, compute the mean of that feature over all cases in the cluster. These means become the features of the new center for the cluster. This is the *means* of K-means.

To say the same thing with math, let C_k be the indexes of the cases in cluster k and let n_k be the number of cases in C_k. For example, looking at Fig 92, the cases in cluster 1 are tcm10y, tcm1y, tcm3y, tcm5y, tcm7y, yen, euro and the indexes that correspond to them are $C_1 = \{67, 68, 69, 70, 71, 72, 73\}$. These number $n_1 = 7$. Compute the means

$$m_{kj} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}$$

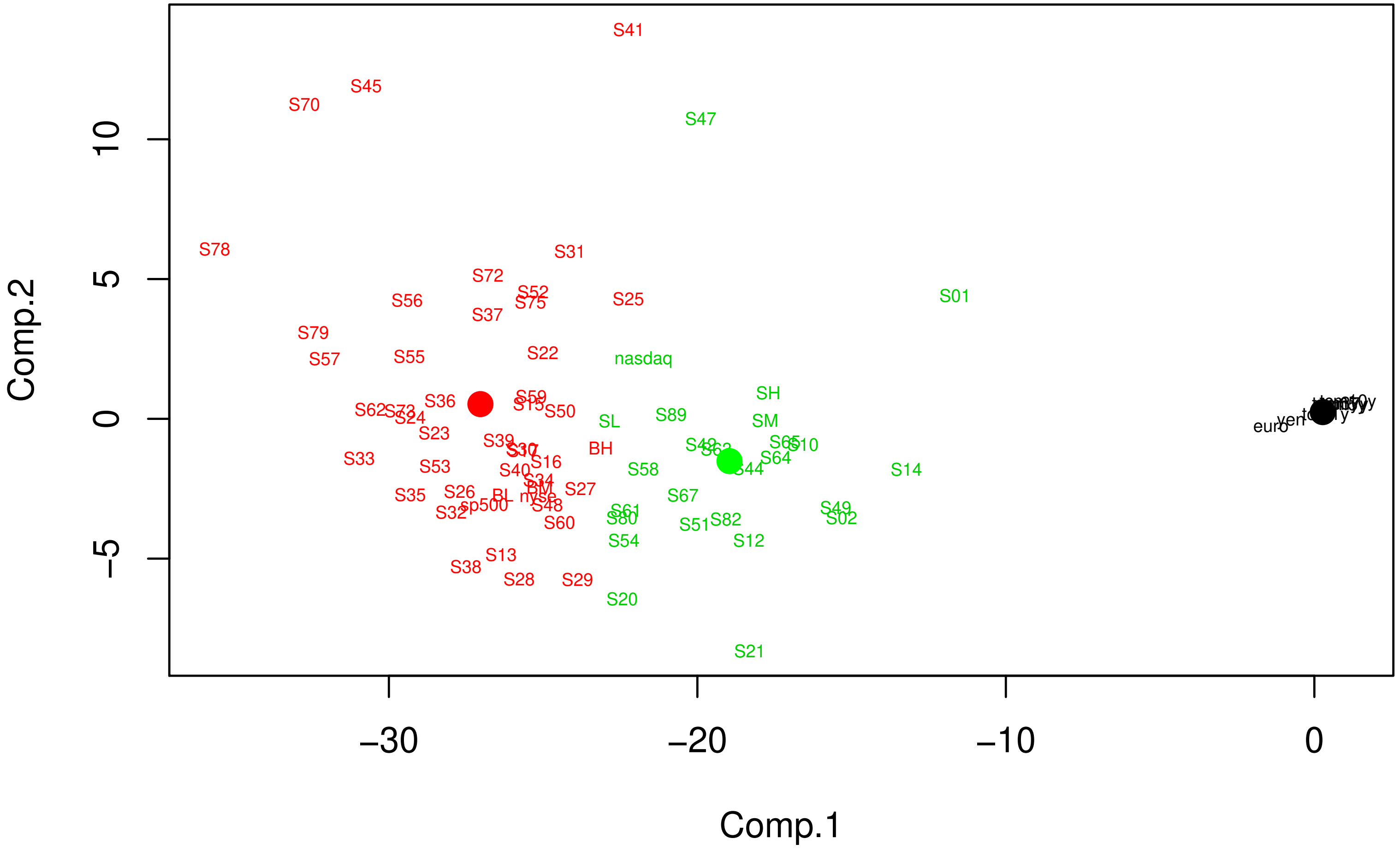
The new cluster center is

$$c_k = (m_{k1}, m_{k2}, \ldots, m_{kp})$$

K-Means, Iter 1

The next figure shows the new cluster centers and the cases assigned to the new cluster centers.

Fig 93. K-Means Iter 2, Crashes



The K-Means Algorithm

The K-means algorithm proceeds in this way

1. Recompute cluster centers

2. Assign cases to clusters

until cases stop moving among clusters.

Clusters for Crashes and Rallies

In our application, convergence is very quick. Fig 93 basically gives the final answer.

Here is that figure repeated without the dots to indicate cluster centers followed by the same thing for rallies.

Fig 94. K-Means Clusters, Crashes

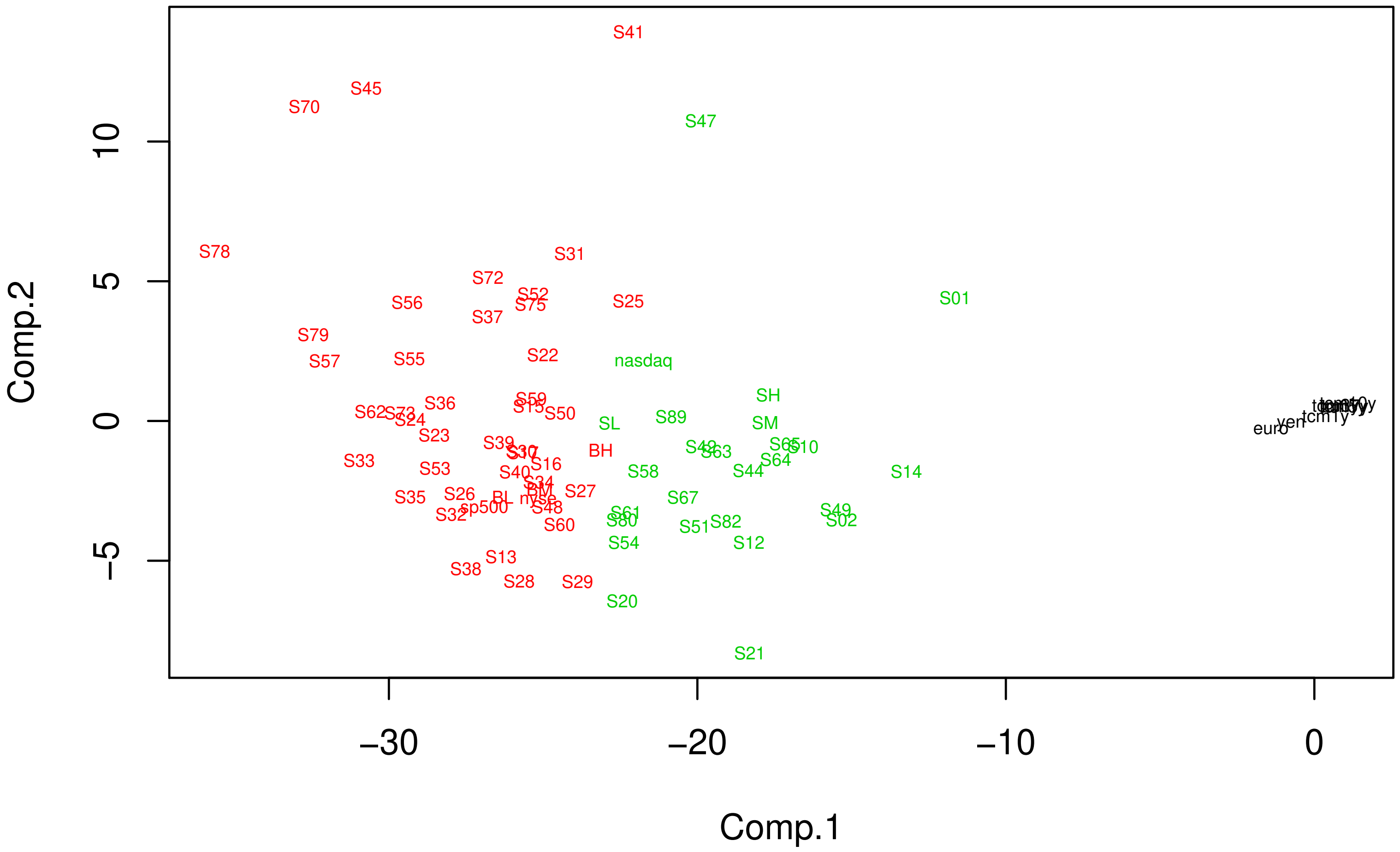
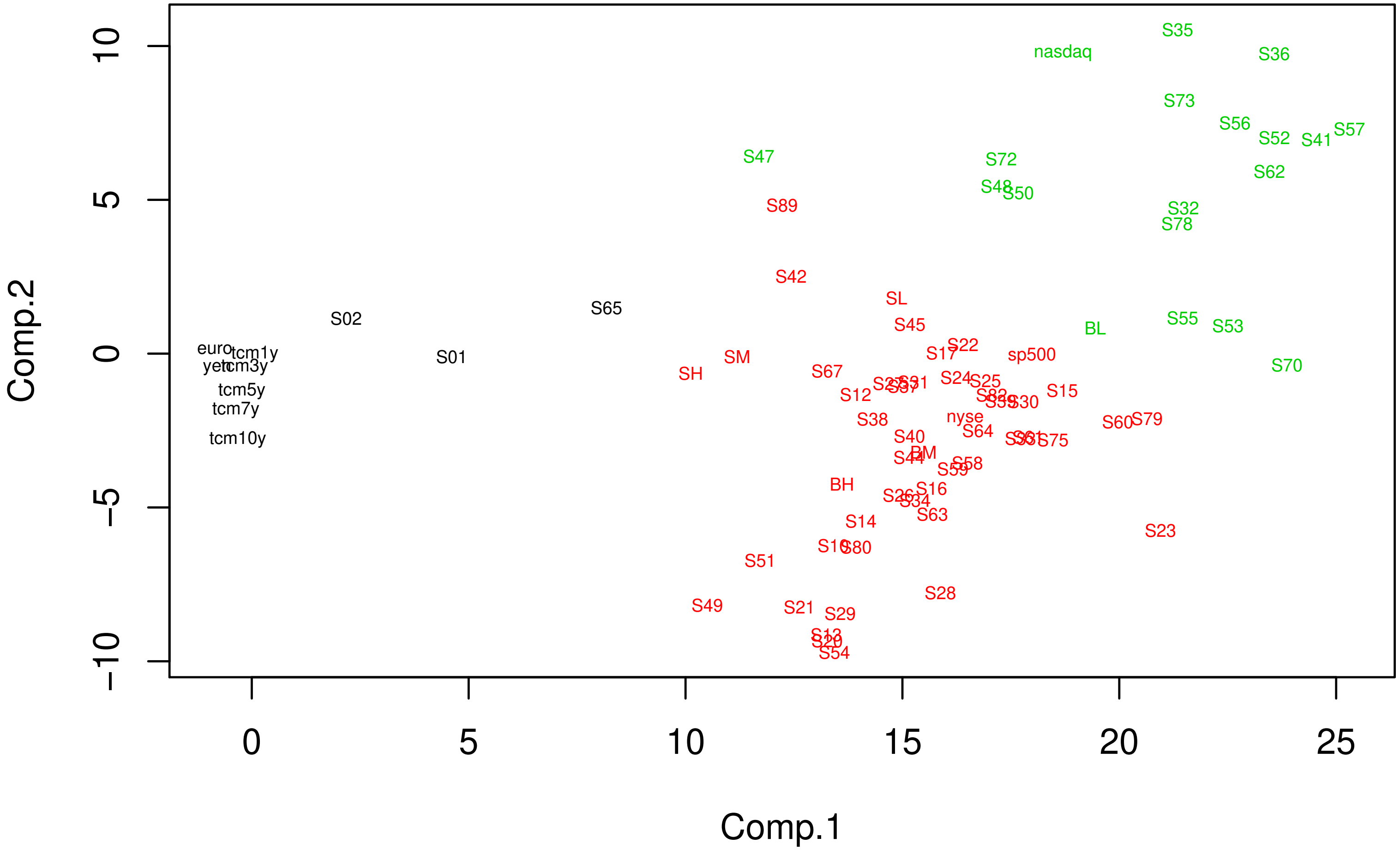


Fig 95. K-Means Clusters, Rallies



Interpretation

1. In crashes the nasdaq cluster beats the S&P500 cluster.

2. The nasdaq index is a poor representer of its cluster. It has not outperformed the S&P500 index in the last four crashes although the same seems not to be true of its cluster; see Table 26

3. In rallies the nasdaq cluster beats the S&P500 cluster.

4. The Fama-French portfolios switch clusters in crashes and rallies.

Implications for Value at Risk

From the point of view of a leveraged investor who does not want to risk either bankruptcy or an involuntarily liquidation of securities in a market crash.

- There is no silver bullet: The only protection in a crash is cash or cash equivalents.

- For a value at risk computation, one can apparently beat the curse of dimensionality by averaging one's securities by cluster thereby reducing to three portfolios and then only bothering with the correlations among these three in crashes.

Details

What is often done is to try to get K-means to find K and the optimal groupings — the task that we used hierarchical clustering to perform.

This is done by running K-means many times with randomly allocated initial clusters. For each K, that configuration that produces the smallest value of the within cluster variance is chosen.

Most software has a mechanism to do this automatically.

My view is that if one has a hierarchical clustering tool, it will do a better job of both initial cluster determination and providing a feel for the data than K-means with random start.

Unsupervised Learning: Synthesis

- Evaluation of unsupervised learning results is subjective.
- The main tools of unsupervised learning are the cluster analysis tools: principal components, hierarchical clustering, and K-means.
- These tools complement each other. Hierarchical clustering helps to get an initial feel for the data and to find likely clusters. This information feeds into K-means for refinement. Principal components helps visualize K-mean results.
- Success depends critically on good feature selection.

Blank page

Blank page

Blank page