Visualizing the Sampling Variability of Plots

Rajiv Menjoge, Google

Roy Welsch, MIT

rwelsch@mit.edu

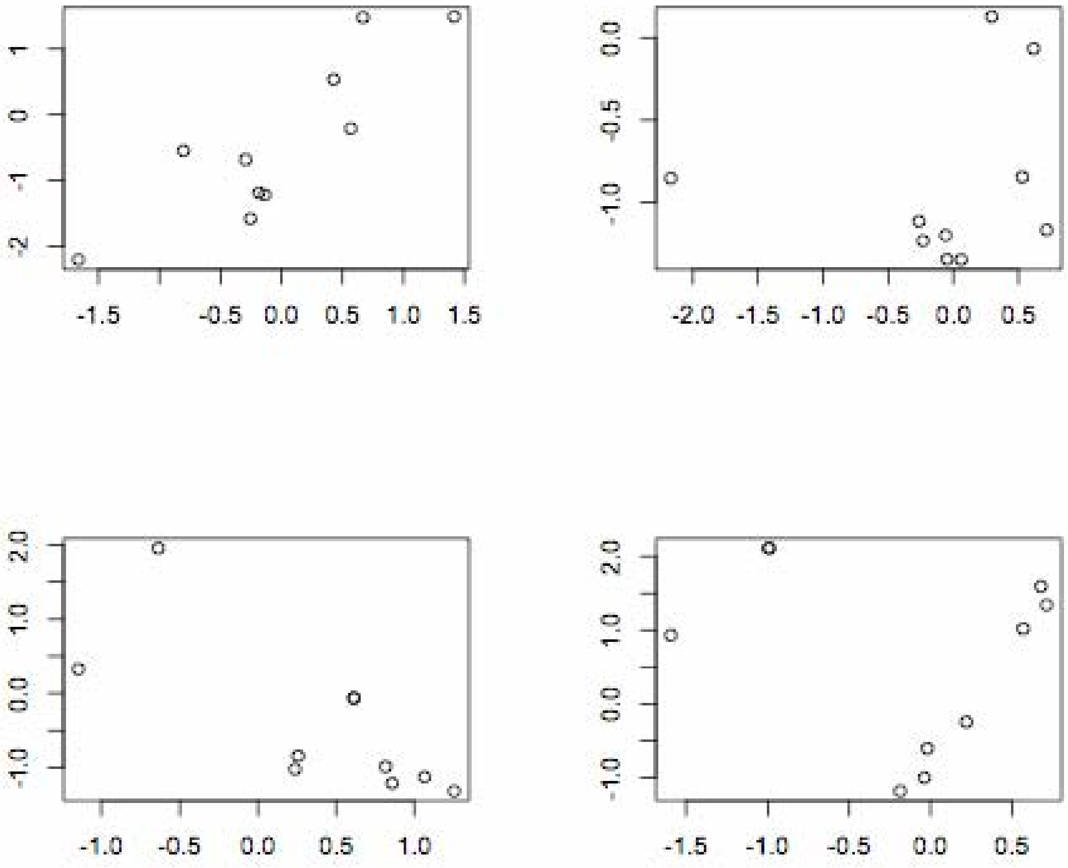
COMPSTAT 2010

24 August 2010

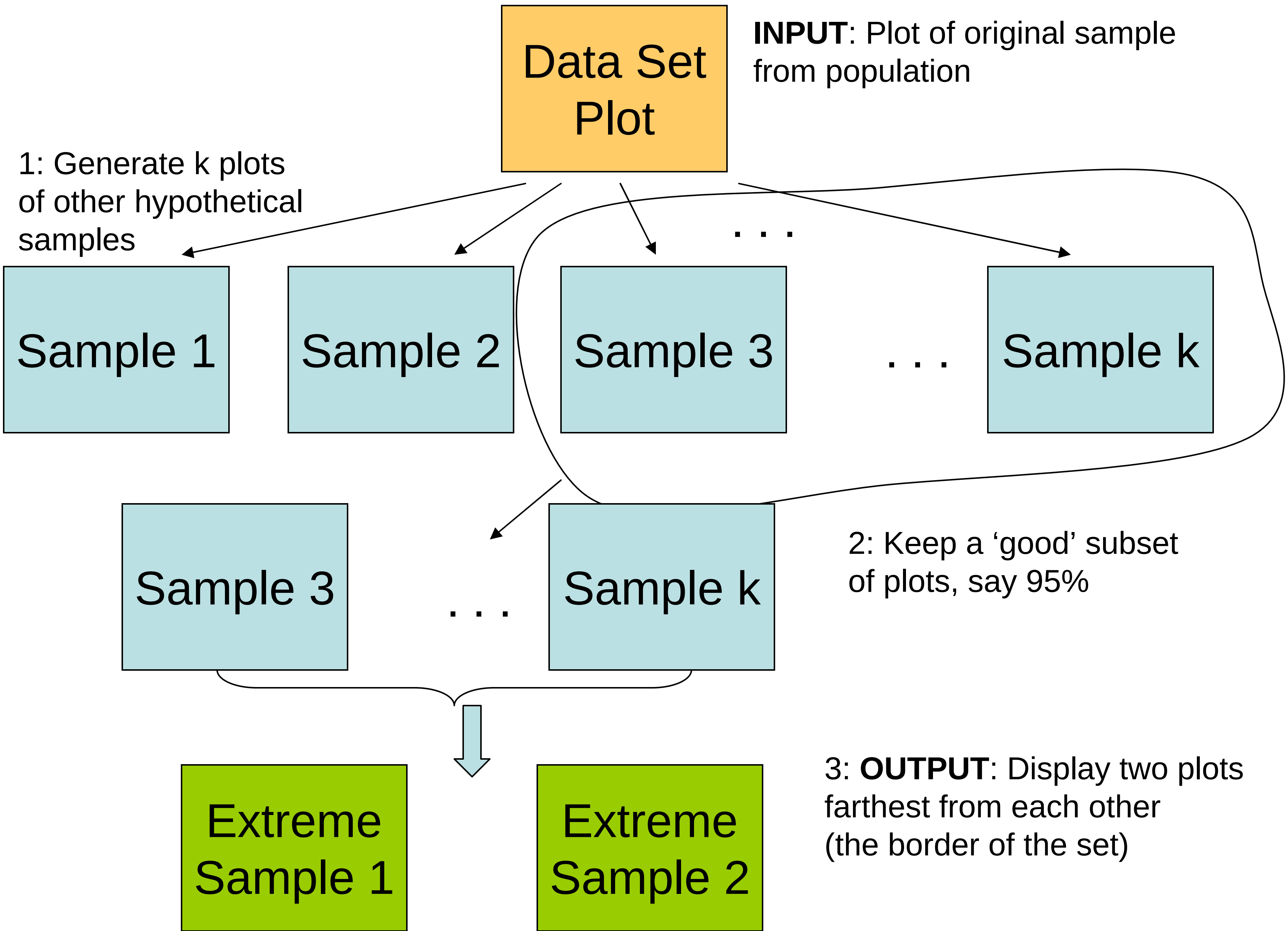
Paris

Motivation: Plots of a data set can look different than the plot of the population they are from!

A simulated illustration: Four data sets, sampled from the same population.



A Picture of the Methodology

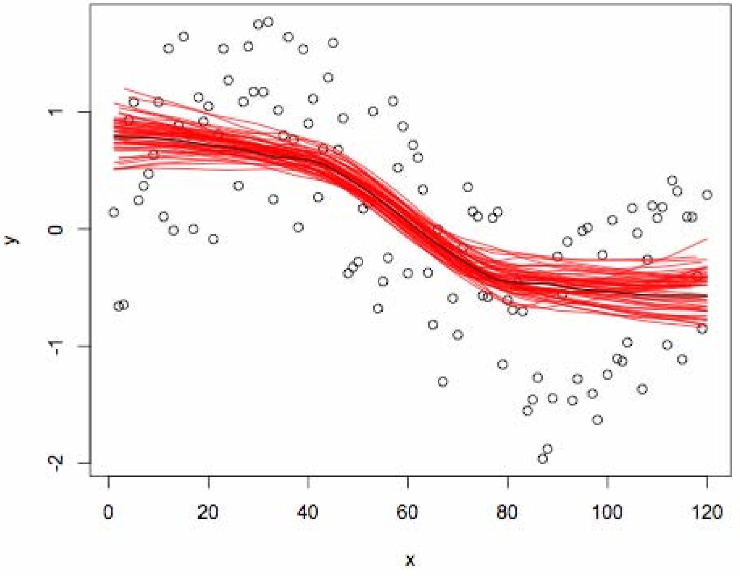


Generating k Plots of Other Hypothetical Samples

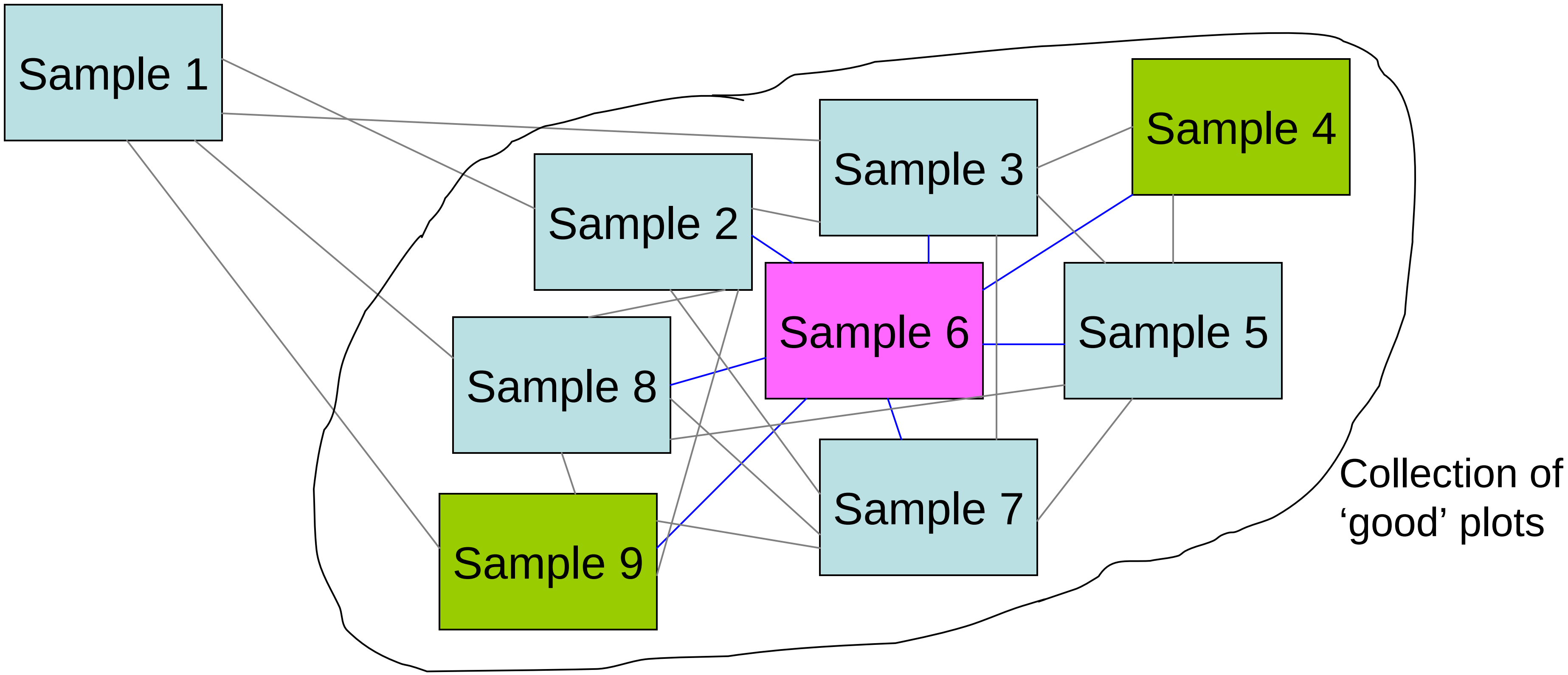
- Use bootstrap methods *[Efron, 1979]*:
  - Resample the data with replacement
  - Compute statistic of interest for this sample
  - Repeat above procedure k times to get its sampling distribution

- An Example of
  the bootstrap:

50 Bootstrapped
Loess Curves for
a simulated data set

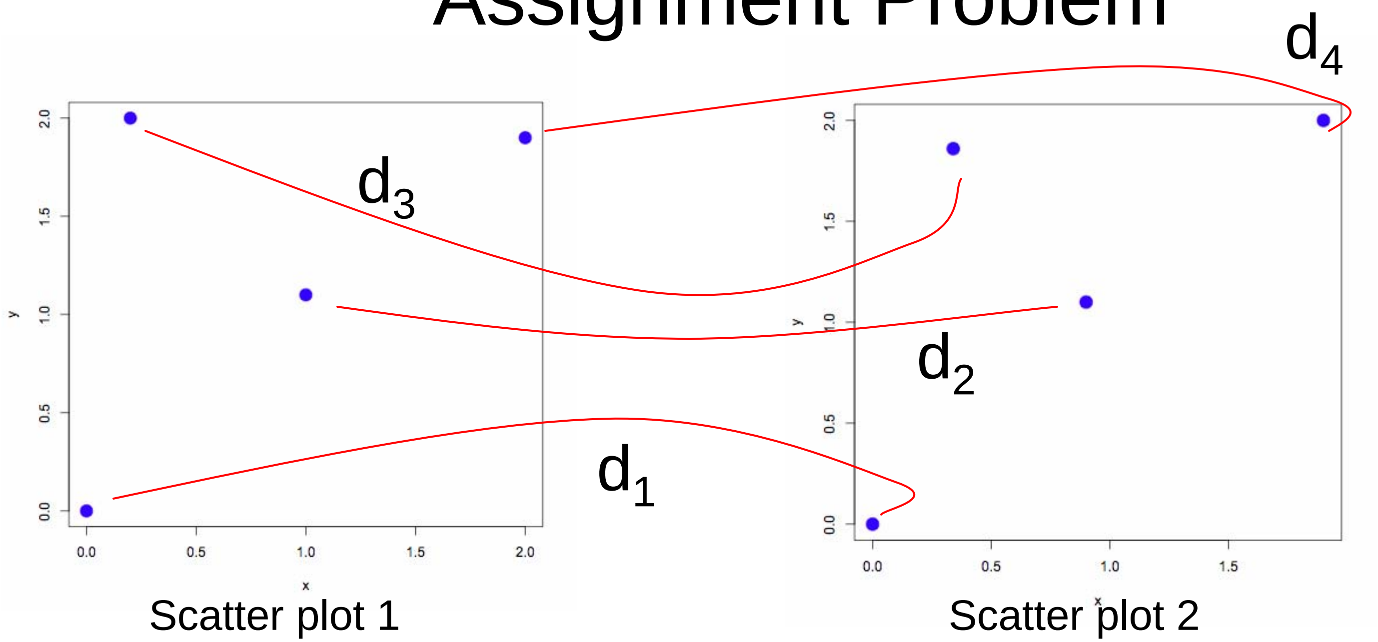


2: Filtering and Summarizing a Group of Plots



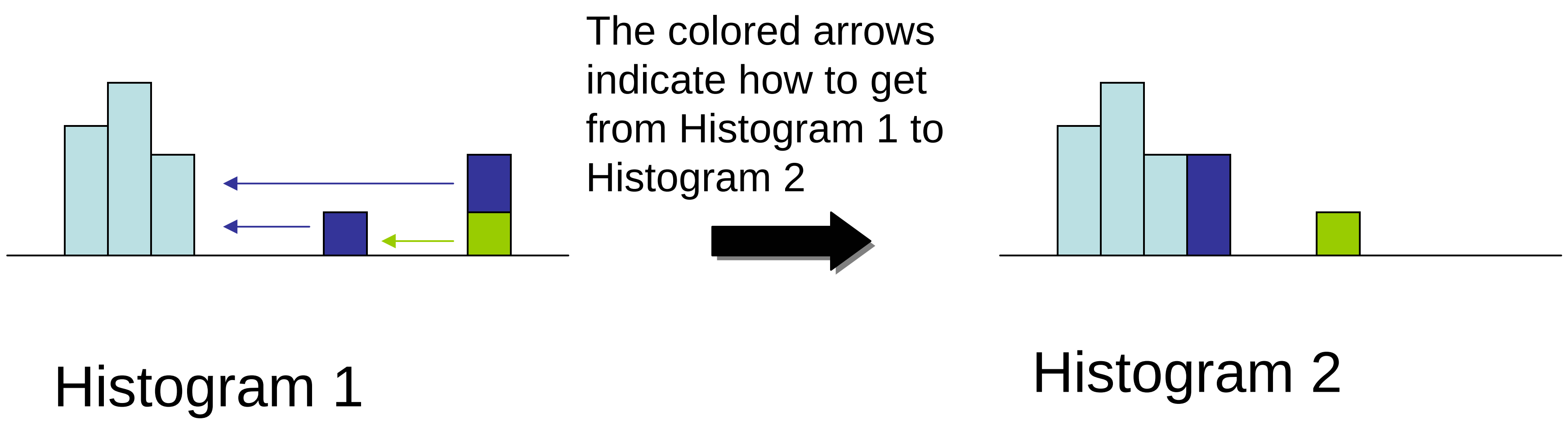
Central plot has minimum summed distance to other plots
'Good' plots are those plots closest to the center, say 95% of them
Border or extreme plots are the two good plots farthest from each other

Distance Between Two Plots is an Assignment Problem



- Assign each point in scatter plot 1 with a corresponding point in scatter plot 2 optimally
- Distance = $d_1+d_2+d_3+d_4$
- This metric is a special case of **Earth Mover's Distance** [*Peleg, Werman, and Rom, IEEE, 1989*]
- Minor modification allows for generalization to other types of plots

The Earth Mover’s Distance

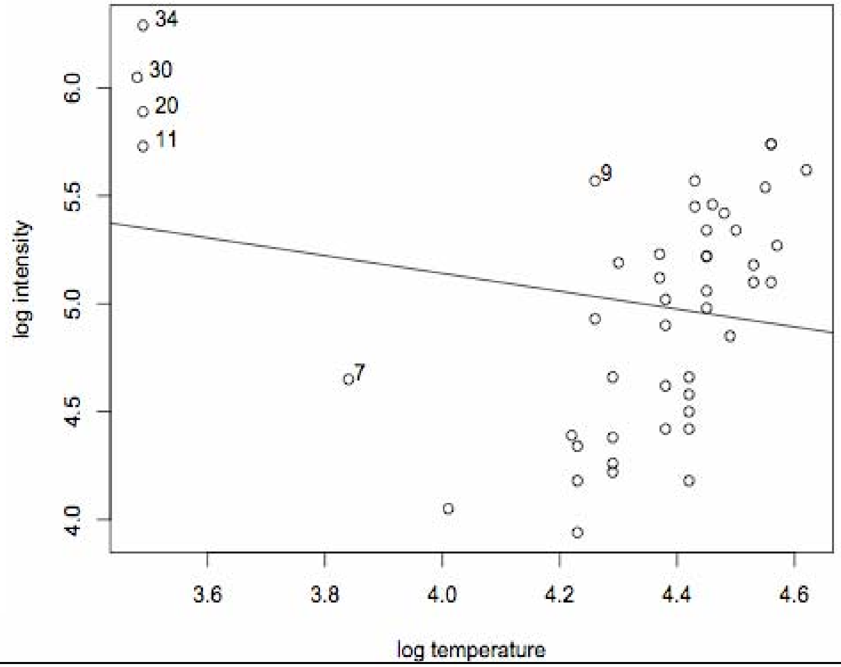


- Histograms are viewed as piles of ‘Earth’ or dirt
- Earth Mover’s distance equals the amount of work ((amount moved) * (distance moved)) required to turn one pile into another pile
- Computing the Earth Mover’s distance requires solving an assignment problem (a network flow problem)
- Earth Mover’s distance generalizes to several types of plots:
  - Scatter plots, parallel coordinate plots, biplots, … etc.
- Ordering plots is related to the traveling salesman problem (Touring a set of cities with smallest total distance.)

Ex. 1: Our method depicts variability of relationships in data

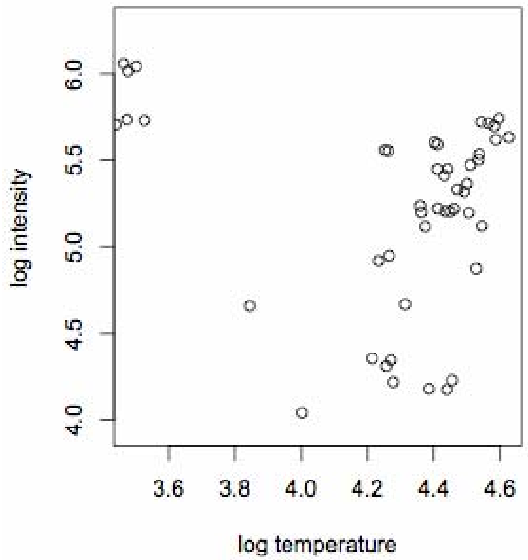
INPUT

Plot of the original data **(Hertzsprung Russell Star Data)**

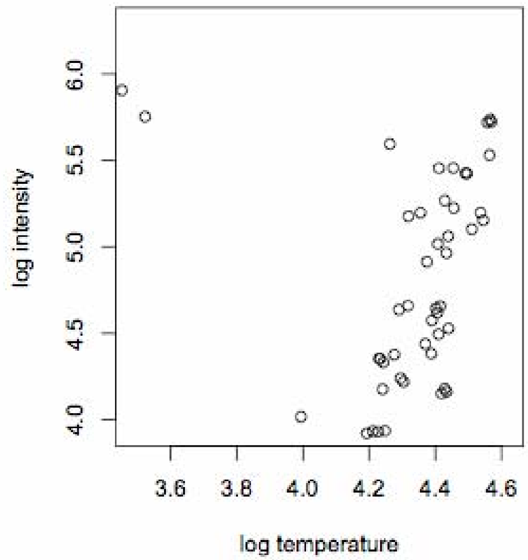


OUTPUT

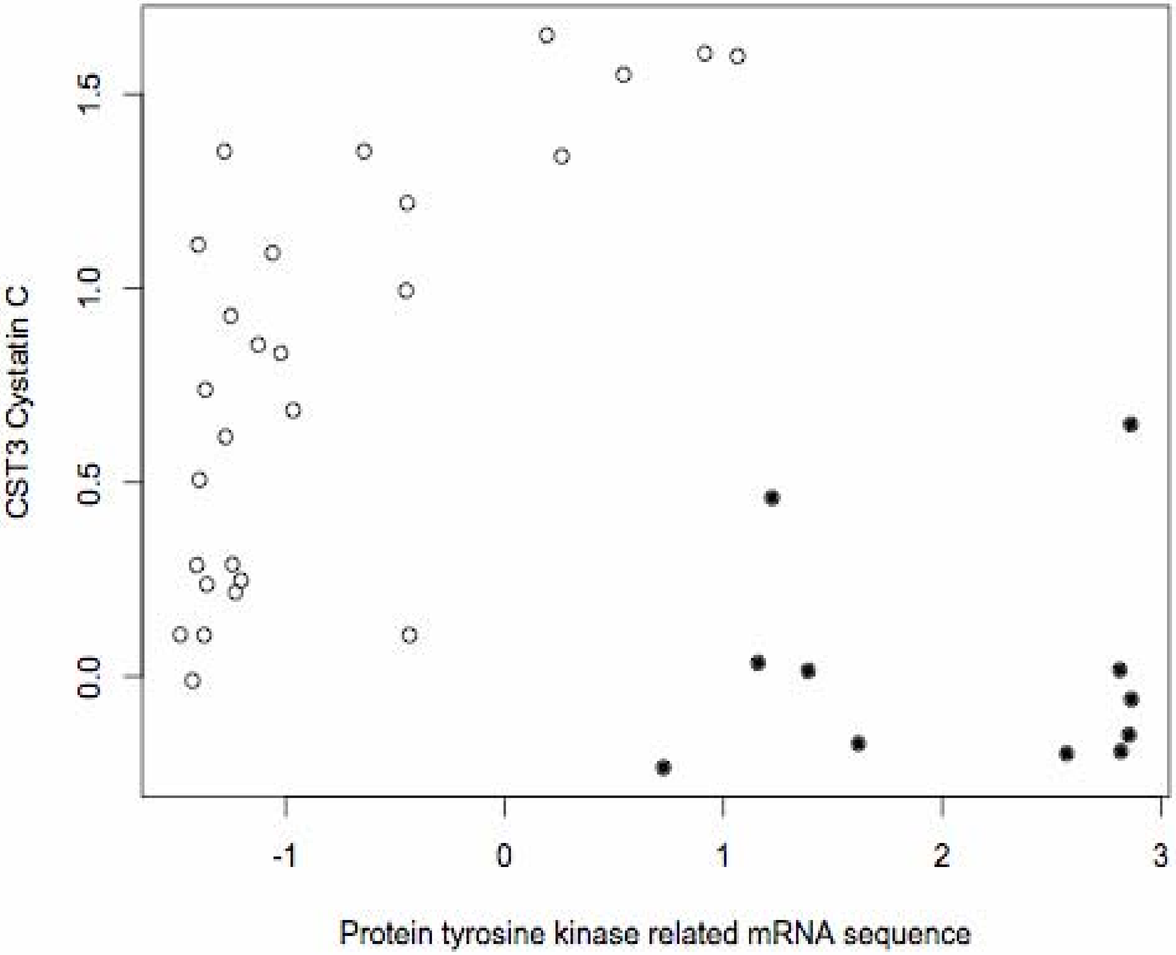
Plot of one extreme



Plot of the other extreme



Ex. 2: Our method captures the optimism created by looking through many plots

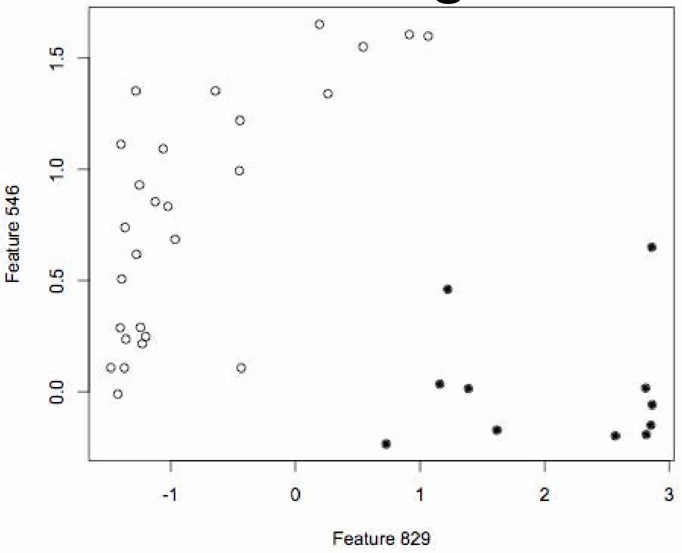


PROBLEM: We searched (forward logistic selection) through 3,051 variables to get a relationship like this!

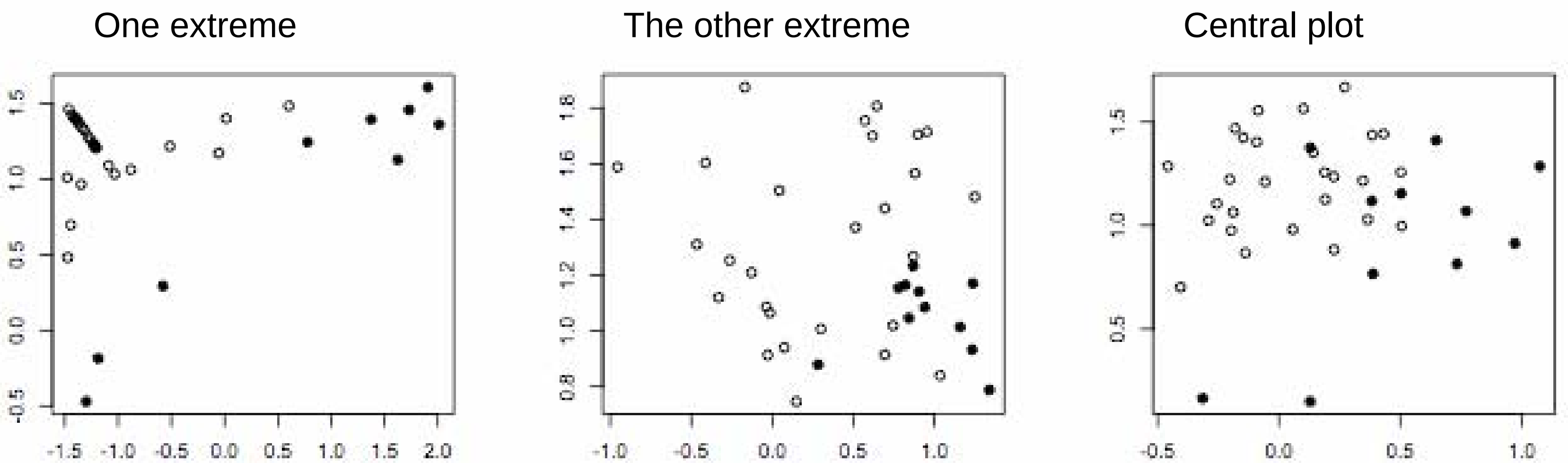
Leukemia patients (not filled in) and healthy patients (filled in) on a scatter plot of the mRNA readings of two genes

Our method shows that the plot *is* optimistic, but still interesting!

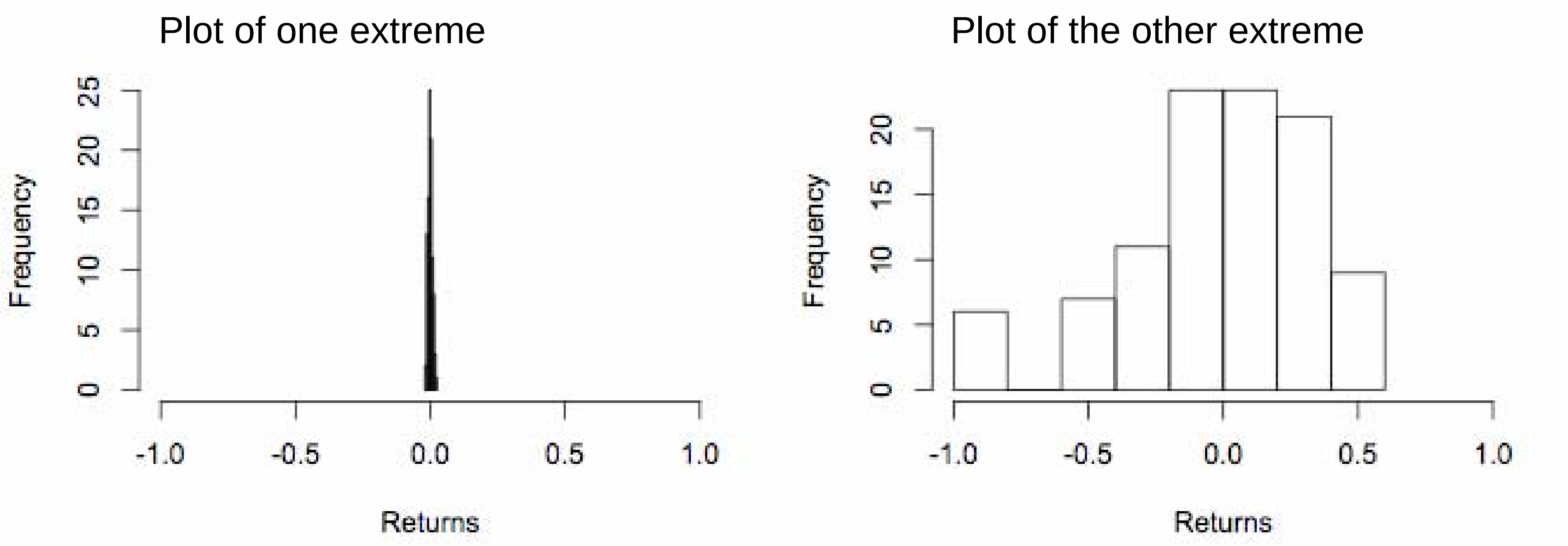
Plot of the original data



Plot of the optimal features when cancer labels were randomly assigned



Ex. 3: Our method demonstrates the highly variable results of portfolio optimization!



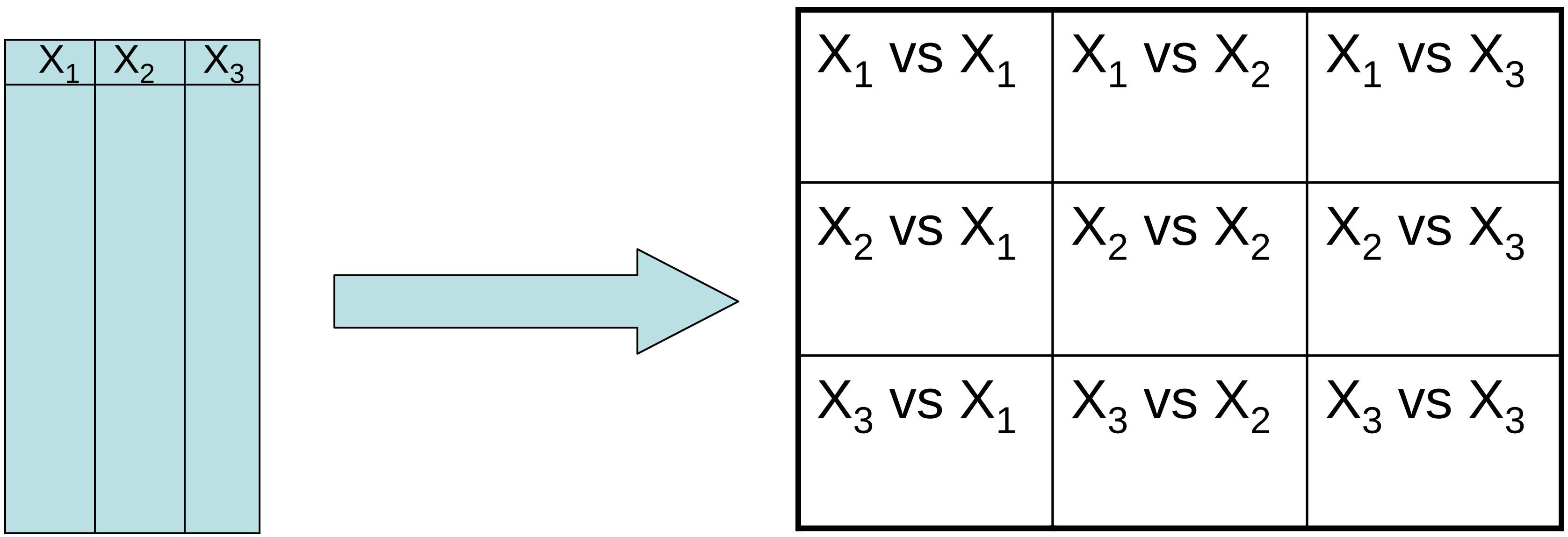
- Data are daily returns on 50 industries among the MSCI US Equity indices 01/03/1995 - 02/07/2005
- Portfolio weights trained on first 100 days, in order to maximize Sharpe ratio
- Object of interest is the histogram of portfolio returns for the next 100 days

Advantages of the approach

- Generalizes to several types of plots
- Only two plots are necessary to convey the message
- Can report the most interesting plots in a data set while remaining statistically sound
- Improves validity of visualization in statistics

Scatter plot matrices are used to visualize multivariate data

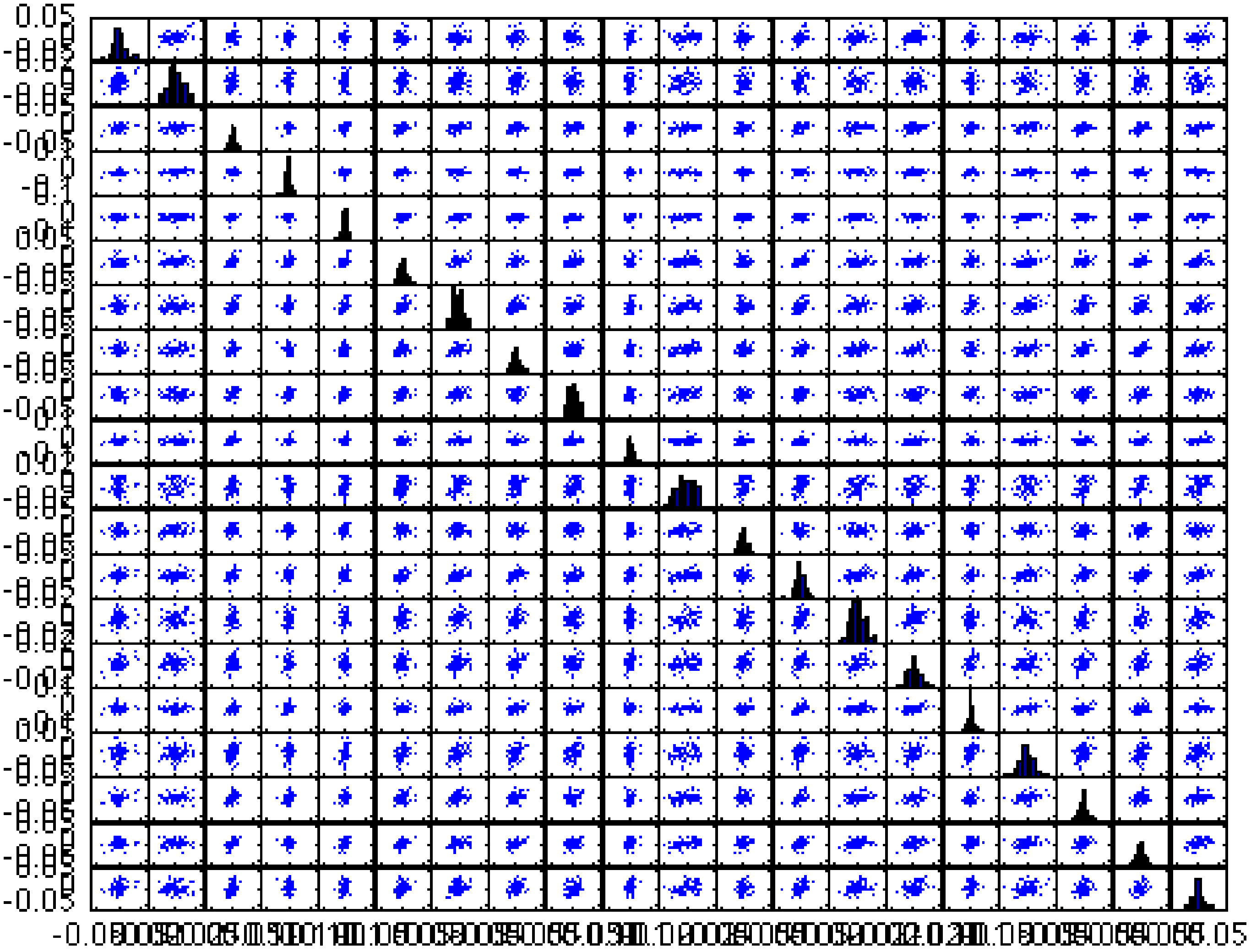
What is a scatter plot matrix?



Data set

Scatter plot matrix

Motivation: Scatter plot matrices can get very complex with many variables!



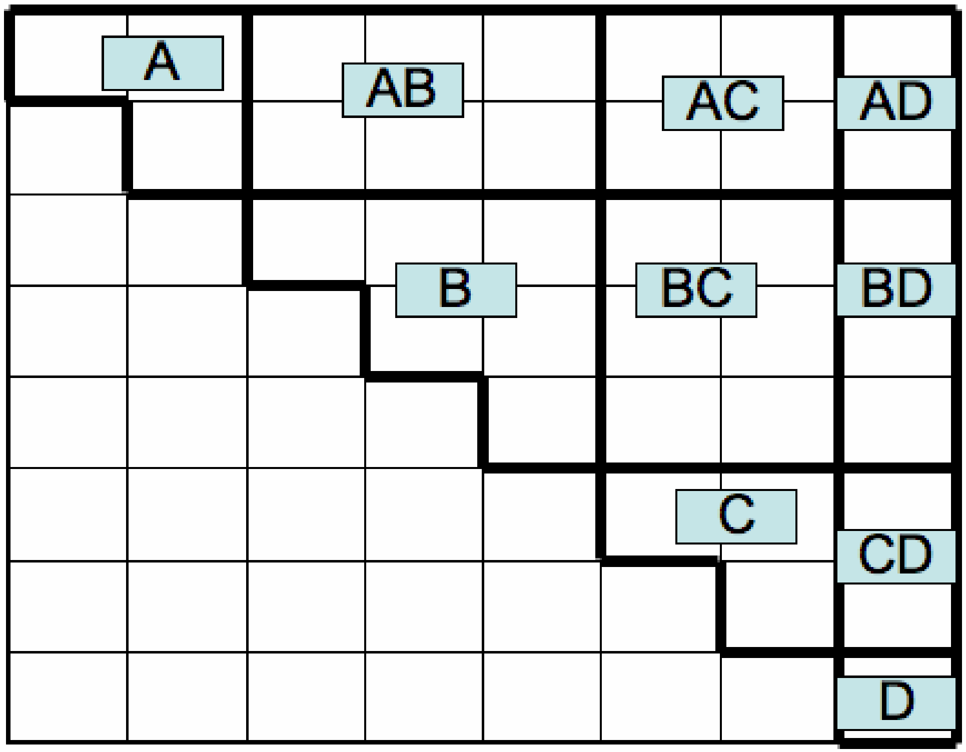
Related Literature

- Reorder variables so prominent plots are on the diagonal *[Hurley, 2004]*
- Principal Component-related methods *[Pearson, 1901]*
- Scagnostics *[Tukey, 1985]*

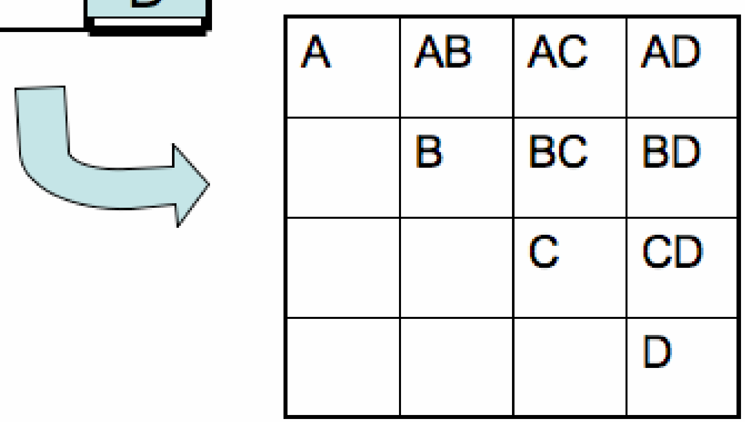
Previous methods either tend to have non-interpretable features, or do not reduce the size of the scatter plot matrix!

Result of our method: In certain cases, we can reduce the size of the scatter plot matrix, while keeping feature interpretability

Methodology Picture



Original scatter plot matrix, with variables reordered so that similar images are near each other

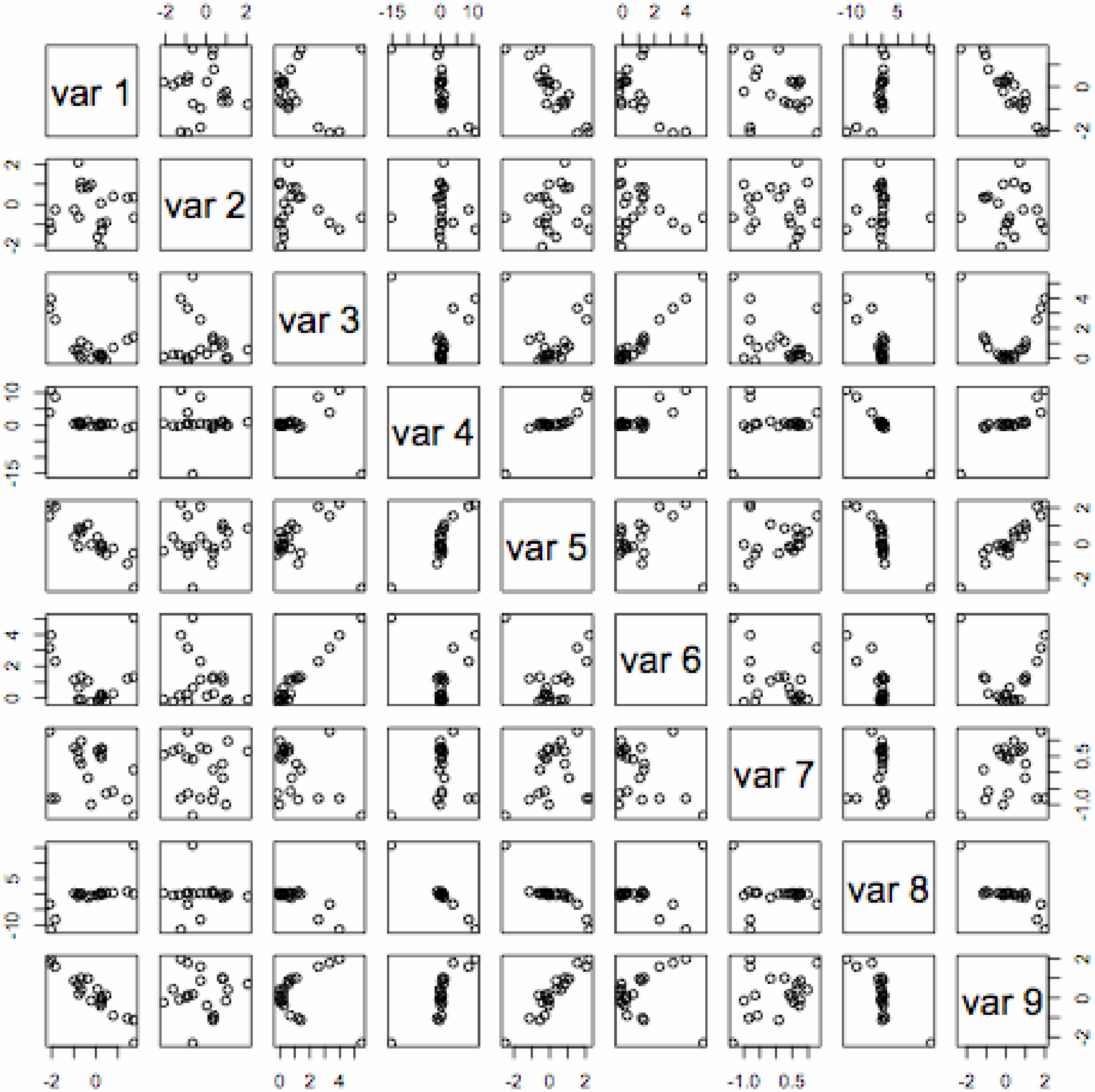


Reduced scatter plot matrix

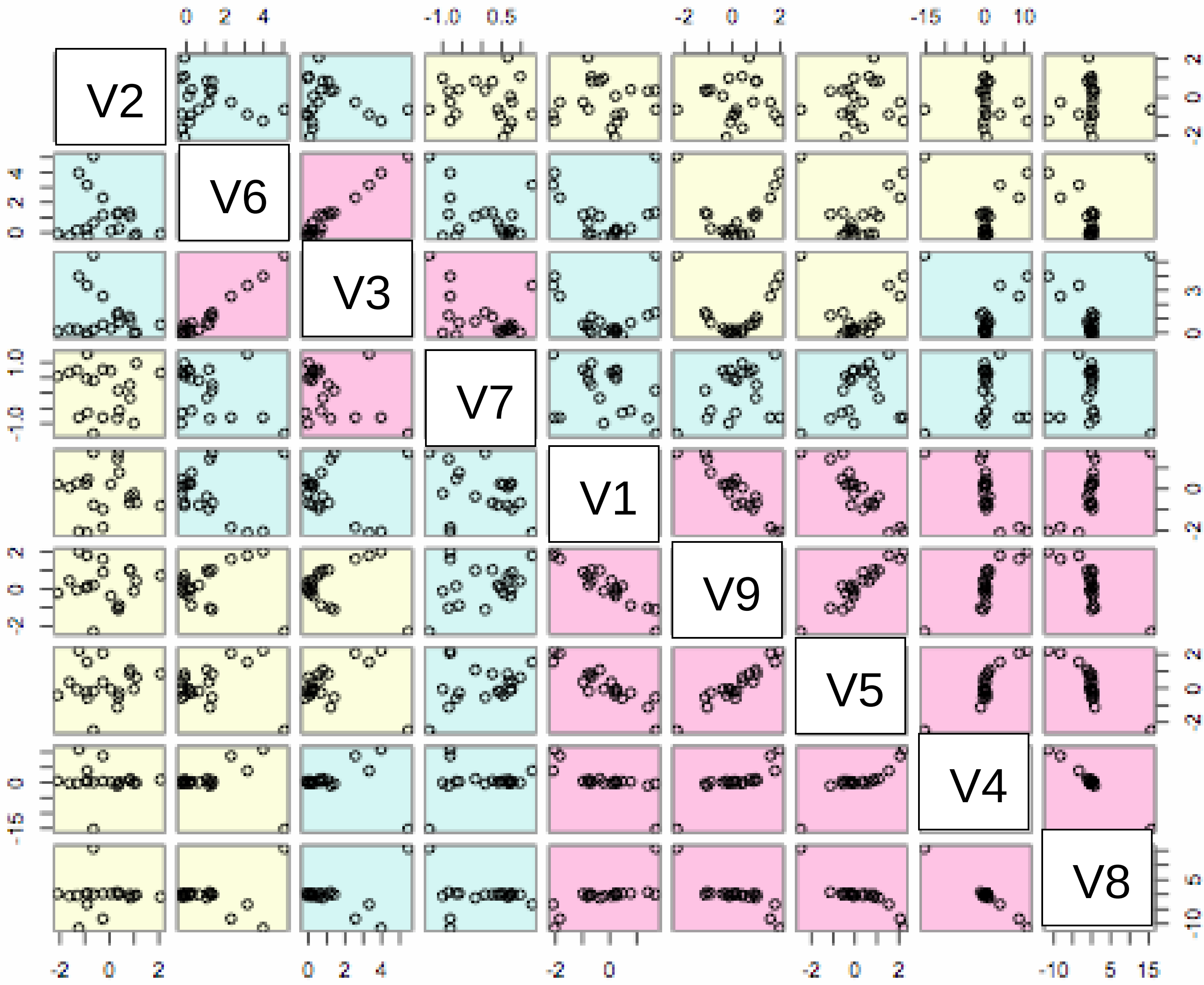
Methodology Description

- **Step 1: Group variables together**
  - Measure dissimilarity between variables
  - Cluster similar variables together (heirarchical clustering)
- **Step 2: Summarize scatter plot collection that each cell contains**
  - Use method of previous section

Example: Simulated dataset with 9 variables (INPUT)

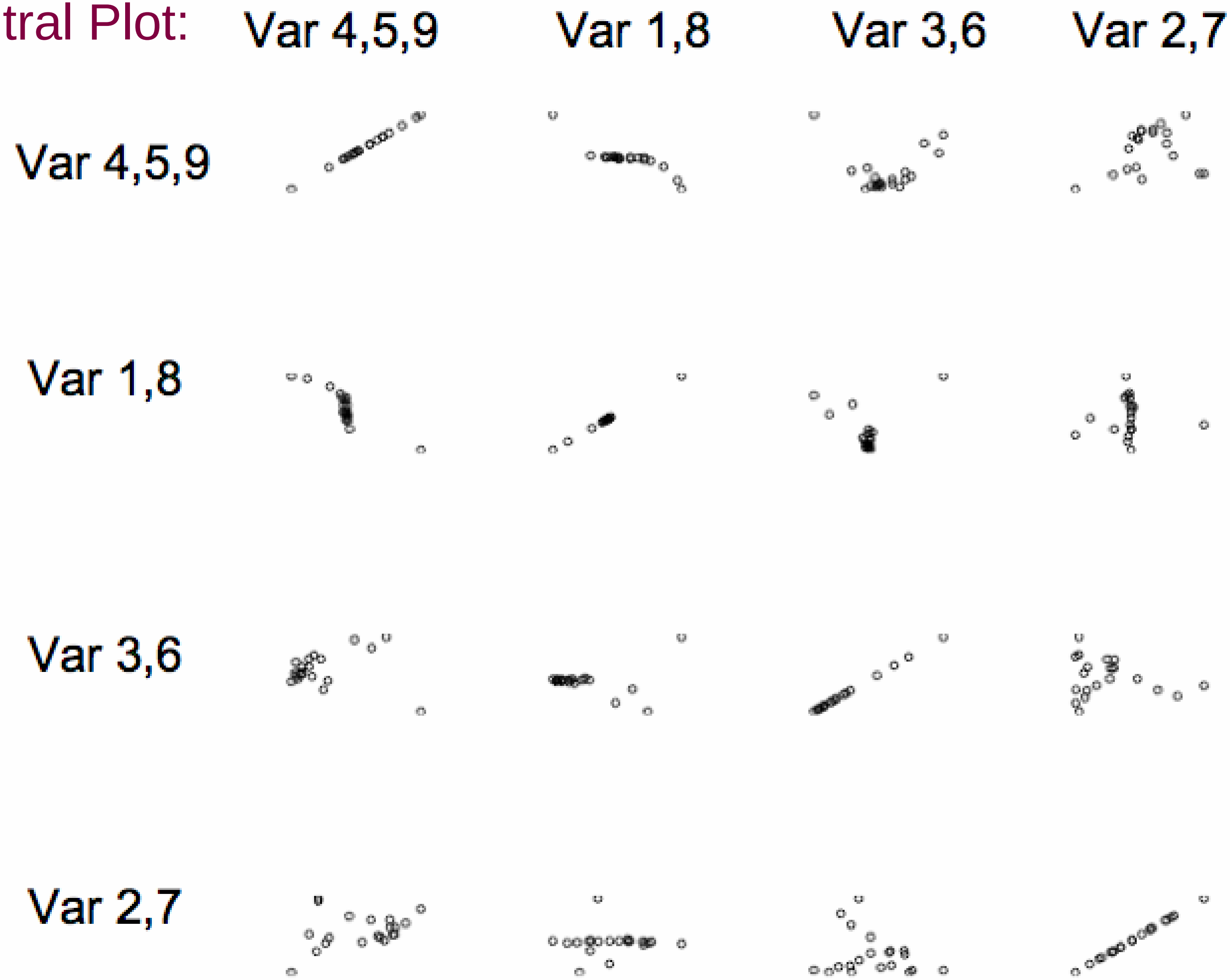


The plot is hard to read even after variables are reordered by previous methods!



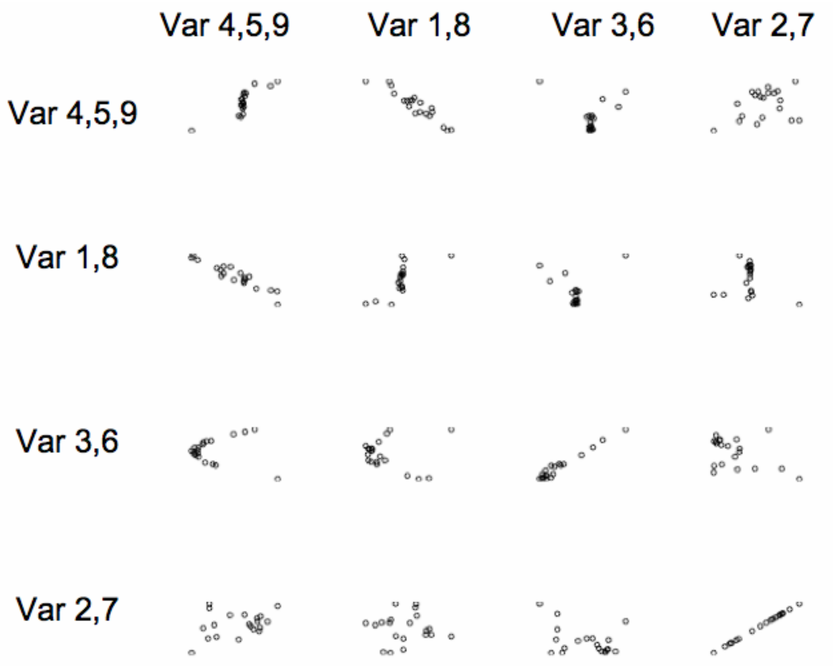
Output of our method shows key relationships!

Central Plot:

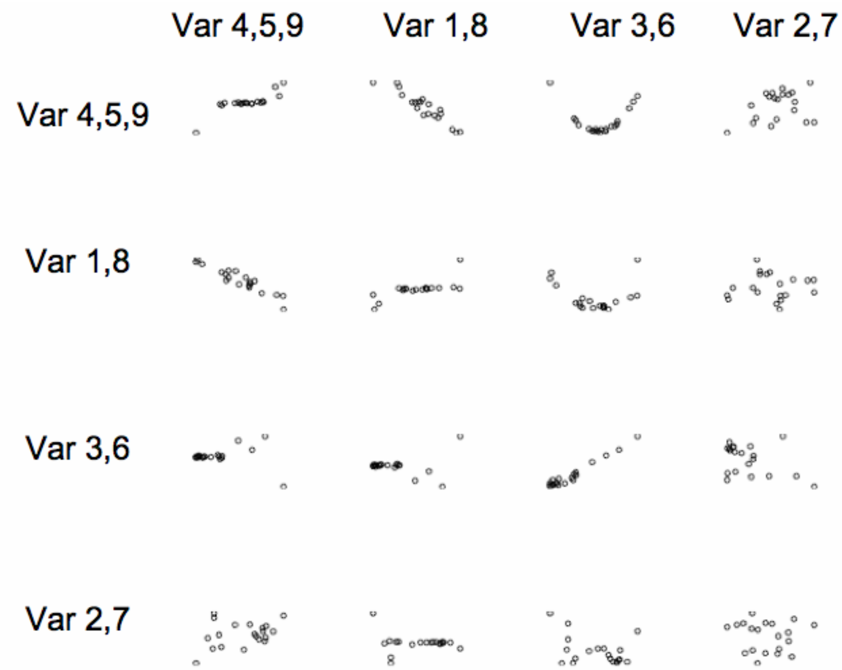


In this case, the reduction doesn't sacrifice much information!

Border of the reduced scatter plot matrix



One Extreme



The Other Extreme

Strengths and Limitations of the Method

- Limitations
  - If it isn’t the case that a few characteristic plots summarize the rest of the plots, the reductions won’t be terrific
- Strengths
  - We can alter feature dissimilarities to do things such as ensuring that certain features are grouped together and certain ones are not
  - We can incorporate sampling uncertainty in this method as well
  - Provides a simplified description of multivariate data

Concluding Remarks

- Contributions
  - Improve the validity of statistical visualization
  - Simplify the visualization of multivariate data
- Future Work
  - Compare to recent work on visual analytics by Buja, et al. (2009)
  - Incorporating prior knowledge in plots

References

1. Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E-K, Swayne, D.F., and Wickham, H. (2009), Statistical Inference for Exploratory Data Analysis and Model Diagnostics, *Royal Society Philosophical Transactions A*, vol. 367, no. 1906, pp 4361-4383.
2. Hurley, C.B. (2004), Clustering Visualizations of Multivariate Data, *Journal of Computational and Graphical Statistics*, 13: 129-133.
3. Menjoge, R. (2010), New Procedures for Visualizing Data and Diagnosing Regression Models, MIT Ph.D. Thesis.
4. Menjoge, R. and Welsch, R.(2010), Visualizing the Sampling Variability of Plots, *Proceedings in Computational Statistics: COMPSTAT 2010.*
5. Tukey, J. and Tukey, P. (1985), Computer Graphics and Exploratory Data Analysis, *Proc. of the 6th Conf. and Exposition of the National Computer Graphics Association*, 773-785.