MANCHESTER 1824

The University of Manchester

Big Variates - Visualizing and identifying key variables in a multivariate world

Or - "Information Theoretic Multivariate Analysis "

Steve Watts, Lisa Crow

School of Physics and Astronomy
The University of Manchester
Manchester, UK

Introduction

In science we visually explore data with histograms (1D) and scatter plots (2D) to find interesting relationships → "Exploratory Visual Data Analysis"

Question:

→ How do you display and identify key variables in a multivariate dataset (D >> 2) ?

Answers:

i) Multivariate visualisation techniques – hugely helped by computers, colour, brushing, linked plots and use of transparency.

ii) Information – theoretic algorithms to guide which 1D and 2D plots to examine.

Tools:

- ✓ **Visual Exploratory Data Analysis**
- ✓ **Identifying variables that matter**
- ✓ **Exploring relationships between variables**
- ✓ **Guided Analysis – precursor to using datamining algorithms**

Example to illustrate the point - dataset – 'wine data'

Relevant Information about the dataset:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Remarks from the data curator:

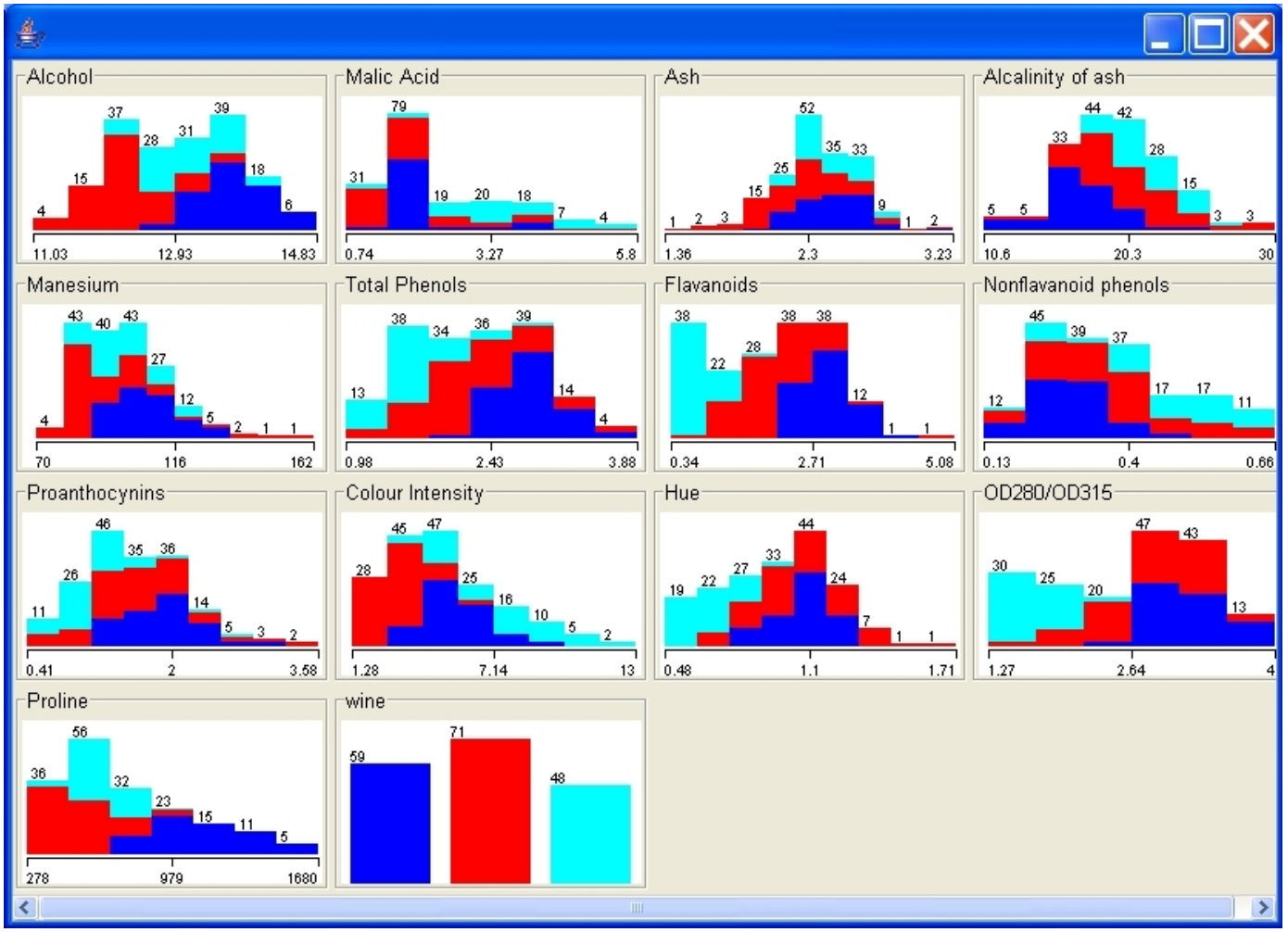
"I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set. " !!!

Number of events: class 1 = 59; class 2 = 71; class 3 =48

Number of Attributes (Variates) = 13

Visualization→ histogram of all the wine dataset

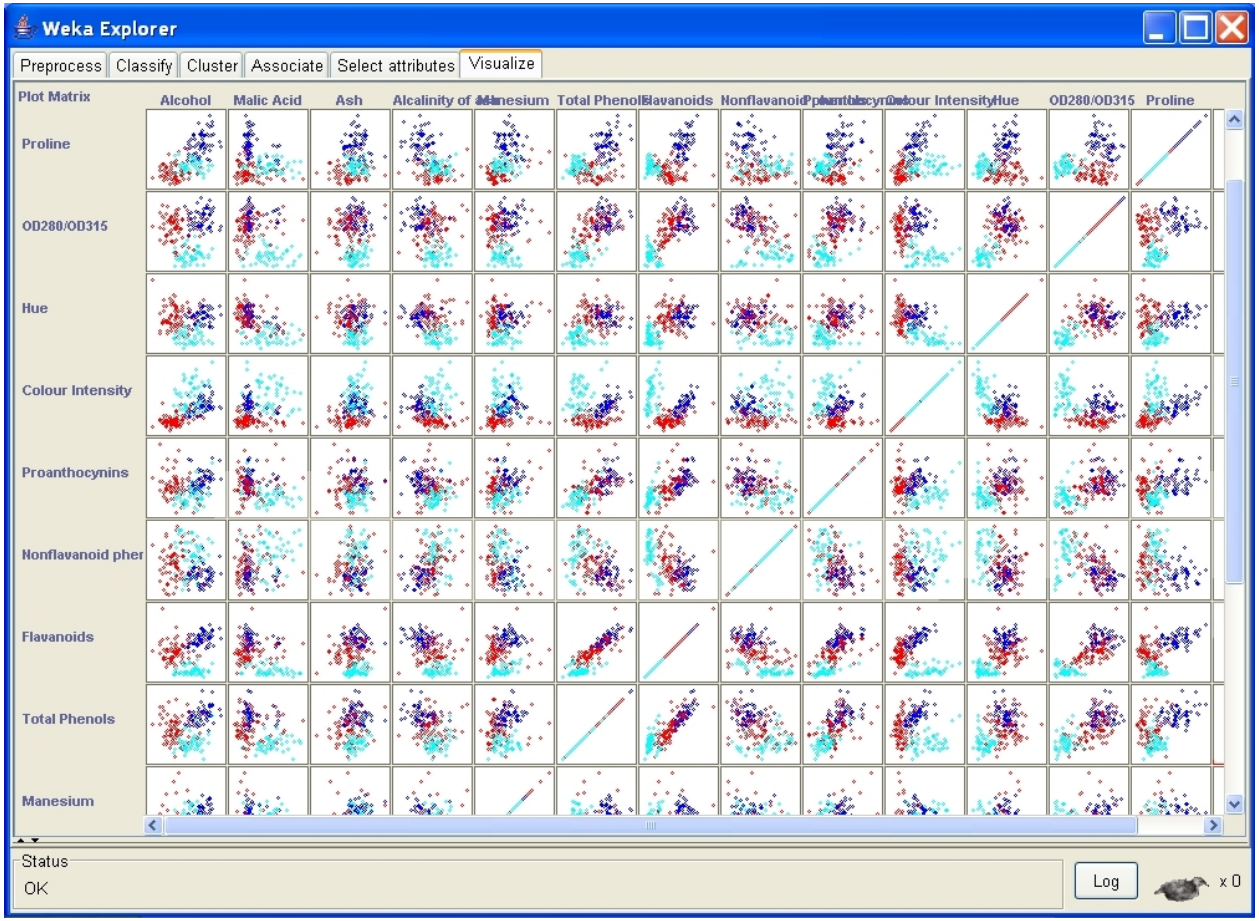
13 variables – so 13 histograms with three classes = 39 images to check!!



Histograms made using open source WEKA software

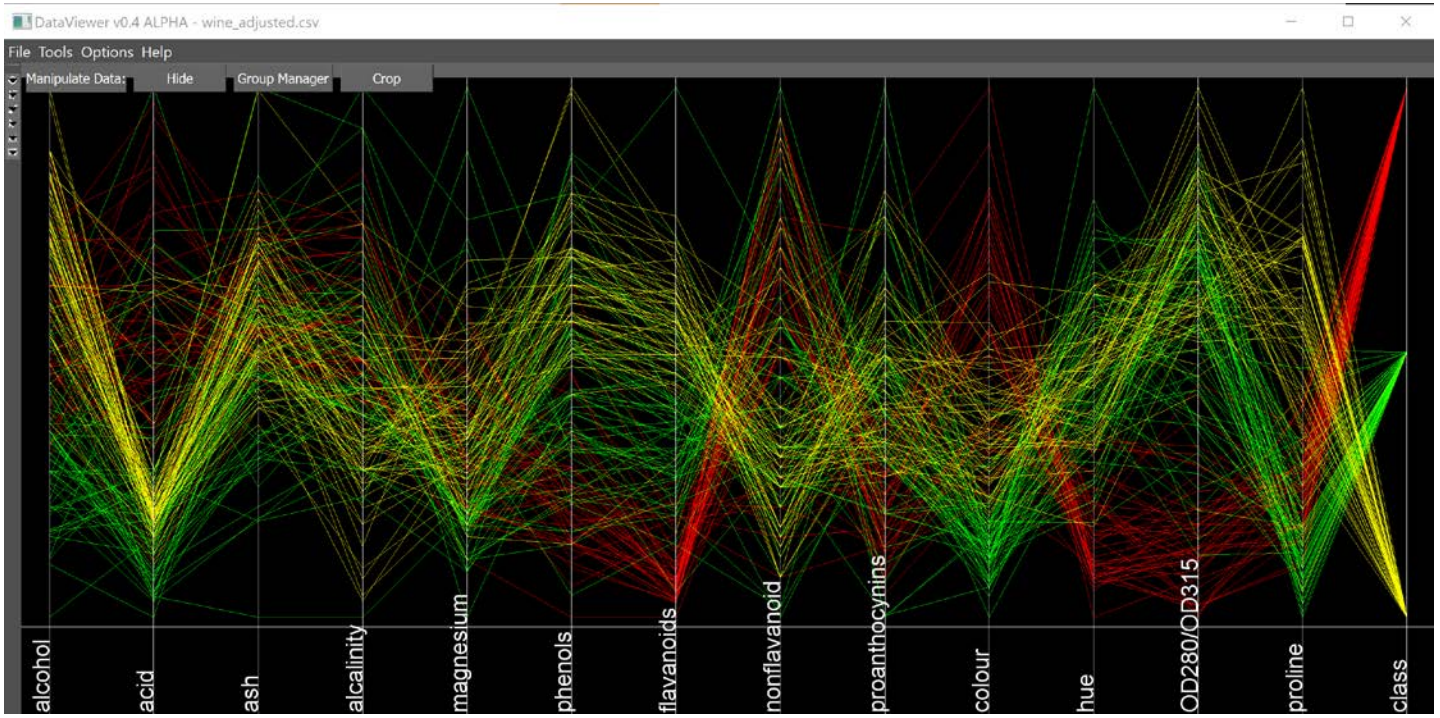
Visualization→ scatter plots of all the wine dataset

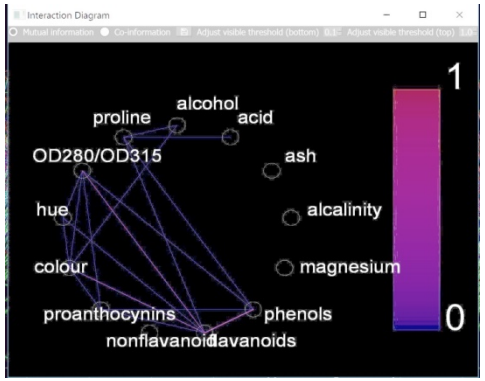
Number of scatter plots = M x (M-1)/2 13x12/2 = 78!!

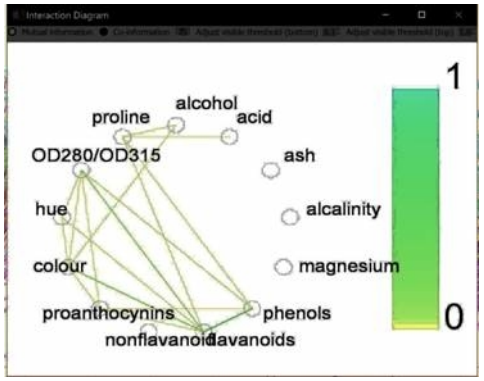


Histograms made using open source WEKA software

Visualization→ parallel coordinates of the wine dataset



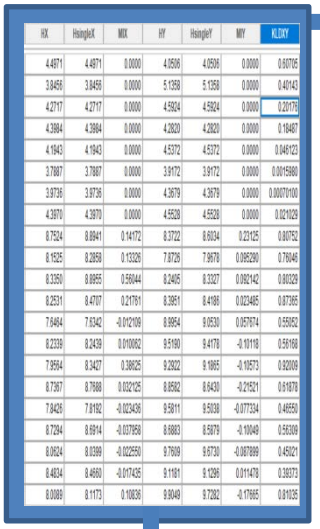




Interaction Diagrams

Information theory view of the problem

Big Variates - Variables matter as much as the sample size or number of events!



M Variables

No of entries – N "events"

Information (Shannon Entropy) ~ 0.5 $\log_2 N$ per variate (continuous)

Total Information = HX + HY + …. – "mutual or shared information"

What if you take more data ?

4 x N implies just one extra bit per variate – Information theoretic equivalent of four times more data means you halve the errors

Increasing the number of variates (if relevant !) faster way to create more information. But it brings its own challenges.

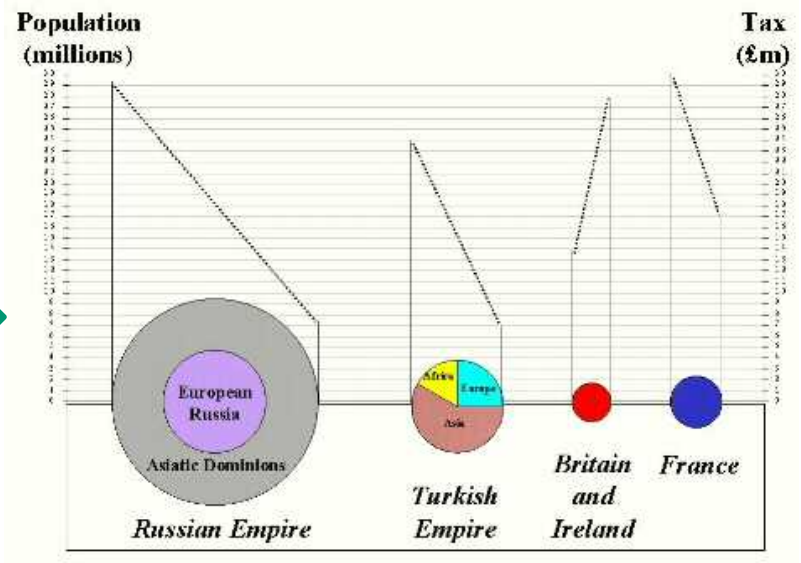
Improved experiment may be more effective than lots more data.

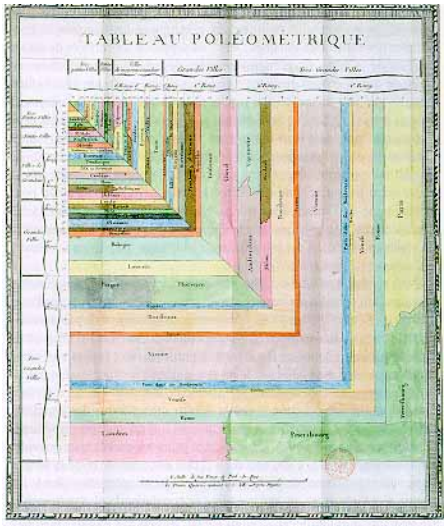
$H = \frac{1}{2}\log_2 N$ *per variate for continuous data*

Some history of data visualization

Luke Howard 1800
1st use of coordinate paper in a research paper

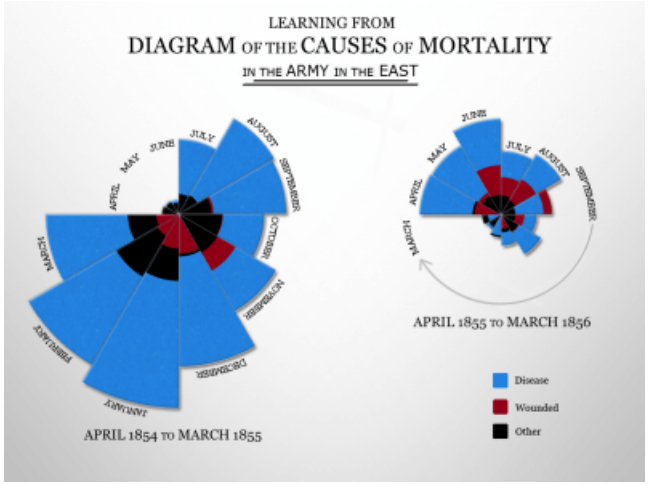
William Playfair's chart
Invented pie-chart. 1801





Charles de Fourcroy 1782
"proportional squares"

Florence Nightingale "coxcomb"
Polar area chart 1857



Milestones in the history of thematic cartography, statistical graphics and data visualisation – M. Friendly and D. Denis

1975 to now High D data visualisation

Some key dates…**selective list ..**

1985 Alfred Inselberg **Parallel Coordinates**
1985 D. Asimov **Grand Tour**
1985 DataDescription Inc. Paul Velleman Cornell - DataDesk
1987 A. Becker and W. Cleveland **Linking and Brushing**
1998 A. Buja, D. Asimov, C. Hurley, J. McDonald XGobi
1990 **E. Wegman Statistical analysis and parallel coord. CrystalVision**.
1991 M. Friendly Mosaic Display and Categorical data
1999 L. Wilkinson "Grammar of Graphics"

Systemization of data and graphs and graph algebras in an OO framework.

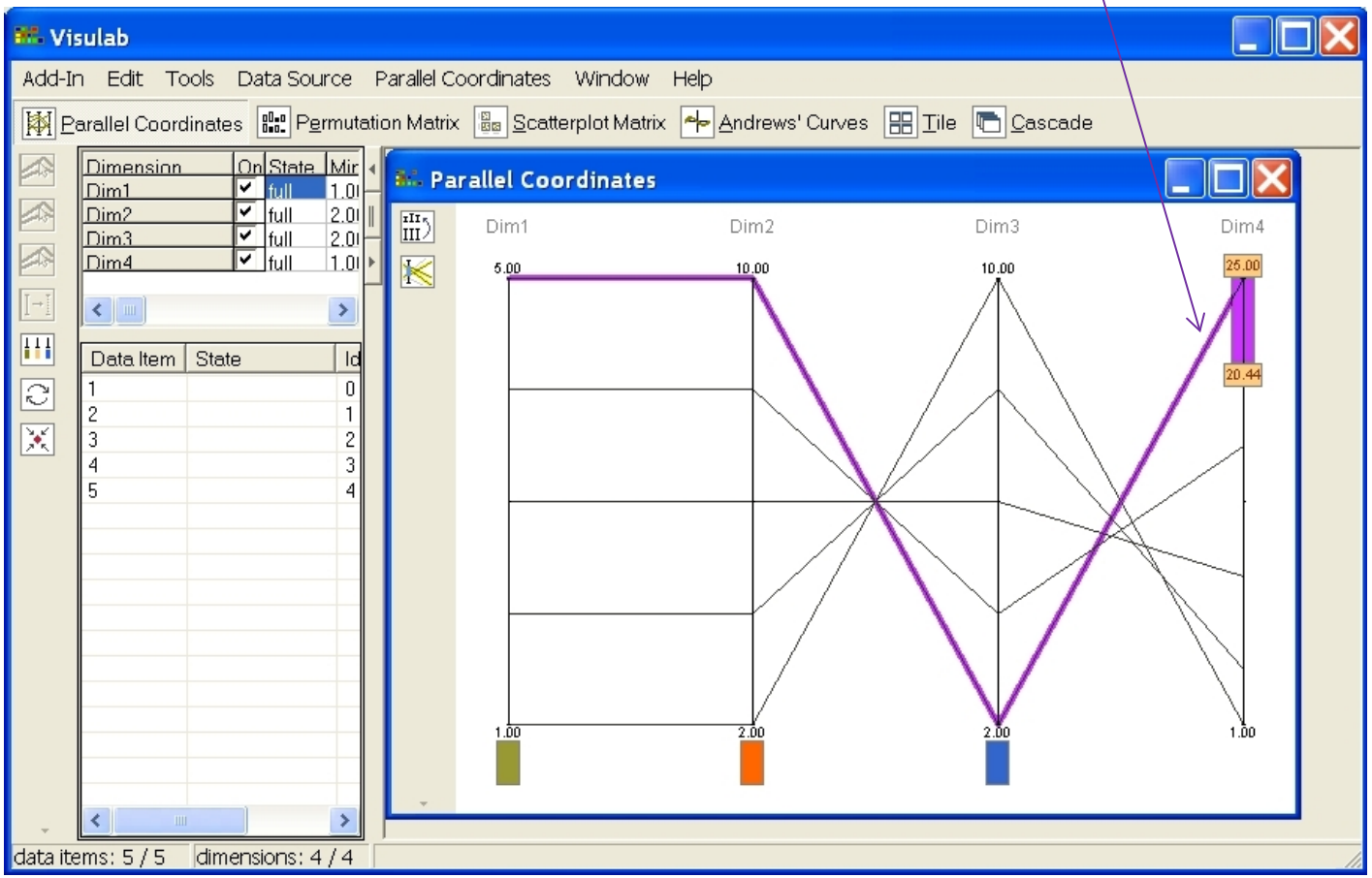
Thanks to Michael Friendly website http://www.datavis.ca/milestones/

Introduction to Parallel Coordinates

DataPoint	Dim1	Dim2	Dim3	Dim4
1	1	2	10	1
2	2	4	8	4
3	3	6	6	9
4	4	8	4	16
5	5	10	2	25

Simple Implementation with EXCEL plugin http://www.inf.ethz.ch/personal/hinterbe/Visulab/

This also shows the idea of brushing!!!



Practical example

E. Wegman has done much on the use of parallel coords.

Hyperdimensional Data Analysis Using Parallel Coordinates

Edward J. Wegman

Journal of the American Statistical Association, Vol. 85, No. 411 (Sep., 1990), 664-675.

Some useful things to note:

- **Clusters**
- **Correlations**

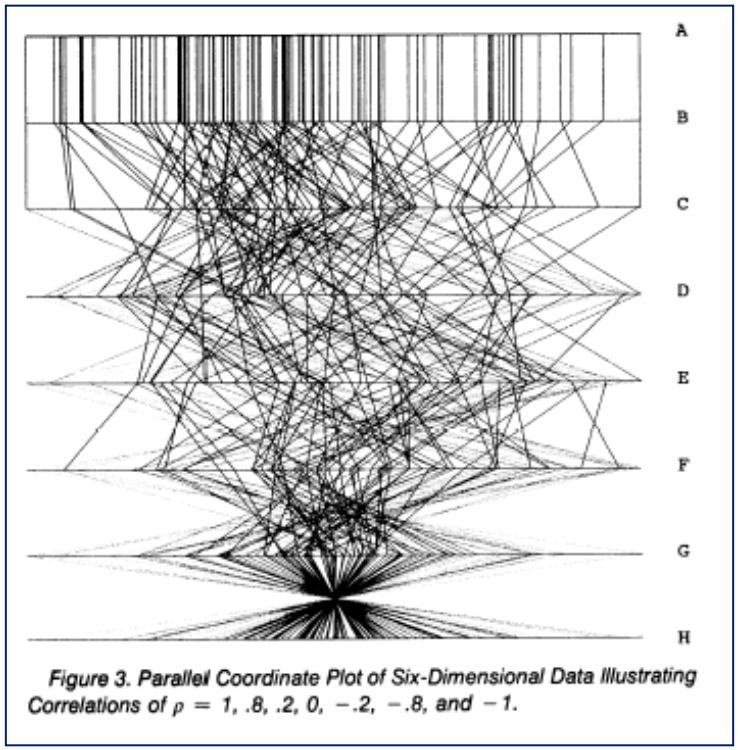


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $\rho = 1, .8, .2, 0, -.2, -.8$, and -1.

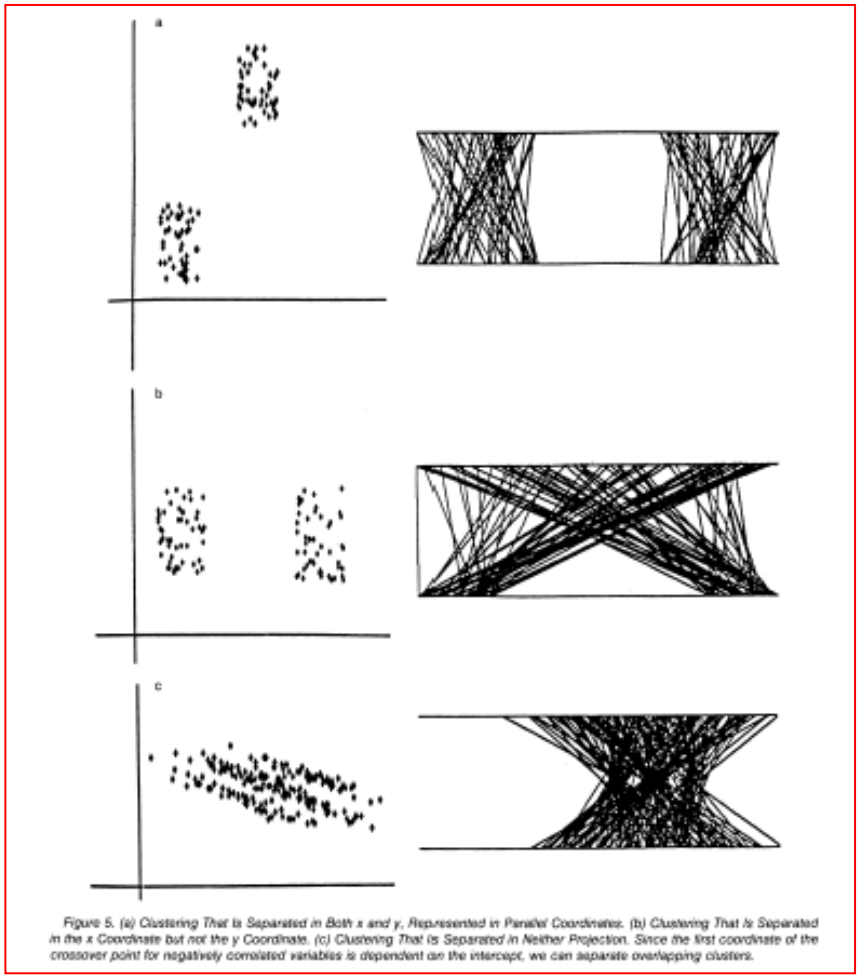
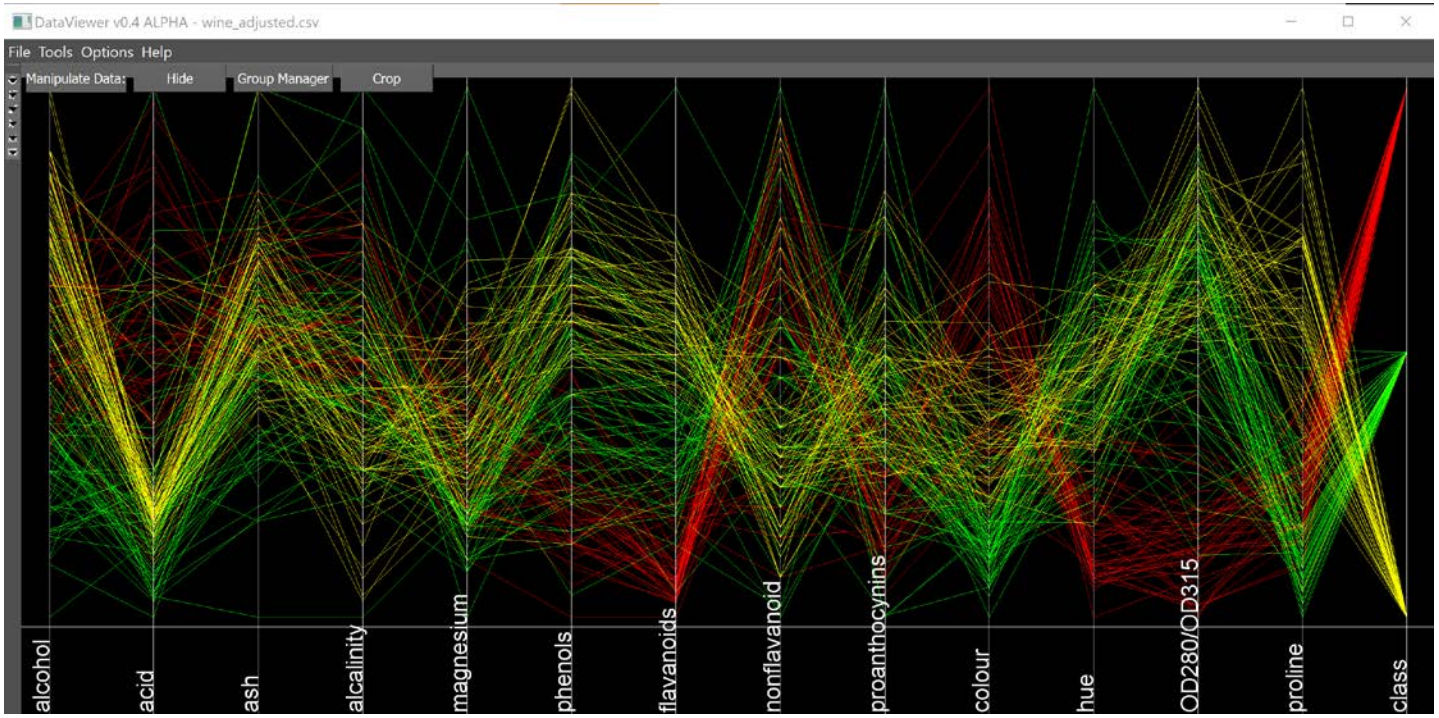
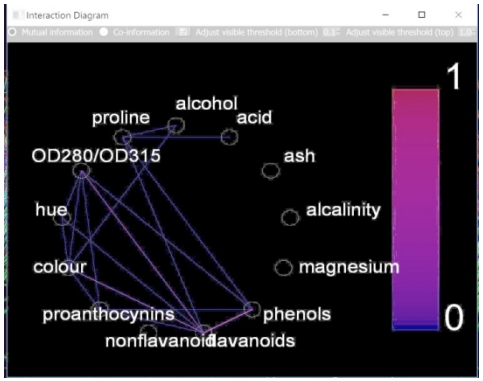
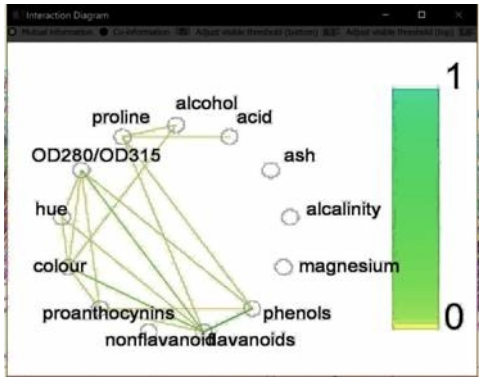


Figure 5. (a) Clustering That Is Separated in Both x and y, Represented in Parallel Coordinates. (b) Clustering That Is Separated in the x Coordinate but not the y Coordinate. (c) Clustering That Is Separated in Neither Projection. Since the first coordinate of the crossover point for negatively correlated variables is dependent on the intercept, we can separate overlapping clusters.

Discussion→ parallel coordinates of the wine dataset







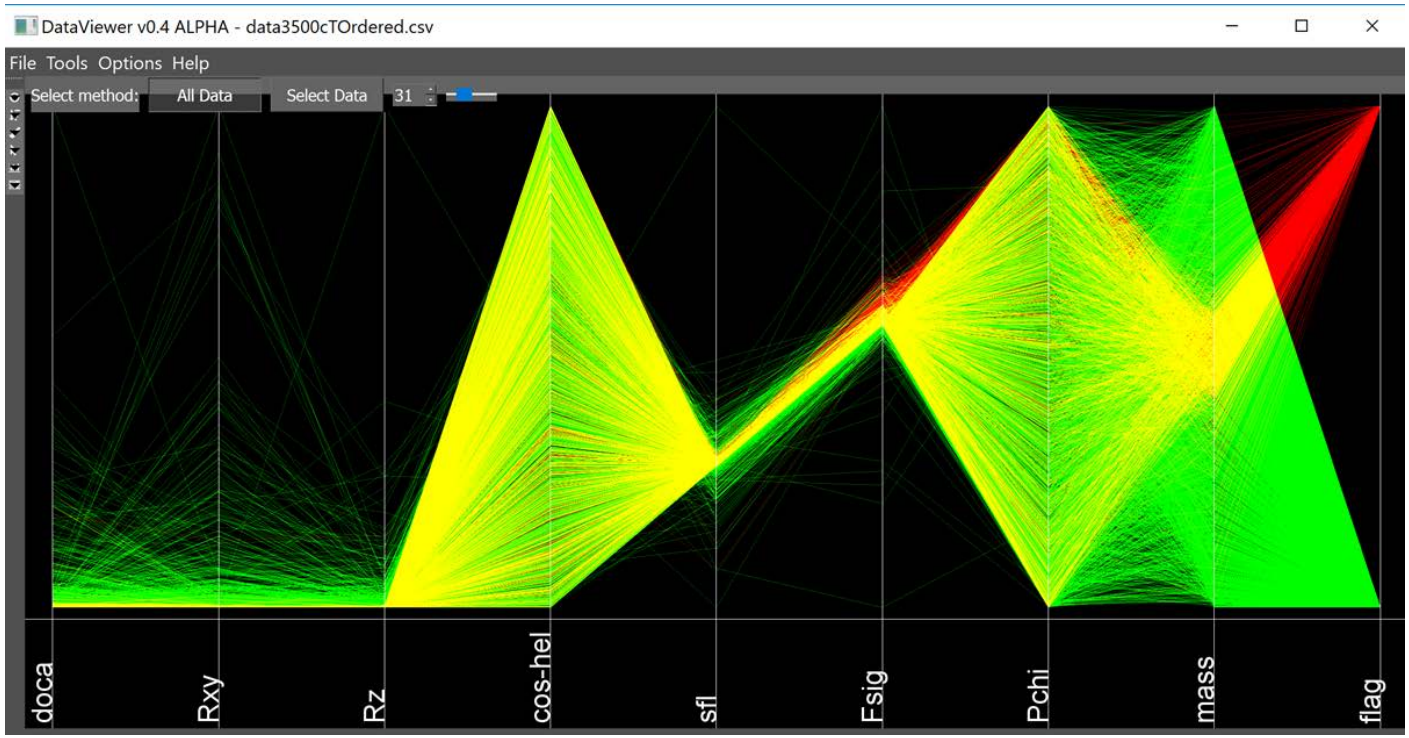
Variables Interaction Diagrams

(they are the same. different colours maps showing different depending on projectors)

Example 2: particle physics Monte Carlo data ‘ppdata’

Decay of K_s to $\pi^+ \pi^-$

1264 Kzero + 3734 background
(and a flag to tell us which is which ! Flag =1 S Flag=0 B



Brush colors:

Signal
Background

variables

Doca = distance of closest approach
Rxy radius of cylinder for interaction region
Rz abs. half length of cylinder defining the IR
Cos_hel abs. Value of cosine of Ks helicity angle
SFL – signed flight length
Fsig stat. Sig. Of Ks flight length
Pchi chisq prob of Ks vertex
Mass – reconstructed mass of the Ks

- **R_{xy} , Doca (and sfl less so) discriminate the background**
- **Only variable where signal can be seen is Fsig.**

Data from Liliana Teodorescu –Brunel University UK

Question: Is there a way to pick out key variables and interesting 1D and 2D plots using data driven algorithms only ?

Visual Exploratory Analysis is vital – but beware personal bias.
Want to avoid any models of the data – "data driven analysis"

$$H = -\sum_1^N p_i \log(p_i)$$ Discrete Shannon Entropy

$$h = -\int p(x) \log(p(x)) dx$$ Differential Entropy

From Shannon's Famous 1948 paper

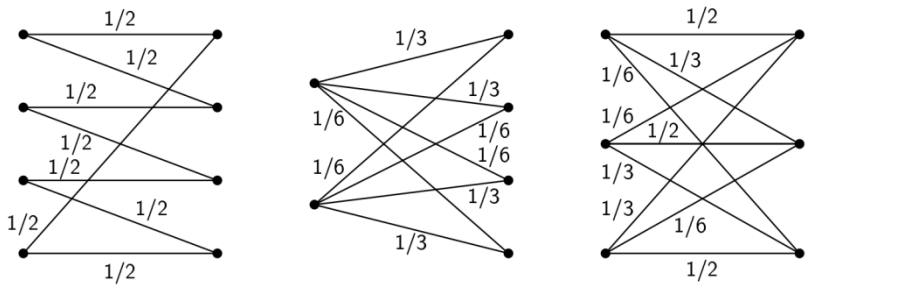


Fig. 12—Examples of discrete channels with the same transition probabilities for each input and for each output.

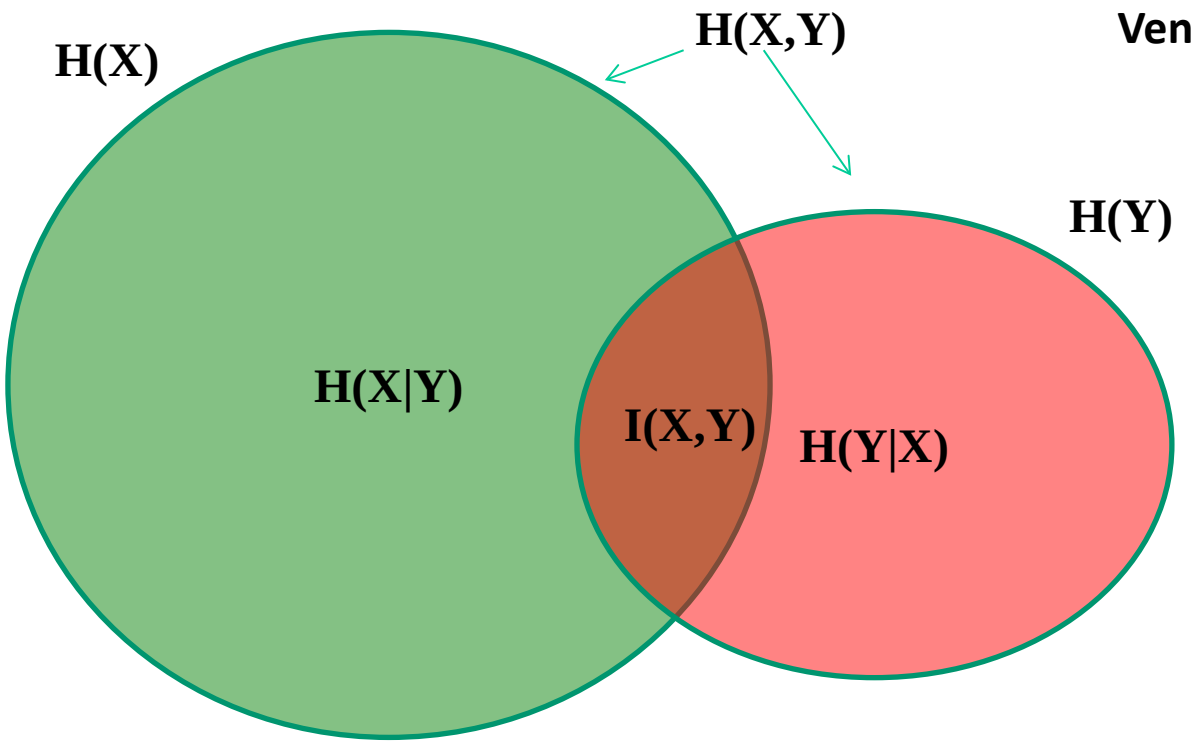
h depends only on the pdf. In nats…

e.g. Gaussian
$h = 0.5\log(2\pi e\sigma^2)$

Uniform
h = log (Range)

Answer Like a parallel coordinates plot!! Use information theory ?

Finding links between variables using shared information



$I(X,Y) = I(Y,X) = H(X) + H(Y) - H(X,Y)$ Mutual Information

Similarity Index, SI = $I(X,Y)/(\text{Min}(H(X), H(Y))$

SI = 1 Variables are completely dependent

SI = 0 Variables are completely independent

To get H for continuous data one must histogram it first.

Histograms: bin size and entropy

H = h – log(D) where D is the bin size……..so H does not look well defined…

- **But, the bin size should be chosen such that the minimum integrated squared error (MISE) is minimized. Basis for Scott's Law, which is often used but need one to know the probability density function (pdf) in advance, which is the reason for the histogram!!**

- **Tackle with a "cost" function → small bins - Poisson Fluctuations are costly**
  →large bins – lose the pdf shape which also costs.

- **H only depends on the number of events – of the form $(1/M)\log_2 N$.**
  →Now the problem is to find the minimal/optimal value of M

- **After a lot of work→ there is an maximum H that ensures that Poisson fluctuations due to over-binning (too many bins) are removed. →This is when H = (1/2)log(N)**

M < 2 Poisson Fluctuations. Over binned
M > 3 Under-binned

" too few bins carry little information, but to many bins lead to too few events per bin" W. T. Eadie, D. Drijard, F. E. James, M. Roos and B. Sadoulet"

"Statistical Methods in Experimental Physics", North Holland, (1971).

Summary - Algorithm for choosing bin width - inspired by asking "what is the entropy of a histogram ?"

$$\Delta = 2^h N^{-\frac{1}{M}}$$

$$H_{Measured} = -\sum p_i \log_2 p_i$$

$$h = -\int p(x) \log_2 (p(x)) dx$$

This algorithm is data driven!!!

Bin width, Δ $2 < M < 3$.
M=2 best for number of events, N >35

Estimate the differential entropy, **h**, using nearest neighbour distance for each point Kozachenko, L. F. and Leonenko derived this non-parametric estimator in 1987.

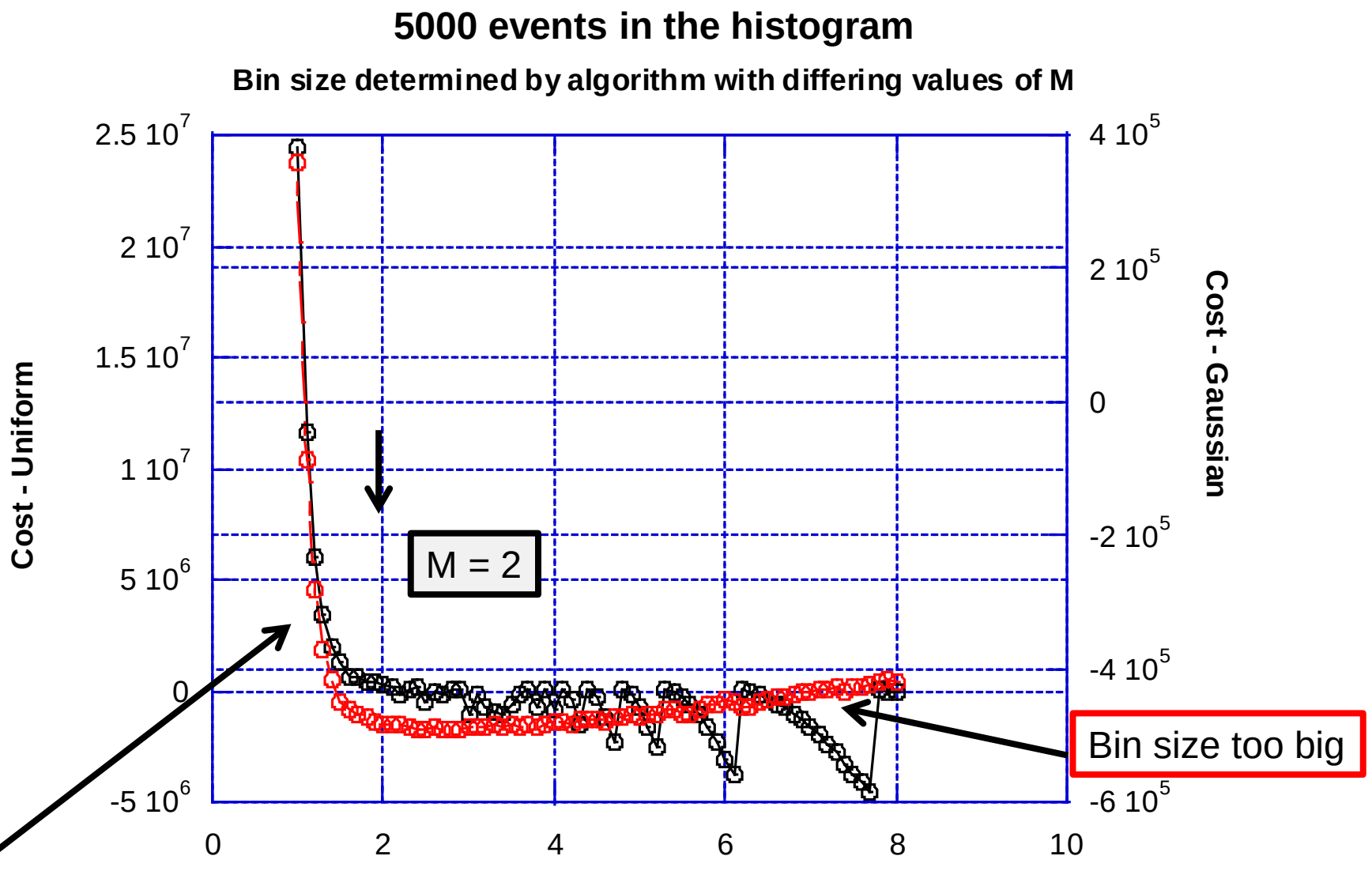
h $$\approx \frac{1}{N}\sum_{j=1}^{N} \log_2 \lambda_j + \log_2[2(N-1)] + \frac{\gamma}{\ln(2)}, \quad (7)$$

where λ_j is the observed distance from x_j to its nearest neighbor.

Binsize, $\Delta = \frac{2^h}{\sqrt{N}}$

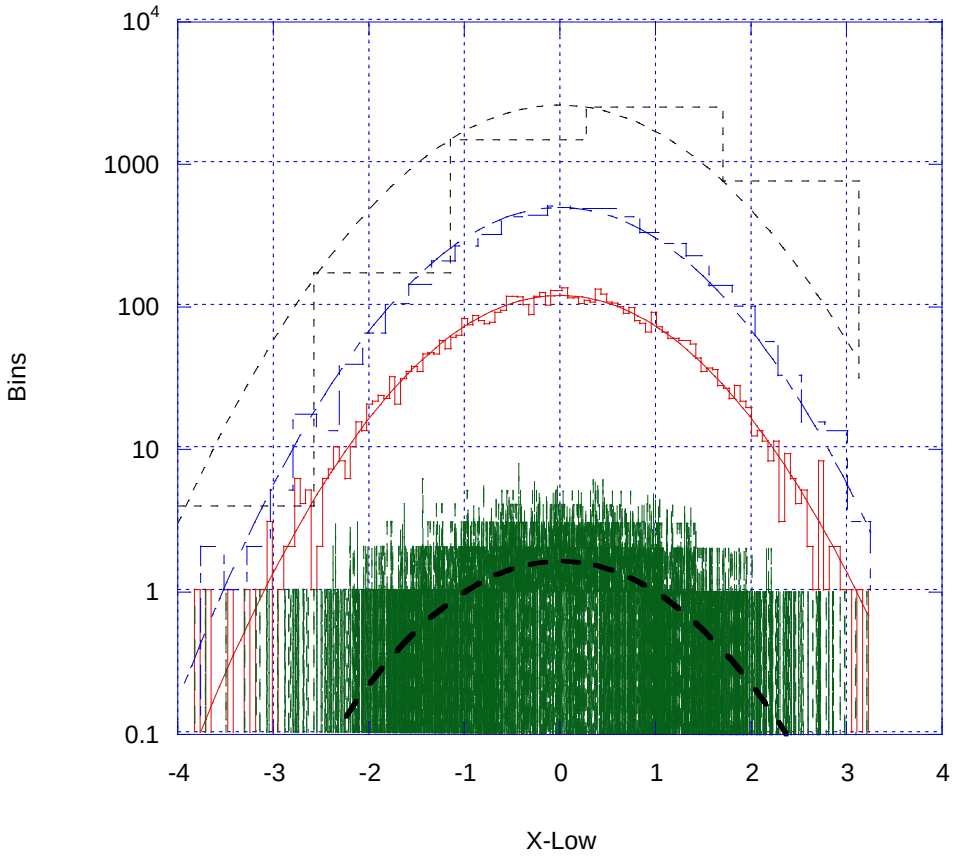
h depends on the data, H is fixed to ½ logN → this defines the histogram bin width!!

Example: Data histogrammed with algorithm for different M and cost function estimated



Cost Function from "A method for selecting the bin size of a time histogram" Neural Computation 19, 1503-1527 (2007) Shimazaki and Shinomoto

Example: Gaussian Distribution



H = (1/M) logN

M = 1 Over binned

M = 2 Optimal

M = 3 Still OK.

M = 8 Under-binned.

M = 2 Optimal for N > 35. Thus - H = ½ logN

Note: for uniform distribution between 0 and 1, h=0 => $\Delta = 1/\sqrt{N}$
This is the algorithm used in Microsoft Excel !!

Discussion on the meaning of histogram entropy

- Entropy means that a minimum of 2^H bins are needed to histogram the data

  e.g. a fair coin with a head or tail, each has a probability of 0.5

  Thus H = 1 and there needs to be 2 bins.

- Note: Unfair coin then H will be between 0.0000X and just short of 0.9999Y

  But the number of bins is integer so a minimum of 2 required !

- Define efficiency, $\varepsilon = 2^H/\text{Nbins}$ →how many does H say compared to what is actually needed.

  Only for a uniform distribution is $\varepsilon = 1$

Efficiency, $\varepsilon \equiv \frac{2^{H(bits)}}{N_{Bins}} = \frac{e^{H(nats)}}{N_{Bins}} = \frac{2^{h(bits)}}{Range} = \frac{e^{h(Nats)}}{Range}$

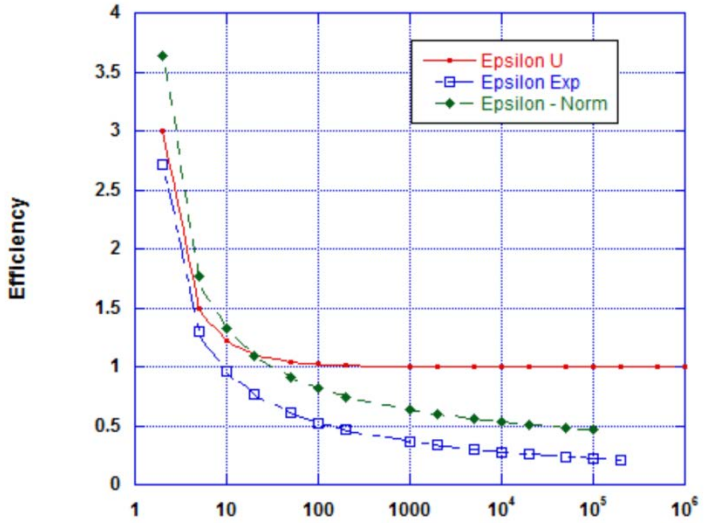
How efficiently are the bins filled or the phase space filled.

Independent dimensions X, Y, Z....

$\varepsilon = \varepsilon_X \times \varepsilon_Y \times \varepsilon_Z \ldots\ldots$

Gets smaller as no of dimensions increase.

Another way of looking at the "curse of dimensionality".



Efficiency for 1D Uniform, Exponential and Gaussian Distributions as function of number of data points.

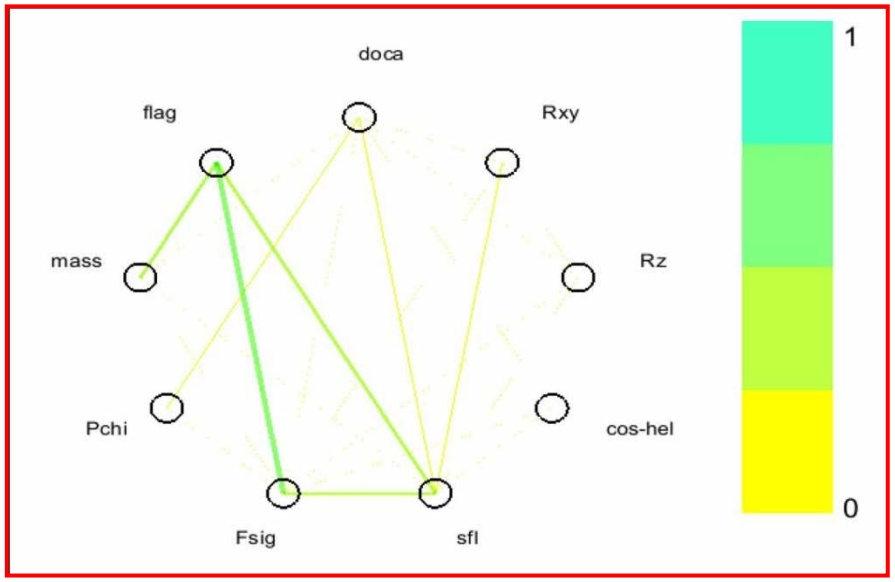
Note: h is a measure of the “spread” of the data, and works in ***any*** number of dimensions.

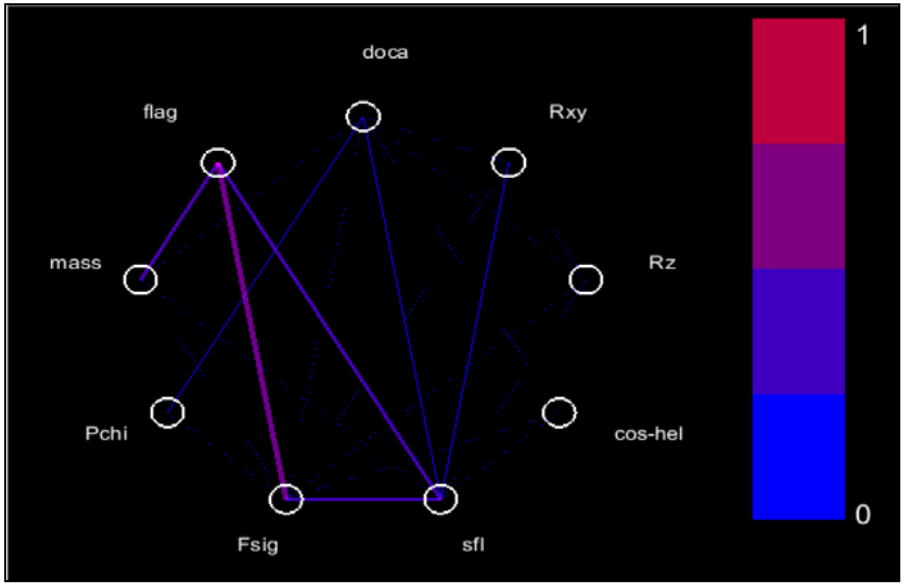
Standard deviation only works in 1D

Interaction diagrams of the ’ppdata’

Now we have made the data discrete, one can calculate the Similarity Index

Shown twice in two colour scales so that one should work with the projector.

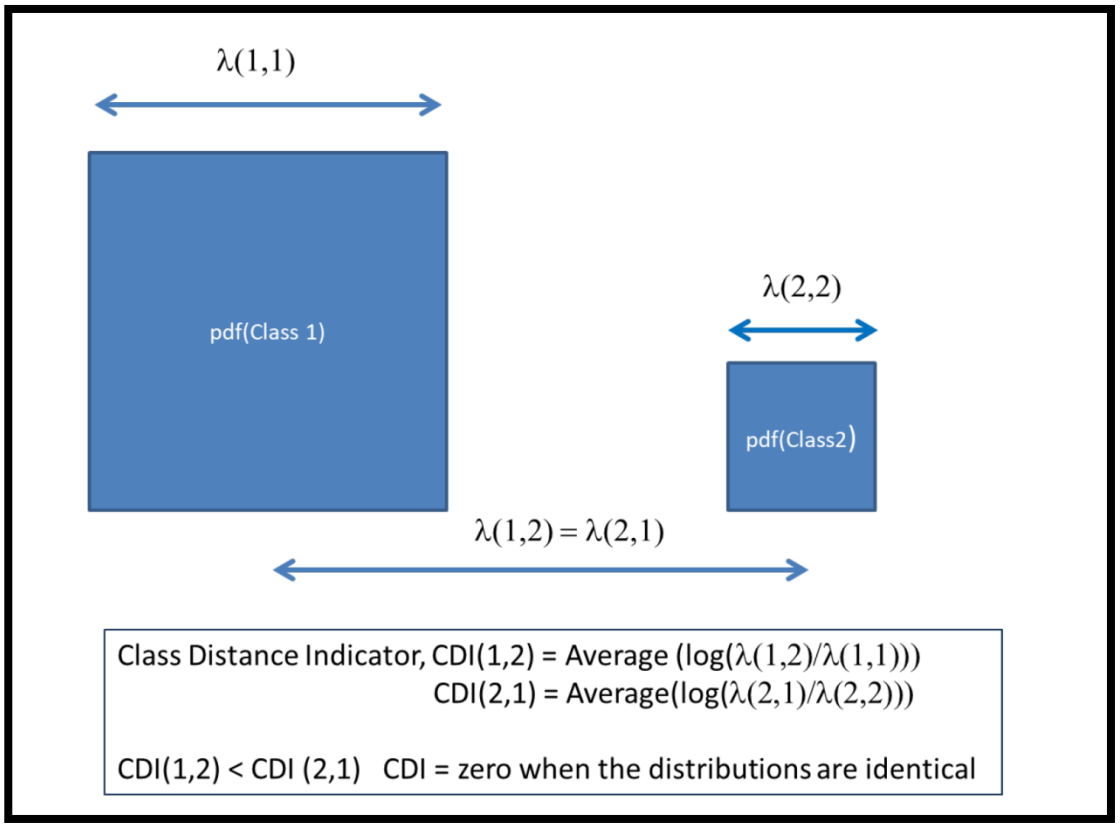




Class Distance Indicator

The Similarity Index tells one how pairs of variables are related. It does not tell one how different one class of events is from another.

Kullback-Leibler Distance or divergence can do this – been around a long time Now there is a simple algorithm to calculate it from data.



$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,dx,$$

where *p* and *q* denote the densities of *P* and *Q*.

it is not really a distance. it is not symmetrc because the two distributions may have different variances.

→Just need to calculate nearest neighbour distances.

Class Distance Indicator

The Class Distance Indicator or Kullback- Leibler Distance (KL) places a limit on how well one can separate two classes of events – rather old result !

Steins Lemma (due to Chernoff in 1956)

Exponential Rates of decay of probabilities using information theoretic distances for large N.

→Prob. False Alarm -> $2^{-KL(S,B)}$

Note this can be different for S and B since KL(S,B) may not be the same as KL(B,S)

Most data mining algorithms tend o optimise the selection of Signal and Background equally.

Better to *S. Sinanović, D.H. Johnson / Signal Processing 87 (2007) 1326–1344*

Refer to

("Parallel Resistor Combination of KL(S,B) and KL(B,S))

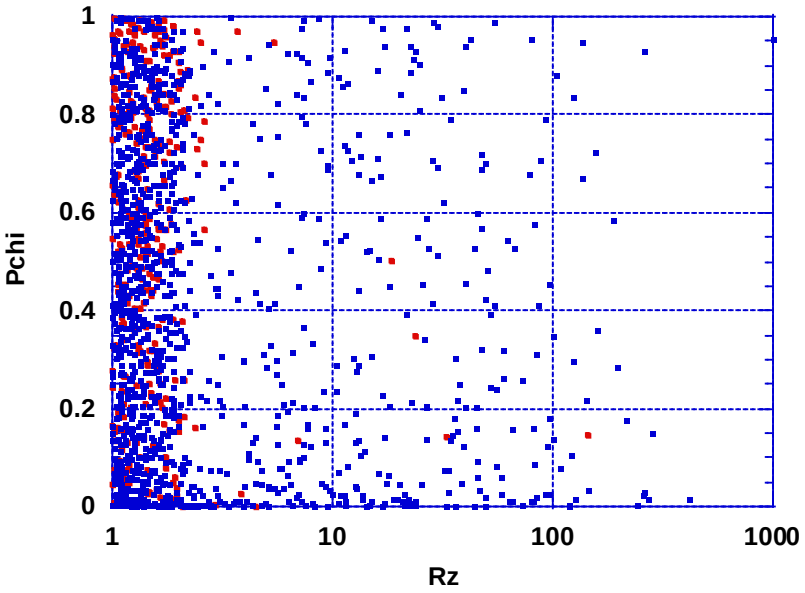
Now lets do this on the PP MC data set shown earlier…..

8 variables but we will not use themass variable – so 7 in total…

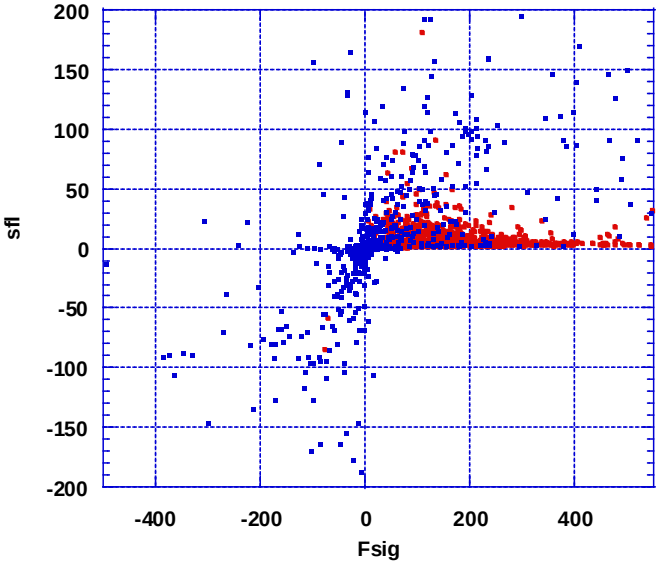
		KL_{SB},	KL_{BS},	KL_{Res}
1	**FSig**	**4.0**	**3.1**	**1.75**
1	**sfl**	**3.56**	**2.75**	**1.55**
1	**Doca**	**0.26**	**0.31**	**0.14**
1	**Rxy**	**0.33**	**0.19**	**0.12**
1	**Rz**	**0.18**	**0.18**	**0.09**
1	**Pchi**	**0.06**	**0.10**	**0.04**
1	**cos-hel**	**0.01**	**0.0**	**0.0**
2	**Rxy FSig**	**4.46**	**5.12**	**2.38**
2	**Doca FSig**	**4.24**	**4.67**	**2.22**
3	**Rxy Fsig sfl**	**4.84**	**6.03**	**2.69**
4	Rxy Fsig sfl cos-hel	5.58	6.79	3.06
5	Rxy Fsig sfl cos-hel Doca	5.96	7.54	3.33
6	**Rxy Fsig sfl cos-hel Doca Pchi**	**6.46**	**7.63**	**3.5**
7	All	4.66	5.96	2.6

Conclusion – Picks out the variables that discriminate signal and background.
Do not expect to have perfect selection.
Error rate around 8 % expected.

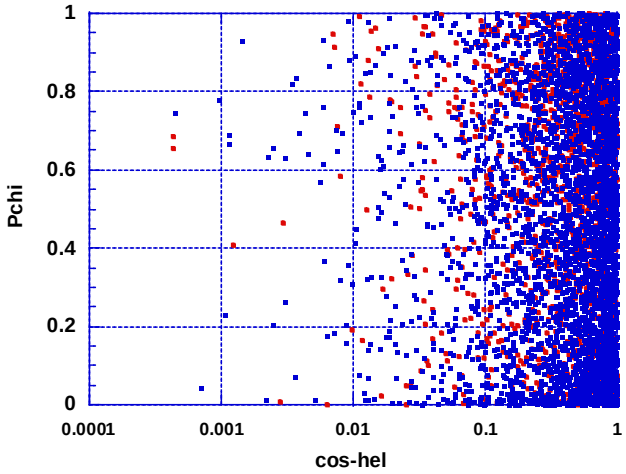
RED = SIGNAL BLUE = BACKGROUND



SI = Zero KL(S,B) = 0.42 KL(B,S) = 0.17



SI = 0.25 KL(S,B) = 4.4 KL(B,S) = 4.7



SI = 0.0 KL(S,B) and KL(B,S) Zero

Combination of the Similarity Index and KL Distance picks out interesting plots !

Plot at bottom left is of no interest !

Use the parallel coord. plot and information theoretic statistics to guide the datamining

Briefly - Use WEKA and just a decision tree (J48) which is optimal for this data

Variables	Success Rate	Confusion Matrix			
		S (Found)	B (Found)		
Fsig	90.3% correct	1165	99	**S**	
	9.7% wrong	386	3350	**B**	**Input**
Fsig/sfl/Rxy	95.3 % correct	1146	118		
	4.7% wrong	119	3619		
All	95.35% correct	1129	135		
	4.64% wrong	98	3638		

SUMMARY AND CONCLUSIONS

- **Information theory allows a consistent description of multivariate data.**
- **Similarity Index can be used to identify independence and interdependence.**
- **Parallel coordinates break the mind set that data lives in a "Cartesian" space.**
- **New histogramming algorithm based on the entropy of the histogram. It does not need to know the underlying pdf of the data to work.**
- **There is a finite amount of information in a data set with finite (N) data points. As experimentalists know – sometimes you just have to collect more data if there is not enough information to get results. Or, find more variables.**
- **Can use data itself to calculate Kullback-Leibler distances which can be used to estimate how well classification algorithms should do and also identify key variables to separate classes of events. Class either from Monte Carlo or user selection of different classes (e.g. healthy/unhealthy individuals).**