Characterizing distributions of data and similarities between di

Feb 23, 2018

Isabelle Guyon and Andrew Yale

We propose means of characterizing and comparing distributions amenable to be use
comparison of datasets with possible covariate shifts of drift or real and simulated datasets.

We limit ourselves to feature representation datasets with no temporal dependencies for the moment.

1) Data format

To be able to easily apply shared tools we need to use a simple generic format. Should we adopt Pandas data frames and R data frames? Are they sufficient for our purpose?

Else: we can use the AutoML format.

2) Characterizing and visualizing single distributions

Before we begin to compare distributions, we may want to represent them in a human-friendly format to get a good impression of them. How do we create a kind of “identity card”.

a. Descriptive statistics and “meta-features” of datasets [TODO]

We need to be able to remove/add meta-features as we see best fit to our purposes.

- AutoML info features
- OpenML meta features (no landmark)
- Nuria Macia meta features
- Kate Smith-Kline meta features
- LDA lisheng-style representations of meta features

We need to also look for DOMAIN EXPERT knowledge giving particular constrains e.g. on distribution support, faction of values etc.

b. Quick look at the overall dataset

At this stage it is difficult to represent the overall datasets without any simplifying preprocessing. I suggest:

- Missing values, NaN, Inf: Replace +Inf by max of that variable value, -Inf by min of that valuable value, NaN by the median of the variable value. If there are lots of values in this case, create matrices with indicator variables with such values. We can later submit such matrices to the same analyzes as the data matrix.
- Missing data for categorical variables: create another category “missing”. After 1-hot encoding, just put all the values to zero.

- Normalization: use **a one-hot encoding of all categorical variables** and **replac** **continuous variables by their “rank”** then **normalize all variables with (x-min** the target to the features. This will lead to a kind of Spearman correlation.

❖ **Heat map with hierarchical clustering (metric = correlation; average link** **want to separate regular features and meta-features.**

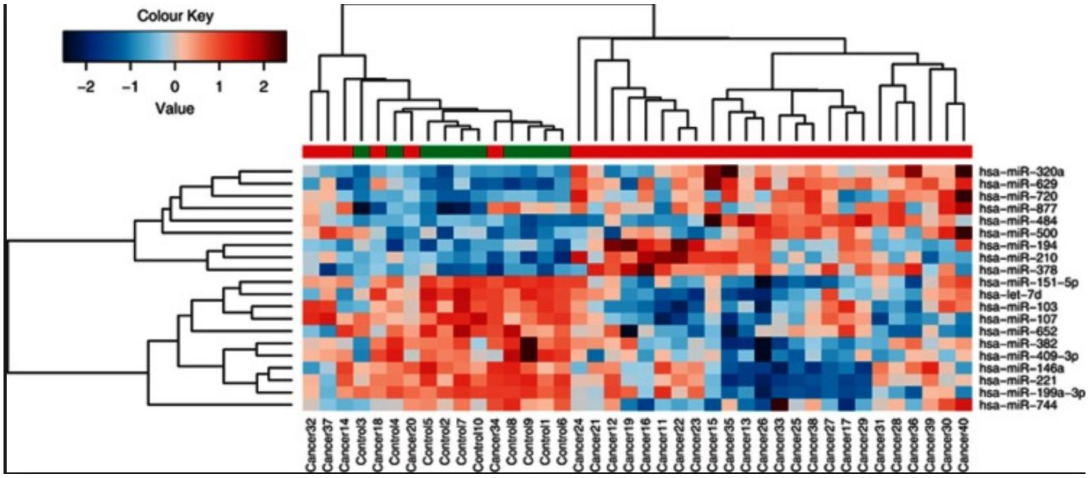


Figure 1: Hierarchical clustering with heatmap of the data matrix. The lines and columns are re-ordered. E.g. http://altanalyze.blogspot.fr/2012/06/hierarchical-clustering-heatmaps-in.html

Compare with the same thing with randomly permuted features.

We can also represent the correlation matrices of the raws and the columns. Show the colorbars.

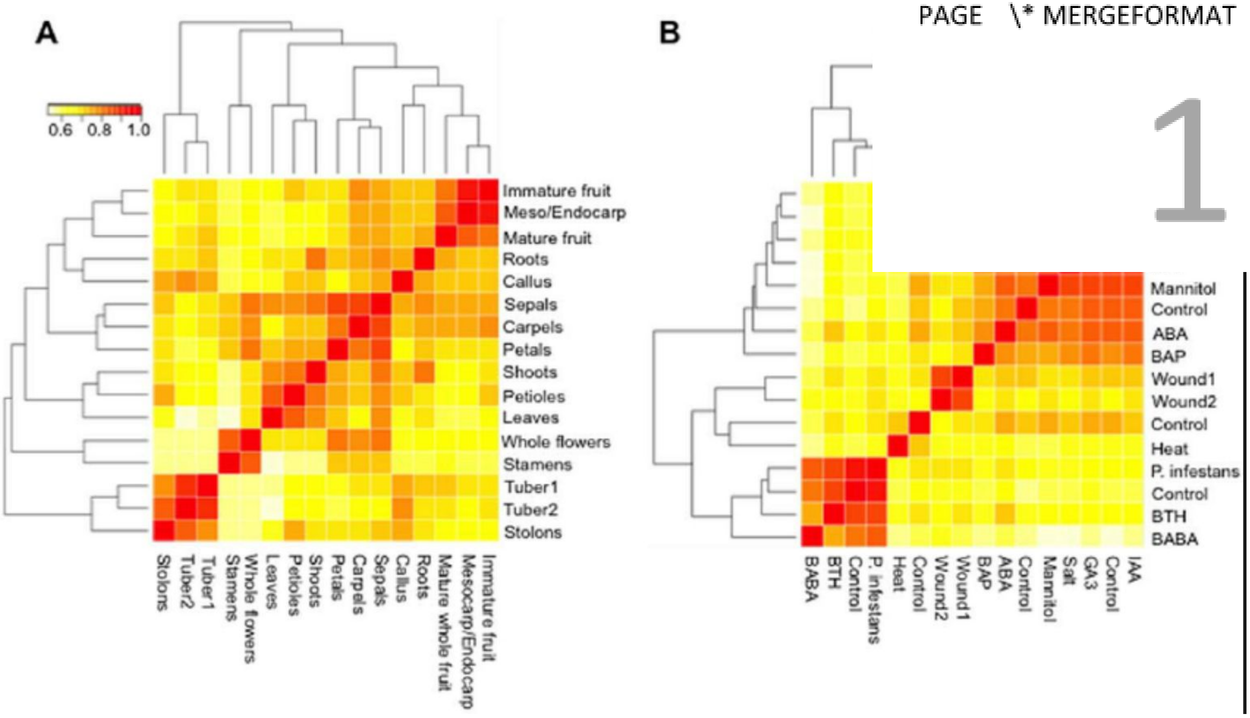


Figure 2: Hierarchical clustering with correlation matrices of lines of columns.

❖ Scatter plot in the 2 dimensions

- **PCA**

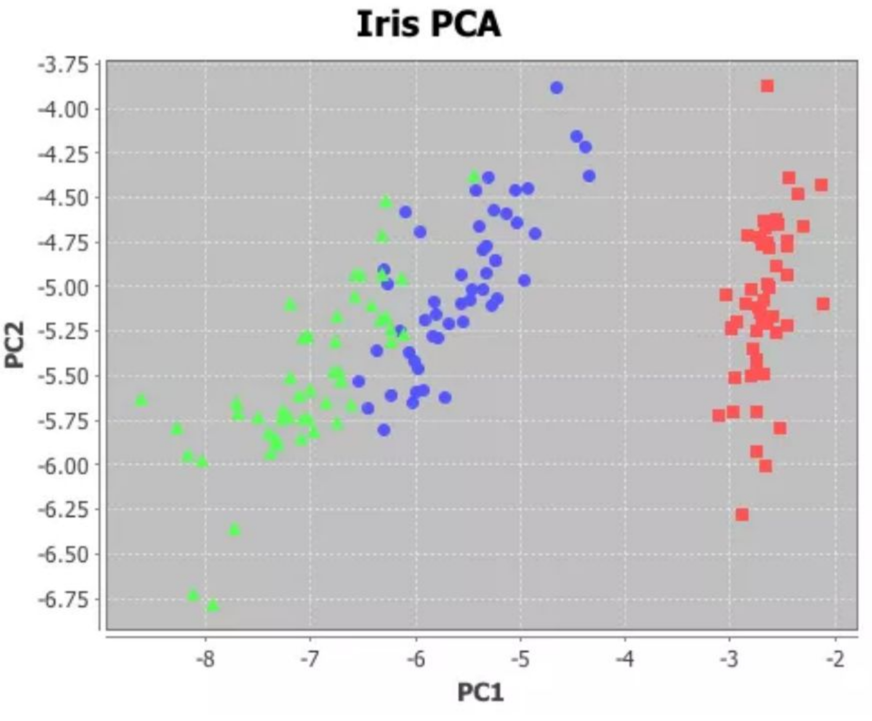


Figure 3: PCA of the Iris data.

Points are the samples and colors the classes (or the target values for regression). We symbol sizes or shapes.

- LDA

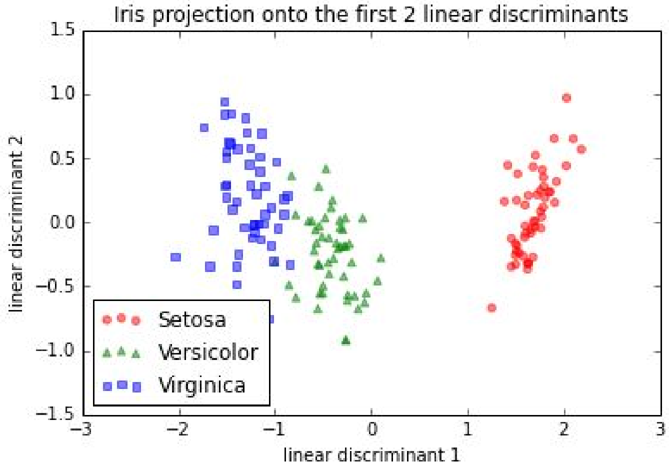


Figure 4: LDA of the Iris data

- t-SNE

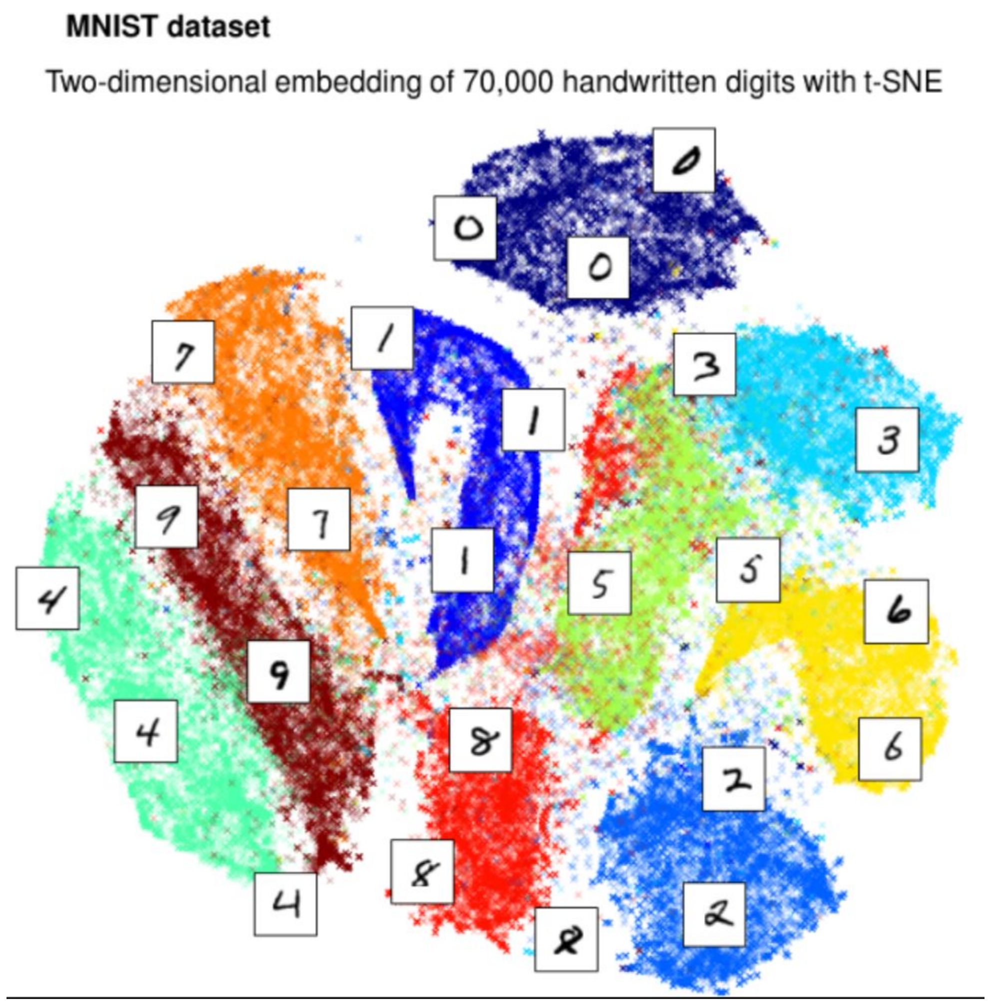


Figure 5: t-SNE of MNIST.

- Pairs of selected features

We can do scatter plots for pairs of selected features. Histograms can be shown on the diagonal.

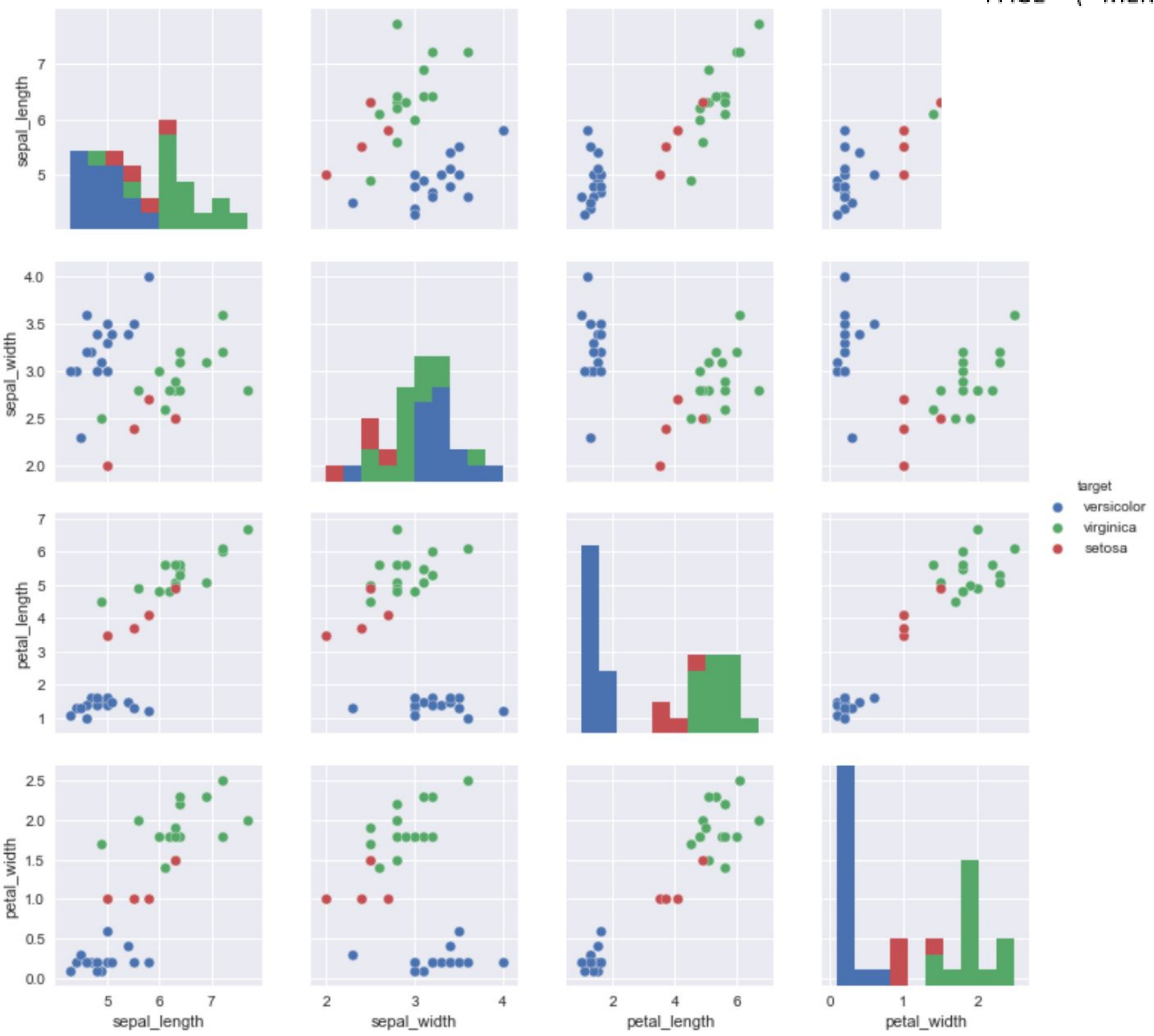


Figure 6: Combination of histograms on the diagonal and scatter plots of pairs of features.

c. Feature selection [TODO]

Because we may have lots of features, it may not be possible to visualize all of them easily so we need to identify those most predictive of the target and eventually remove redundant features.

d. Individual variable distributions

We go back to the un-preprocessed data. We check a few features we selected in the previous step.

For each of the suggestions below, we should be able to give a list of variables to v the graphs nicely layed out.

- Histograms, with or without coding the target variable(s).
    - Continuous variables: A continuous color spectrum can coc target (regression). Two (or more) colors can code for the c categorical target.
    - Categorical variables: we can use a segmented bar graph.

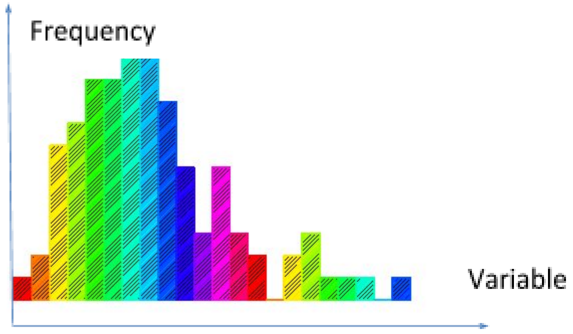


Figure 7: Histogram of one variable showing the continuous target as color shadings.

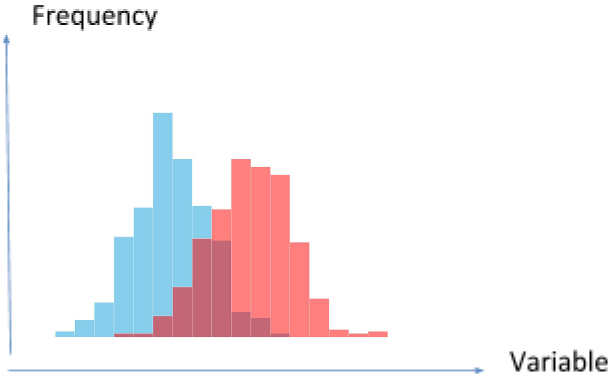


Figure 8: **Histogram of one variable showing the classes as color shading** (binary classification problem)

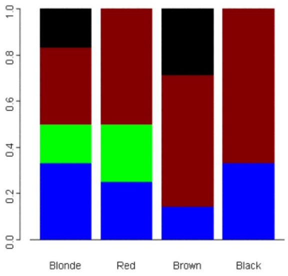


Figure 9: Segmented bar graph of a categorical variable. Categorical target values (claases) are represented with the different colors. This is a 4-class problem.

- Box plots and violin plots (continuous variables)

Here we can represent several features at a time:

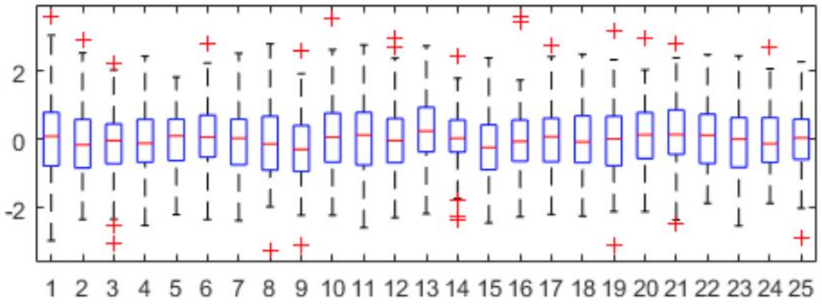


Figure 10: Box plot simultaneously showing the distribution of 25 variables.

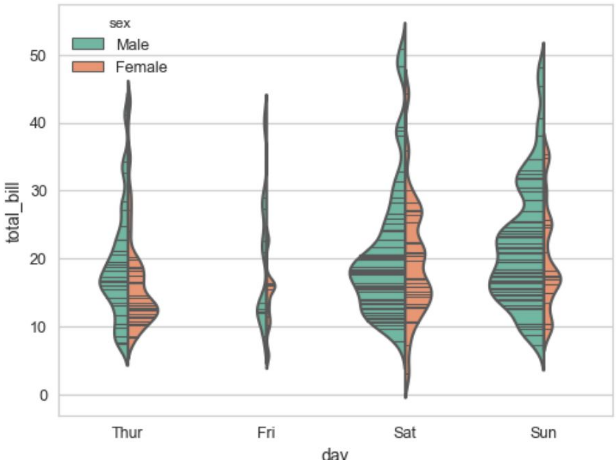


Figure 11: Violin plot simultaneously showing the distribution of 4 variables. Here the distribution of one class is shown on the left and of the other class on the right.

Violin plots are kind of a synthetic way of representing distributions and we can show the target variable (left green=1st class; right orange = 2nd class). See seaborn.violinplot

When we have very few samples, it is useful to show the all:

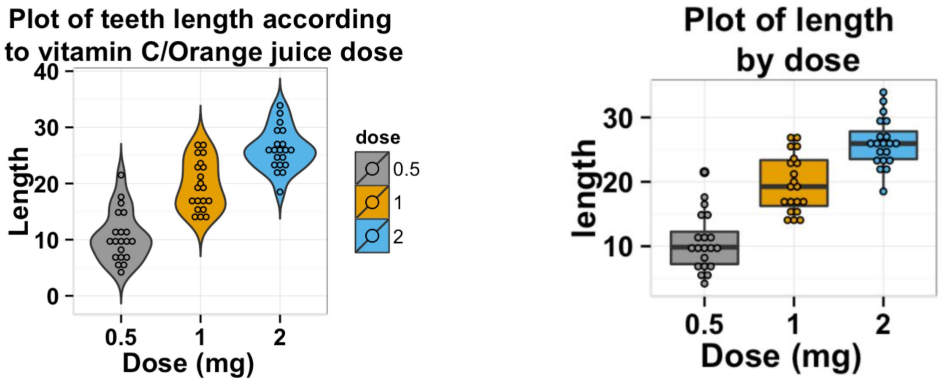


Figure 12: Violin plots overlayed with 1d scatter plots.

This resource (in R) seems pretty good: http://www.sthda.com/french/wiki/ggplot2-violin-plot-guide-de-demarrage-rapide-logiciel-r-et-visualisation-de-donnees . See also R pirate plots https://www.r-bloggers.com/the-pirate-plot-an-r-pirates-favorite-plot/

We can also use this with continuous targets (regression) coding the target values with colors or gray shadings:

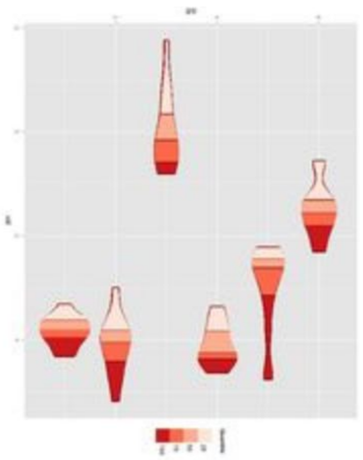


Figure 13: Violin plots showing as color shadings the (continuous) target value.

- Deviation to normality {?]

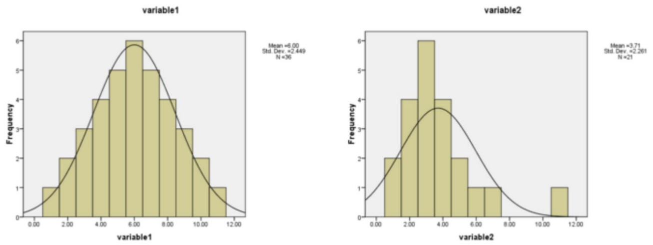


Figure 14: Histograms overlayed with fitted Normal distribution.

We can test the normality of the distribution of continuous variables (with or without segmenting with the class label) and apply some statistical tests like Kolmogorov-Smirnov http://www.psychwiki.com/wiki/How_do_I_determine_whether_my_data_are_normal%3F. We can apply some standard non-linear transforms to see whether we can restore normality. This would be useful to simplify the data. We can later invert the transform in the synthetic data.

Eventually learn a mapping (continuous monotonic increasing) to normal distribution.

We also need to compute **individual variable summary statistics**, e.g.:

- Fraction of 0 and 1 for binary variables (column or row-wise)
- Fraction of each class for categorical variables (column or row-wise)
- Mean and variance for continuous variables (column or row-wise)

- Pairwise differences in distributions/distributional clustering. [TODO]

We can apply distributional clustering in pairs of variables.

e. Multivariate dependencies [TODO]

f. Prediction of the target [TODO]

3) Characterizing and visualizing distribution similarities and differences

Now that we have "well" characterized distributions, we can compare the different representations to characterize the similarities. We can also compute overall statistics.

a. Global statistics

We can use various metrics to compare the two distributions "globally".

- MMD http://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf http://alex.smola.org/teaching/iconip2006/iconip_3.pdf loss https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox/blob/7120fcf3d2073f224680b50a8432817d4cd389e7/cdt/utils/Loss.py

- Other (including Wasserstein) https://arxiv.org/pdf/1509.02237.pdf

Those are also used as objective functions in GANs.

We can also compare the differences in all our descriptive statistics. We need to look t "exaplinability": these 2 distributions are different BECAUSE xxx. Some differences ma others.

b. Overall dataset representations

Distance between correlation matrices. Display the absolute differences of correlation matrices.

This will seek to compare pairwise joint distributions. We can also compare pairwise MI.

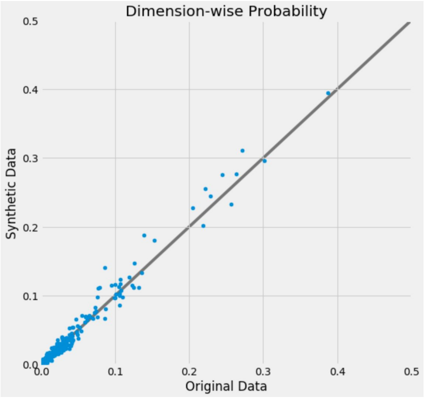
Other perspective: we may want to retain the "type" of variable distribution without matching the variables themselves. In that case, we might want to look at features of the dentrogram? For later.

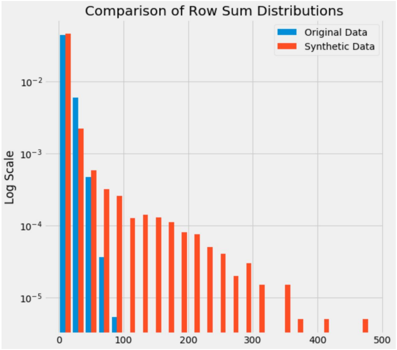
c. Individual features/variables

Overlay histograms real and simulated.

Simple scatter plots real versus simulated for the summary statistics of all variables.

Examples in Andrew's notebook.





QQ plots (for continuous variables)

d. Multivariate dependencies

e. Prediction of the target