

Introduction to Data Science

**DATA UNDERSTANDING: BIVARIATE
STRUCTURE**

BRIAN D'ALESSANDRO

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.

TYPES OF STRUCTURE

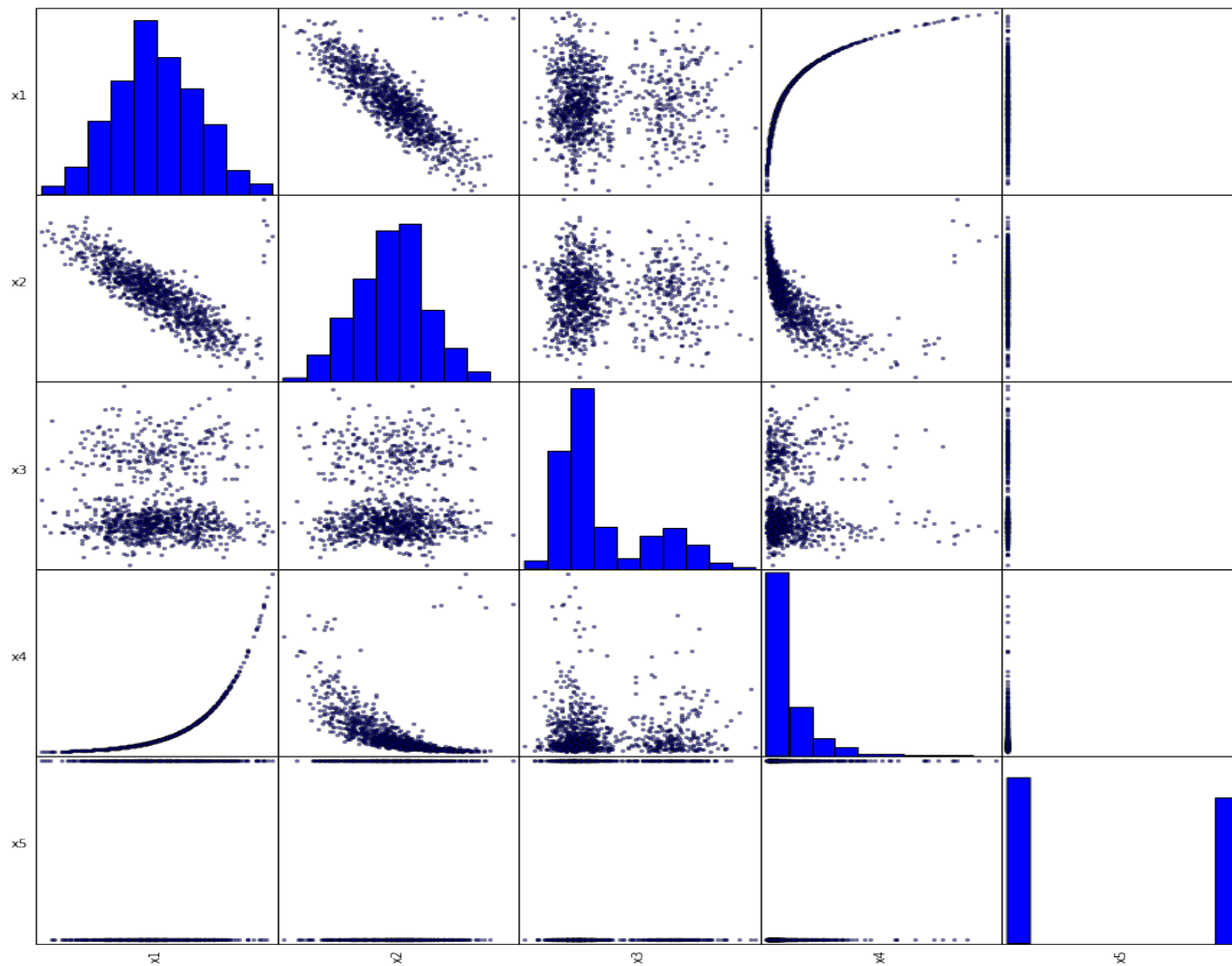
Univariate – does a single variable follow a predictable pattern?

Pairwise (i.e., between 2 variables)

Global (i.e., across a matrix, or within a set of variables)

Supervised (technically includes the above, but with special emphasis on a single target variable)





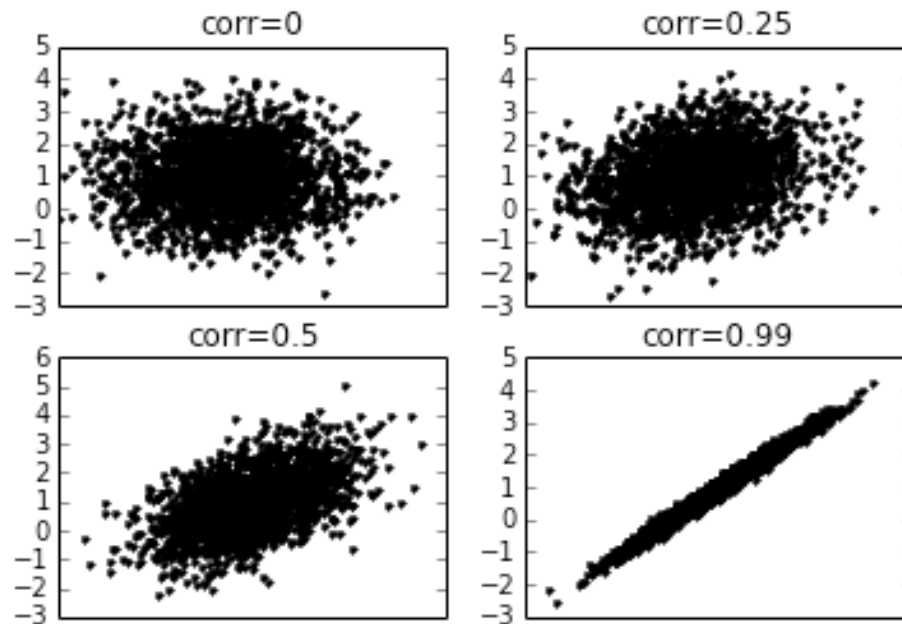
The intuitive explanation
of “structure” in data:

*When you look at
univariate or bivariate
plots, can you describe
its shape?*

COVARIANCE AND CORRELATION

These are probably the most used and thought of metrics when considering pairwise structure. These are statistical quantities and have a fairly intuitive geometric interpretation.

Scatter of Bi-Variate Normally Distributed Variables with various Correlations



COVARIANCE

Covariance

$$\begin{aligned}\sigma(x, y) &= E[(x - E[x])(y - E[y])] \\ &= E[xy - xE[y] - E[x]y + E[x]E[y]] \\ &= E[xy] - E[x]E[y] - E[x]E[y] + E[x]E[y] \\ &= E[xy] - E[x]E[y].\end{aligned}$$

Sample Covariance (between X_j and X_k)

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Note: I have decided to pull equations often from wikipedia, as I feel wp represents a crowd-sourced vote on notation standardization: <http://en.wikipedia.org/wiki/Covariance>

CORRELATION

Aka: Pearson-Product Moment Correlation Coefficient.

This is just the covariance normalized by the variance of each variable.

This scales correlation to the interval $[-1,1]$, which makes it a very intuitive tool for analysis and reporting.

Correlation

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Sample Correlation

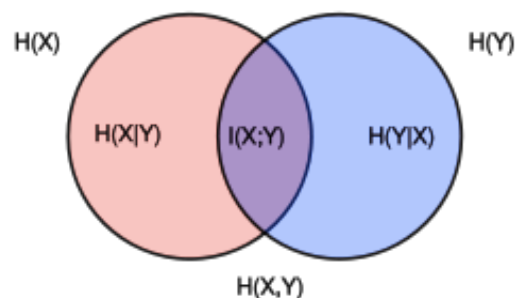
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

Almost any programming language has standard functions for these formulas, but it doesn't hurt to understand them!

Equation Source: <http://en.wikipedia.org/wiki/Covariance>

MUTUAL INFORMATION

This comes from Information Theory, is used often for feature importance ranking and is related to important aspects of Decision Tree algorithms.



Mutual Information

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right),$$

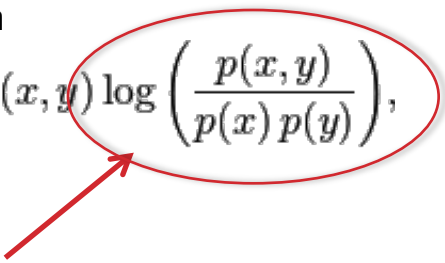
Image and formula source: http://en.wikipedia.org/wiki/Mutual_information

NYU – Intro to Data Science
Copyright: Brian d'Alessandro, all rights reserved

MUTUAL INFORMATION

Let's break this down

Mutual Information

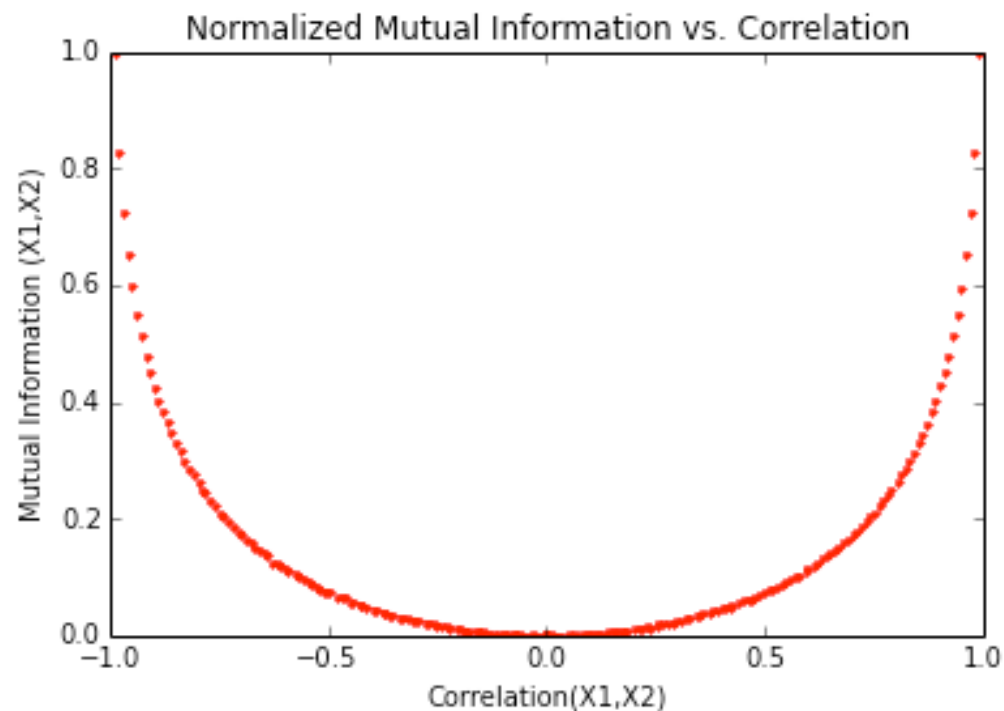
$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right),$$


This is a quantity with the following properties:

- If X,Y are independent, F=0
- If X,Y are completely dependent, F is at a maximum
- F is symmetric (F(x,y)=F(y,x))

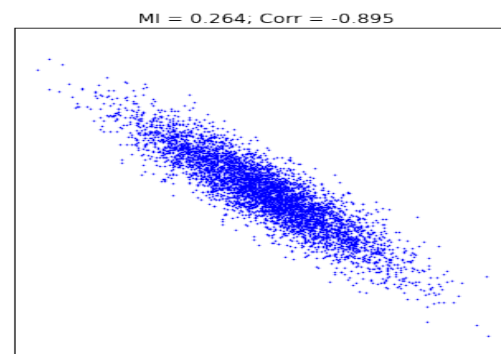
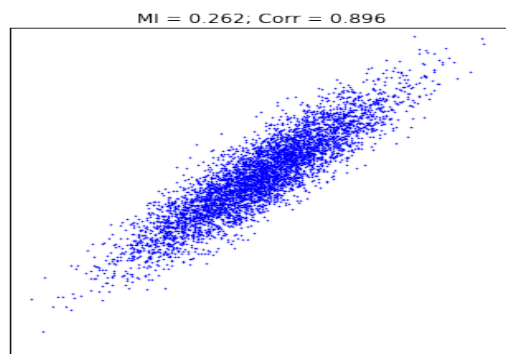
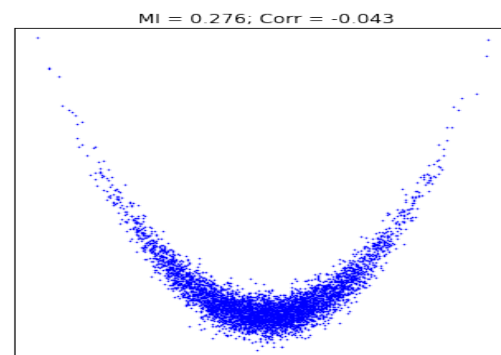
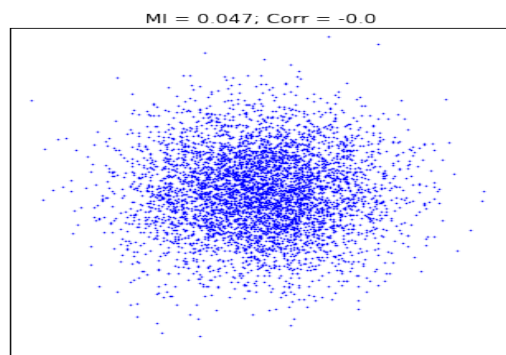
MI VS CORRELATION

Scikit-learn has functions for MI and normalized Mutual Information. We can see that MI and correlation are monotonically related concepts, though MI is strictly positive so does not indicate negative dependencies.



MI VS CORRELATION

We can see the difference by looking at both quantities for different types of variable dependencies.



MI VS CORRELATION

Mutual Information

- Can capture non-linear dependencies better
- Works naturally with categorical data

Correlation Coefficient

- Expresses negative dependencies
- Well understood and intuitive (easy to communicate)

PUTTING THESE TO USE

- **General understanding of dependencies in data**
- **Validating assumptions for statistical modeling**
- **Feature ranking and selection**
- **Decision Tree Algorithms**