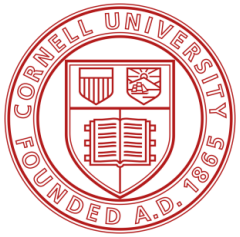


Codependence

Prof. Marcos López de Prado
Advances in Financial Machine Learning
ORIE 5256

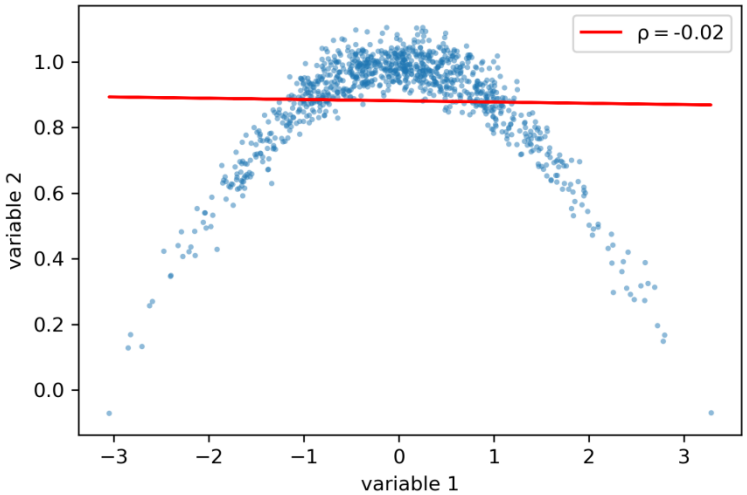
Key Points

- Two random variables are codependent when knowing the value of one helps us determine the value of the other
  - This should not me confounded with the notion of causality
  - Y may tell us something about Z because both Y and Z are caused by X
- Correlation is perhaps the best known measure of codependence in Econometric studies
  - There is no theoretical reason to justify this choice
- **Despite its popularity among economists, correlation has many known limitations in the contexts of financial studies**
  - **Correlations are often unstable due to wrongly assuming linearity (misspecification error)**
- In this seminar we will explore more modern measures of codependence, based on Information Theory, which overcome some of the limitations of correlations

Correlation

Limitations of Correlation

- Consider two random vectors $\{X, Y\}$, and a correlation estimate $\rho[X, Y]$, with the only requirement that $\sigma[X, Y] = \rho[X, Y]\sigma[X]\sigma[Y]$, where $\sigma[X, Y]$ is the covariance between the two vectors, and $\sigma[.]$ is the standard deviation
  - Pearson's correlation is one of several correlation estimates to satisfy this requirement
- Correlations present three important caveats:
  1. Correlation quantifies the *linear* codependency between two random variables. It neglects non-linear relationships
  2. Correlation is highly influenced by outliers
  3. The application of correlations beyond the multivariate Normal case is questionable. We may compute the correlation between any two real variables, however that correlation is typically meaningless unless the two variables follow a bivariate Normal distribution



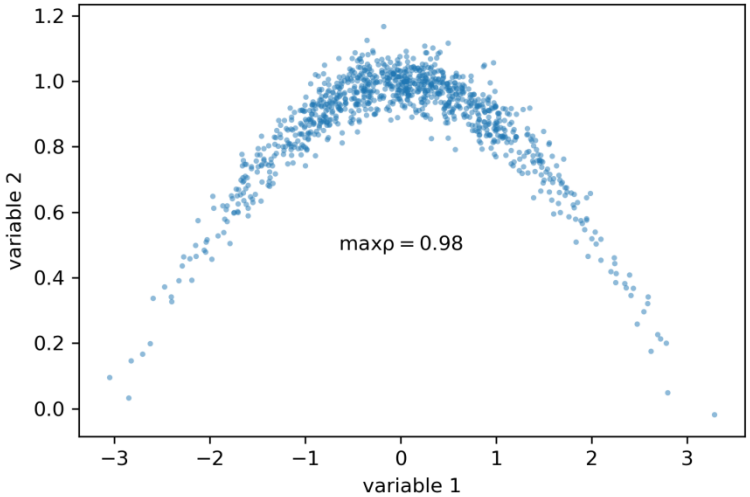
Correlation is a flawed measure of financial codependence. Many financial relationships are non-linear, and correlation fails to recognize them

Maximal Correlation (1/2)

- In 1959, mathematicians Hirschfeld, Gebelein and Rényi proposed a non-linear generalization of Pearson's correlation:

$$\rho_{max}[X,Y] = \sup_{\substack{f:\mathcal{X}\to\mathbb{R}, g:\mathcal{Y}\to\mathbb{R} \\ \mathrm{E}[f[X]]=\mathrm{E}[g[Y]]=0 \\ \mathrm{E}[f^2[X]]=\mathrm{E}[g^2[Y]]=1}} \mathrm{E}[f[X]g[Y]]$$

- This definition has the properties:
  - $0 \le \rho_{max}[X,Y] \le 1$
  - $\rho_{max}[X,Y] = 0$ if and only if X and Y are independent
  - $\rho_{max}[X,Y] = 1$ if there exist functions such that $f[X] = g[Y]$
  - $\rho_{max}[X,Y] = |\rho[X,Y]|$ if X and Y are jointly Gaussian
- The ACE algorithm finds $f[.], g[.]$



Maximal correlation extends the notion of Pearson correlation non-linear relationships. One drawback is that its estimation is relatively computational expensive

Maximal Correlation (2/2)

```
import numpy as np
from ace import model # https://pypi.org/project/ace/
#--------------------------------------------------------------------------
def max_correlation(x:np.array,y:np.array)->float:
    # Get max correlation using ace package
    # https://mlfinlab.readthedocs.io/en/latest/implementations/codependence.html
    ace_model=model.Model()
    ace_model.build_model_from_xy([x],y)
    return np.corrcoef(ace_model.ace.x_transforms[0],
                       ace_model.ace.y_transform)[0][1]
```

Python implementation for the estimation of maximal correlation.

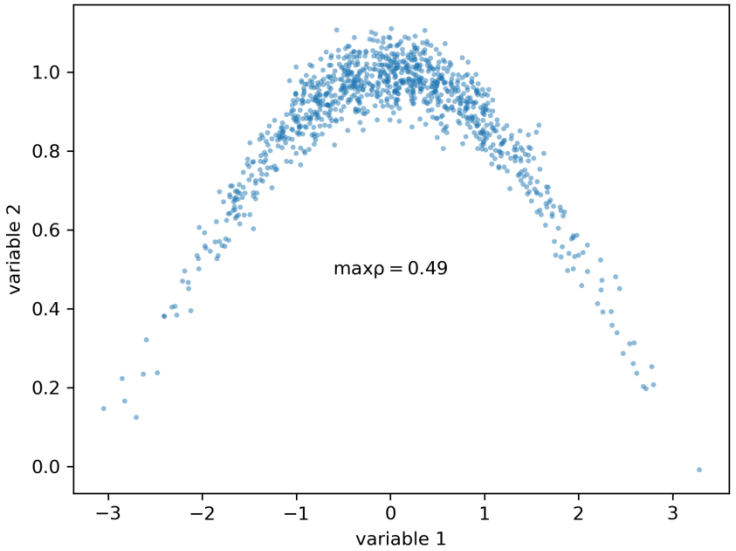
Distance Correlation (1/2)

- In 2005, Gábor Székely introduced the notion of distance correlation, also as a non-linear generalization of Pearson's correlation:

$$\rho_{dist}[X,Y] = \frac{dCov[X,Y]}{\sqrt{dCov[X,X]dCov[Y,Y]}}$$

where $dCov[X,Y]$ can be interpreted as the average Hadamard product of the doubly-centered Euclidean distance matrices of X, Y

- Then
  - $0 \leq \rho_{dist}[X,Y] \leq 1$
  - $\rho_{dist}[X,Y] = 0$ if and only if X and Y are independent



Distance correlation is another non-linear extension of Pearson's correlation. Like Maximal correlation, its estimation can be computational expensive

Distance Correlation (2/2)

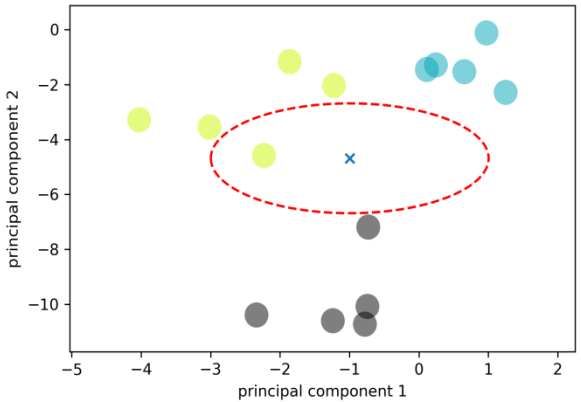
```
import numpy as np,copy
from scipy.spatial.distance import pdist, squareform
#-------------------------------------------------------------------------------
def distcorr(Xval, Yval, pval=True, nruns=500):
    # https://gist.github.com/wladston/c931b1495184fbb99bec
    X,Y=np.atleast_1d(Xval),np.atleast_1d(Yval)
    if np.prod(X.shape) == len(X):X = X[:, None]
    if np.prod(Y.shape) == len(Y):Y = Y[:, None]
    X,Y=np.atleast_2d(X),np.atleast_2d(Y)
    n=X.shape[0]
    if Y.shape[0] != X.shape[0]:raise ValueError('Number of samples must match')
    a,b=squareform(pdist(X)),squareform(pdist(Y))
    A = a - a.mean(axis=0)[None, :] - a.mean(axis=1)[:, None] + a.mean()
    B = b - b.mean(axis=0)[None, :] - b.mean(axis=1)[:, None] + b.mean()
    dcov2_xy = (A * B).sum() / float(n * n)
    dcov2_xx = (A * A).sum() / float(n * n)
    dcov2_yy = (B * B).sum() / float(n * n)
    dcor = np.sqrt(dcov2_xy) / np.sqrt(np.sqrt(dcov2_xx) * np.sqrt(dcov2_yy))
    if pval:
        greater = 0
        for i in range(nruns):
            Y_r = copy.copy(Yval)
            np.random.shuffle(Y_r)
            if distcorr(Xval, Y_r, pval=False) > dcor:
                greater += 1
        return (dcor, greater / float(nruns))
    else:
        return dcor
```

Python implementation for the estimation of distance correlation.

When pval=True, the function also returns the *p*-value associated with the estimated distance correlation, based on the number of simulations set by nruns.

Comparing Correlations

- Codependence attempts to measure how closely associated two random variables are
- However, **correlation is not a metric**, because it does not necessarily satisfy two conditions:
  - non-negativity: $-1 \leq \rho[X,Y] \leq 1$
  - subadditivity: $\rho[X,Z] \nleq \rho[X,Y] + \rho[Y,Z]$
- Comparing non-metric measurements of codependence can lead to rather incoherent outcomes
  - For instance, the difference between correlations $(0.9,1.0)$ is the same as $(0.1,0.2)$, even though the former involves a greater difference in terms of codependence
- We cannot cluster directly on correlations. We can either:
  a) define a metric based on correlation
  b) apply a metric on correlations (e.g., compute the Euclidean distance on observed correlations)



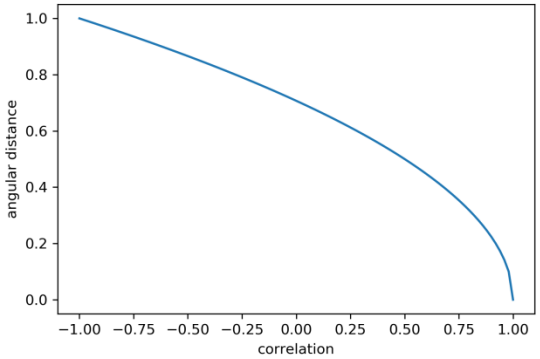
Metric functions are important because they induce an intuitive topology on a given set, which allows us to apply notions of "proximity" or "similarity"

Correlation-based Distance (1/2)

- For Pearson's correlation $\rho[X,Y]$, consider the measure

$$d_\rho[X,Y] = \sqrt{\frac{1}{2}(1-\rho[X,Y])}$$

- This measure is known as the angular distance. It is a metric, because it is a linear multiple of the Euclidean distance between the vectors $\{X,Y\}$ (after standardization)
  - See López de Prado [2016] for a proof
- The metric $d_\rho[X,Y]$ is normalized, $d_\rho[X,Y] \in [0,1]$, because $\rho[X,Y] \in [-1,1]$

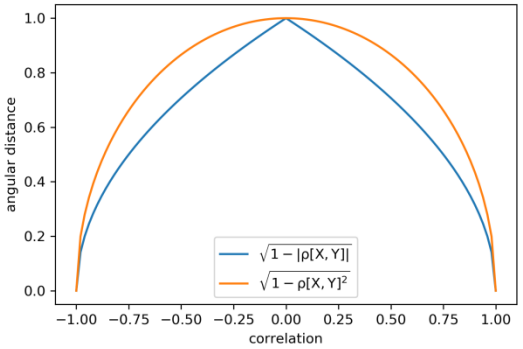


The angular distance satisfies all the conditions of a true metric

Correlation-based Distance (2/2)

- The metric $d_\rho[X,Y]$ deems more distant two random variables with negative correlation than two random variables with positive correlation, regardless of $|\rho[X,Y]|$
- This property makes sense in many applications. For example, we may wish to build a long-only portfolio, where holdings in negative-correlated securities can only offset risk, and therefore should be treated as different for diversification purposes
- In other instances, like in long-short portfolios, we often prefer to consider highly negatively-correlated securities as similar, because the position sign can override the sign of the correlation
- In those cases, we can define alternative normalized correlation-based distance metrics, such as

$$d_{|\rho|}[X,Y] = \sqrt{1-|\rho[X,Y]|}, \text{ or } d_{\rho^2}[X,Y] = \sqrt{1-\rho[X,Y]^2}$$



In some financial applications, it makes more sense to apply a modified definition of angular distance, such that the sign of the correlation is ignored

Information-Theoretic Codependence

Entropy

- Let X be a discrete random variable that takes a value x from the set S_X with probability $p[x]$. The entropy of X is defined as

$$H[X] = -\sum_{x \in S_X} p[x] \log\big[p[x]\big]$$

- A few observations:
  - The value $\frac{1}{p[x]}$ measures how surprising an observation is, because surprising observations are characterized by their low probability
  - Entropy is the expected value of those surprises, where the $\log[.]$ function prevents that $p[x]$ cancels $\frac{1}{p[x]}$ and endows entropy with desirable mathematical properties
  - Accordingly, entropy can be interpreted as the amount of uncertainty associated with X. Entropy is zero when all probability is concentrated in a single element of S_X. Entropy reaches a maximum at $\log[\|S_X\|]$ when X is distributed uniformly, $p[x] = \frac{1}{\|S_X\|}, \forall x \in S_X$

Joint Entropy

- Let Y be a discrete random variable that takes a value y from the set S_Y with probability $p[y]$. Random variables X and Y do not need to be defined on the same probability space. The joint entropy of X and Y is

$$H[X,Y] = -\sum_{x,y \in S_X \times S_Y} p[x,y] \log\big[p[x,y]\big]$$

- In particular, we have that $H[X,Y] = H[Y,X]$, $H[X,X] = H[X]$, $H[X,Y] \geq \max\{H[X], H[Y]\}$, and $H[X,Y] \leq H[X] + H[Y]$
- It is important to recognize that Shannon's entropy is finite only for discrete random variables
  - In the continuous case, one should use the limiting density of discrete points (LDDP), or discretize the random variable, as explained later on (Jaynes [2003])

Conditional Entropy

- The conditional entropy of X given Y is defined as

$$H[X|Y] = H[X,Y] - H[Y] = -\sum_{y \in S_Y} p[y] \sum_{x \in S_X} p[x|Y=y] \log\left[p[x|Y=y]\right]$$

  where $p[x|Y=y]$ is the probability that X takes the value x conditioned on Y having taken the value y

- Following this definition, $H[X|Y]$ is the uncertainty we expect in X if we are told the value of Y
- Accordingly, $H[X|X] = 0$, and $H[X] \geq H[X|Y]$

Kullback-Leibler Divergence

- Let p and q be two discrete probability distributions defined on the same probability space. The Kullback-Leibler (or KL) divergence between p and q is

$$D_{KL}[p\|q] = -\sum_{x\in S_X} p[x]\log\left[\frac{q[x]}{p[x]}\right] = \sum_{x\in S_X} p[x]\log\left[\frac{p[x]}{q[x]}\right]$$

  where $q[x] = 0 \Longrightarrow p[x] = 0$.
- Intuitively, this expression measures how much p diverges from a reference distribution q
- The KL divergence is *not* a metric: Although it is always non-negative ($D_{KL}[p\|q] \geq 0$), it violates the symmetry ($D_{KL}[p\|q] \neq D_{KL}[q\|p]$) and triangle inequality conditions
- Note the difference with the definition of joint entropy, where the two random variables did not necessarily exist in the same probability space
- KL divergence is widely used in variational inference

Cross Entropy

- Let p and q be two discrete probability distributions defined on the same probability space. Cross entropy between p and q is

$$H_C[p\|q] = -\sum_{x \in S_X} p[x]\log\big[q[x]\big] = H[X] + D_{KL}[p\|q]$$

- Cross entropy can be interpreted as the uncertainty associated with X, where we evaluate its information content using a wrong distribution q rather than the true distribution p
- Cross entropy is a popular scoring function in classification problems, and it is particularly meaningful in financial applications (López de Prado [2018], section 9.4)

Mutual Information (1/3)

- Mutual information is defined as the decrease in uncertainty (or informational gain) in X that results from knowing the value of Y,

$$I[X,Y] = H[X] - H[X|Y] = H[X] + H[Y] - H[X,Y] = \sum_{x \in S_X} \sum_{y \in S_Y} p[x,y] \log \left[\frac{p[x,y]}{p[x]p[y]} \right]$$

$$= D_{KL}[p[x,y] \| p[x]p[y]] = \sum_{y \in S_Y} p[y] \sum_{x \in S_X} p[x|y] \log \left[\frac{p[x|y]}{p[x]} \right]$$

$$= \mathrm{E}_Y[D_{KL}[p[x|y] \| p[x]]] = \sum_{x \in S_X} p[x] \sum_{y \in S_Y} p[y|x] \log \left[\frac{p[y|x]}{p[y]} \right]$$

$$= \mathrm{E}_X[D_{KL}[p[y|x] \| p[y]]]$$

- From the above we can see that $I[X,Y] \geq 0$, $I[X,Y] = I[Y,X]$, and that $I[X,X] = H[X]$

Mutual Information (2/3)

- When X and Y are independent, $p[x, y] = p[x]p[y]$, hence $I[X, Y] = 0$
- An upper boundary is given by $I[X, Y] \leq \min\{H[X], H[Y]\}$
- However, mutual information is *not* a metric, because it does not satisfy the triangle inequality: $I[X, Z] \not\leq I[X, Y] + I[Y, Z]$
- An important attribute of mutual information is its grouping property,

$$I[X, Y, Z] = I[X, Y] + I[(X, Y), Z]$$

  where (X, Y) represents the joint distribution of X and Y.
- Since X, Y and Z can themselves represent joint distributions, the above property can be used to decompose mutual information into simpler constituents
  - This makes mutual information a useful similarity measure in the context of agglomerative clustering algorithms and forward feature selection

Mutual Information (3/3)

```
import numpy as np,scipy.stats as ss
from sklearn.metrics import mutual_info_score
#-------------------------------------------------------------------------
def mutualInfo(x,y,bXY,norm=False):
    cXY=np.histogram2d(x,y,bXY)[0]
    iXY=mutual_info_score(None,None,contingency=cXY) # mutual information
    if norm:
        hX=ss.entropy(np.histogram(x,bXY)[0]) # marginal
        hY=ss.entropy(np.histogram(y,bXY)[0]) # marginal
        iXY/=min(hX,hY)
    return iXY
```

Python implementation for the estimation of $I[X,Y]$:

- Pass the number of bins as bXY (needed for discretizing x and y)
- Pass norm=True if for the normalized mutual information (bounded between 0 and 1)

Variation of Information (1/3)

- Variation of information is defined as

$$VI[X,Y] = H[X|Y] + H[Y|X] = H[X] + H[Y] - 2I[X,Y] = 2H[X,Y] - H[X] - H[Y]$$

- This measure can be interpreted as the uncertainty we expect in one variable if we are told the value of another
- It has a lower bound in $VI[X,Y] = 0 \Leftrightarrow X = Y$, and an upper bound in $VI[X,Y] \leq H[X,Y]$
- Variation of information is a metric, because it satisfies the axioms: (a) non-negativity, $VI[X,Y] \geq 0$; (b) symmetry, $VI[X,Y] = VI[X,Y]$; and (c) triangle inequality,$VI[X,Z] \leq VI[X,Y] + VI[Y,Z]$
- Because $H[X,Y]$ is a function of the sizes of S_X and S_Y, $VI[X,Y]$ does not have a firm upper bound
  - This is problematic when we wish to compare variations of information across different population sizes

Variation of Information (2/3)

- The following quantity is a metric bounded between zero and one for all pairs (X, Y),

$$\widetilde{VI}[X,Y] = \frac{VI[X,Y]}{H[X,Y]} = 1 - \frac{I[X,Y]}{H[X,Y]}$$

- Following Kraskov et al. [2008], a sharper alternative bounded metric is

$$\widetilde{\widetilde{VI}}[X,Y] = \frac{\max\{H[X|Y], H[Y|X]\}}{\max\{H[X], H[Y]\}} = 1 - \frac{I[X,Y]}{\max\{H[X], H[Y]\}}$$

where $\widetilde{\widetilde{VI}}[X,Y] \leq \widetilde{VI}[X,Y]$ for all pairs (X, Y).

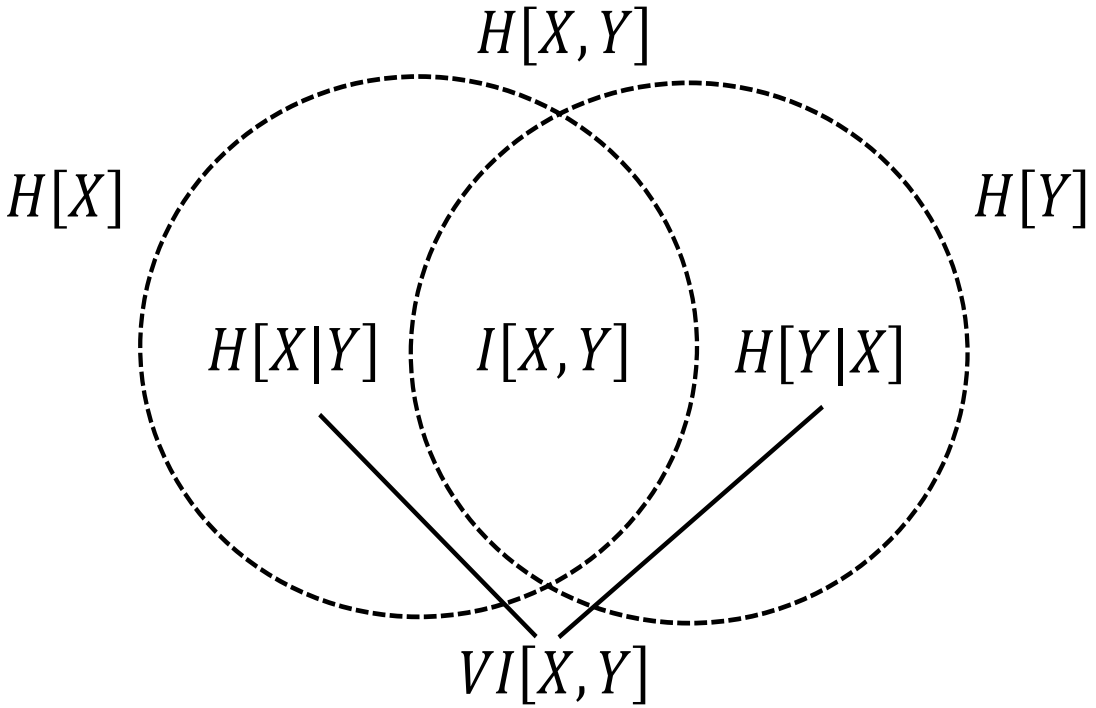
Variation of Information (3/3)

```
def varInfo(x,y,bXY,norm=False):
    cXY=np.histogram2d(x,y,bXY)[0]
    iXY=mutual_info_score(None,None,contingency=cXY) # mutual information
    hX=ss.entropy(np.histogram(x,bXY)[0]) # marginal
    hY=ss.entropy(np.histogram(y,bXY)[0]) # marginal
    vXY=hX+hY-2*iXY # variation of information
    if norm:
        hXY=hX+hY-iXY # joint
        vXY/=hXY # normalized variation of information
    return vXY
```

Python implementation for the estimation of $VI[X,Y]$:

- Pass the number of bins as bXY (needed for discretizing x and y)
- Pass norm=True if for $\widetilde{VI}[X,Y]$ (bounded between 0 and 1)

Information Correspondences



The correspondence between joint entropy, marginal entropies, conditional entropies, mutual information and variation of information

Discretization

Discretization (1/4)

- Throughout this section, we have assumed that random variables were discrete
- For the continuous case, we can quantize (coarse-grain) the values, and apply the same concepts on the binned observations
- Consider a continuous random variable X, with probability distribution functions $f_X[x]$. Shannon defined its (differential) entropy as

$$H[X] = -\int_{-\infty}^{\infty} f_X[x] \log[f_X[x]] dx$$

- The entropy of a Gaussian random variable X is $H[X] = \frac{1}{2}\log[2\pi e\sigma^2]$, thus $H[X] \approx 1.42$ in the standard Normal case

Discretization (2/4)

- One way to estimate $H[X]$ on a finite sample of real values is to divide the range spanning the observed values $\{x\}$ into B_X bins of equal size Δ_X, $\Delta_X = \frac{\max\{x\} - \min\{x\}}{B_X}$, giving us

$$H[X] \approx -\sum_{i=1}^{B_X} f_X[x_i] \log[f_X[x_i]] \Delta_X$$

  where $f_X[x_i]$ represents the frequency of observations falling within the *i*th bin.
- Let $p[x_i]$ be the probability of drawing an observation within the segment Δ_X corresponding to the *i*th bin
- We can approximate $p[x_i]$ as $p[x_i] \approx f_X[x_i]\Delta_X$, which can be estimated as $\hat{p}[x_i] = \frac{N_i}{N}$, where N_i is the number of observations within the *i*th bin, $N = \sum_{i=1}^{B_X} N_i$, and $\sum_{i=1}^{B_X} \hat{p}[x_i] = 1$

Discretization (3/4)

- This leads to a discretized estimator of entropy of the form

$$\widehat{H}[X] = -\sum_{i=1}^{B_X} \frac{N_i}{N} \log\left[\frac{N_i}{N}\right] + \log[\Delta_X]$$

$$\widehat{H}[X,Y] = -\sum_{i=1}^{B_X}\sum_{j=1}^{B_Y} \frac{N_{i,j}}{N} \log\left[\frac{N_{i,j}}{N}\right] + \log[\Delta_X \Delta_Y]$$

- From the estimators $\widehat{H}[X]$ and $\widehat{H}[X,Y]$, we can derive estimators for conditional entropies, mutual information and variation of information
- As we can see from these equations, results may be biased by our choice of B_X and B_Y

Discretization (4/4)

- For the marginal entropy case, Hacine-Gharbi et al. [2012] found that the following binning is optimal,

$$B_X = \text{round}\left[\frac{\zeta}{6} + \frac{2}{3\zeta} + \frac{1}{3}\right]$$

$$\zeta = \sqrt[3]{8 + 324N + 12\sqrt{36N + 729N^2}}$$

- For the joint entropy case, Hacine-Gharbi and Ravier [2018] found that the optimal binning is given by,

$$B_X = B_Y = \text{round}\left[\frac{1}{\sqrt{2}}\sqrt{1 + \sqrt{1 + \frac{24N}{1 - \hat{\rho}^2}}}\right]$$

where $\hat{\rho}$ is the estimated correlation between X and Y.

Python Implementation

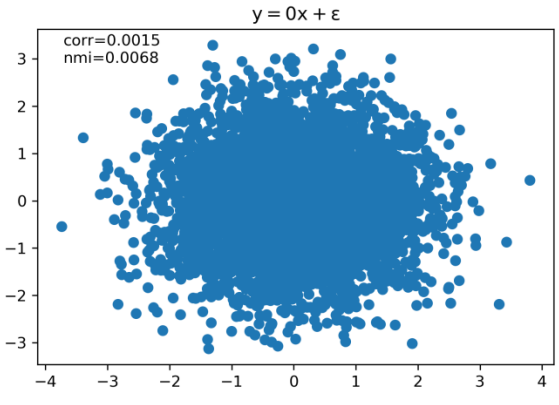
```
def numBins(nObs,corr=None):
   if corr is None: # univariate case
      z=(8+324*nObs+12*(36*nObs+729*nObs**2)**.5)**(1/3.)
      b=round(z/6.+2./(3*z)+1./3)
   else: # bivariate case
      b=round(2**-.5*(1+(1+24*nObs/(1.-corr**2))**.5)**.5)
   return int(b)
#-------------------------------------------------------------------------
bXY=numBins(x.shape[0],corr=np.corrcoef(x,y)[0,1])
```

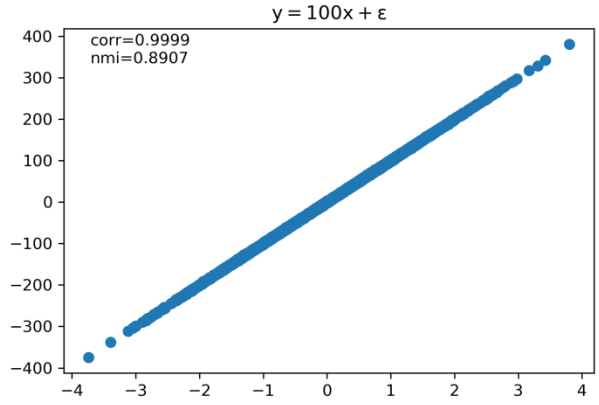
Use this function to estimate the bXY argument required by mutualInfo() and varInfo().

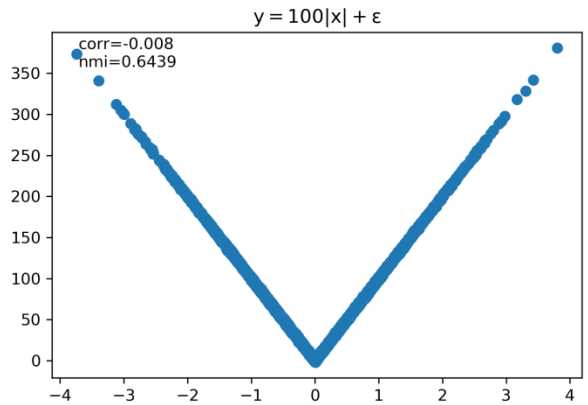
Numerical Examples

Does It Work?

- The normalized mutual information is the information theoretic analogue to linear algebra's correlation coefficient
- Consider two arrays x and e, of random numbers from a standard Gaussian distribution. We evaluate the normalized mutual information (NMI) and correlation between x and various y

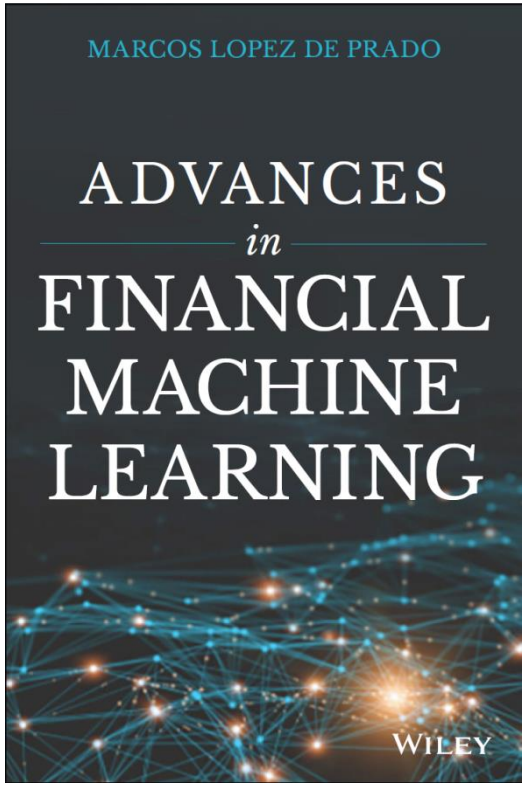






Correlations fail to recognize the strong relationship that exists between x and y when that relationship is non-linear. In contrast, mutual information recognizes that we can extract a substantial amount of information from x that is useful to predict y, and *vice versa*.

For Additional Details



The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org