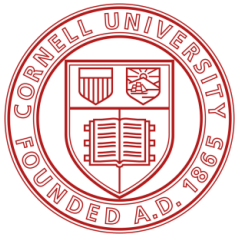


Clustered Feature Importance

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

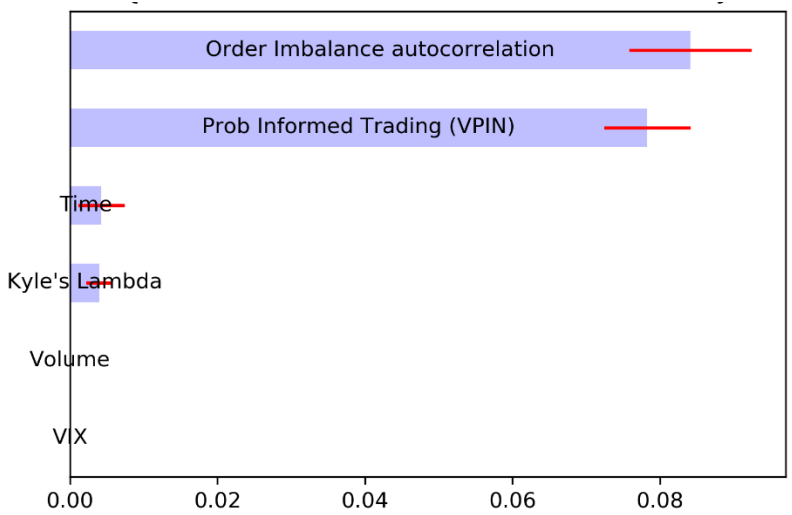
Key Points

- A primary goal of empirical research is to identify the variables involved in a phenomenon
  - The identification of these variables is a prerequisite to the formulation of a theory
- In classical statistics (e.g., Econometrics), the significance of variables is established through *p*-values
- *p*-values suffer from multiple flaws, which have led to the acknowledgement that most discoveries in finance are false
- In this seminar, we explain how Machine Learning (ML) methods overcome the flaws of *p*-values, and facilitate the discovery of new theories
- **This application demonstrates that ML is *not* necessarily a "black box", contrary to popular perception**

What is Feature Analysis?

The Role of Feature Analysis

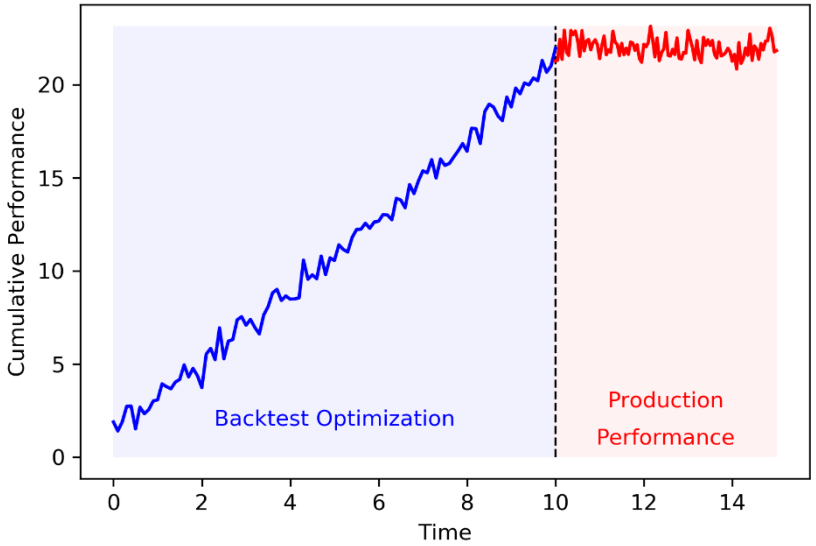
- With the help of ML, feature analysts find which variables are able to predict a particular outcome
  - ML algorithms decouple the specification search from the variable search
- The patterns identified by feature analysts do not imply causation
  - Strategists/Economists must hypothesize the cause-effect mechanism that underlies the pattern uncovered by feature analysts
  - That hypothesis can then be tested, in order to discriminate **causation** from mere **codependence**



In this feature analysis, researchers found that volatility bursts can be predicted with the help of two microstructural variables. Other variables were shown to have little or no predictive power. This finding could lead to the formulation of a hypothesis, which can then be backtested.

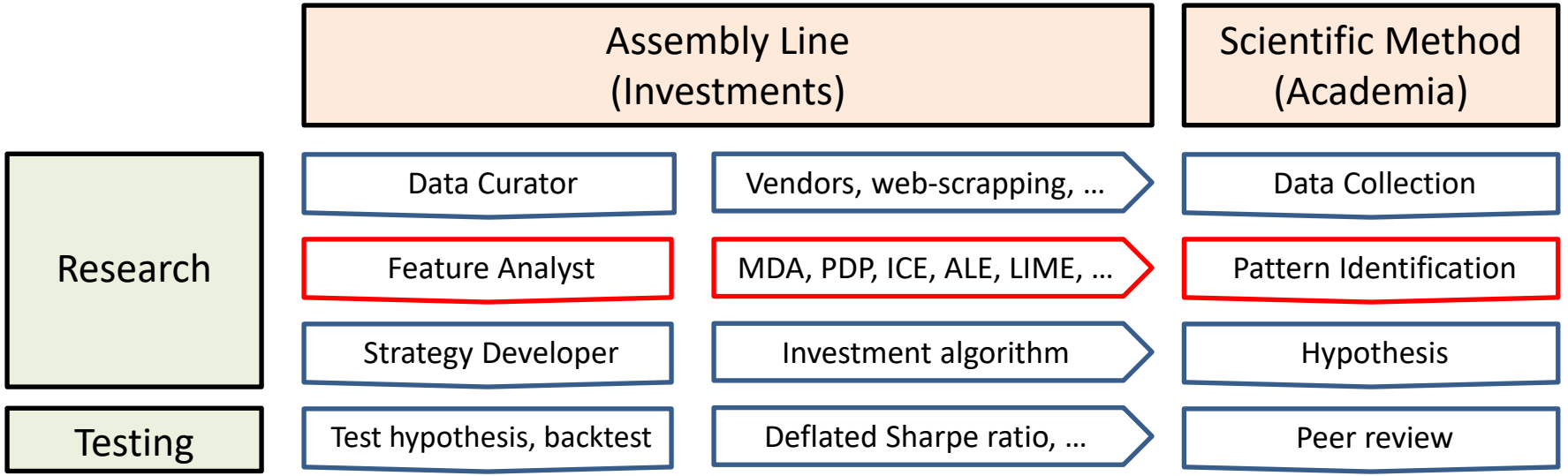
Feature Analysis vs. Backtesting

- **Backtesting is not a research tool**. It is a validation tool
  - By the time we backtest, research should have concluded
  - Building a hypothesis through backtesting leads to selection bias and false discoveries
- The goal of backtesting is not to form a hypothesis
  - On the contrary, one goal of backtesting is to **deconstruct a hypothesis, and to prove the researcher wrong** through counter-examples
- To form a hypothesis, first we must find the variables involved in a phenomenon
  - This is the key role of Feature Analysis



Backtest overfitting results from confounding Research with Validation. The failure of most quantitative funds can be traced back to this basic misunderstanding of the scientific method.

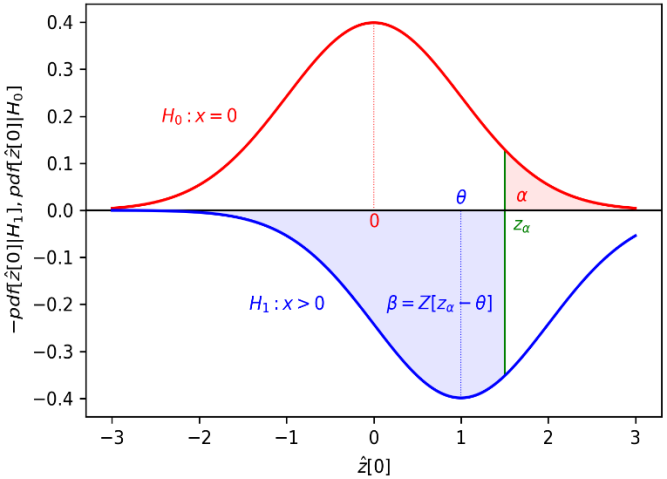
How Feature Analysis Fits in the Scientific Method



Classical Feature Importance: *p*-Values

Hypothesis Testing

- The null hypothesis (H_0) represents the absence of a phenomenon (a negative)
- The alternative hypothesis (H_1) represents the existence of a phenomenon (a positive)
- A test rejects a null hypothesis H_0 with confidence $(1-\alpha)$ when the test's statistic exceeds a value τ that, should H_0 be true, could only occur with probability α
- Two possible errors:
  - Type I: Reject a true H_0 (false positive probability, α)
  - Type II: Reject a true H_1 (false negative probability, β)
- ***p*-value is the α associated with an observed τ**



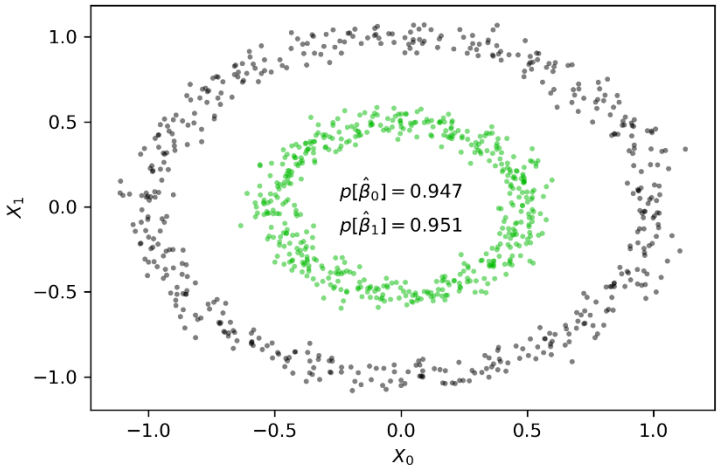
Example of a right-sided test:

$P[\hat{x} \geq \tau | H_0 : x = 0] \leq \alpha$ (red area)

$P[\hat{x} \leq \tau | H_1 : x = \theta] \leq \beta$ (blue area)

Pitfall #1: Specification-Significance Entanglement

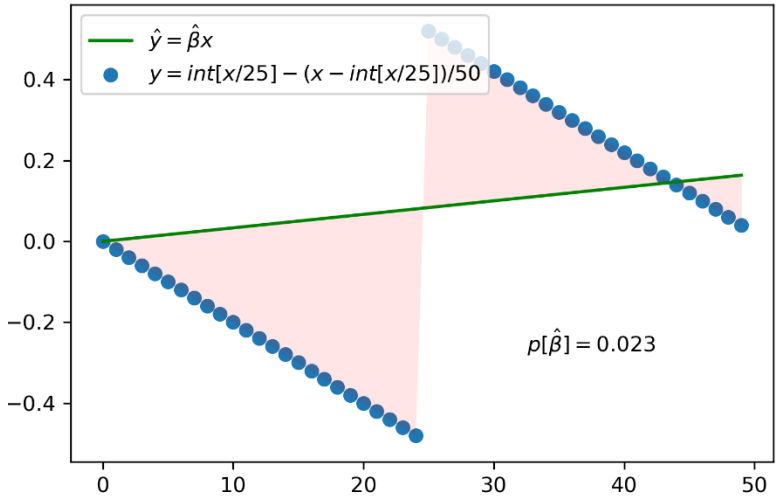
- *p*-values are typically computed on the estimated coefficients of a given model
- When the *p*-value is less than 5%, the variable associated with that coefficient is deemed *statistically significant*, under the assumption that the model is correctly specified
- Thus, *p*-values cannot tell us whether a variable is significant *per se*. They cannot decouple the specification search from the significance search
- In finance, where systems are so complex that researchers can only guess the correct specification, *p*-values are likely to lead to false conclusions



Features X_0 and X_1 can be jointly used to perfectly separate these two classes. However, a logistic regression deems both features uninformative (a false negative).

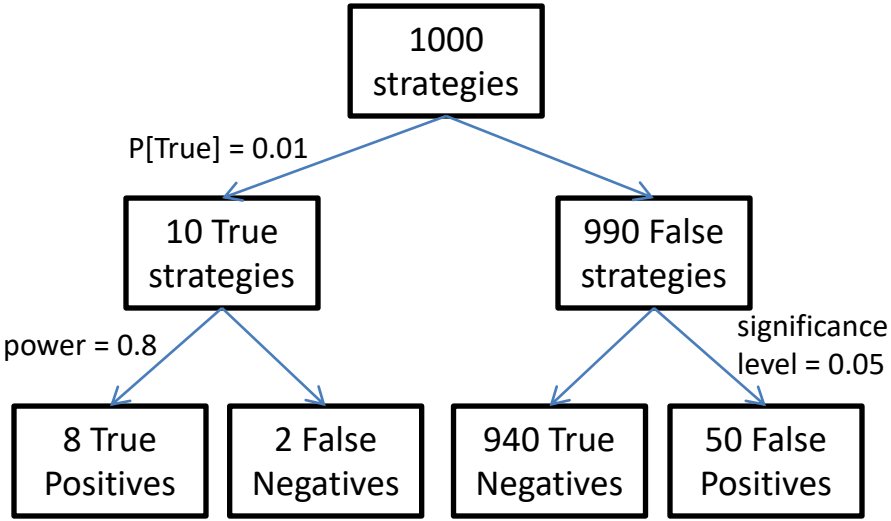
Pitfall #2: In-Sample Estimate

- OLS regressions attempt to adjudicate variance in-sample, not forecast
  - **Informational leakage: Each observation is used for fitting the model and evaluating the model**
- *p*-values are derived from *estimation* errors (in-sample), not *generalization* errors (out-of-sample)
  - A variable that appears to be significant ($p < .05$) may indeed have little predictive power
- In this example, an OLS regression (green line) attempts to minimize the in-sample error (red area)
  - The *p*-value is low, even though the model gets the sign of the relationships wrong



The relationship appears to be linearly ascending (in-sample). Cross-validation would have shown that the model performs poorly on unseen data

Pitfall #3: A Dubious Probability



Suppose that the probability of a backtested strategy being profitable is 1%.

Then, at the standard thresholds of 5% significance and 80% power, researchers are expected to make 58 discoveries out of 1000 trials, where 8 are true positives and 50 are false positives.

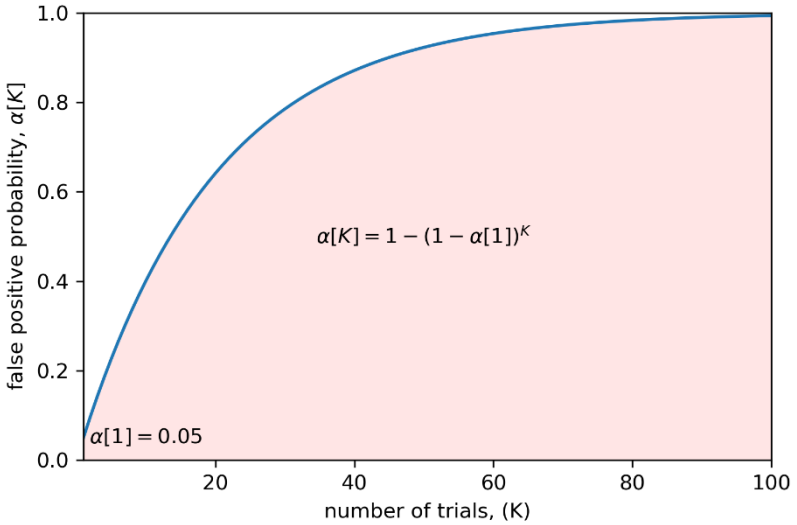
Under these circumstances, **a *p*-value of 5% implies that at least 86% of the discoveries are false!**

The problem is, *p*-values evaluate a somewhat irrelevant probability: $\mathrm{P}[X > x|H_0]$.

What we really care about is a different probability: $\mathbf{P}[\boldsymbol{H_1}|\boldsymbol{X > x}]$.

Pitfall #4: *p*-Hacking

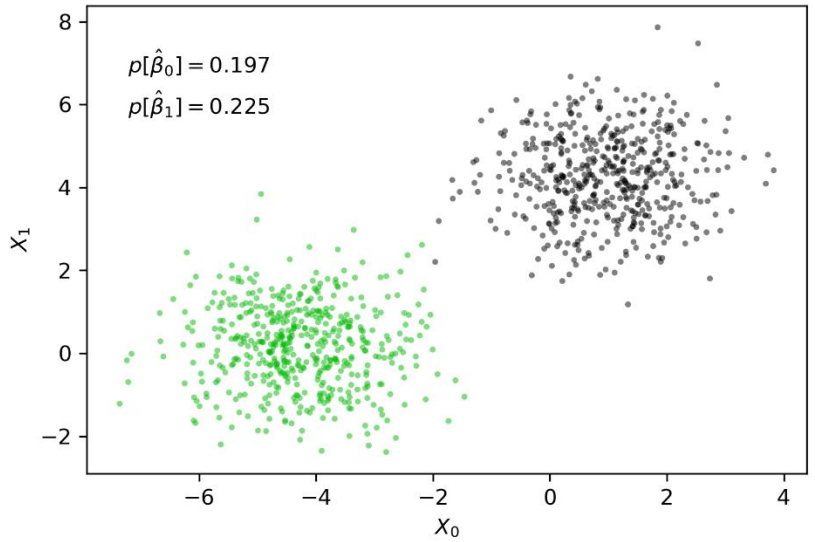
- If hypotheses are tested more than once on a given dataset, the probability of a false positive exceeds α
- Selecting the model with lowest *p*-values out of multiple trials leads to **selection bias**
- In finance, where datasets are limited, it is highly likely that a researcher reuses a datasets multiple times
- Virtually no paper published in financial academic journals controls for the number of trials involved in a discovery (K)
- **It is highly likely that published *p*-values are artificially low (false positives)**



The false positive probability quickly rises after the first trial. Journal articles present findings as if they had been the result of a single trial. Because that is almost never the case, most discoveries in finance are false.

Pitfall #5: Substitution Effects

- In addition to correct model specification, *p*-values require (among other assumptions):
  - Uncorrelated regressors (no multicollinearity)
  - White noise residuals
  - Normally-distributed residuals
  - No outliers
- In particular, *p*-values are not robust to multicollinearity (linear substitution effects)
- Because of this lack of robustness, important variables can be assigned high *p*-values, and be wrongly dismissed as irrelevant



These two blobs can be easily separated with either X_0 or X_1. When we pass both features to a logistic regression, both are deemed insignificant, as a result of substitution effects.

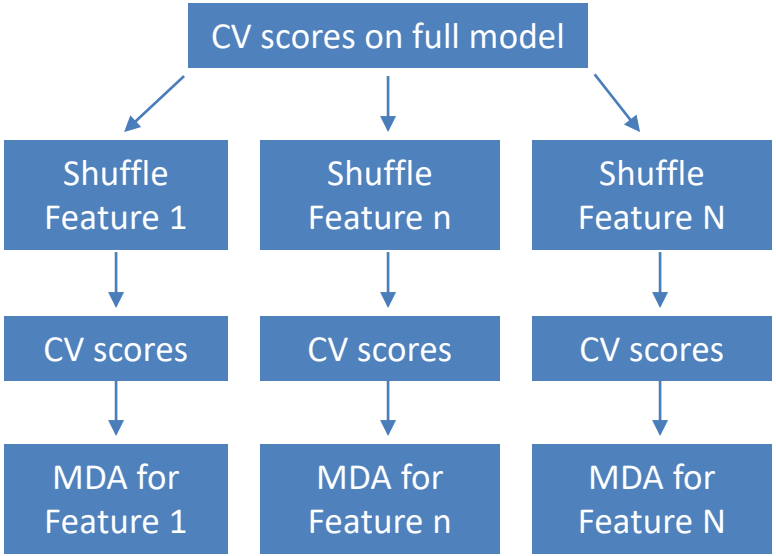
Machine Learning Feature Importance: Mean Decrease Accuracy (MDA)

Mean Decrease Accuracy (MDA)

- MDA is one particular feature analysis:

1. Fit a model using all features, and cross-validate the prediction's scores
2. For each feature
   a) Shuffle the feature, refit the model and cross-validate the new prediction's scores
   b) Compute the distribution of the difference between scores (2.a) and (1)

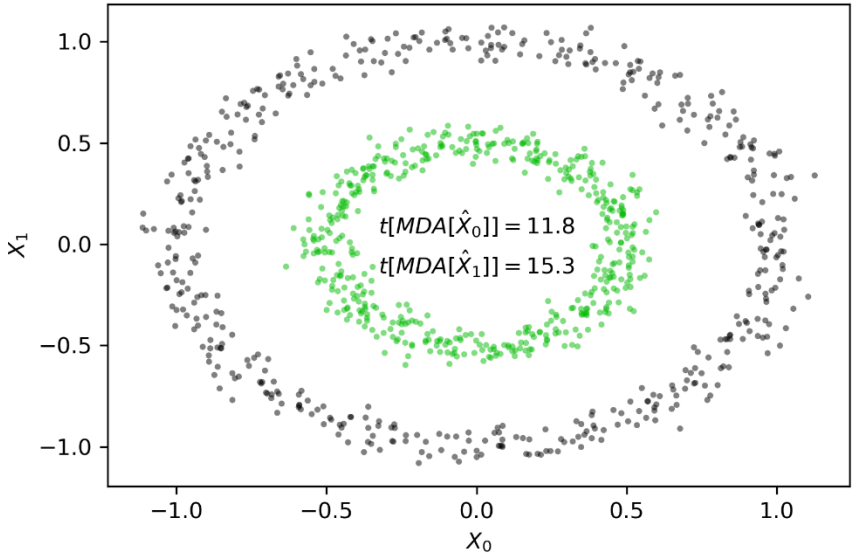
- Shuffling an important feature causes a large loss in model performance
- MDA is also known as *permutation importance*, because it can be derived on any score (not only accuracy)



We can compute K losses in performance in a K-fold cross-validation. That allows us to bootstrap the distribution of the generalization error.

Specification-Significance Disentanglement

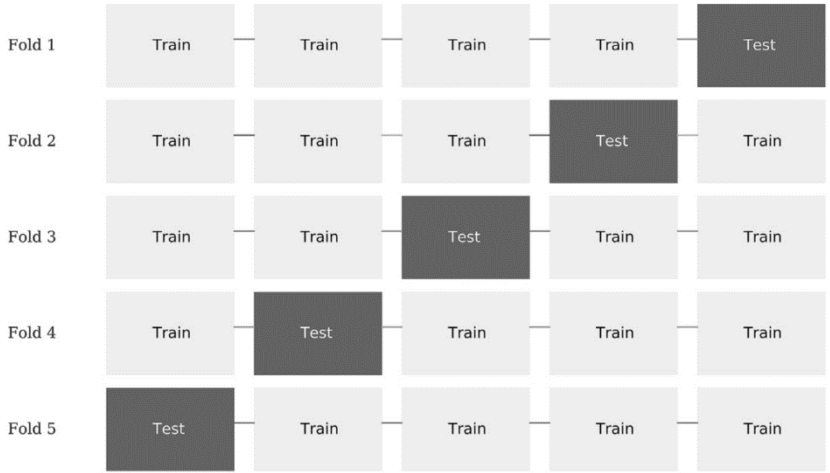
- MDA relies on a particular algorithm to determine the loss in performance (e.g., a random forest), however these algorithms can fit a very wide range of specifications
  - Interestingly, a black-box ML algorithm can be a tool for insight and model interpretability!
- Unlike *p*-values, MDA determines the importance of a feature irrespective of the specification
- Once we know the variables involved in a phenomenon, we can search for the correct specification, and formulate a hypothesis



An MDA analysis of the two concentric classes correctly concludes that both features are extremely informative (above, t-values computed as the ratio bean mean[MDA] and std[MDA]).

Out-of-Sample Measurement

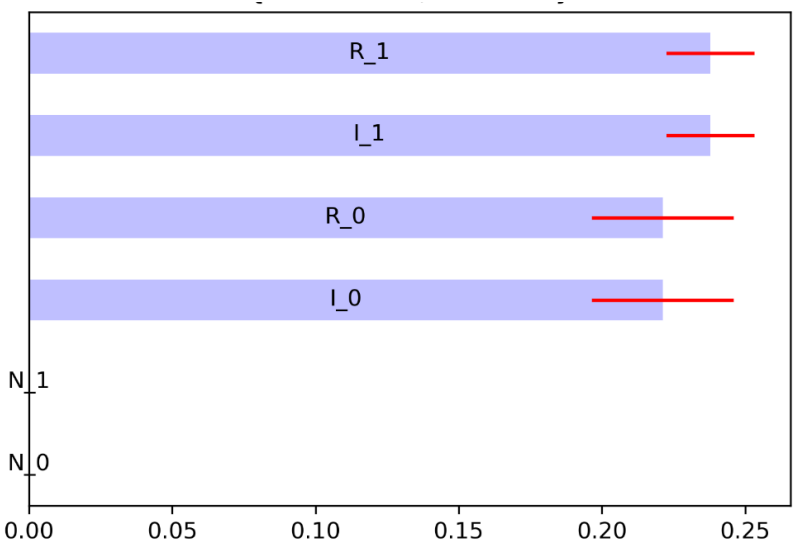
- Unlike *p*-values, MDA does not evaluate the importance of a feature in-sample
  - Observations used to fit the algorithm are not used by MDA to evaluate its performance
- Instead, MDA relies on cross-validation to determine a feature's importance
  - The importance is derived from the increase in the *generalization* error (not the *estimation* error) that results from shuffling a feature
- In the event that X exhibits serial dependence and y is formed on overlapping periods, the train set must be **purged** and **embargoed** (see CPCV)



Data splits in a K-Fold cross-validation experiment. Regardless of the particular cross-validation approach used, MDA always derives out-of-sample importance.

Interpretability

- MDA estimates the drop in performance (out-of-sample) that would result from sampling a variable randomly
- Unlike *p*-values, MDA has a clear interpretation, which
  - does not rely on strong assumptions
  - can be extended to any scoring function
  - is model agnostic
- This is an intuitive, general and flexible approach, that enables the comparison among very different models



MDA's flexibility is a consequence of its experimental nature. Rather than relying on strong (and potentially unrealistic) inferential assumptions, the generalization error is estimated via computational methods.

Avoidance of Selection Bias

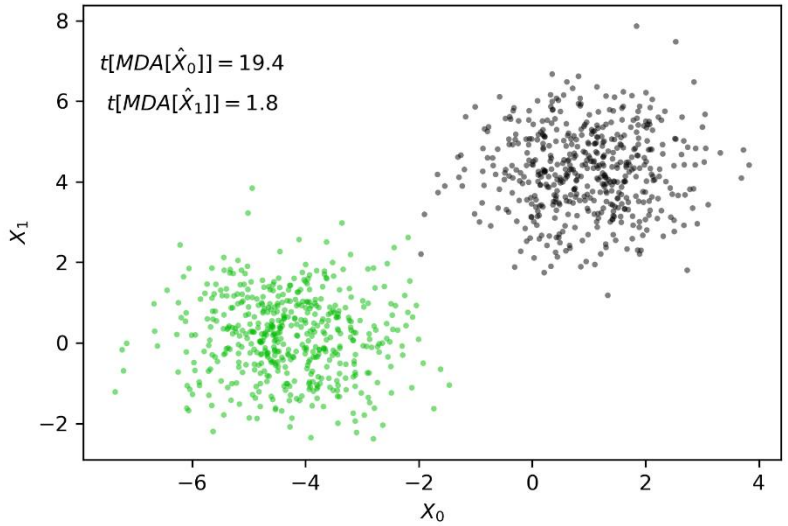
- When implemented correctly, MDA can prevent selection bias
- The key is to avoid reusing a single test set multiple times
- For instance
  - Repeated CV: K-Fold can be applied on shuffled rows of (X, y), after purging, thus preventing that a model is selected out of a particular split scheme
  - Monte Carlo CV: Cross-validation can be conducted on subsampled rows of (X, y) (random sampling without replacement)
  - Leave-p-out CV: This is an exhaustive cross-validation, where all possible combinations of p-sized test sets are evaluated

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	Paths
G1	x	x	x	x	x											5
G2	x					x	x	x	x							5
G3		x				x				x	x	x				5
G4			x				x			x			x	x		5
G5				x				x			x		x		x	5
G6					x				x			x		x	x	5

Combinatorially-purged cross-validation (CPCV) is a variation of *leave-p-out*, where the train set in purged and embargoed.
For example, by holding 2 out of 6 folds, we can form 15 different train-test splits, resulting in 5 full paths (where the 6-Fold path is only one of them). The number of paths can be raised, by increasing the value of *K* or the value of *p*. The end effect is an arbitrarily small chance of overfitting.

Substitution Effects

- When two important features share information, shuffling one may not result in a material reduction in model performance
  - MDA may wrongly dismiss one or both as uninformed
- Typically MDA is more robust to substitution effects than *p*-values (see right plot), however this is a potential vulnerability
  - One possibility is to shuffle together all features with mutual information
- The rest of the presentation explains how **Clustered MDA addresses substitution effects**



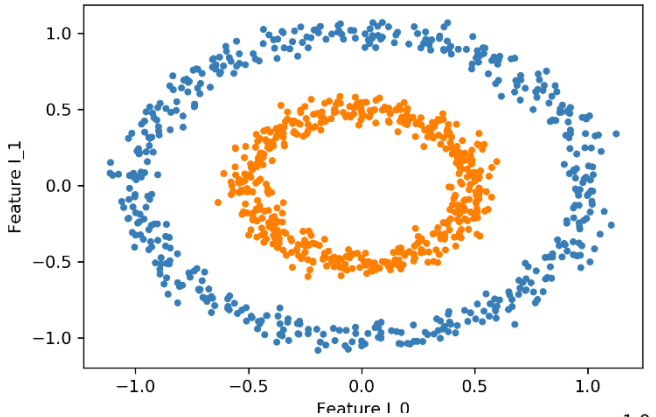
These two blobs can be easily separated with either X_0 or X_1. MDA shows a high t-value for both, however we can appreciate a substitution effect whereby the importance of X_1 is undermined by X_1.

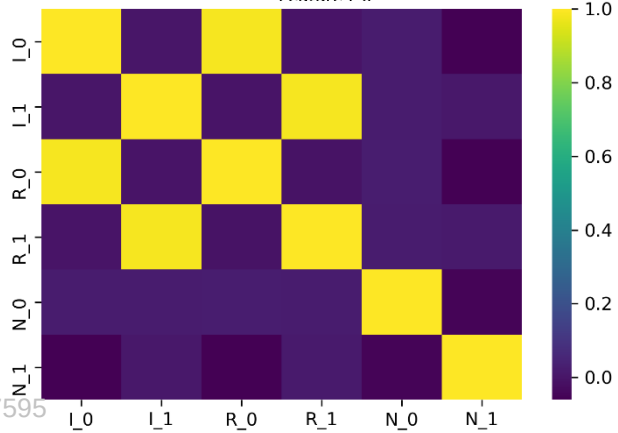
Clustered Feature Importance under Linear Substitution Effects

Linear Substitution Effects

- Consider 1,000 binary labels that can be clearly separated as a function of two informed features, I_0 and I_1
  - This is not an artificial example. Credit markets display this type of concentric patterns
- We add two redundant features:
  - R_0 is a linear function of I_0
  - R_1 is a linear function of I_1
- We add two noise feature, N_0 and N_1

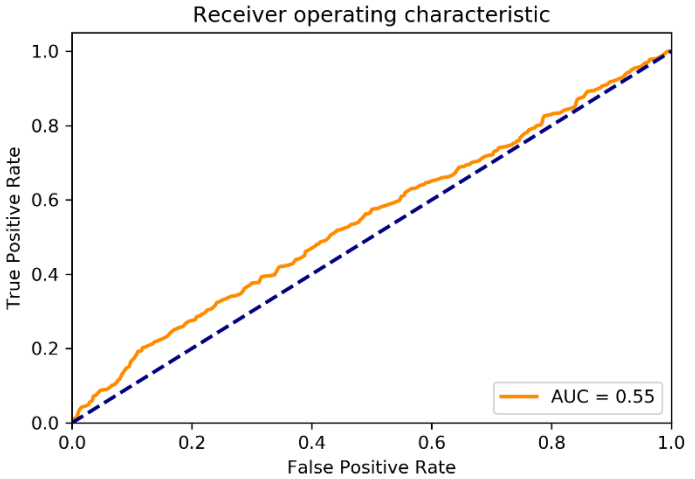
- The correlation matrix evidences that the system is multicollinear
  - Notice the off-diagonal blocks (I_0,R_1) and (I_1,R_1)



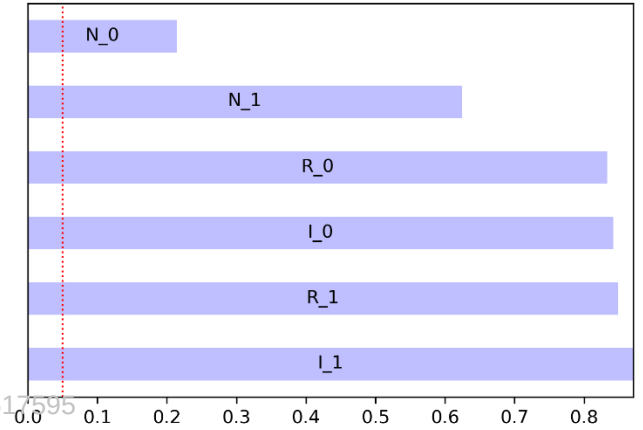


p-Values

- All *p*-values exceed the 5% threshold
- *p*-values mislead us into concluding that *all* features are noise (a false negative)
- Moreover, N_0 and N_1 exhibit lower *p*-values than the informed and redundant features



- Why do *p*-values mislead us?:
  - The model is misspecified
    - AUC > 0.5, but it should have been 1, because the labels are perfectly separable
  - The system is multicollinear
  - *p*-values do not disentangle the specification search from the significance search

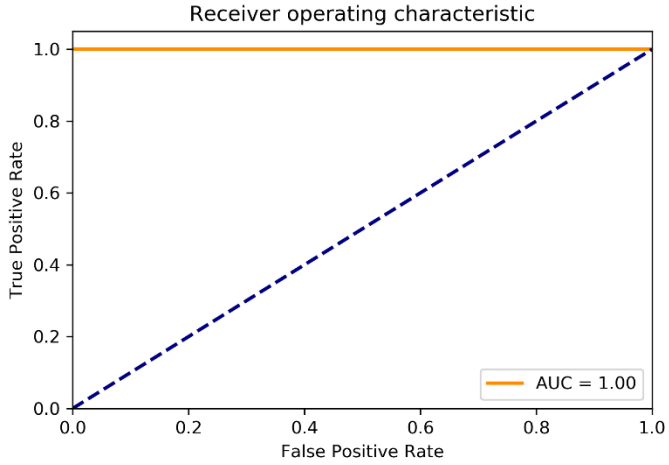


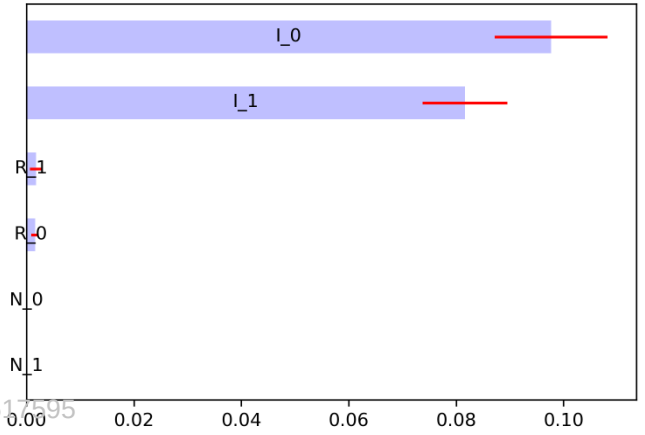
MDA on Random Forest

- MDA correctly identifies I_0 and I_1 as informed
- However, MDA incorrectly dismisses R_0 and R_1 as noise

- MDA (partially) failed because of multicollinearity
  - (I_0, I_1) prevent a reduction of AUC when (R_0, R_1) are shuffled

- Note that, unlike with *p*-values, misspecification is not an issue: AUC=1



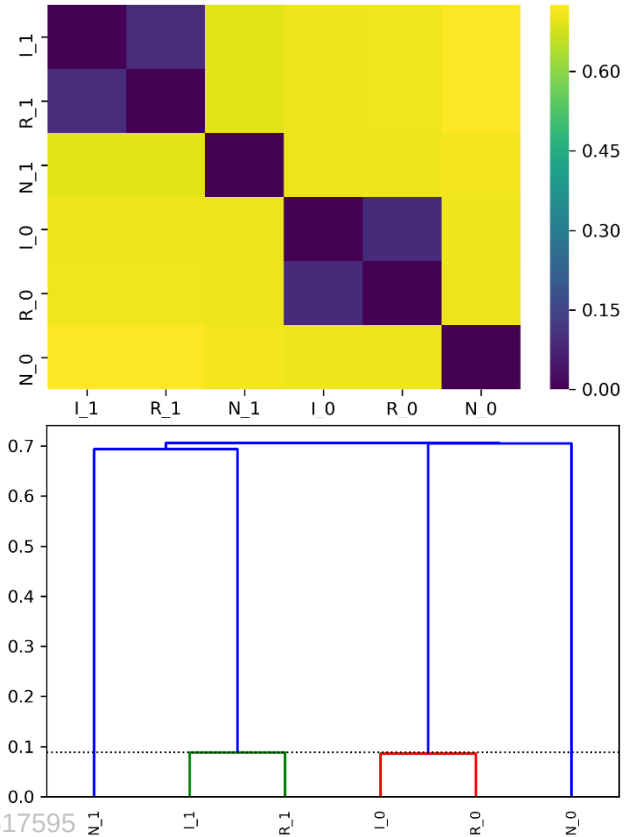


Features Clustering

- Apply a single-linkage agglomerative clustering algorithm on a distance matrix

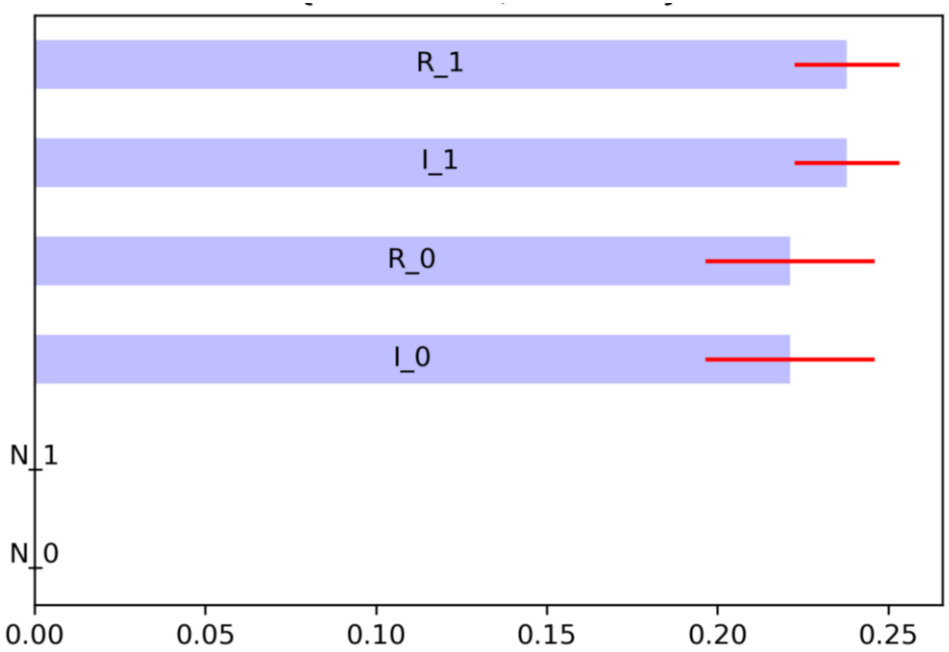
$$d_{i,j} = \sqrt{\frac{1}{2}(1 - \rho_{i,j})}$$

- We base the distance on correlation, because the substitution effect is linear
- The algorithm recognizes that
  - The optimal number of clusters is 4
  - R_0 is redundant to I_0
  - R_1 is redundant to I_1
- **The system formed by the clusters is not multicollinear (no off-diagonal blocks)**



Clustered MDA on Random Forest

- Instead of shuffling each variable individually, we shuffle together all variables within a cluster
- Clustered MDA gives the right answer:
  - All informed and redundant features are important
  - N_0 and N_1 have zero contribution to the model's performance
- Why did Clustered MDA work?
  - MDA decouples the specification search from the significance search
  - The clusters are not multicollinear

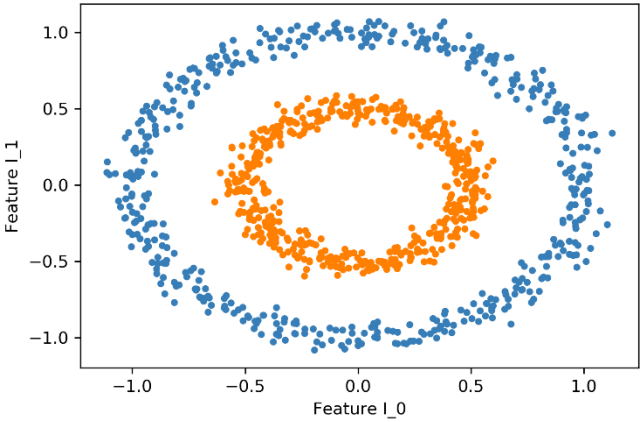


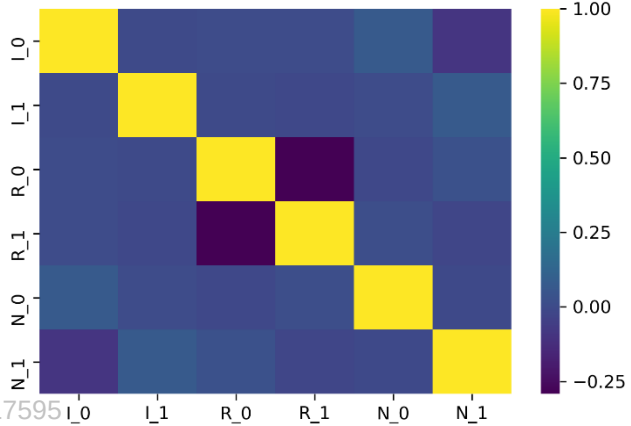
Clustered Feature Importance under Non-Linear Substitution Effects

Non-Linear Substitution Effects

- Consider 1,000 binary labels that can be clearly separated as a function of two informed features, I_0 and I_1
- We add two non-linear redundant features:
  - R_0=cos[I_0]
  - R_1=cos[I_1]
- We add two noise feature, N_0 and N_1

- The system is not multicollinear, however there are strong non-linear substitution effects between the features





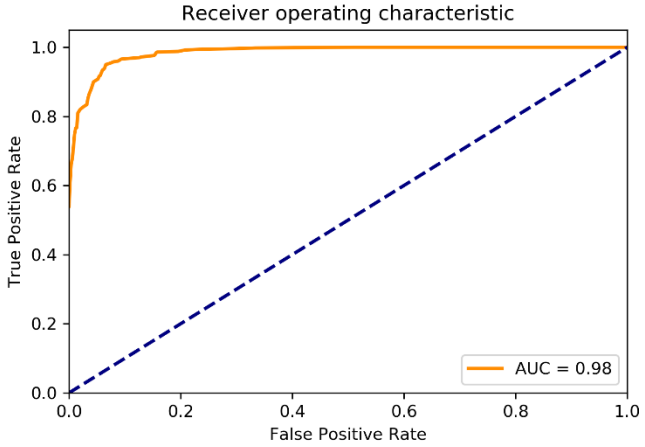
p-Values

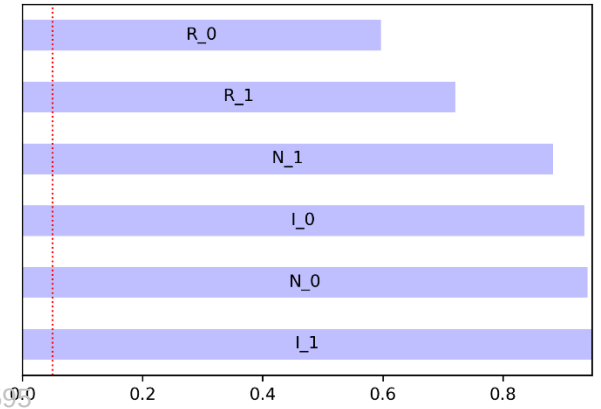
- The model is correctly specified
  - The redundant features allow logit to separate the labels well, with an ROC of almost 1

- Despite of the high ROC, *all* *p*-values exceed the 5% threshold. Why?

- The reason is, the substitution effects do not allow logit to estimate the parameters robustly

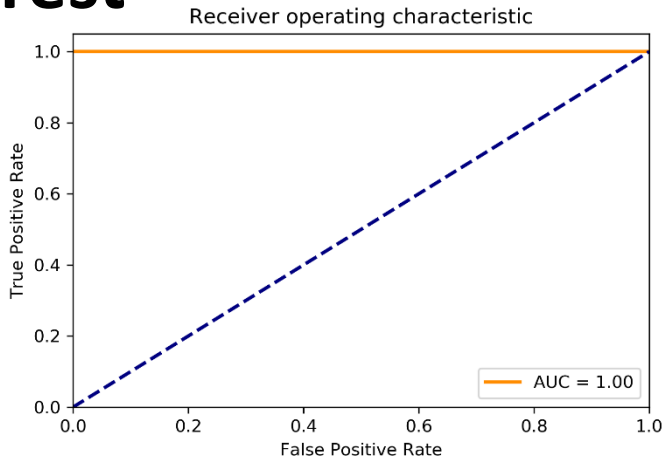
- One again, *p*-values are unable to select the important features

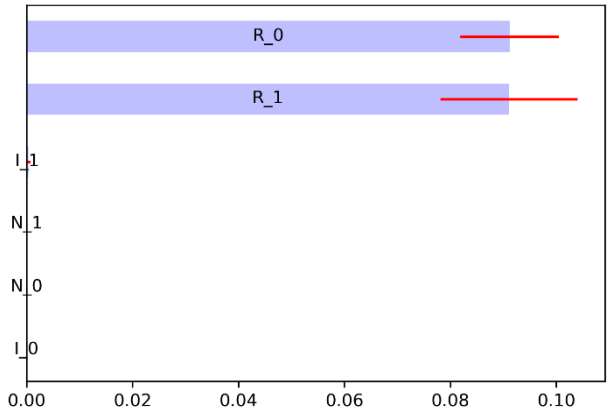




MDA on Random Forest

- At AUC=1, the model is correctly specified
- MDA correctly identifies R_0 and R_1 as important
- However, MDA incorrectly dismisses I_0 and I_1
- MDA (partially) failed because of the non-linear substitution effects
  - (R_0, R_1) prevent a reduction of AUC when (I_0, I_1) are shuffled

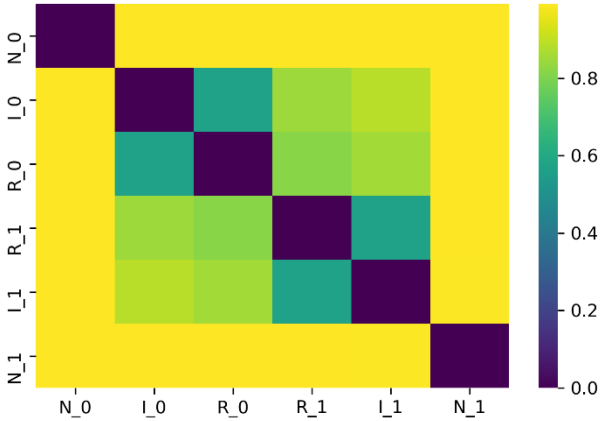


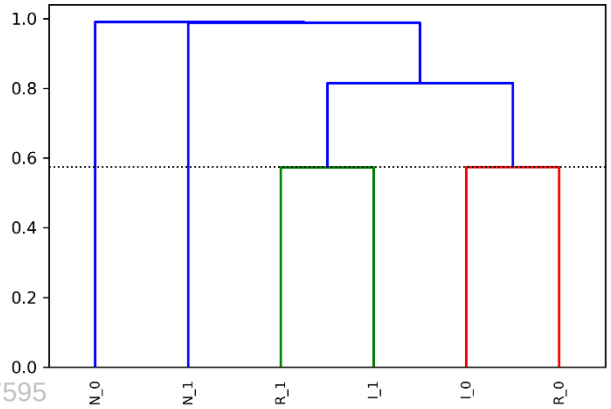


Features Clustering

- Apply a single-linkage agglomerative clustering algorithm on a **variation of information (VI) matrix**
- VI can effectively measure linear as well as non-linear codependence
- The algorithm recognizes that
  - The optimal number of clusters is 4
  - R_0 is redundant to I_0
  - R_1 is redundant to I_1

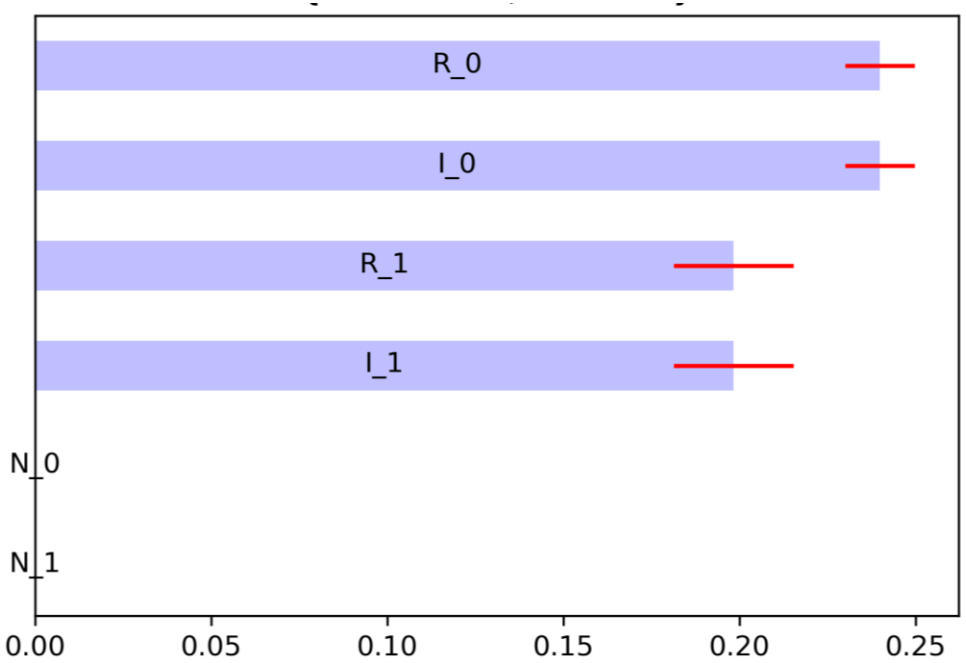
- **The system formed by the clusters does not exhibit substitution effects (no off-diagonal blocks)**





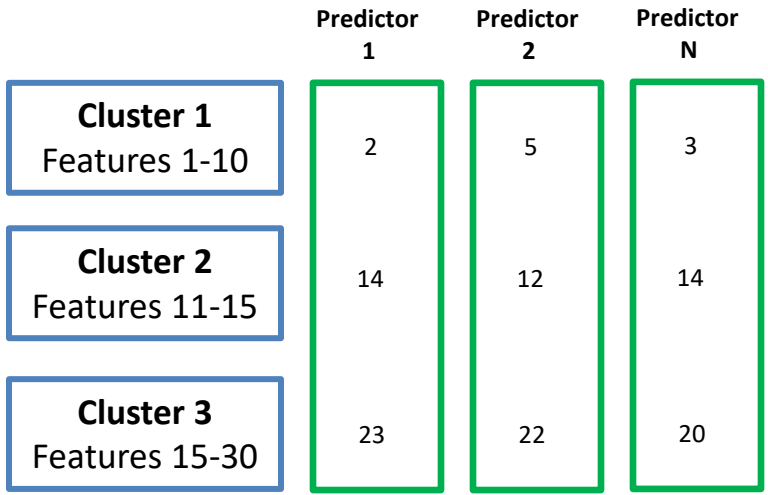
Clustered MDA on Random Forest

- Instead of shuffling each variable individually, we shuffle together all variables within a cluster
- Clustered MDA gives the right answer:
  - All informed and redundant features are important
  - N_0 and N_1 have zero contribution to the model's performance
- Why did Clustered MDA work?
  - MDA decouples the specification search from the significance search
  - Clusters concentrate most of the mutual information, muting substitution effects

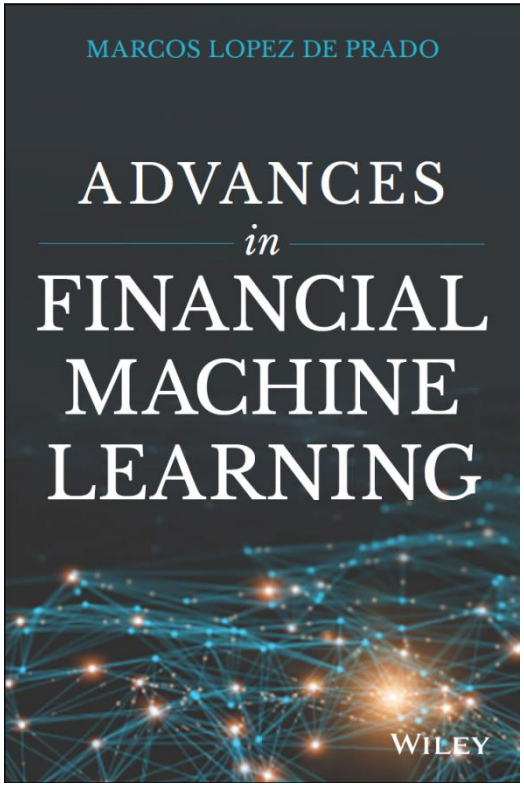


Uses of Clustered MDA

- Clustered MDA identifies the variables involved in a phenomenon
  - With that knowledge, we can hypothesize a particular cause-effect mechanism that binds those variables together
- Clustered MDA avoids substitution effects
  - It provides an intuitive form of regularization
- Moreover, Clustered MDA also helps us fit better ensemble models
  - Each individual predictor is fit by randomly drawing one feature per cluster
  - As a result, the individual predictors exhibit lower correlation, thus reducing the variance of the ensemble predictions



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