

---

# Feature Selection in Finance

---

**Benjamin A. Schifman**

**Justin J. Siekmann**

Department of Electrical and Computer Engineering

University of Arizona

Tucson, AZ 85719

bschifman@email.arizona.edu

jsiekman@email.arizona.edu

## Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

## 1 Introduction

There exist technical analysis indicators traditionally used by analysts to evaluate and predict market and equity performance, as they “can provide a unique perspective on the strength and direction of the underlying price action” (<–what is this quote from??).

Feature selection is used to determine relevant indicators while identifying those that are irrelevant and redundant. Different implementations of algorithms based on these indicators could be used to predict performance of individual equities, sectors, or overall markets. They could also be used to classify and identify the correlation and interdependencies between equities, sectors, and markets. Our goal is to implement various approaches to determine efficacy of technical indicators as enablers to financial analysis. A challenge in the implementation of this project was to develop an accurate testing method.

Possibly reword and keep/move:

From this project we hope to deepen our understanding of the usage cases for applying specific machine learning algorithms as well as expanding upon our technical analysis of the stock market and which indicators play a role in successful market analysis.

## 2 Related Work

This is optional. If wanted, save til last.

## 3 Methods/Approach

The following subsections present details and explanations of the methods and functions implemented as part of this project.

### 3.1 Normalization

As each technical analysis indicator produces values applicable based on the way the indicator was calculated, normalization of the indicators makes correlations between them during feature selection more accurate and applicable. Each value in a specified time period is normalized using Equation (1).

If there is extra data not consisting of a full time period, the extras are thrown out at the beginning of the data as the data near the end may be more relevant.

insert norm equation (1)

### 3.2 Maximal Information Coefficient

The Maximal Information Coefficient (MIC) is "a measure of two-variable dependence designed specifically for rapid exploration of many-dimensional data sets" - <http://www.exploredata.net/>. A benefit to MIC correlations between two variables is that it can be described regardless of linear or non-linear relationships. The MIC yields a single value  $0 \leq MIC \leq 1$  with a value closer to 1 representing that the variables are more closely correlated, and a value near 0 indicates statistically independent variables that have neither linear nor nonlinear relationships. The *minepy* library was used in python to rank the features according to their MIC with the target variable. The MIC was calculated for each feature in each ticker, and then a final MIC value for each feature was calculated by taking the mean of the values.

## 4 Results

## 5 Conclusion

### References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to small (9 point) when listing the references. **Remember that you can use a ninth page as long as it contains only cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauero, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.