An information theory based measure of structural break

Alexandre P. Borentain

April 3, 2024

Abstract

Structural change is a central topic in modeling. We want to learn the mechanisms that make the structure of the data generating process underlying a set of observations change. In finance, structural breaks are mostly defined according to two categories: CUSUM tests and Explosiveness tests. Both of these methods provide insights into structural changes, however, with the rise of nonlinear modeling, it becomes important to supplement these tests with the more general and natural measure of association that is mutual information (MI) and a measure of complexity like entropy. The procedure outlined is intended to be used in the context of machine learning predictions and consists of comparing the information given by the features of a model to the complexity of the order flow. This text first outlines the goal of generalizing structural changes to information theory, then derives a measure directly applicable to finance, and finally tests the model on both generated and real world data. We show that our new measure improves the prediction accuracy of ML models compared to using CUSUM test.

1 Introduction

When modeling a time series, we need to take into consideration the fact that the relationship between the variables that we observe can change abruptly. We define structural breaks as the periods when the structure of the data generating process underlying a set of observations changes, the consequence being the change in the relationship between our variables Darbellay and Wuertz [2000]. When using equation free modeling like machine learning models, the assumption is that the model will learn those breaks on its own, but this rarely happens accurately and it is always preferable to build features on which the model can rely to make those predictions.

In finance, there are currently two categories of tests that can serve as such features:

- CUSUM tests, which test whether the cumulative forecasting errors significantly deviate from white noise Ploberger [1992]. Here we will use the Chu-Stinchcombe-White CUSUM Test Homm [2011]
- Explosiveness tests, which test whether the process exhibits exponential growth or collapse Guo [2019].

The following example outlines the limitations of the CUSUM test: Imagine a process that follows the relationship $y = 0x + \epsilon$ over a certain time interval. The correlation will be 0.01. Over the next time interval, the process now follows the relationship $y = 100|x| + \epsilon$ with a very similar correlation coefficient. There was a structural change between the two time intervals, and in fact, the variable x went from having no association with y to having a strong one. Yet the CUSUM test would not be able to show it.

Another limitation is that the the CUSUM test does not explore which of our variables was affected by the change. A first method already used when making predictions with any machine learning model is to look at feature importance scores over different periods of time. SHAP Values measures the influence a certain feature had toward driving model decision El Mokhtari [2019]. If the model has near perfect accuracy, this is essentially measuring which feature influences a certain outcome the most.

When modelling asset prices however, we cannot possibly have a consistently high accuracy. This is because there is a varying amount of randomness associated with an asset price and any change in feature importance we observe needs to be put in relation to the change in randomness of the process.

Even without this issue, our model will rarely be accurate enough. We need to measure the maximum possible accuracy that our features could obtain when predicting asset prices.

2 Method

Some considerations: this work is based on the assumption that we have a set of features X that has a large enough predictive power on a set of observations y. If X is uninformative about y, then we cannot infer any structural change in the process underlying y from X

2.1 Estimation of the randomness of a process

Market microstructure theory tells us that as prices carry more information, they become more random. In fact, the Efficient Market Hypothesis implies that if an asset price reflects all available information, then its behavior is completely random Fama [1970].

For an asset price, the process can be better approximated by looking at trade data, more specifically, estimating the unpredictability of the order flow imbalance (OI) will reveal how much of the mechanism underlying the price of an asset is random Darbellay and Wuertz [2000]. We can find the complexity in the process underlying the OI by looking at the entropy of the process and comparing it to the maximum entropy that process could have. This will give us an approximation of how unpredictable the OI is. Shannon defined entropy for a discrete random variable X with possible values $\{x_1, x_2, ..., x_n\}$ and a probability mass function $p(x)$, as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i)) \tag{1}$$

For a sequence of volume bars $\tau = 1, \ldots, N$, of size V, the buyer initiated volume is computed with the bulk volume classification algorithm as follows:

$$V_\tau^B = V_\tau \times t\left(\frac{\Delta p_\tau}{\sigma_{\Delta p}}, df\right) \tag{2}$$

where V_t^B is the estimated buyer-initiated volume during bar τ, V_τ is the aggregated volume, Δp_τ is the price change between consecutive bars, $\sigma_{\Delta p}$ is the standard deviation of price changes, and t is the cumulative distribution function of the Student's t-distribution, with df degrees of freedom.

The portion of buyer-initiated volume during bar τ is $v_\tau^B = \frac{\hat{V}_\tau^B}{V_\tau}$, where $v_\tau^B \in [0, 1]$. We measure the entropy in v_τ^B by quantizing. calculate the q-quantiles of the buy volume series $\{v_\tau^B\}$, which segment it into q exclusive groups, denoted as $K = \{K_1, ..., K_q\}$. Next, we construct a function that assigns each buy volume v_τ^B to a corresponding group, defined as $f : v_\tau^B \mapsto \{1, ..., q\}$, such that $f(v_\tau^B) = i$ if v_τ^B is in K_i, where i ranges from 1 to q. Subsequently, we discretize the series $\{v_\tau^B\}$ by attributing to each v_τ^B the identifier of the group to which it is allocated by f. This process transforms the set of buy volumes $\{v_\tau^B\}$ into a sequence of integers $X = [f(v_1^B), f(v_2^B), ..., f(v_N^B)]$. Lastly, the entropy $H[X]$ of this discrete sequence is computed using the method developed by Kontoyiannis, which is based on the Lempel-Ziv complexity. If the sequence X was random, it could take each value of q with equal probability. The Shannon entropy of that process would be: $\hat{H}[X] = -\log q$. Therefore, we estimate the randomness in the process underlying a set of asset price as $R = \frac{H[X]}{\hat{H}[X]}$

2.2 Maximum predictive power

Given our set of explanatory variables x we want to find the best achievable classification accuracy for an observation y_i with $y_i \in \{1, \ldots, q\}$ Fano's inequality theorem Fano [1949] states:

$$\bar{F}(P_{y_i,x}) \leq \bar{h}_q^{-1}(h(y_i|x)),$$

where $\bar{F}(P_{y_i,x})$ is the highest classification accuracy achievable, and $\bar{h}_q^{-1}$ is the inverse of the function

$$\bar{h}_q(a) := -a \log a - (1-a) \log \frac{1-a}{q-1}, \quad a \in \left[\frac{1}{q}, 1\right].$$

We can rewrite $\bar{F}(P_{y_i,x})$ as:

$$\bar{F}(P_{y_i,x}) = \bar{h}_q^{-1}(h(y) - I(y_i; x)),$$

and use the Difference-of-Entropies (DoE) Estimator Poole [2019] to find the bounds on $I(y_i; x)$ and estimate it along with an estimate of the error.

Our intuition is that $\bar{F}(P_{y_i,x})$ should be dependent on the randomness of the process underlying the data generating process y and the level of influence that our variables x exert on that process.

2.3 Intra-Feature relations

If we want to measure the influence of each feature on the process we are modeling, it is not enough to measure the best achievable classification accuracy for each feature individually. This is because a set of N features where $N > 1$ can be largely uninformative when each feature in the set is taken individually to predict an outcome, but very informative when the taken jointly to predict the same outcome.

For our purposes of measuring the change in influence of each variable, we want to measure the added maximum accuracy that our feature provides by comparing the maximum accuracy of all of our features together to the maximum accuracy of all our features minus the one we want to test. Note that this isn't a proper feature importance test. This is meant to be used with a set of features which all provide some information not redundant to all other features.

Let $X = \{X_1, X_2, \ldots, X_N\}$ denote a set of N features, and let y represent the target variable we aim to predict. For each feature X_i within this set, we seek to assess its contributory value or influence on the model's predictive capability when utilized jointly with the other features. This assessment involves comparing the highest achievable classification accuracy using all features collectively against the highest achievable classification accuracy using all features except X_i.

Formally, for each feature X_i, the ratio R_i is defined as:

$$R_i = \frac{\bar{F}(P_{y,\{X_1,\ldots,X_N\}})}{\bar{F}(P_{y,\{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_N\}})}$$

where:

- $\bar{F}(P_{y,\{X_1,\ldots,X_N\}})$ denotes the highest classification accuracy achievable with the complete set of features $\{X_1, X_2, \ldots, X_N\}$ for predicting the outcome y.
- $\bar{F}(P_{y,\{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_N\}})$ denotes the highest classification accuracy achievable with the set of features $\{X_1, X_2, \ldots, X_N\}$ excluding X_i, for predicting the outcome y.

The ratio R_i quantifies the added value or influence of the feature X_i within the context of the entire feature set, illustrating how the inclusion of X_i impacts predictive accuracy relative to its exclusion.

We measure R_i for each successive observation. An abrupt change in this ratio over a set of observations is indicative of a structural change.

2.4 The measure of structural break

Although they will both be close, we use the entropy of the order flow imbalance instead of the entropy of the returns for the estimation of randomness because the order flow imbalance more closely approximate the process over the entire volume bar.

To quantify structural breaks in the context of our predictive modeling, we define a measure that captures the changes in the information provided by the features relative to the randomness of the process. This measure takes into consideration both the predictive power of the model and the inherent unpredictability of the asset price movements. Each new volume bar τ, we define the structural break measure $INSB_\tau$ as the ratio of the classification accuracy $A(P_{x,y})$ to the randomness R over the last 100 bars:

$$INSB_\tau = \frac{\bar{F}(P_{x_k,y_k})}{R_k} \tag{3}$$

Here, $\bar{F}(P_{x_k,y_k})$ is the classification accuracy of our model over the past 100 bars ending at the k-th bar, and R_k is the estimated randomness in the process underlying the order flow imbalance for

the same interval bar. The measure S_τ thus reflects the ratio of predictive power to randomness, over the most recent 100 bars.

The result is that as $INSB_\tau$ increases, out of all the features that influence the process we are modeling, our features take a less important place. I will call this an Expansive structural break. The inverse is that as $INSB_\tau$ decreases, out of all the features that influence the process we are modeling, our features take a more important place. I will call this a Constrictive structural break.

2.5 Market data testing

We follow the recommendations of Lopez de Prado [2019] and declare that testing was conducted only once and only on MSFT trade data obtained from ThetaData (thetadata.net).

We use the methods from López de Prado [2018] and use a primary XGBoost model to predict the sign of the returns and a secondary XGBoost model to predict the size, more specifically, the second model predicts whether the side predictions of the initial model is correct or not. Because of the nature of the measure $INSB$ and its relation to entropy, we decided to implement it as a feature of the size predicting model. We note however that it could be used for both as structural breaks are an important aspect of predicting the side of future returns. We implement two size predicting models, one given $INSB$ and another CUSUM test statistics, with every other feature being the same. We test the out-of-sample performance with combinatorial purged cross-validation (CPCV) and measure accuracy, precision, recall, and negative log-loss.

3 Results

We use the trade and quote data of Microsoft from 2019 to 2021. Our data is from ThetaData (thetadata.net). We make sure to preventing look-ahead bias. We observed distinct patterns in the calculated entropy of the BVC, the approximated maximum predictive power $A(p_{x,y})$, and the resulting ratio $INSB$. The model using $INSB$ outperforms the one using CUSUM in all metrics: accuracy, precision, recall, F1 score, and negative log-loss.

3.1 Analysis

The entropy of BVC over time presents a fluctuating pattern, reflecting varying degrees of randomness in buyer-initiated volumes. Peaks in entropy suggest periods of high uncertainty in trading behavior, while troughs indicate more predictable patterns.

The calculated $A(p_{x,y})$, which estimates the predictive power of our features towards the returns, fluctuated during the period but stayed above 0.62 for 90% of the duration tested, which validates our initial condition of having a set of features X that has a large enough predictive power on a set of observations y.

$A(p_{x,y})$ showed a correlations of -0.47 with R. Periods of high entropy often coincided with lower values of $A(p_{x,y})$, suggesting that increased market randomness diminishes the predictive capability of the features. Our measure $INSB$ captures the difference between them.

3.2 Model performance

In this section, we evaluate and compare the performance of two predictive models: Model 1, which incorporates our novel structural break measure S, and Model 2, a baseline model that does not use the measure S and uses the CUSUM test instead. The performance of both models was assessed using the following metrics: accuracy, precision, recall, F1 score, and negative log-loss.

Table 1 presents a summary of the performance metrics obtained for each model. The results clearly indicate that Model 1 significantly outperforms Model 2 across all evaluated metrics. Notably, Model 1 achieves a higher accuracy of 0.734 compared to 0.656 for Model 2. This improvement suggests that incorporating the structural break measure enables the model to more accurately predict outcomes.

Table 1: Comparison of Model Performance

Metric	Model 1	Model 2
Accuracy	0.734	0.656
Recall	0.704	0.615
Precision	0.838	0.778
F1	0.765	0.687
Negative Log-Loss	-0.25	-0.30

Model 1 uses measure S and Model 2 uses CUSUM. This table provides a comparison between Model 1 and Model 2 across several performance metrics. Model 1 shows superior performance in terms of accuracy, precision, and recall, while having a higher negative log-loss value.

References

Georges A Darbellay and Diethelm Wuertz. The entropy as a tool for analysing statistical dependences in financial time series. *Computational Statistics & Data Analysis*, 34(4):371–390, 2000.

K. El Mokhtari. Interpreting financial time series with shap values. *CASCON '19: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pages 166–172, 2019.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25 (2):383–417, 1970.

R. M. Fano. The transmission of information. *RESEARCH LABORATORY OF ELECTRONICS*, 65, 1949.

G. Guo. Testing for moderate explosiveness. *The Econometrics Journal*, 22(1):73–94, 2019.

U. Homm. Testing for speculative bubbles in stock markets: A comparison of alternative methods. *2011Journal of Financial Econometrics*, 10(1):198–231, 2011.

M. López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.

M. Lopez de Prado. Confidence and power of the sharpe ratio under multiple testing. *available at SSRN: https://ssrn.com/abstract=3193697 or http://dx.doi.org/10.2139/ssrn.3193697*, 2019.

W. Ploberger. The cusum test with ols residuals. *Econometrica*, 60(2):271–285, 1992.

Ben. Poole. On variational bounds of mutual information. *Proceedings of the 36th International Conference on Machine Learning,*, pages 5171–5180, 2019.