Symmetric Information but Asymmetric Trust?
The revealed and tacit knowledge of markets under pandemic risk

Percy Venegas¹ and Tomas Krabec²

¹ Economy Monitor, Costa Rica
² University of Economics, Prague, Czech Republic
v.percy@economymonitor.com

Abstract. Schneider and Trojani's work on Divergence and the Price of Uncertainty (2019) sees divergence as tradeable in the context of an options portfolio, where realized divergence measures the distinct realized moments associated with time-varying uncertainty. At a macro level, we could identify the special instances where asset prices and a given indicator move in opposite directions during events where survival is at stake (e.g. in the transition from epidemic to pandemic). The true information content of the indicator is not always apparent since the information is seemingly accessible to market participants at a level playing field - nonetheless, divergent anomalies are present. So, how to explain that strange behavior? We propose a nonlinear approach using evolutionary algorithms to capture the dynamics given the nature of the process, as Norman, Bar-Yam, and Taleb, describe in their recent research note Systemic risk of pandemic via novel pathogens – Coronavirus (2020): "We are dealing with an extreme fat-tailed process owing to an increased connectivity, which increases the spreading in a nonlinear way ". The biology-inspired heuristic is based on Kotancheck's implementation of genetic programming.

Keywords: Epidemic, Pandemic, Coronavirus, COVID-19, Precautionary Principle, Genetic programming, Futures pricing, Information Asymmetry, Trust Asymmetry.

'Symmetry' indicates a dynamic relationship or connection between objects, and it is all-pervasive. –Marcus du Sautoy

1 Background

1.1 Literature

Authors have [1] studied the relationship between economic activity and the spread of viral diseases using high frequency data from the real and financial economies. However, problems arising include transparency and reliability –for instance, even “gold standard” datasets such as Johns Hopkins CSSE is known to sometimes disagree with numbers found in other sources [2]. Moreover, there is implicit
knowledge encoded in the structure of networks [3], that is not readily accessible even if information is considered to be symmetrical among market participants.

1.2 Genetic programming modeling metrics

In the present paper we demonstrate metrics computed using symbolic regression via genetic programming [4], that were introduced by Kotanchek in the DataModeler software and have direct application to data symmetry problems:

Data Strangeness: Comparing the strangeness of each of the data records during Outlier Analysis allows calculation of the outlier Distance of each data record which provides a ranking and assessment of the difficulty of modeling that particular data record.

Divergence function: which returns the response consensus behavior (as measured by the Ensemble Divergence Function) of the supplied model ensemble. A Model Ensemble is a special form of model since it is trustable. The measure of that trustability is the Ensemble Divergence Function which measures the spread in constituent model predictions.

We evaluate these metrics using empirical datasets available from Statista, Morningstar, and, the Wolfram Data Repository (Epidemic Data for Novel Coronavirus COVID-19). The end goal is to detect speculative behavior by economic agents, that may encode hidden information not released by governments or institutions.

2 Methodology

Our data set contains daily time series for 500 websites visited by users of the World Health Organization website; the parameter under study is bounce rate (the percentage of users who visited only 1 page in the website before leaving), which provides a proxy for interest on the information content of the site and its change over time. The target variable is the CBOE VIX Volatility Index, which measures the expectation of stock market volatility over the next 30 days implied by S&P 500 index options (a score below 10 is considered to be low, while a score above 20 is considered high [5]). The period of study is from November 1st, 2019 (just before the coronavirus associated with Covid-19 was discovered) to April 30th, 2020.

Figure 1 shows a graphical representation of the dataset, note how some information sites (or new sections within existing sites) that were created specifically for the purpose of addressing issues related to the pandemic raise from 0 visits (yellow) to high levels of traffic (dark blue).
The first round of modeling using symbolic regression produces the more prevalent variable combinations across models, as seen in Table 1. Counterintuitively, official sites with high traffic that were present in the data set did not make it to the list of important variables, one example is the Center for Disease Control—the usage of scientific information journals such as The Lancet depicted better the climate of uncertainty and expectation of volatility.

**Table 1.** Variable association table

<table>
<thead>
<tr>
<th>Rank</th>
<th>Var 1</th>
<th>Var 2</th>
<th>Var 1 %</th>
<th>Var 2 %</th>
<th># Models</th>
<th>% of Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>thelancetcom</td>
<td>usnewscom</td>
<td>55.2</td>
<td>79.4</td>
<td>492</td>
<td>43.2</td>
</tr>
<tr>
<td>2</td>
<td>thelancetcom</td>
<td>worldometersinfo</td>
<td>50.3</td>
<td>94.7</td>
<td>449</td>
<td>39.5</td>
</tr>
<tr>
<td>3</td>
<td>thelancetcom</td>
<td>naturecom</td>
<td>49.1</td>
<td>90.1</td>
<td>438</td>
<td>38.5</td>
</tr>
<tr>
<td>4</td>
<td>thelancetcom</td>
<td>weforumorg</td>
<td>48.2</td>
<td>83.2</td>
<td>430</td>
<td>37.8</td>
</tr>
<tr>
<td>5</td>
<td>thelancetcom</td>
<td>healthnswgovau</td>
<td>42.7</td>
<td>82.8</td>
<td>381</td>
<td>33.5</td>
</tr>
<tr>
<td>6</td>
<td>thelancetcom</td>
<td>uniceorg</td>
<td>41.7</td>
<td>77.2</td>
<td>372</td>
<td>32.7</td>
</tr>
<tr>
<td>7</td>
<td>weforumorg</td>
<td>uniceorg</td>
<td>65.6</td>
<td>70.3</td>
<td>339</td>
<td>29.8</td>
</tr>
<tr>
<td>8</td>
<td>thelancetcom</td>
<td>unorg</td>
<td>36.9</td>
<td>91.4</td>
<td>329</td>
<td>28.9</td>
</tr>
<tr>
<td>9</td>
<td>usnewscom</td>
<td>weforumorg</td>
<td>48.7</td>
<td>59.6</td>
<td>308</td>
<td>27.1</td>
</tr>
<tr>
<td>10</td>
<td>usnewscom</td>
<td>worldometersinfo</td>
<td>48.1</td>
<td>62.9</td>
<td>298</td>
<td>26.2</td>
</tr>
<tr>
<td>11</td>
<td>usnewscom</td>
<td>naturecom</td>
<td>45.5</td>
<td>56.0</td>
<td>282</td>
<td>24.8</td>
</tr>
<tr>
<td>12</td>
<td>naturecom</td>
<td>worldometersinfo</td>
<td>57.4</td>
<td>58.9</td>
<td>279</td>
<td>24.5</td>
</tr>
<tr>
<td>13</td>
<td>worldometersinfo</td>
<td>healthnswgovau</td>
<td>55.1</td>
<td>56.7</td>
<td>261</td>
<td>22.9</td>
</tr>
<tr>
<td>14</td>
<td>weforumorg</td>
<td>naturecom</td>
<td>46.8</td>
<td>49.8</td>
<td>242</td>
<td>21.3</td>
</tr>
</tbody>
</table>

Figure 2 shows a rendering of variable associations at a 0.2 significance level.
3 Analysis

3.1 The breaking of symmetric correlation

If we isolate the timeseries of the most interesting variable associations (figure 3) we can appreciate that bounce rates at most sites, with the exception of worldometers.info, move in tandem with the VIX. Furthermore, we notice how large spikes in that variable tend to occur at discontinuities of the VIX (e.g. during weekends).
A closer inspection shows that in fact that variable is highly uncorrelated with the VIX and most other variables, as seen in figure 4.
In fact, that seemingly unofficial information channel that offers real time indicators on many economic variables, created a popular section to track Covid-19 stats https://www.worldometers.info/coronavirus/

### 3.2 Trustable ensemble

During the second round of modeling (using only associated variables) we obtained an ensemble of quality $R^2 0.948425$, encompassing 48 models and 8 variables, with average model complexity of 31.8542. As expected, even when the ensemble of models performs better than any individual model, higher levels of volatility are difficult to predict, as seen in figure 5 (in red).
Fig. 5. Ensemble prediction plot

However, for the variable that breaks symmetry, low values (low bounce rate, indicating more interest on the content during uncertain times) are actually what becomes difficult to model. Figure 6 shows the residuals.
3.3 Metrics

The response comparison in figure 7 demonstrates how the ensemble prediction behaves across the values of each variable. Regions with large yellow ribbons are more uncertain: for instance, traditional news sources and scientific information sites are less trustworthy to predict extreme low or high volatility. Here again, our variable of interest not only moves in the opposite direction of volatility but is also more trustable (narrow ribbons), less divergent.

Finally, for the computation of outliers this approach performs nonlinear outlier detection under the presumption that the supplied models do a good job of capturing the overall response behavior. The returned result is a list of the records in the supplied data ordered by difficulty in modeling that data point. The Strangeness Metric is a pure function that evaluates the model residuals on a record-by-record basis. This function should produce a scalar with larger absolute values (either positive or negative) indicating more difficulty. The pure function should be of the form pureFunc[#1, #2]& where the first slot is the list of ModelResidual for that record and the second slot is the observed response associated with that record [6]. Table 2 shows the values computed.
The higher ranking outlier (row 133) occurred on March 11th, that day the bounce rate in most sites tended to be high, signaling low levels of engagement with the information content, while the value of the VIX that day showed that in fact more interest was merited: the level of volatility was skyrocketing.

4 Conclusions

We have demonstrated how trustability and outlier detection metrics derived using genetic programming reveal hidden information not accessible when following traditional information symmetries approaches. This is especially relevant in the context of pandemics, where official information channels might be not the most trusted sources.

For future work we recommend exploring complementary metrics, such as Neighborhood Asymmetry [6], which is an information metric needed to measure the information content of each data sample. One of the valid choices is the Average Neighborhood Distance which simply returns the average distance between the data record and the other data records implicitly specified by the data matrix. The Neighborhood Asymmetry function sums the vectors from the data record to the neighbors implicitly defined by the supplied data matrix and returns the length of this resulting neighborhood directionality normalized by the number of neighbors. Thus, this metric is primarily concerned with the symmetry of the neighbor distribution but also contains a contribution from the distance to each of the neighbors.

References