**Information Theory and Ensemble Models**

**How should we ensemble time-series forecasts better?**

Coming out of the pandemic, statisticians have run into a slew of geo-political complications further augmenting the inability to forecast business variables with accuracy. Did Ukraine drive retail prices higher in 2022 or was it the quantitative easing in 2021 that was the culprit? Some models say one thing and some say another, making it hard to *accurately* forecast inflation.

At the very foundation of econometrics, we rely on minimizing distance (MSE, RMSE, etc) between two points (forecast, actual) in the same domain (time, frequency). All of these have served the community immensely to improve the accuracy of forecasts. There are a few reasons why these metrics are popular:

1. They are **non-parametric**. Meaning models developed using different assumptions and structures can all be compared since the end output is the same — accuracy.
2. Historically, it has been true that these metrics **offer enough variation in residual distribution** for us to classify different models and cluster their performance. For instance, they help us answer if a class of ARMA models better suits the data or a class of state space models. More generally, it helps classify best fit algorithms.
3. Often, they sit in **euclidean geometries** and have nice properties that make building newer and more sophisticated methods on top of them much easier. The topology of measurement does not change. This unlocks the potential for making transformations and representations of the data, the model, or both to extract deeper relationships. Modern day ML models are built by exploiting this key property. WLOG, we’ll focus on simple econometric models for the purposes of this article.

These benefits have helped improve statistical packages since the late 40s so much so that today we have more general packages that are able to select best fit models without the user ever having to assume a structure.

However, as is true for any situation, continuing to optimize the same metrics offers fewer and fewer improvements with every iteration. Metrics measuring distance can no longer offer enough separation between optimized models making it hard to rank performance.

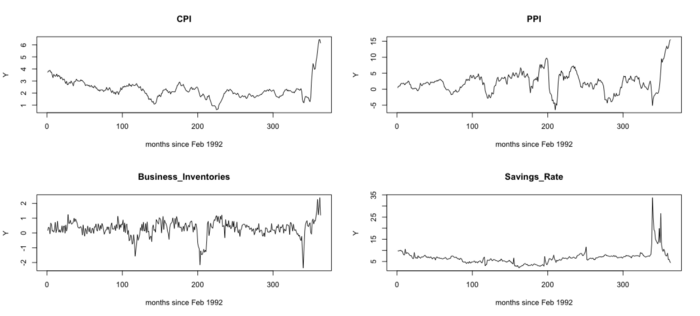
So what do we do?

Potential solutions can come from 2 different paths:

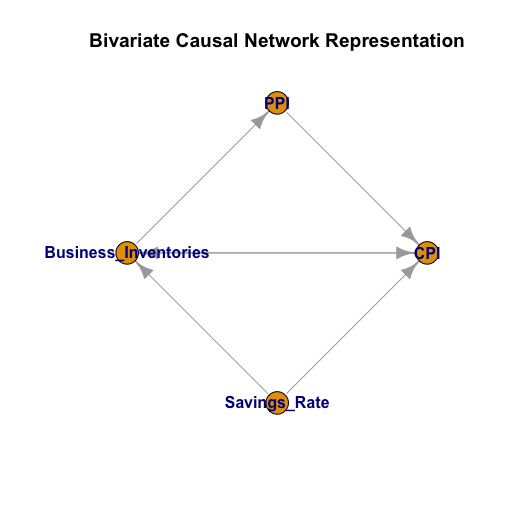
1. **Coming up with new metrics within the same topology** that can improve best bit classification — the community has come up with and extensively uses new metrics like AIC, AICc, BIC, BICc, and so on but these metrics are often model specific in that they might allow us to rank ARMA models within each other but cannot compare ARMA vs. ETS models, for instance.
2. **Coming up with new geometries and topology** for improved methods — I explore a version of this idea through granger causal networks which is very much a work in progress but continues to show immense potential (perhaps I’m biased given my interest in the topic). Interested readers can find a [primer on my approach here](https://towardsdatascience.com/granger-causal-networks-and-indirect-feedback-676549ba99e).

Let’s first start by contextualizing the arguments so far on actual data. I’m going to analyse inflationary trends and variables and how accurately we can forecast CPI using existing models. Just about every econometrician is bound to use the following variables when considering an inflation model:

1. **Consumer Price Index YoY Growth** — measure of inflation for consumers; demand side
2. **Producer Price Index YoY Growth** — measure of inflation for producers; supply side
3. **Savings Rate** — % of earned income that is saved; measures friction on the demand side
4. **Business Inventories MoM Growth** — excess stock of goods that businesses have every month; measures friction on the supply side

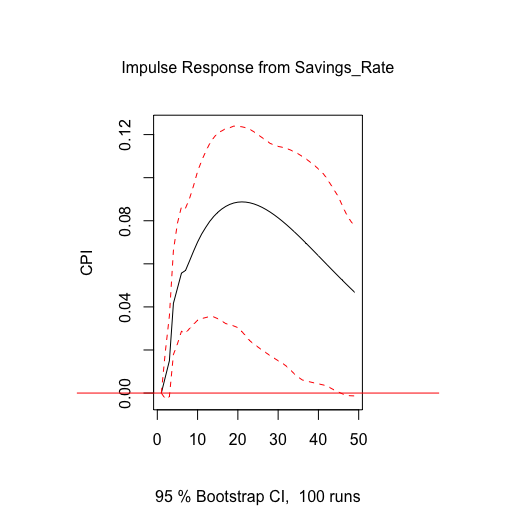


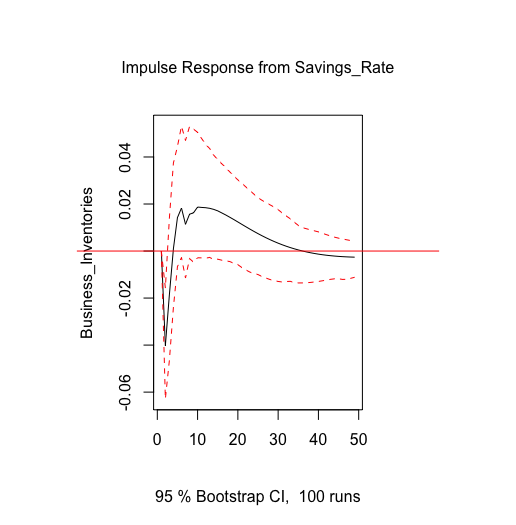
Below is a Bivariate Granger Causal graph of the 4 variables:

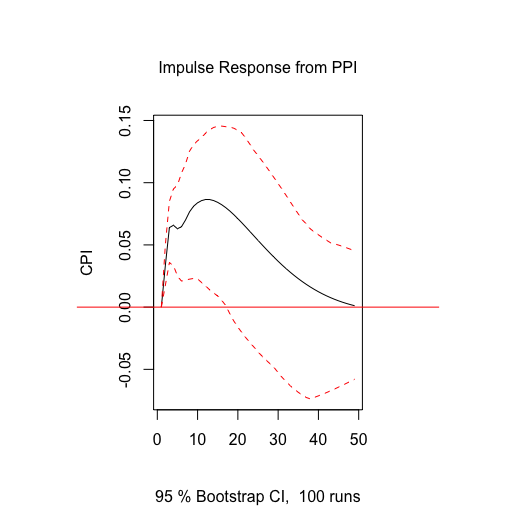


Already, the causal network helps explain a few key details about the data. First, savings rate can influence both the demand side and the supply side of the economy. Second, inflation for producers means some feedback into consumers. These are good sanity checks to make sure we’re considering the right variables.

Getting a little deeper, let’s fit a generic non-parametric VAR model and consider some sensitivities to non-causal & randomly induced shocks:







Plots show impulse-responces on on a fit VAR model.

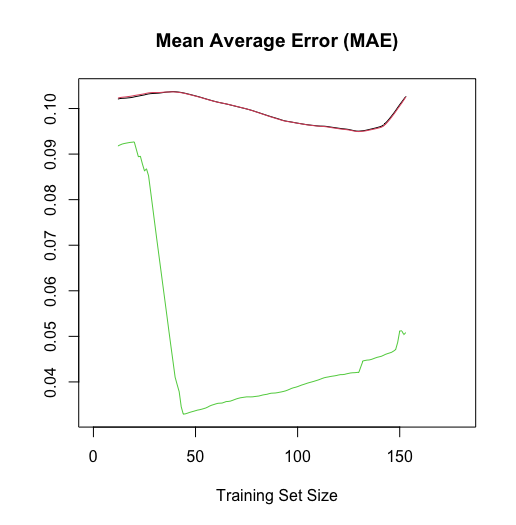
On a tangential note, the first plot confirms the infamous *Permanent Income Hypothesis*, conceptualized by Milton Friedman in 1957: An unexpected increase in people’s savings today, say because of a decision by the government to hand out stimulus checks, leads to those excess savings being pumped into the economy in the following periods leading to demand pull inflation. It also causes the stock of business inventories to compress until they normalize to new levels again.

*We cannot persuade people to save more tomorrow, even if we increase their disposable income, if no explicit reason compels them to do so*. *Future expectations of consumers play a crucial role in policy judgements.*

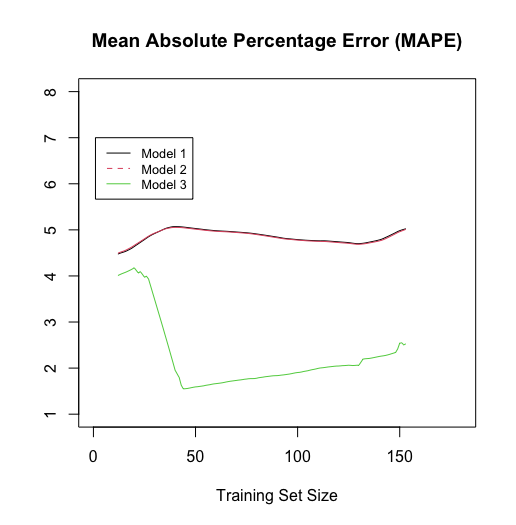
Similarly, as producers experience higher costs of productions, evident through increases in PPI, they pass it onto the consumers leading to supply driven inflation as shown in the last plot.

At least directionally, the VAR model confirms some economic theories. Unfortunately, the 21st century requires accuracy not direction. Even if we’re confident about the qualitative takeaways, econometricians have the task of fine tuning decisions and policies. Why does the Fed increase rates by 75 bps every quarter and not by 65 bps?

How can we make precise and accurate measurements using multiple models applied to the same data? Fitting 3 different forecasting models to our inflation data, below are a few metrics that measure the accuracy of out-of-sample forecasts. It’s clear that we cannot effectively differentiate the performance between 2 out of the 3 models.







**Model Ensembling**

The need of forecast ensembles is born out of the fact that since our traditional metrics for accuracy do not offer enough separation of performance, we can’t ever, with concrete confidence, say that one type of model can capture the true and causal dynamics between a set of time series. So if no model seems to come out on top, can we weigh the outputs of all the models in some fashion for a combined ensemble?

The task then becomes to come up with the best way to weight the forecasts which again requires measuring the difference in performance in some form.

Back to square one.

**Here is where we return to the idea of conceptualizing a new metric.**

The problem of ensemble modelling has inspired many solutions in recent literature and all the solutions have had one trait in particular — they draw inspiration from the simple idea that when a problem is looked at from a different lens, we may learn more about the problem and adapt our ways to accommodate the added intricacies.

In the same spirit, I would like to propose a framework, albeit nascent and very much open to correction, of ensemble modeling informed by information theory. I will spare the reader the mathematical framework and try to form explanations solely out of pragmatic intuition.

*Intimidating greek symbols have already scared away many minds from studying new disciplines and I do not intend to contribute to the loss. I would rather that* ***ideas*** *spread across readers and not* ***jargon****. For readers who would like an engineering layout of the framework, Fazlollah M. Reza’s “Information Theory” is an excellent source to get started and draw intuitive parallels to problems in econometrics.*

**Information Theory**

Classical forecasting methodologies include ARIMAX, Exponential Smoothing, Polynomial Regressions, Harmonic Regressions, State Space Models, and each of these require an assumption on the structure of the time series itself. In *information theory* terms —

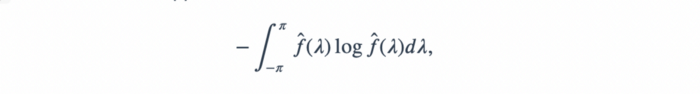
A. Each of these methodologies observe a *source* transmitting *information.*

B*.* Having understood the nature of the information, they try to predict future signals that can come from the source.

C. If these methodologies understand the dynamics of the source perfectly, they will be able to forecast it better in the future with no information going to waste other than some stochastic *noise* in the transmission.

If we can quantify this measure of information, or its lack of, perhaps we can start building quantitative ensemble schemes to be optimized.

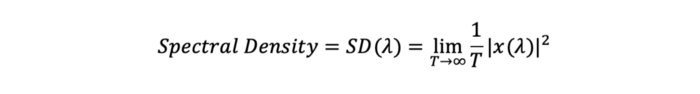
Let’s first define this measure of information:

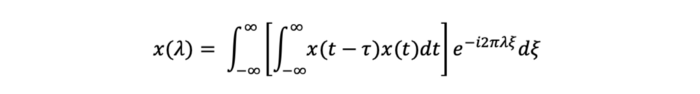


Definition of Shanon Entropy

where 𝑓̂(𝜆) is an estimate of the *spectral density* of the data. This metric might look complicated but is a natural ouput of combinatorial logic expertly highlighted by Fazlollah M. Reza.

The spectral density is a metric defined in the frequency domain and describes how densely a signal, carrying information, might be distributed over different frequencies:

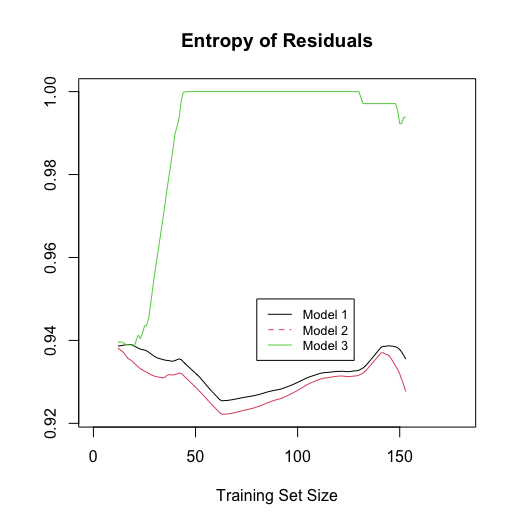




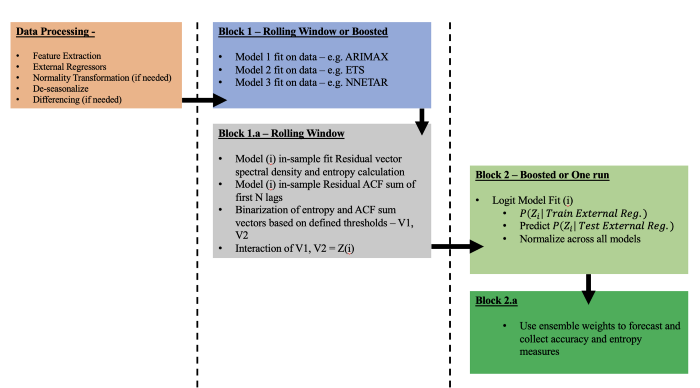
Just like metrics that measure distance between two points, entropy also carries the properties that make it a useful metric to consider. For instance:

1. **It’s non-parametric.** Entropy is bounded by [0,1]. The closer the data is to being white noise, aka having no discernible information, the closer the entropy value will be to 1. If we’re good at forecasting the information, the residuals should look closer to white noise.
2. **It does not change the topology**. A mere Fourier transform is needed which is an injective transformation (i.e. it just reshapes the information in the data but does not bend or twist it).

This only leaves the critical property of *offering separation* in performance between models to be tested.



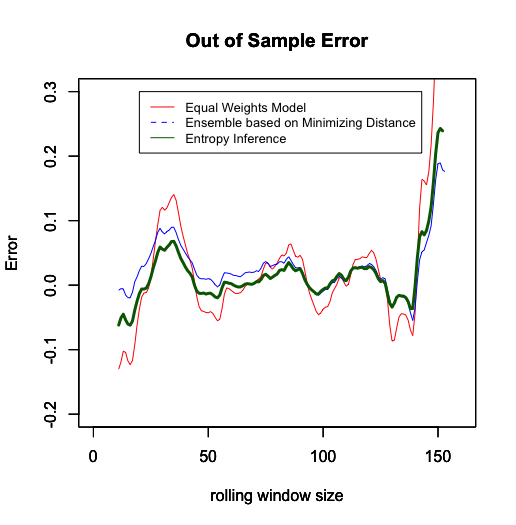
Voila! A difference in performance between all 3 models. Below is a rough scheme that introduces entropy based inference to estimate ensemble weights between the 3 models.

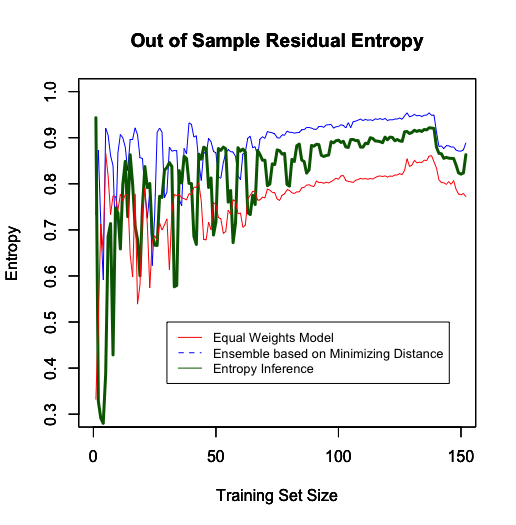


It’s important to note that the above is an *inference* scheme and not an *optimizing* scheme. The user specifies what entropy thresholds the model weights should satisfy while an Optimization ensemble would find weights that maximize the entropy of residuals; this would require an objective function with suitable constraints to ensure solutions.

*Entropy optimization is a natural next step but out of the scope of this article.*

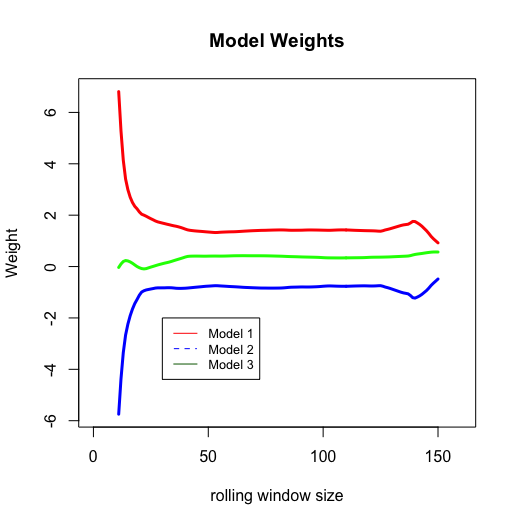
The intuition of this scheme is as follows: we’re estimating the probability with which we can we say that model X is a good representation of our source that emits information. The measure of that *information* is *entropy*. Putting our scheme to work, below are some results from the inflation data:



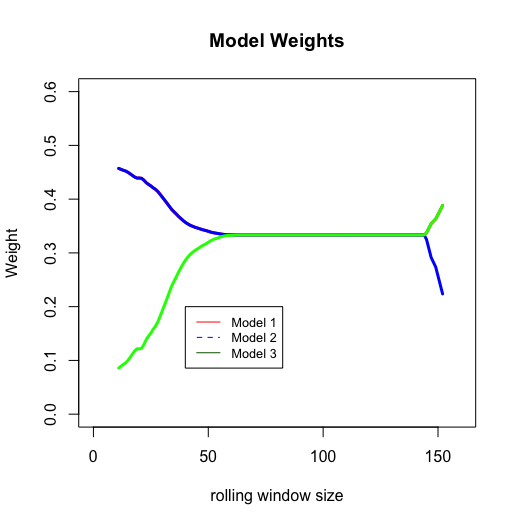


For the entropy inference ensemble, I set the minimum entropy threshold to 0.75 and interestingly, the out-of-sample entropy of the forecasted residuals is much higher and the accuracy performance is at par with the distance based ensemble. However, its entropy is still below a distance based ensemble meaning there is still information in the residuals that our ensemble has not parsed out. There are a few reasons why our new ensemble comes in inferior:

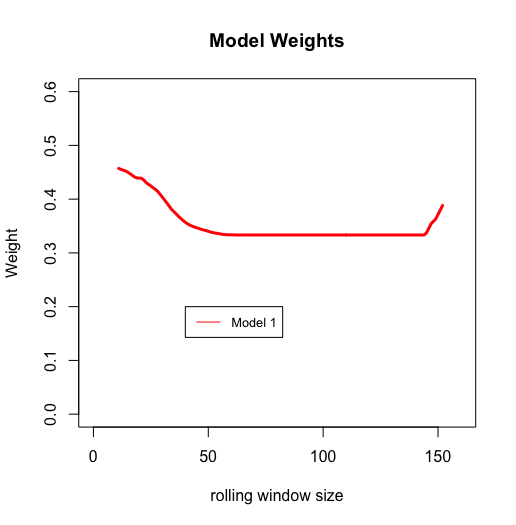
1. **Lower entropy threshold for training**. Since this an inference based model, it would be no surprise that different combinations of training thresholds could produce a higher out-of-sample residual’s entropy.
2. **The distance based ensemble is unbounded** **while the entropy ensemble is bounded.** The ensemble weights that the distance model assigns can, for this project, become negative or blow up and might lead to overfitting.
3. **Not enough unique models considered to ensemble**. This exercise looked at only 3 models to ensemble together; systemic entropy related inferiority of these models will also add inferiority to the ensemble. Fortunately, the ensemble scheme can be extended to N models.



Distance based ensemble

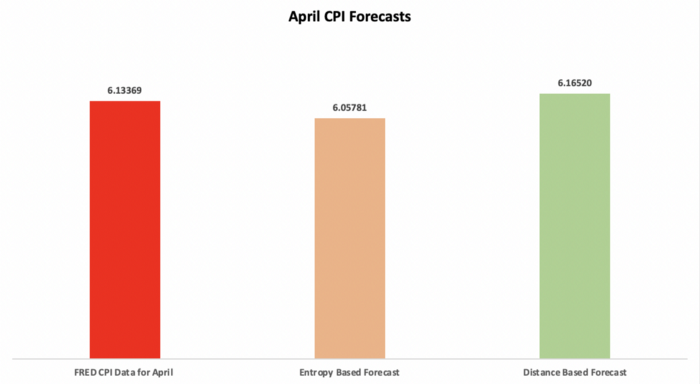


Entropy based ensemble — Model 1 shares equal weights as model 2 and later equal weights as model 3. Consequently it’s weights get masked in the plot



Weights for Model 1 in entropy based ensemble

Below are inflation forecasts the 2 ensembles put out in relation to the St. Louis Federal Reserve using their own data.



Overall, the community’s need to step back and look at forecasting problems under newer lenses can offer up many advancements to both how we model the solution and also how efficient our current solutions are. Entropy is one such approach and still has ways to go.

I hope the readers are able to draw parallels within the problems they’re solving and assess if a change in topology or a new metric like entropy can help them get closer to the solution.

*All images by author unless otherwise noted.*

*Vedant Bedi is an Analyst at Mastercard working on the NAM portfolio development team. He holds a Bachelor’s degree in Mathematics and Economics from NYU and holds an avid interest in data science, econometrics and its many applications in finance.*

*Vedant is also an inducted member of Phi Beta Kappa (NYC chapter) — the oldest academic honors society in the United States.*