Interpretable vs Black-box AI in Action

Jin-Chuan Duan*

(January 21, 2025)

Abstract: Combinatorial optimization enables machine-learning stable parsimonious conventional models that are naturally interpretable. This interpretable AI approach expands the realm of possibilities in generating better performing models to meet the needs of managerial usage and policy analysis in those fields bound by traditions and/or regulations. We demonstrate the working of interpretable models on two datasets (i.e., hedonic regression on house prices and vector autoregression on seven macroeconomic time series). Although interpretability distinguishes our approach, its performance can still be comparable to, if not better than, black-box AI models on the same data. We benchmark the performance of the hedonic house pricing model against a random-forest regression and two feed-forward neural networks.

***The author is the Chairman of ADBIZA and Criat. He is a professor emeritus of the National University of Singapore and currently serves as an adjunct chair professor in the College of Global Banking and Finance, National Chengchi University. Email: bizdjc@nus.edu.sg**

ADBIZA (https://adbiza.com) specializes in advanced business analytics, providing solutions that are premised on interpretable AI. Product/service offerings include topic-focused media sentiment analysis, bespoke real estate marking-to-market models, digital patients for training medical professionals, and general advisory services on business analytics.

I. Interpretable AI

Artificial Intelligence (AI) has been a branch of scientific undertaking for many decades. The underlying technology is centered around building neural networks to mimic how human brains function. It is fair to say that until the emergence of Google's BERT model for natural language processing in 2018, AI's power and potential were not quite appreciated even within the scientific community. Open AI's release of ChatGPT in 2022 has wowed the general populace and ushered in a new era. The impact has gone way beyond the scientific community, and AI is now a household term.

While our imagination is flying high and the debate on what GenAI (Generative AI) or AI in general can deliver continues, the current AI approach in managerial applications and policy analyses has already hit a wall, so to speak. It is primarily due to the black-box nature of neural network models. Interpretability reigns supreme for a model to be deployed for managerial decision making. Expecting models to be interpretable by decision makers is hardly surprising and quite understandable because they are held accountable to various stakeholders for the outcomes of a decision regardless of how it has been reached.[1] As a safeguard, financial regulations typically dictate that AI models in use must be interpretable. Beyond regulations, demanding interpretability has been ingrained in our tradition-bound systems in health care, finance and other highly regulated industries. For example, a medical prescription prompted by a black-box model that goes wrong will have little chance of surviving the legal scrutiny of the doctor responsible in the event of litigation.

In the AI domain, some experts differentiate explainability and interpretability, and others use them interchangeably. Instead of dwelling on the subtleties, it suffices to say that achieving interpretability is a taller order than giving a model some explainability.[2] Beyond any abstract description, what is the interpretability of a model anyway? In this author's view, we shall consider interpretability either as theory or common sense based.

For theory-based interpretability, classical mechanics is a good example for illustrating the point. An object being exerted with an initial force is expected to travel in a parabolic curve and due to gravity land in a predictable distance. This is well understood by and interpretable to people with exposure to elementary physics. Added real-life complexities such as friction may cause the prediction to be off for objects of different sizes and shapes, and thus the model requires further tuning. If the prediction has undergone a data-based adjustment on the same physics framework using data collected through experiments, the refined model will remain

[1] The European Union's Artificial Intelligence Act (Article 13), for example, states that "High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations …".

[2] Explainable AI provides the user with how and why the AI model reaches a prediction. In contrast, interpretable AI offers a transparent decision process, and the user can appreciate with his/her background knowledge how a prediction is made and whether it is sensible. Random forest vis-à-vis classical decision tree offers a case in point. The former is explainable but not interpretable, because one can explain to the user how a random forest works, but its decision recommendation cannot be interpreted due to randomization. The latter is, on the other hand, both explainable and interpretable.

interpretable to those who have knowledge of classical mechanics. Were the model built entirely on data with a neural network, it would be viewed as a black box and becomes uninterpretable.

As to the common-sense interpretability, one can appreciate it using a hedonic regression model. Economists often use such a model to describe how the many attributes of a composite good together determine the total price. The typical use of hedonic regressions on real estate prices similarly follows this line of thinking to view a housing unit as a composite good whose location, size, condition, view, and many other features together command its transaction price. With a hedonic pricing model built on prices on many transacted properties, one can mark-to-market those many more off-market properties that have not been changed-hands lately. To economists and lay people alike, a hedonic regression model becomes interpretable when the appearance of a feature in the regression jells well with their intuition. The magnitude and the sign of its regression coefficient reflect the extent and direction of the response to a change in the value of this feature. Moreover, the magnitudes of the regression coefficients convey a sense of market weightings in the overall importance for pricing.

Interpretability depends on the user's knowledge background; for example, the classical mechanics model mentioned earlier makes little sense to people without some level of physics knowledge. It also changes over time because something completely foreign to people today may become common sense years down the road. In short, interpretability is not absolute, which evolves over time and depends on the user's educational background, professional training and life experience. Industry standards of the day and the stage of technological development also play a role. It is therefore conceivable that we may one day reach the point at which neural networks are no longer viewed as black boxes.

Building AI or machine-learning models for managerial usage and policy analysis will likely be more productive if we set out to enhance the conventional interpretable models with modern analytical tools. Instead of upsetting the conventional understanding, why not take advantage of familiarity to ensure a model's interpretability to the target users? In this paper, we will showcase two concrete examples in economics to demonstrate how this can be accomplished, and such interpretable AI models can be competitive in performance with black-box approaches.

II. Interpolation vs extrapolation usage

Interpolation means to generate a predicted value where the input values corresponding to the features of a model all fall within the range of data in the sample used to train the model. Neural networks, for example, have shown their mighty interpolation power in natural language and imagine processing, among others. Due to their mathematical structure, however, they will behave unpredictably in extrapolation; that is, a prediction faces an input value for some feature that goes outside the range of the data used in training the model.

It is well known in statistical literature that the prediction made by all nonparametric curve fitting techniques deteriorates when an input value gets closer from within to the data boundary. For predictions made with input values beyond the boundary, they become unreliable, and one should therefore refrain from applying them in the extrapolation context. This is hardly surprising because flexibility of a modeling technique such as neural networks inevitably comes with a price, and the lack of guidance from theory or intuition renders its inability to predict something based on a pattern beyond what has already been revealed in the data.

Interpretability of a model is a source of persuasive power which enables the user to feel comfortable in extrapolation with the model. This is essential in managerial applications, policy analyses and other areas where conventional expert judgements play a critical role. Take again the hedonic house pricing model as an example. When the model indicates that the size of a property determines the overall price with a particular positive multiplying factor, i.e., a positive regression coefficient, the user will deem it in line with his/her intuition and feel comfortable to extrapolate with it on a property that is way larger than any property in the training data. If the model also suggests that the property price's response magnitude to the size depends on which district is located, i.e., an interaction term of the size and the district indicator, one will find such a model implication to be interpretable. After all, location, location, location is the mantra of real estate.

Interpretability can lead to extrapolation with a sense of ease, but one needs to be mindful of its limitations. The specific usage context should also be factored in. Again, we can use the interpretable hedonic pricing model to elaborate. Suppose the model has included among other features the year in which a transaction took place to capture the real estate boom-and-bust cycle, and the model has been trained on the transaction prices up to the end of 2024. This model would make sense to specialists and lay people, and it would thus be interpretable. However, this model could not be used for extrapolation into 2025 and beyond, but that may be the period for which the model has been intended. If such extrapolation for the off-market properties in 2025 is indeed a model's intended purpose, one must avoid including a feature like "year". A simple solution is to substitute it with a real estate price index as an alternative way to reflect the boom-and-bust cycle. In short, interpretability is necessary for the extrapolation purpose but by no means sufficient. Other model design elements may also be important.

III. Machine learning interpretable models

The two machine-learned interpretable models to be shown in this section will serve as our examples of interpretable AI in action. Their interpretability is naturally endowed because they are conventional models for tabular data familiar to the users in their respective domains. Our machine learning via combinatorial optimization simply facilitates the building of a better performing model on the familiar conceptual framework. The common challenge is to find the best subset of variables/features out of a very large set of possible combinations so that the conventional model can perform well and stably in and out of the training sample. In short, this line of machine learning highly depends on a flexible combinatorial optimization technique made possible by sequential Monte Carlo sampling.

Ideally, the chosen best subset of features and the directions of response for the key features of the model are theoretically justified and/or intuitively sensible. Moreover, the final model is parsimonious, meaning that the number of chosen features is manageably small so that the model can stay operationally nimble. We will also show through the first example that the interpretable AI model is competitive with two black-box methods if not better.

A. A stable optimized hedonic house pricing model

We use the dataset on house prices in Ames, Iowa as described in De Cock (2011) to showcase the working of an interpretable AI model. In this demonstration, we also compare its performance with two commonly used black-box methods (random forest and neural network) in machine learning. Our demonstration further refines what has been studied in Duan, *et al* (2022), which deployed the same dataset. We utilize the stable combinatorially-optimized feature selector (SCOFS) of Duan (2024) as well as the stable optimized decision tree (SODT) of Duan and Li (2024). The former is an improved algorithm over its earlier version used in Duan, *et al* (2022) by engaging a cross-validated target function to achieve a more stable out-of-the-sample performance. The latter is a method of finding a stable decision tree where a cross-validated performance target is optimized over all possible tree configurations.[3]

There are 80 variables in the Ames housing dataset where one is the transaction price and the remaining 79 are the features describing the property. Readers can find a description of these 79 features at http://jse.amstat.org/v19n3/decock/DataDocumentation.txt. Many of these 79 features are categorical data, for example, 28 neighborhoods within Ames city limits. Some features are ordinal such as slope of property (gentle, moderate and severe) whereas some are numerical, for example, gross living area. The dataset covers the period from 2006 to 2010 with 2,930 transactions, and among them 2,269 have no missing values. We will use these 2,269 data points to conduct the comparison analysis.[4]

We adopt a two-stage model building process. The aim of the first stage is to simplify many categorical and ordinal features by consolidating them into some composite group dummy variables. The fact that there are 28 neighborhoods as mentioned earlier explains why this approach is a good idea. It is hard to imagine that all 28 neighborhoods are distinct for the pricing purpose. Neither are they homogeneous to the point that buyers would view them all as comparable. Hence, the possible combinations to consider are numerous; for example, two neighborhoods are equally preferrable and distinct from all others. The total number of combinations to consider for, say, two and eight neighborhoods alone are 378 and over 3 million, respectively. The number of possibilities rapidly rises when more neighborhoods are

[3] We deploy *i*Select, a proprietary software of ADBIZA, to run SCOFS and SODT. Both rely on sequential Monte Carlo combinatorial optimization devised in the respective papers.

[4] Missing values present no problem to our building of interpretable hedonic pricing model because *i*Select handles them by directly treating missing values as a distinct class for categorical, ordinal and numerical variables. For the Ames dataset, the hedonic pricing models with and without missing value cases perform comparably. In this demonstration, we do not report the result because the comparison models (random forest and neural network) would need to first engage some data imputation, which would inevitably complicate the comparison.

grouped together. It is a tall order to exhaust all possibilities even with a powerful computer unless deploying an intelligent algorithm.

This is where we engage SODT to come up with the ten composite groups. The choice of ten groups is for ease of interpretation, and the software can produce the optimal number of groups through its optimal decision-tree regression function. Out of the resulting ten composite groups, the top-price category turns out to have the average house price at $406,894 whereas the bottom-price group faces $108,675 on average. The rules that define the top-price group are three criteria – (1) the overall quality is rated 8 or above in the scale of 1 to 10, (2) the basement square footage is larger than 1,718, and (3) the location is in one of the select five neighborhoods out of 28 in total. On the other hand, the bottom-price group consists of properties with the quality rated 4 or below plus two other conditions (size of the basement and the type of dwelling). This sort of division into composite groups gives rise to an easily interpretable model typical of a commonly seen scoring system in finance.

The first-stage division of the properties yields an R^2 of 72.96% on the testing data. The model may even suffice for some managerial usage where further refinement is deemed unnecessary. If the usage context turns out to be more demanding of pricing precision, we can move on to the next stage to build a more powerful hedonic pricing model. In that case, we add these ten group indicators as additional features for the second-stage model construction. As stated earlier, the choice of ten composite groups is for ease of interpretation and using ten, eleven or a larger number of groups does not materially change the performance of the final hedonic regression model, but using a too small number of groups may adversely affect its performance.

Deploying categorical variables directly in a model makes little sense because the different categories of such a variable are distinct and without any meaningful ordering relationships. Per usual, we create binary indicators out of each categorical variable; for example, the 28 neighborhoods are converted into 28 binary indicators to record whether a property is in a particular neighborhood. After completing this type of conversion, the original 79 features have been turned into 106 variables.[5] Further adding the ten composite group indicators gives rise to 116 feature variables. In economics or social science, interaction terms are often considered for their natural interpretability. We therefore add to the 116 features in the first order with those non-redundant interaction terms in the second order to yield 6,094 potential features in total.[6]

A hedonic regression being directly estimated on so many potential features is not only technically infeasible but would also become intuitively silly as a model. Selecting a good performing manageable subset of features is obviously the essence of building a practical and interpretable hedonic pricing model. We use a three-fold cross-validated sum of squared residuals as the minimization target function when applying SCOFS on the training sample of

[5] A binary indicator for a category will not be created when it constitutes less than 5% of the data instances.

[6] Redundancy occurs, for example, squaring a binary indicator yields the same binary variable.

1,701 data instances, which is a randomly sampled dataset (three quarters of the whole dataset). The remaining 568 data points are saved for the testing purpose.

The final hedonic pricing model selects 25 features out of 6,094 potential ones. By design, each of the selected features is highly statistically significant. The model's good performance is visually reflected in Figure 1 where the results for both the training and testing data are presented. The horizontal axis represents the model's predicted price whereas the vertical axis is the corresponding transaction price of a house. The plotted points scatter around the 45-degree line for both samples, which indicates the model's good and stable performance.

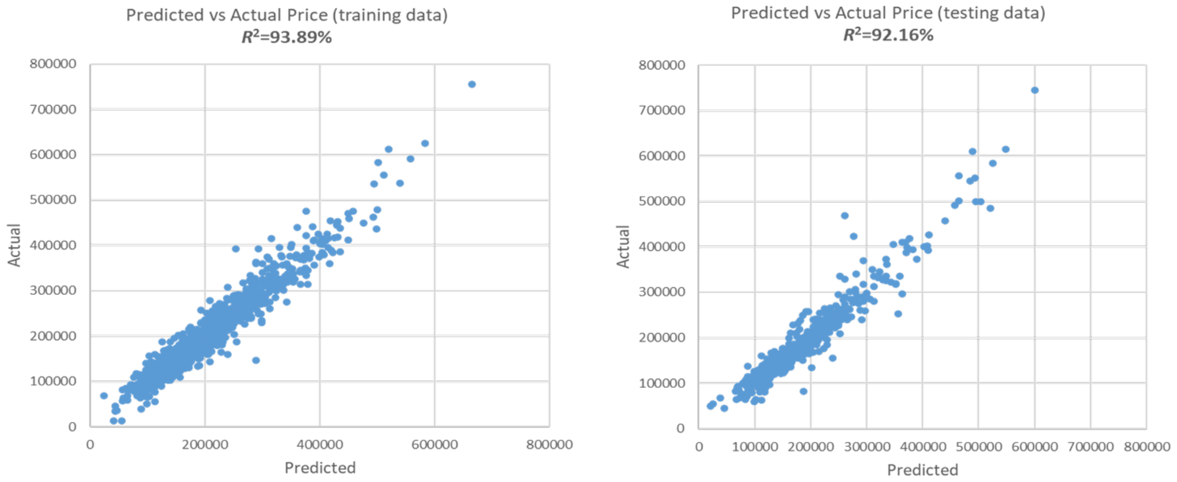

Figure 1: The predicted values by the hedonic pricing model vs the actual transaction prices for the training and testing datasets on a sample of Ames, Iowa house prices.

It is interesting to note that this chosen hedonic regression model does not need an intercept term. Among the selected variables, many are interaction terms. After algebraically rearranging those interaction terms by anchoring on a set of 18 features, a highly interpretable version of the model emerges with variable coefficients on the anchoring features as shown in Table 1. Corresponding to the first anchoring feature, “Gross Living Area”, for example, its variable coefficient has a positive constant of 47.84 which indicates that the house price increases with “Gross Living Area”, and every 1,000 square feet commands additional $47,840. To appreciate the results, we note that the average housing unit size in this sample is 1,505 square feet. In addition to the constant, the second term “Year Built_demeaned” in this variable coefficient suggests that a house one year newer will fetch an additional $173.5 per 1,000 square feet. The second anchoring feature appearing in Table 1 is “Bsmt Qual_demeaned”, whose negative coefficient implies a lower price if the unit's basement has above average quality, a rather counterintuitive result that calls for an explanation. It turns out that “Bsmt Qual” and “Gross Living Area” are negatively correlated at -0.35, and their interaction term thus serves to offset somewhat the positive price effect of “Growth Living Area”. The rearrangement work continues until exhausting all 25 selected features.[7]

[7] The sequence of anchoring features is part of the subjective preference of the user; for example, one may find more appealing to anchor the first feature on “Lot Area” followed by “Gross Living Area” and so on. All

Table 1: An interpretable version of the hedonic pricing model by casting interaction terms as variable regression coefficients

Anchoring Variable	Variable Coefficient
Gross Living Area	47.84 + 0.1735*Year Built_demeaned - 2.4*Bsmt Qual_demeaned - 1.8*Bsmt Exposure_demeaned - 4.98*Kitchen Qual_demeaned
Lot Area	0.9428 + 0.6331*Half Bath_demeaned + 0.3803*Garage Cars_demeaned
1st Floor SF	51.62 + 13.09*Overall Qual_demeaned - 0.02516*1st Flr SF_demeaned - 158.59*Group8
Garage Area	19.39 + 12.32*Full Bath_demeaned
Garage Cars	1317.29 + 78.47*Screen Porch_demeaned
Overall Cond	6934.65
Basement Finished SF Type 1	21.11*Condition 1_Norm
Masonry Veneer Area	39.55*Foundation_PConc
Full Bath	- 7934.86*Sale Condition_Abnorml
Fireplaces	28216.03*Group10
……	

This interpretable hedonic pricing model yields a median pricing error rate[8] of 6.23% for training data and 6.85% for the testing sample, respectively. Using the typical statistics for measuring the performance of a regression, this model has delivered an R^2 of 93.89% and 92.16% for the training and testing data as reported in Table 2, indicating that the model's out-of-the-sample performance is only marginally off when compared to the training-data R^2.

As a quick benchmark, we present the selection result obtained on the same dataset with the popular Lasso regression of Tibshirani (1996) and re-estimate the selected model with a post-selection OLS regression to remove any downward bias (in the magnitude of regression coefficients) caused by the Lasso penalty. This post-selection step is commonly performed these days. For comparability, we also use the three-fold cross validation for Lasso. Table 2 reveals that Lasso has grossly over-selected the features, yielding 59 variables in contrast with 25 by SCOFS, and many of the regression coefficients, 26 out of 59 selected features, are statistically insignificant. Evidently, Lasso is not a reliable way to obtain a parsimonious

interpretable versions share the same model but with different algebraic rearrangements to cater to user preferences.

[8] The median pricing error rate is computed as the sample median of all individual property's error rates defined by |transaction price – predicted price|/transaction price.

regression model.[9] Even with an over-selected model, its training and test data performances reported in Table 2 are only comparable to that of the hedonic pricing model produced by SCOFS.

Table 2: Compare the performance of the interpretable hedonic regressions obtained by SCOFS and Lasso, and the two black-box approaches

	Hedonic Regression		**Neural Network**		**Random Forest**
	106 converted features + 10 group indicators		106 converted features		
	1st + 2nd order terms		1 hidden layer (64 nodes)	2 hidden layers (64+32 nodes)	# trees = 180 max depth = 12
	Total # features = 6094				
	SCOFS	Lasso + OLS			
# features selected	25	59	-	-	-
# significant features (5%)	25	33	-	-	-
R^2 (training data)	93.89%	93.81%	89.87%	98.74%	98.21%
R^2 (testing data)	92.16%	92.05%	86.13%	91.29%	91.23%

We now compare the hedonic pricing model's performance with two black-box approaches where one is based on the random-forest regression and the other deploys two feed-forward neural networks. In the former case, we engage the greedy search and the three-fold cross validation to determine the final tree for each of 180 randomized samples with the maximum depth of 12. For the latter, the neural network models are constructed according to the standard practice of using a validation sample (one-third of the training data) to control overfitting. We consider two feed-forward networks – (1) one hidden layer with 64 nodes and (2) two hidden layers with 64 and 32 nodes in the first and second layers. Since these two modeling approaches by design handle nonlinear relationships, we leave out the ten composite group indicators and the interaction terms to use the 106 converted input features as described earlier, which in principle does not compromise the information content in the feature space.

Table 2 shows that either one of the two alternative approaches yields an overfitting result with their training-data performance clearly better than that on the testing data. In fact, the neural network's performance on the testing data is worse than that of the hedonic pricing model chosen by SCOFS, yielding an R^2 of 86.13% for the model with one hidden layer and 91.29% with two hidden layers. Compared to their training-data counterparts (89.87% and 98.74%), the neural network approach vis-à-vis the hedonic regression shows a more

[9] Lasso's tendency to grossly over-select features has been documented in, for example, Duan (2024) through a simulation study, and the mathematical reason for such a tendency has been provided.

pronounced overfitting. The random forest model fares no better either with its training- and testing-data R^2 at 98.21% and 91.23%, respectively.

B. A stable parsimonious vector autoregression

A vector autoregression (VAR) of moderate dimension and/or long lag structure contains many parameters. A priori, most of these parameters are likely inconsequential and statistically insignificant if estimation is even feasible. Placing many zeros on a VAR either by theory and/or intuition has its limitations and practical difficulty. Here, we show how deploying the stable combinatorially-optimized feature selector (SCOFS) of Duan (2024) can effectively simplify the model structure and thus identify a stable parsimonious VAR that is interpretable to typical users in the concerned fields. In essence, we conduct a combinatorial optimization using the cross-validated likelihood function corresponding to the seemingly unrelated regression[10] formulation of the VAR to find the best performing stable parsimonious VAR.

We now develop a parsimonious VAR on the quarterly time series of seven macroeconomic variables studied in Smets and Wouters (2007) but expanded to the period from 1947:Q2 to 2020:Q1. These seven variables are (1) Consumption Growth, (2) Investment Growth, (3) Output Growth, (4) Hours Worked, (5) Inflation, (6) Wage Growth, and (7) Fed Funds Rate. For details on their definitions, one can read Smets and Wouters (2007).

The data sample spanning over 73 years has likely experienced some structural breaks. It is therefore desirable to introduce some time-period indicators to allow the algorithm to discover a structure that better fits the data. In line with this consideration, we introduce three period indicators: (1) $P_t^{(1)}$ for Golden Age (1947:Q2-1979:Q4), (2) $P_t^{(2)}$ for Post-Golden Age (1980:Q1-2008:Q2), and (3) $P_t^{(3)}$ for Post-Global Financial Crisis (2008:Q3-2020:Q1), and each of which takes the value of 1 if a quarter falls in that period and 0 otherwise. We interact these three period indicators with the seven macroeconomic variables lagged up to four quarters. Together, the VAR model faces 812 potential coefficients; that is, 4×(7 intercepts + 4×49 lagged coefficients). In addition, the VAR model has 28 residual variance-covariance terms to be estimated. If one is in doubt of the above time-period division, more sets of period indicators can be added to the potential model, and SCOFS will then identify the best performing parsimonious VAR in the expanded space.

To face up to the large VAR model of this type, researchers typically adopt a Bayesian approach with the Minnesota-type prior of Sims and Zha (1998). Putting aside whether such a prior is justifiable, having a prior is essential to the estimation of such an overparameterized VAR model in the first place. Applying combinatorial optimization to simplify the VAR structure allows us to return to the frequentist view or remain as a Bayesian. For the former, SCOFS simply targets the likelihood function of the VAR whereas for the latter, the target becomes the likelihood function times the prior, a Minnesota-type or not.

[10] Seemingly unrelated regression is a well-known econometric technique. Unfamiliar readers are referred to any standard econometrics book.

For ease of presenting the results, we introduce some mathematical notations. The VAR model considered is

$$\boldsymbol{X}_t = \boldsymbol{A}_t + \sum_{j=1}^{4} \boldsymbol{B}_{j,t}\boldsymbol{X}_{t-j} + \boldsymbol{\epsilon_t}$$

where $\boldsymbol{X}_t$ is the vector of the seven macroeconomic variables (in their stated order) at quarter t. $\boldsymbol{A}_t$ is the vector of seven corresponding intercepts, and $\boldsymbol{B}_{j,t}$ is the 7×7 coefficient matrix for lag j. Their subscript t is meant to say that these coefficients may depend on the time for which they are intended, and it is of course due to our allowance for structural breaks with the periods defined earlier. Finally, $\boldsymbol{\epsilon_t}$ is the seven-dimensional residual term.

$\boldsymbol{B}_{1,t}$							
Consumption Growth	0	0	0	0	-0.3380	0	0
Investment Growth	0.5598	0.3095	0	0	0	0	0
Output Growth	0.4156	0	0	0	0	0	0
Hours Worked	0.2843	0	0.1940	0.9759	0	0	0
Inflation	0	0	0	0	0.9301 -0.4521×$P_t^{(1)}$	0	0.4467×$P_t^{(1)}$
Wage Growth	0	0	0	0	0	-0.4685×$P_t^{(3)}$	0
Fed Funds Rate	0.0517	0.0465×$P_t^{(2)}$	0	0	0	0	1.0045 -0.2129×$P_t^{(2)}$

$\boldsymbol{B}_{2,t}$							
Consumption Growth	0	0	0	0	-0.8744×$P_t^{(3)}$	0	0
Investment Growth	0	0	0	0	0	0	0
Output Growth	0	0	0	0	0	0	0
Hours Worked	0	0	0	0	0	0	0
Inflation	0	0	0	0	0	0	0
Wage Growth	0	0	0	0	0	0	0
Fed Funds Rate	0	0	0	0	0.3606×$P_t^{(2)}$	0	0

With a five-fold cross-validated likelihood function as the target, SCOFS selects a stable parsimonious VAR that comprises 21 non-zero coefficients, a sharp reduction from 812 potential ones. We summarize the selected VAR model as follows. First, there are four selected non-zero intercepts, i.e., $\boldsymbol{A} = [0.7724, 0, 0.2377, -0.2584, 0, 0.4422, 0]'$, and there is no structural break in these intercepts.[11]

As to the autoregression coefficients, there are no selected terms at lag 3 or higher. The two tables provide a consolidated view of the 17 selected lag terms governing the structure of this parsimonious VAR. Although their standard errors are not presented to conserve space, all lag coefficients are highly significant.

These results offer simple interpretations and likely make a great deal of sense to economists. For example, the second row of $\boldsymbol{B}_{1,t}$ implies that either a positive consumption growth or investment growth in the previous quarter will increase the investment growth of the current quarter, reflective of their positive coefficients in (row 2, column 1) and (row 2, column 2). The seventh row suggests that the Fed Funds rate of the current quarter will positively respond to a rise in the consumption growth in the previous quarter. In addition, the Fed Funds rate will increase in response to a positive investment growth in the previous quarter if the current quarter falls in the Post-Golden Age. Furthermore, the Fed Funds rate of the current quarter will positively react to itself revealed in the previous quarter, and the magnitude of response will be dampened somewhat if the current quarter is in the Post-Golden Age. The impacts from the two quarters ago are, however, only limited to two macroeconomic variables (i.e., Consumption Growth and Fed Funds Rate) as shown in $\boldsymbol{B}_{2,t}$, and they are only relevant to specific periods.

To avoid over-displaying the information that is not critical to our appreciation of the parsimonious VAR, we have omitted the reporting of the estimated residual covariance matrix. But it suffices to say that all residuals are correlated to indicate contemporaneous relationships.

In summary, the above parsimonious model has a structure that characteristically differs from the typical Bayesian VAR, and it is far easier to interpret this model. Although a performance comparison on a testing data has not been conducted, it will be hardly surprising to find this parsimonious VAR model to outperform its competitor, because this model has been constructed with a cross-validated target likelihood function and it is free from the bias introduced by the Bayesian prior.

IV. Concluding remarks

Black-box AI models such as neural networks have already proven their power in natural language and image processing and some other settings. However, the lack of interpretability expected for managerial and/or policy usage has obviously limited their applicability and calls for a different approach to unleashing computing power for those applications.

In this paper, we have demonstrated interpretable AI by two examples of machine learning conventional interpretable models. Sequential Monte Carlo combinatorial optimization

[11] "Hours Worked" has a negative intercept because this variable is a logarithmically transformed index value.

enables us to expand the realm of possibilities through finding a stable parsimonious representation out of many potential features. The resulting model retains the interpretability, but its performance has markedly improved to the point of being comparable to, if not better than, those black-box models such as neural network and random forest. Instead of forcing interpretability on black boxes, it seems to make more sense to utilize machine learning to improve conventional models when interpretability is utmost important.

In principle, the combinatorial optimization approach adopted for our two examples can work on all conventional models to enhance their performance in addressing the fast-increasing data footprints. The practical challenge rests with the computing power that the model developer needs to muster when a conventional model becomes increasingly complex. However, the algorithm for sequential Monte Carlo combinatorial optimization is fundamentally parallel which simply means a need for the model developers to network more multicore computers to complete increasingly more demanding tasks.

References

1. De Cock, D., 2011, Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, *Journal of Statistics Education*, 19(3):1-15. https://doi.org/10.1080/10691898.2011.11889627
2. Duan, J.-C, 2024, Stable Combinatorially-Optimized Features Selection via Sequential Monte Carlo, National University of Singapore working paper.
3. Duan, J.-C. and S. Li, 2024, Stable Optimized Decision Tree in a Random Forest, National University of Singapore working paper.
4. Duan, J.-C., S. Li, and Y. Xu, 2022, Sequential Monte Carlo Optimization and Statistical Inference, *Wiley Integrative Reviews: Computational Statistics*, e1598. https://doi.org/10.1002/wics.1598
5. Sims, C.A. and T. Zha, 1998, Bayesian Methods for Dynamic Multivariate Models, *International Economic Review*, 39(4):949-968.
6. Smets, F. and R. Wouters, 2007. Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach, *American Economic Review*, 97(3):586-606.
7. Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), 58(1):267-288.