IN-SAMPLE ANALYSIS OF BAYESIAN SYMBOLIC REGRESSION APPLIED TO CRUDE OIL PRICE

Krzysztof Drachal, Faculty of Economic Sciences, University of Warsaw, +48 22 55 49 111, kdrachal@wne.uw.edu.pl

Overview

This research analyses in-sample crude oil price forecasting with Bayesian Symbolic Regression (BSR). It is found that this new method (Jin et al., 2019) has a great potential for further studies and can serve as an interesting forecasting tool. This method except forecasting applications is able to deal with variable selection (feature selection) issues in econometric modelling. The in-sample analysis is performed, because the aim of this research is to primary focus on comparing the new econometric method with its more simple predecessor, i.e., symbolic regression with genetic programming (Koza, 1992). In other words, this research is focused on two aspects. The first one is to analyse the initial parameter selection of BSR and comparing various BSR models between themselves with respect to fitting the observed data. The out-of-sample forecasting performance analysis is postponed for further studies. In particular, the current research aim is to study whether there are any evidences that the new method, i.e., BSR, can have some modelling potential over the already used methods, and whether its initial features with its internal parameter specification can impact the results.

Methods

Monthly data between 1989 and 2021 are analysed (CBOE, 2021; EIA, 2021; FRED, 2021; MSCI, 2021; Stooq, 2021; The World Bank, 2021). In particular, the dependent variable is WTI spot price. Besides, for robustness check also Brent and Dubai prices are taken. The independent variables are world production of crude oil including lead condensate, OECD refined petroleum products consumption, U.S. ending stocks excluding Strategic Petroleum Reserves of crude oil and petroleum products, MSCI World stock market index for developed markets, VXO index, Kilian index of global economic activity (Kilian, 2009) and real narrow effective exchange rate for U.S. Additionally, Chinese stock market index is taken. This index is constructed by the suitable gluing of Hang Seng index (before 1991) and SSE Composite index (after 1991).

Independent variables are taken in their 1-st lags. Oil prices, stock market indices and exchange rate are taken in logarithmic 1^{st} differences. Production and consumption data and stocks data are taken in logarithmic 12^{th} differences due to seasonality patterns. The data obtained after such transformations are standardized (Coulombe et al., 2021). However, the obtained forecasts are analysed as raw time-series representing price levels, i.e., not as differences representing price changes values, i.e., after backward trasformations.

Between 1 and 9 components in BSR are tested (Jin et al., 2019). As well as, 7 various sets of functions for symbolic regression are tested (Nicolau and Agapitos, 2012). The forecasts from these models are compared between themselves with Model Confidence Set (MCS) testing procedure. Besides, Dynamic Model Averaging, Bayesian Model Averaging, Dynamic Model Selection and Dynamic Model Averaging methods are considered as competitive models (Raftery et al., 2010). These models are considered in various specifications. Moreover, LASSO and ridge regressions, both in conventional and Bayesian versions are considered, as well as, elastic net, least-angle regression and time-varying parameters regressions (Friedman et al., 2010). Besides, auto ARIMA (Hyndman and Khandakar, 2008), historical average and the naïve (no-change) forecasts are taken. The forecasts from the benchmark models are also analysed with MCS procedure. The forecasts obtained from superior models are next compared with MCS procedure again. The significance level of 5% is assumed.

The forecasts from several BSR models are also compared with forecasts from symbolic regression with genetic programming with the Diebold-Mariano test with 5% significance level (Diebold and Mariano, 1995).

The forecast accuracy is measured by Root Mean Square Error (RMSE).

Results

It is not easy to find a certain pattern between RMSE of BSR forecasts and the number of components or the function set used. In case of benchmark models all of them generate forecasts with smaller RMSE than the naïve method. However, only Bayesian ridge regression and least-angle regression generated forecasts with smaller RMSE than the auto ARIMA method. Interestingly, a certain BSR model and auto ARIMA model happenes to be the superior models due to MCS procedure. If Brent oil price is considered, then the superior models are also a certain BSR model and Dynamic Model Averaging. If Dubai oil price is considered, then the superior models are several BSR models, Dynamic Model Averaging and auto ARIMA.

Finally, in majority of cases, if the function set is fixed, then BSR generates significantly (according to the Diebold-Mariano test) more accurate forecasts than the corresponding symbolic regression with genetic programming.

Conclusions

The obtained results suggest that the newly proposed method, i.e., BSR, has a great potential for modelling crude oil spot price when there exists variable uncertainty. However, the suitable choice of internal parameters (i.e., the number of components and the function set) can have an important impact on in-sample fitting. As a result, for out-of-sample forecasting purposes, it seems that pre-selection of initial parameters should be carefully done. For example, over some in-sample (testing) period. This is worth to be analysed deeper in the further studies.

Secondly, BSR seems to generate significantly more accurate forecasts than symbolic regression with genetic programming. Therefore, the novelty of this method, indeed, improves symbolic regression modelling approach.

Acknowledgements

Research funded by the grant of the National Science Centre, Poland, under the contract number DEC-2018/31/B/HS4/02021.

References

CBOE, 2021, VIX Historical Price Data, https://www.cboe.com/tradable_products/vix/vix_historical_data

Coulombe, P. G., Leroux, M., Stevanovic, D., Surprenant, S., 2021, Macroeconomic data transformations matter, International Journal of Forecasting 37, 1338-1354.

Diebold, F. X., Mariano, R. S., 1995, Comparing predictive accuracy, Journal of Business and Economic Statistics 13, 253-263.

EIA, 2021, U.S. Energy Information Administration, https://www.eia.gov

FRED, 2021, Economic data, https://fred.stlouisfed.org

Friedman, J., Hastie, T., Tibshirani, R., 2010, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33, 1-22.

Hansen, P. R., Lunde, A., Nason, J. M., 2011, The Model Confidence Set, Econometrica 79, 453-497.

Hyndman, R. J., Khandakar., Y., 2008, Automatic time series forecasting: the forecast package for R, Journal of Statistical Software 26, 1-22.

Jin, Y., Fu, W., Kang, J., Guo, J., Guo, J., 2019, Bayesian symbolic regression, https://arxiv.org/abs/1910.08892

Kilian, L., 2009, Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market, American Economic Review 99, 1053-1069.

Koza, J. R., 1992, Genetic Programming, Cambridge, MA: MIT Press.

MSCI, 2021, End of Day Index Data Search, https://www.msci.com/end-of-day-data-search

Nicolau, M., Agapitos, A., 2021, Choosing function sets with better generalisation performance for symbolic regression models, Genetic Programming and Evolvable Machines 22, 73-100.

Raftery, A. E., Karny, M., Ettler, P., 2010, Online prediction under model uncertainty via Dynamic Model Averaging: Application to a cold rolling mill, Technometrics 52, 52-66.

Stooq, 2021, Quotes, https://stooq.com

The World Bank, 2021, Commodities Markets, https://www.worldbank.org/en/research/commodity-markets