Symbolic Regressions: Opening the Black Box of Equity Premium Predictions

Igor Kuznetsov

London Business School

Latest Version: July 31, 2025

Abstract:

This paper investigates equity premium predictability using Deep Symbolic Regression (DSR), a method that identifies sparse and interpretable functional forms in the data. Unlike traditional opaque machine learning models, DSR allows explicit capture of nonlinear economic relationships. The paper introduces a novel regularization parameter within the DSR methodology, ensuring robust model selection. Extensive simulations validate the effectiveness of this approach. Empirical analysis using monthly U.S. stock market data (1927–2021) demonstrates that DSR consistently outperforms benchmarks such as linear regression and random forests in forecasting accuracy. The findings highlight significant nonlinear dynamics in market returns, particularly during periods of economic stress, thereby providing a transparent and economically insightful framework for equity premium prediction.[1]

[1]This project has received support from the Google Cloud Research Credits Program.

Introduction

In recent years, there has been a growing interest in the use of machine learning tools in various contexts in finance. For example, machine learning methods have been applied to predict the market returns (Kelly and Pruitt 2013; Feng, He, and Polson 2018; Rapach and G. Zhou 2020; Kelly, Malamud, and K. Zhou 2022b), the cross-sectional predictability of returns (Cong et al. 2022; Feng, He, Polson, and Xu 2018; Y. Han et al. 2022; Kelly, Malamud, and K. Zhou 2022a), measuring asset risk premium (Gu, Kelly, and Xiu 2020), and the behavior of stock analysts (Bew et al. 2019; Qiu, Zhewei Song, and Z. Chen 2022; Bianchi, Büchner, and Tamoni 2021; Binsbergen, X. Han, and Lopez-Lira 2020). These studies highlight the flexible nature of the machine learning methods and their ability to capture nonlinearities, interactions, and high-dimensional features in the data. However, they also share a common limitation: they act essentially like black box prediction tools, and it is very difficult to bridge the gap between the flexible reduced-form predictions produced by the methods and the structural models that should generate the underlying relationship.

In this paper, I argue for applying the advantage of symbolic regression methodology to the classical problem of finance: predictability of market excess return. Symbolic regression is a method that allows the recovery of flexible nonlinearities and - crucially - retains the functional interpretation of the resulting relationship. In a nutshell, symbolic regression uses genetic programming to search for analytical expressions that best fit the data, imposing minimum prior on the functional form or parameter restrictions. The method can handle complex and noisy data, discover hidden patterns and interactions, and provide interpretable and parsimonious models.

The main contribution of this paper is twofold. First, I examine the properties of the method in simulated data - both in small and large samples, with high and low signal-to-noise ratios that proxy realistic data-generating processes. I propose a new penalty parameter that allows symbolic regression to recover the key foundational dependencies in the data and the true level of predictability. The search procedure, predicated on a progressive easing of the

regularization constraint, facilitates the recovery of the true data-generating process even in high-noize environments of R^2 at 5% when the sample size is sufficiently large.

I then apply the method to real monthly data from the US stock market over a sample period from 1927 to 2021. I compare the performance of symbolic regression with various benchmark models, such as linear regression and random forest. I find that symbolic regression outperforms the benchmark models in terms of out-of-sample forecasting accuracy and, in the majority of cases, gives a better in-sample fit. This approach allows attaining a timing Sharpe ratio of 0.17 on an annual basis compared to the Sharpe ratio of 0.04 from the linear model. I also analyze the functional forms and economic interpretations of the symbolic regression models and show that they capture some well-known stylized facts as well as some novel nonlinearities and interactions in market return predictability.

The question of market excess return predictability is grounded in the market efficiency hypothesis (MEH). It states that financial markets are "informationally efficient", that is, that "prices 'fully reflect' all available information" (Fama 1970). The hypothesis suggests that it is impossible to consistently achieve returns in excess of average market returns, given the information that is publicly available at the time of investment. In the present-value relationship between returns, asset prices, and their cash flows, the gross return on a stock or a stock market is defined as:

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} \tag{1}$$

where P_t and D_t are prices and dividends of a given asset at time t. Traditionally, this relationship is log linearized with Campbell and Shiller (1988) decomposition to generate a linear relationship. This approximation together with documented no dividend growth predictability has led the literature to the conclusion that returns must be predictable with dividend-to-price ratios (Lettau and Van Nieuwerburgh 2008a; Cochrane 2011; Van Binsbergen and Koijen 2010).

The point of departure for this study is Welch and Goyal (2008) paper (GW) that compre-

hensively reexamines the performance of variables that have been suggested by the academic literature to be good predictors of the equity premium. GW finds that most suggested up-to-date models are unstable or even spurious, no longer significant even in-sample (IS), and most of them fail simple regression diagnostics. For many models, any previous apparent statistical significance was often based only on years up to and especially on the Oil Shock years of 1973-1975.

The current work builds upon an estimation setting in GW,[2] investigating functional forms that link the monthly market excess stock returns to the six variables used in GW to construct key theoretical predictors. GW and similar studies assume a linear relationship between log returns and predictors (eg. the log of the dividend-to-price ratio). Unlike the GW study, I do not impose ex-ante linear dependence or any predefined functional form. Instead, I use Deep Symbolic Regression methodology (DSR) that searches through the space of functional forms and identifies expressions that not only explain in-sample variation well but also generate better OOS predictions. It allows to identify relatively parsimonious functional forms that outperform the linear model in 3-month prediction horizons in both IS and OOS estimates, although R^2OOS still being negative and not statistically significant. The linear model suggests an R^2OOS of -0.34% for a 75-year prediction window while DSR estimates it at -0.15%.

As a response to GW, Campbell and Thompson (2008) propose restricting the signs of coefficients and return forecasts and steady-state valuation models as a way to improve out-of-sample predictability of excess market return. Van Binsbergen and Koijen (2010) and Kelly and Pruitt (2010) suggest latent variables approach for better out-of-sample predictions of the market equity premium. Rapach and G. Zhou (2013) demonstrate that combining model individual forecasts leads to significant out-of-sample predictability relative to the historical average consistently over time. Most of these approaches achieve higher OOS predictability in terms of R^2OOS than identified in this study. Contrasting with this direction of the literature, however, I do not use any other variables, data structure features, or coefficient

[2]Throughout this paper, I refer to the linear regression of log excess market returns on the predictors in GW as a linear model.

sign restrictions, and limit the model to the information set of time-series excess market returns as in GW. Documented market return predictability in this paper stems directly from identifying non-linear relations between returns and explanatory variables in the most simplistic setup.

This study offers an alternative perspective on the "virtue of complexity" phenomenon, initially explored by Kelly, Malamud, and K. Zhou (2022b). The referenced work provides a theoretical foundation and empirical validation demonstrating that in the context of U.S. market return predictability, model efficacy – both in terms of expected out-of-sample forecast accuracy and portfolio performance – tends to enhance with increased model complexity.[3] This increment in complexity is achieved by augmenting the number of random Fourier features derived from a predetermined group of initial predictors and applying a ridge-type regularization mechanism.

In this paper, the exploration of model complexity is approached through a variation of functional forms, particularly focusing on potential non-linear interactions among predictors. Rather than approximating the functional form with high-dimensional linear expansions or a neural network (Hornik, Stinchcombe, and White 1990; Jacot, Gabriel, and Hongler 2018; Hastie et al. 2022; Allen-Zhu, Li, and Zhao Song 2019), DSR enables me to rigorously investigate the precise functional form underlying the data-generating process. The space of these functional forms is defined by the range of permissible operations within the framework and an ex-ante-selected set of predictors. Introducing a regularization parameter, conceptualized as the maximum permissible number of tokens – encompassing variables, operations, or weights – serves to regulate the extent of linear and non-linear transformations applied to predictors during the training phase. Both simulation exercises and the empirical analysis shows that there is an optimal amount of allowed model complexity for DSR that leads to the highest out-of-sample predictability and portfolio performance. Beyond the optimal level, the accuracy starts to decay.

DSR demonstrates particular advantage in the period of significant economic events like the

[3]Kelly, Malamud, and K. Zhou (2022a) shows the advantage of this approach for predicting returns in other asset classes.

Great Recession showing improved return predictability during such periods. This observation collaborates with the prior findings that the predictability of returns predominantly occurs during periods of economic recession (Rapach, Strauss, and G. Zhou 2010; Henkel, Martin, and Nardari 2011; Dangl and Halling 2012). It potentially indicates the ability of the DSR method to approximate well the magnitudes in the steady-state shifts of the economy mean during and after big economic shocks that is known to be challenging in the current literature (Lettau and Van Nieuwerburgh 2008a).

On the theoretical side, this paper introduces a new regularization parameter for the cost function that enables the identification of the true sparse model that is behind the data generating process (DGP) and that ensures a valid model selection.[4] The paper shows through extensive simulations that DSR performs well in scenarios with low signal-to-noise ratios similar to common applications in finance. The out-of-sample performance improves with larger samples and optimal regularization parameters and approaches the true R^2 of the DGP. These results hold even with high noise and allow DSR to recover the exact structure of simple and complex functional forms with large enough samples.

Several recent studies have successfully adopted ML models for out-of-sample return predictions. Using lagged returns from many countries as predictors, Rapach, Strauss, and G. Zhou (2013) apply the ENet to multiple predictive regressions for monthly country stock returns and discover that the US market return has a leading role in other countries' returns. Rapach and G. Zhou (2022) adopt elastic net for forecasting US market excess return. Dong et al. (2022) use 100 anomaly portfolios as predictors to show that a multiple predictive regression with ENet estimation can forecast the US market excess return. Gu, Kelly, and Xiu (2020) demonstrate that investors can achieve large economic benefits by using machine learning forecasts, in some cases doubling the performance of leading regression-based strategies from the literature. They identify trees and neural networks as the best-performing methods and

[4]The idea of limiting the complexity of functional forms by itself is not new. Traditionally, the complexity constraints are used for the construction of the accuracy-complexity Pareto frontier ex-post (eg. see Petersen et al. (2021), Udrescu and Tegmark (2020), and Landajuela et al. (2022)). This paper introduces complexity constraints in situ as a regularization parameter to identify the unique "true" function behind the DGP in a high-noise environment.

attribute predictive gains to allowing nonlinear interactions among predictors that other methods are not able to capture. Incorporating economic intuition like no-arbitrage conditions can allow neural networks to further improve the out-of-sample predictability of excess returns (L. Chen, Pelger, and Zhu 2023).

Although these ML techniques show meaningful improvements in terms of out-of-sample return predictability, the mechanism that lies behind the fundamental relationship between asset prices and conditioning variables remains largely unexplored in these studies. All these methods still exhibit a black box problem. Symbolic regressions, unlike other ML techniques that are popular in economics and finance, aim to identify mathematical expressions that best fit a given data and it is able to discover compact and generalizable expressions that capture the underlying patterns or relationships allowing to narrow down the exact structural relationship between predictive characteristics and an outcome variable. In the context of excess return predictability, DSR models steadily recover the linkage between market excess returns to dividends, lagged prices, earnings, and stock variance suggesting unusual nonlinear equations that involve exponential-term relations of predictive variables for explaining log market excess returns.

There are only a few applications of symbolic regressions in economics and finance literature at this point. Alvarez-Diaz and Alvarez (2003) employ genetic algorithms to find mathematical models that can predict the weekly fluctuations of six exchange rates. Claveria, Monte, and Torra (2017) studies the effect of the Great Recession on agents' expectations using symbolic regression methodology. Claveria, Monte, and Torra (2018) propose a novel data-driven method to construct an economic indicator from survey data using evolutionary computation and symbolic regression. Claveria, Monte, and Torra (2022) use a soft computing method to create economic sentiment indicators for 19 European countries based on business and consumer surveys. They show that these indicators can predict GDP growth rates better than traditional time-series models. To the best of my knowledge, the current paper is the first to apply symbolic regressions to the market return predictability problem.

Symbolic Regressions and Predictability

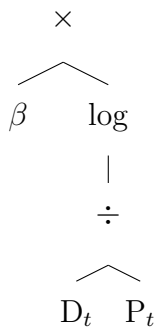
Functional Forms and Symbolic Regressions

In this paper, I employ Deep Symbolic Regression (DSR) methodology (Petersen et al. 2021; Landajuela et al. 2022), a novel machine learning technique that aims to identify an underlying mathematical expression that best describes a relationship between an outcome variable and its explanatory variables. Symbolic regression is different from traditional regression methods because it does not assume a predefined form of the model, but rather searches for the optimal structure and parameters of the model from a set of simple base functions. Symbolic regression can discover hidden nonlinear relationships between variables and produce explicit models that are interpretable and robust.

Any mathematical expression can be written as a pre-order traversal of a symbolic expression tree. For example, $f(D_t, P_t)$ in the right-hand side of an expression like

$$r_{t+1} = \underbrace{\beta \times log\left(\frac{D_t}{P_t}\right)}_{f(D_t,P_t)} + u_{t+1}$$

can be transformed into a symbolic expression tree



and the pre-order traversal of this expression would be $\{\times, \beta, \log, \div, D_t, P_t\}$. The traversal is generated one element at a time by drawing from a library of accepted operations and

variables based on the probability distribution conditioning on a previously drawn element and a sibling element if available.[5] The library can contain a large set of mathematical operations including sigmoid and harmonic transformations as well as indicator functions. In this paper, however, I stick to a simple set of operations that are common in financial models. The library of accessible operations consists of $\{+, -, \times, \div, exp, log, \sqrt{}, [vars]\}$. The library can also include placeholders for numerical values akin to coefficients in linear regressions that I refer to these coefficients as constants following the literature convention. Initial priors on drawing elements are flat given the constraints of the equation form.[6] For example, the traversal has to start with a mathematical operator and end with a variable input.

After the initiation of a pool of expressions, each one of them is evaluated based on a loss function. If the library also contains the constant operator, generated constants are optimized for each equation in the candidate set. Then a top percentile of expressions is selected based on a risk-seeking policy. This percentile threshold is set to top 5% in our case. The traversal structure of selected expressions is then used to update conditional probabilities of drawing operations and variables. The procedure repeats up until the discovery of an equation that gives exact representation of the data or up until the point of convergence.

Prediction, Noise, and Out-of-Sample Forecasts

All symbolic regression methods including DSR are usually tested in environments with no noise or low levels of noise with $R^2 \geq 90\%$ which is common in natural science studies. In finance, noise levels that are equivalent to R^2 of 5% are considered to be the norm. In this paper, I propose a new penalty structure for the DSR loss function that allows to recovery of true functional forms in data with high levels of noise that are common to economics and finance datasets.

For a given loss function $L(.)$, the goal is to find a true function f^* that is behind the data

[5] A sibling conditioning arises in operations that require two inputs like $+$ or $\times$; it is not symmetric meaning that only the second drawn element has a sibling. In the example expression only $\times$ and $R^e_{m,t}$ would be drawn conditioning on a sibling.

[6] One can also set non-flat priors although it is beyond the scope of this project.

generating process (DGP) from a function space F such that

$$f^* = \arg\min_{f \in F} L(f) \tag{2}$$

The function space F_Λ that is accessible to an econometrician is a set of all functions spanned by variables and mathematical expressions in the library. Throughout simulation exercises, it is assumed that $f^* \in F_\Lambda$, i.e. the true function is in the accessible space. Although F_Λ contains functions of infinite size (infinite traversal length), it is further assumed that f^* has a finite although unknown traversal length.

For any given sample of data $H = \{X, y\}$ where X is m by n matrix of inputs and y is m by 1 outcome variable, the true functional form f^* is such that:

$$y = f^*(X) + \epsilon \tag{3}$$

where ϵ_i is a random noise. Without loss of generality, I assume that $\epsilon_i \sim N(0, \sigma^2)$. Then, f^* is estimated as:

$$\hat{f}_\Lambda = \arg\max_{f \in F_\Lambda} \mathbb{E}[L(f(X), y)] \tag{4}$$

A sufficiently long and complex function could fit training data perfectly. However, while this function would have a perfect fit on a given sample, it would be overfitting on any other sample drawn from the same distribution and perform poorly in out-of-sample predictions. This issue of overfitting is not unique to this specific method but common to all ML methods. ML's appeal lies in its ability to handle high dimensionality and fit varied data structures with flexible functional forms. However, this flexibility also means that simply choosing the best in-sample function can lead to poor results. Therefore, the goal should be to minimize:

$$Err_H = \mathbb{E}_H[L(\hat{f}_\Lambda(X'), y')] \tag{5}$$

where $H' = \{X', y'\}$ is a different subset of data than the one used to estimate $\hat{f}_\Lambda$, i.e. a test data. To be precise, the goal should be to minimize $E[Err_H]$, the expected prediction error, rather than test errors for a specific dataset.

One of the most popular tools for minimizing Err_H and addressing overfitting is regularization. When fitting a DSR model, one could select the best-performing function among those with a certain maximum length of the traversal instead of choosing the overall best function. A shorter traversal may have a worse in-sample fit because individual observations may not fit well. However, this also means that overfitting is reduced because the noise from individual observations is averaged out. Restricted traversal length is an example of a regularizer that measures the complexity of a function. As one decreases regularization (increase allowed traversal length), she would improve the ability to approximate in-sample variation but at the cost of increasing the difference between in-sample and out-of-sample performance. By choosing the appropriate level of regularization, one can benefit from flexible functional forms without being overwhelmed by overfitting. So the constrained estimate of f^* is:

$$\hat{f}_{\Lambda,\lambda} = \arg\min_{f \in F_\Lambda} \mathbb{E}_H[L(f(X), y)] + R_\lambda(f) \tag{6}$$

where

$$R_\lambda(f) = \begin{cases} \infty, & \text{if } ||f|| \geq \lambda \\ 0, & \text{otherwise} \end{cases}$$

Here, with some abuse of notation, I denote $||f||$ as the length of the function traversal.

To determine the optimal traversal length or level of regularization, I can utilize empirical tuning. The challenge with overfitting is that the goal for model selection is to achieve the highest prediction performed on the out-of-sample data but models only fit it on in-sample data. Empirical tuning involves creating an out-of-sample experiment within the original sample by fitting the model on one part of the data and evaluating its performance on another

part using different levels of regularization (Mullainathan and Spiess 2017). Cross-validation can increase the efficiency of this process by randomly partitioning the sample into equally sized subsamples or folds. Instead of using one omitted sample for out-of-sample performance as it is usually done in cross-validation, I use a bootstrap. Specifically, for a collection of subsamples $\{H_1, H_2, ..., H_K\}$ for some sufficiently large K, every in-sample equation derrived from H_i sample is tested on all H_{-i} samples and the results are averaged. This process is repeated for each fold and the functional form with the best average performance is chosen. The tuning parameter is selected based on the minimum achievable $E[Err_H]$.

Computational Complexity

The space of functional forms is in L^∞ requiring searching through discrete space of model representations and continuous space of parameters. It is an NP-hard combinatorial optimization problem (Lu, Ren, and Wang 2016).

Earlier solutions to the symbolic regression problem were based on Genetic Programming (GP) and similar combinatorial optimization methods (Koza 1994; E. (Vladislavleva 2008; E. J. Vladislavleva, Smits, and Hertog 2009; M. Schmidt and Lipson 2009; Dabhi and Vij 2011). Using operations like selection, crossover, and mutation, GP-based symbolic regression modifies a population of mathematical expressions to improve a fitness function. This approach is prone to high computational cost and overly intricate output expressions, and the solution varies with the initial value. (Korns 2011). There have been many attempts to tackle these challenges by utilizing semantics and multi-objective formulation (Huynh, Singh, and Ray 2016), proposing a problem-simplification tool for symbolic regression (Udrescu and Tegmark 2020), incorporating grammar variational autoencoder into generative model (Kusner, Paige, and Hernández-Lobato 2017), developing a Bayesian framework (Jin et al. 2020) or NeuroEvolution of Augmenting Topologies (Trujillo et al. 2016).

DSR technique offers an efficient and yet tractable solution to this problem. This approach uses recurrent neural networks (RNN) to generate a distribution over tractable mathematical expressions and it trains the neural network using a risk-seeking policy gradient. DSR is

considered to be the state-of-the-art approach among symbolic regression techniques in real-world data tests. It has a rigorous methodological foundation while offering both flexibility and maintaining computational efficiency.

Simulations

In order to understand the empirical properties of the DSR method, I test it on simulated data for various sample sizes with simple and complex DGPs and different levels of noise levels. The goal is to verify the ability of this method to recover the original DGP by converging to the true functional form as $N \to \infty$. I consider a simple simulation setup with two explanatory variables. Specifically, the following equations are used for DGP: [7]

$$Y = X_1 + X_2 + X_1 * X_2 + \epsilon \tag{7}$$

and

$$Y = \log\left(\frac{X_1}{X_2}\right) + X_1^2 + \epsilon \tag{8}$$

where X_1 and X_2 are two independent random variables drawn from a uniform distribution with a support $[0, 1]$ and ϵ is drawn from a standard normal distribution. For a lack of a better word, I refer to DGP in equation 7 as a simple DGP or DGP 1 and the one generated in equation 8 as a complex DGP or DGP 2. In essence, DGP 1 is meant to represent an uncomplicated linear relationship and DGP 2 mirrors intricate nonlinear dependence, that may resemble the kind of predictability observed in financial markets (e.g. with a logarithmic price-to-dividend ratio, or in cross-sectional predictability with a logarithmic characteristic

[7] In all simulation exercises, I set the intercept to 0 and all parameter coefficients to 1 intentionally to speed up estimations. By design, DSR optimizes constants in an inner loop for each draw of $\hat{f}_{\Lambda,\lambda}$ so the computation time increases substantially with constants. Optimizing constants would add some additional uncertainty to estimations that would shrink asymptotically as $N \to \infty$. On the other hand, the algorithm seems to be able to pick and approximate coefficients well enough through functional representations at the expense of the traversal length. For example, $\frac{1}{3}$ can be expressed by DSR as $\frac{X}{X+X+X}$.

ratio).

For sample sizes of $[50, 100, 250, 500, 1000, 10000]$ and for levels of noise corresponding to $R^2 \in [1, 0.95, 0.7, 0.3, 0.05]$ of a true functional form,[8] I generate 101 random samples for each cell in the grid of parameters. Each DSR model is estimated on the training sample $\{X, y\}_i$, and its out-off-sample performance is then evaluated on remaining $\{X, y\}_{-i}$ samples. In this section, regularization parameter λ is set to one of the values in $[5, 7, 8, 10, 15]$ with the minimum complexity level of DGP 1 being 7 and of DGP 2 being 8 based on the selected library of operations. It should be noted that although linear models are considered to be the most common and tractable relations in the traditional regression context, there is no particular advantage for DSR and symbolic regressions in general when the true DGP is linear.

Densities of in-sample and average out-of-sample R^2 from these simulations are presented in Figures 1 and 2 and the summary statistics for out-of-sample R^2 are reported in Table 1. All of the estimates shown are based on DGPs with noise $R^2 = 5\%$ for the true equations. Estimates with $R^2 \in [1, 0.95, 0.7, 0.3]$ are in Appendix.[9]

These figures show that there is a substantial variation in both R^2IS and R^2OOS. In estimates on small samples as in left columns in Figures 1 and 2, the in-sample R^2 appear to be centered around 0 with very wide distributions and long left tails reaching negative values across all specifications. Small sample sizes also lead to substantial variability of R^2OOS estimates that are centered at 0 for both DGP 1 and DGP 2 regardless of the regularisation level. In this case, mean $R2OOS$ distributions have long left tails with some very negative estimates of below -10.

[8] Noise levels are set to $R^2 \in [1, 0.95, 0.7, 0.3, 0.05]$ on average for a given sample group of 101 samples. The actual noise level may vary somewhat among samples within the same noise level group due to the randomness of DGP. As expected, larger samples have smaller variations due to the central limit theorem.

[9] Although advantageous relative to other symbolic regression methods in terms of computation speed, DSR is still a highly demanding procedure. Simulations take a lot of time to compute. At the moment, a complete set of simulations is available only for noise levels with true $R^2 \in [0.05, 0.30, 1]$. No-noise estimates are not reported because results are trivial, DSR recovers true equations in 100% of the cases when the regularization parameter allows that.

Mean values[10] of R^2OOS and even right bounds of 95% confidence intervals are in a negative zone as it can be observed in row 1 of Panels A and B in Table 1. For all levels of complexity and both DGPs, distribution means of $R2OOS$ start to increase approaching the true level of noise with $R^2 = 5\%$ and standard deviations shrink accordingly. At the maximum sample size of 10,000 obs as in the right columns in Figures 1 and 2, distributions of mean R^2OOS normalize and their variance reduces substantially. The highest density of mean R^2OOS estimates is achieved when the regularizer parameter λ is set to the true minimum traversal length for both simple and complex DGPs. When DSR model complexity is assumed to be low (strong regularization), mean R^2OOS values are centered below the true R^2 level. The same happens when functional forms are allowed to vary beyond the minimum traversal length of the true equation although to a lesser degree. The variance of the estimates also appears to be smaller when the regularizer parameter is closer to the true level.

Table 2 reports the average rate of the exact convergence of DSR models to true functional forms when DGPs have high-level noise with true $R^2 = 5\%$. The exact convergence is defined as a perfect structural match of the equation estimated by DSR to the true equation behind a specific DGP. Tables with convergence rates for DGPs with true R^2 at levels $[0.7, 0.3]$ are in the Appendix.

In both simple and complex cases, there is no convergence to the true functional form when λ is smaller than the minimum traversal length of the true function. On the other hand, when λ is set exactly to the minimum traversal length of the true function, the DSR starts to recover true functional forms and the rate of recovery also increases with the sample size. Thus, with sample size of 50 obs in DGP 2 and up to 250 in DGP 1, DSR is not able to recover the original equations behind these DGPs. As the sample size increases, the recovery rate reaches 100% for the DGP 1 and to 89.1% for the DGP 1 at 10,000-obs level. Further, the convergence rate starts to decline as the regularization parameter is relaxed falling to 0 for both DGPs at $\lambda = 15$ even with the largest sample sizes. For larger levels of λ, the DSR algorithm starts to overfit in the sample capturing not only the structural relationship but also fitting the noise which reduces its out-of-sample performance. What is also evident

[10] Technically, reported mean values in 1 are mean of means estimated from the cross-validation procedure.

Figure 1: R^2IS and Average R^2OOS of DSR Estimates of DGP 1 with High Noise

Figure 1 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 1 for samples sizes of 100 and 10,000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 7, 15$ that referred to as Low, True, and High correspondingly. All samples have 5% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

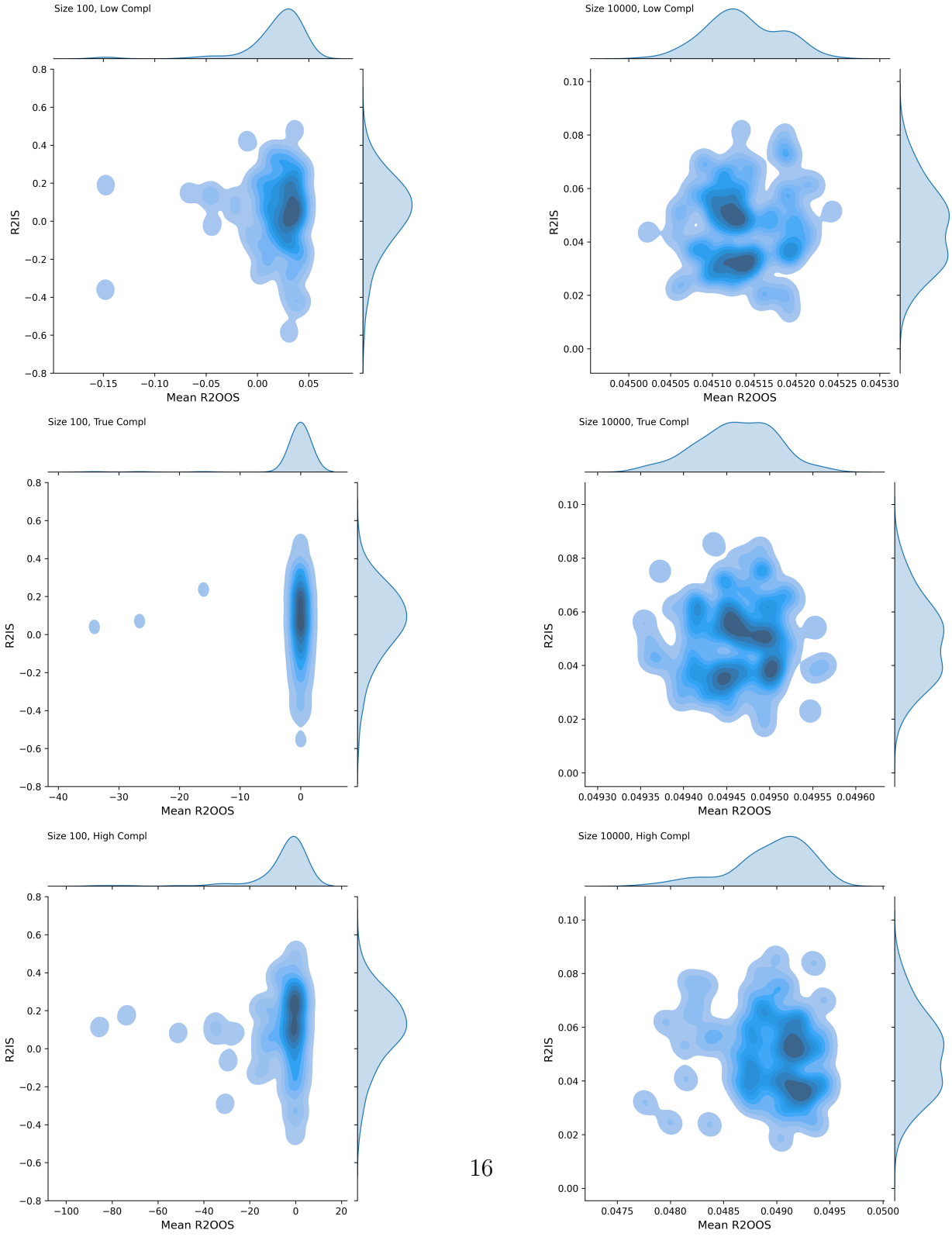


Figure 2: R^2IS and Average R^2OOS of DSR Estimates of DGP 2 with High Noise

Figure 2 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 2 for samples sizes of 100 and 10,000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 8, 15$ that referred to as Low, True, and High correspondingly. All samples have 5% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

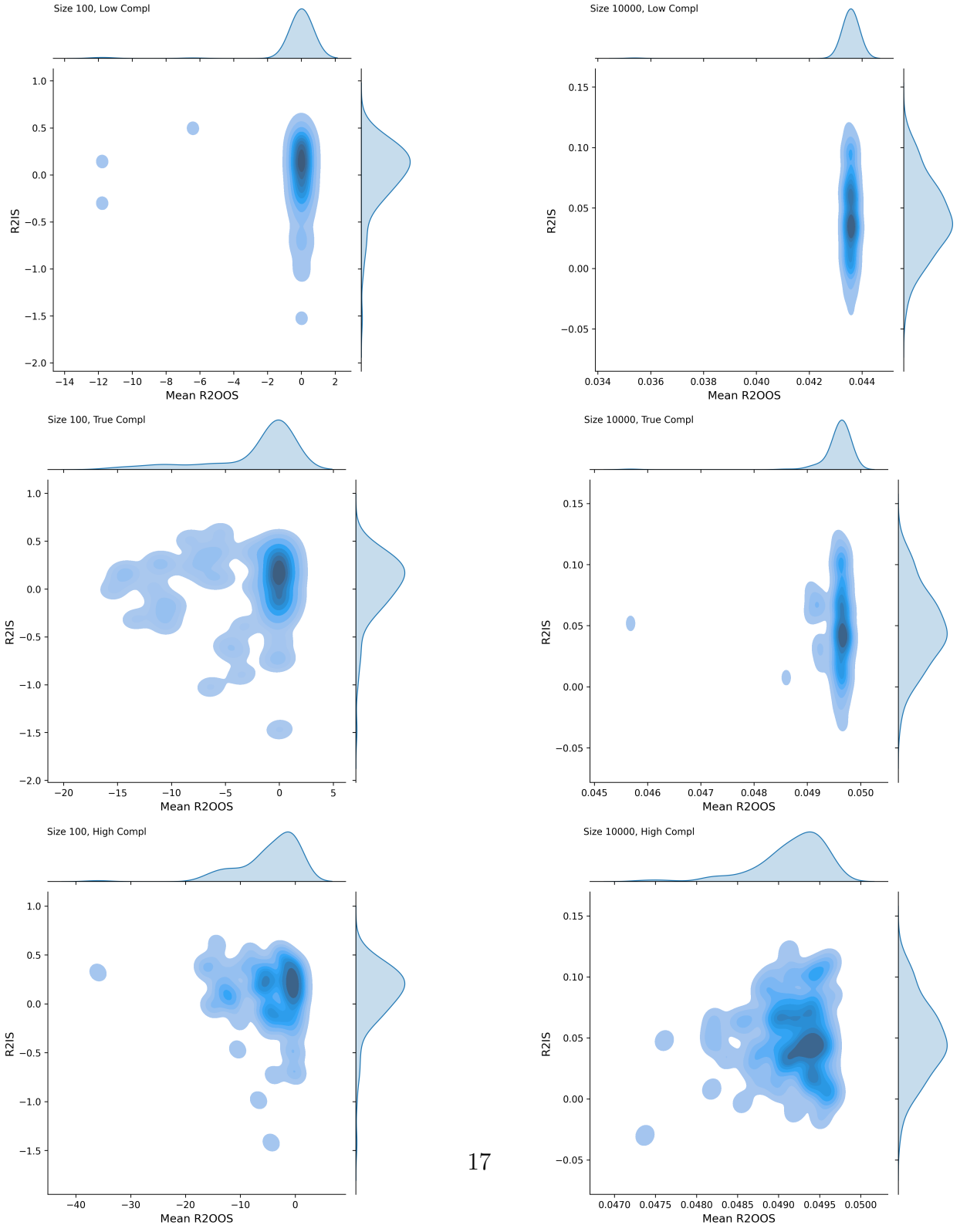


Table 1: Summary Statistics of Mean R^2OOS measures in DSR Simulations

Table 1 reports summary statistics of average R^2OOS in OOS predictions by DSR models estimated over simulated data generated by simple and complex DGPs. All samples have 5% signal in terms of true R^2. Sample size varies from 50 to 10,000 obs. In each specification cell, there are 101 samples for DSR model estimation. Models estimated on one sample are then tested on the remaining 100 samples. Mean R^2OOS is obtained by averaging R^2 from test samples. The regularization parameter can be one of the values $\lambda \in [5, 7, 8, 10, 15]$. Extreme values of $R^2 < -100$ are truncated. It affects $< 0.5\%$ of obs in the most demanding specifications.

Panel A: Simple DGP

Panel A reports summary statistics for simulations of DGP 1. The underlying true equation is $Y = X_1 + X_2 + X_1 * X_2 + \epsilon$. The minimum length of the traversal is 7.

	Max Complexity				
Sample Size	**5**	**7**	**8**	**10**	**15**
50 (n=505)	-0.04	-0.49	-0.40	-1.46	-3.36
(SD)	(0.02)	(0.14)	(0.13)	(0.33)	(0.37)
(95% CI)	[-0.08,-0.01]	[-0.76,-0.22]	[-0.67,-0.14]	[-2.10,-0.82]	[-4.09, -2.64]
100 (n=505)	0.02	-0.10	-0.13	-0.14	-1.44
(SD)	(0.00)	(0.06)	(0.06)	(0.05)	(0.27)
(95% CI)	[0.01,0.02]	[-0.22,0.01]	[-0.25,-0.01]	[-0.24,-0.03]	[-1.96,-0. 92]
250 (n=505)	0.04	-0.06	-0.02	-0.08	-0.34
(SD)	(0.00)	(0.07)	(0.05)	(0.08)	(0.12)
(95% CI)	[0.03,0.04]	[-0.19,0.07]	[-0.11,0.07]	[-0.23,0.07]	[-0.57,-0.10]
500 (n=505)	0.04	-0.05	-0.06	-0.03	-0.23
(SD)	(0.00)	(0.09)	(0.09)	(0.05)	(0.16)
(95% CI)	[0.04,0.04]	[-0.24,0.13]	[-0.24,0.12]	[-0.14,0.07]	[-0.55,0.08]
1000 (n=505)	0.04	0.04	0.04	0.04	-0.06
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)
(95% CI)	[0.04,0.04]	[0.04,0.04]	[0.04,0.05]	[0.04,0.04]	[-0.19,0.07]
10000 (n=505)	0.05	0.05	0.05	0.05	0.05
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Panel B: Complex DGP

Panel B of reports summary statistics for simulations of DGP 2. The underlying true equation is $Y = \log\left(\frac{X_1}{X_2}\right) + X_1^2 + \epsilon$. The minimum length of the traversal is 8.

	Max Complexity				
Sample Size	**5**	**7**	**8**	**10**	**15**
50 (n=505)	-0.27	-1.05	-2.03	-3.00	-3.58
(SD)	(0.09)	(0.19)	(0.26)	(0.27)	(0.28)
(95% CI)	[-0.44,-0.09]	[-1.42,-0.68]	[-2.55,-1.51]	[-3.52,-2.47]	[-4.12,-3.04]
100 (n=505)	-0.11	-0.20	-1.09	-1.01	-2.22
(SD)	(0.08)	(0.07)	(0.20)	(0.16)	(0.27)
(95% CI)	[-0.26,0.04]	[-0.34,-0.07]	[-1.48,-0.69]	[-1.31,-0.70]	[-2.74,-1.70]
250 (n=504)	0.03	0.01	-0.01	-0.04	-0.32
(SD)	(0.00)	(0.00)	(0.02)	(0.03)	(0.09)
(95% CI)	[0.03,0.03]	[0.01,0.02]	[-0.04,0.03]	[-0.10,0.02]	[-0.50,-0.13]
500 (n=505)	0.04	0.03	0.03	-0.01	-0.04
(SD)	(0.00)	(0.00)	(0.00)	(0.04)	(0.04)
(95% CI)	[0.04,0.04]	[0.03,0.04]	[0.03,0.04]	[-0.08,0.07]	[-0.11,0.03]
1000 (n=504)	0.04	0.05	0.05	0.05	0.02
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
(95% CI)	[0.04,0.04]	[0.04,0.05]	[0.04,0.05]	[0.04,0.05]	[-0.02,0.06]
10000 (n=505)	0.04	0.05	0.05	0.05	0.05
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table 2: Convergence Rates of DSR estimates to True Functional Forms.

Table 1 reports mean rates of the exact convergence of DSR to true functional forms with simple and complex and DGPs. The exact convergence is defined as the perfect structural match of the equation estimated by DSR to the true equation behind a specific DGP. Convergence rates are on a 0 to 1 scale. All samples have 5% signal in terms of R^2. Sample sizes vary from 50 to 10,000 obs. In each specification cell, there are 101 samples for DSR model estimation. The regularization parameter can be one of the values $\lambda \in [5, 7, 8, 10, 15]$.

Panel A: Simple DGP

Panel A shows DSR convergence rates for simulations of DGP 1. The underlying true equation is $Y = X_1 + X_2 + X_1 * X_2 + \epsilon$. The minimum length of the traversal is 7.

	Max Complexity				
Sample Size	5	7	8	10	15
50	0.000	0.000	0.000	0.000	0.000
100	0.000	0.000	0.000	0.000	0.000
250	0.000	0.000	0.000	0.000	0.000
500	0.000	0.139	0.020	0.000	0.000
1000	0.000	0.287	0.010	0.000	0.000
10000	0.000	1.000	0.653	0.257	0.000

Panel B: Complex DGP

Panel B shows DSR convergence rates for simulations of DGP 1. The underlying true equation is $Y = \log\left(\frac{X_1}{X_2}\right) + X_1^2 + \epsilon$. The minimum length of the traversal is 8.

	Max Complexity				
Sample Size	5	7	8	10	15
50	0.000	0.000	0.000	0.000	0.000
100	0.000	0.000	0.030	0.000	0.000
250	0.000	0.000	0.099	0.000	0.000
500	0.000	0.000	0.248	0.000	0.000
1000	0.000	0.000	0.376	0.050	0.000
10000	0.000	0.000	0.891	0.475	0.000

is that the recovery rate increases with the sample size even when λ is raised beyond the minimum traversal length of the true DGP, at least in some vicinity of its optimal level.

This simulation exercise highlights several important properties of the DSR methodology. The OOS performance as measured by average R^2OOS in cross-validation seems to improve with larger sample sizes approaching the level of R^2 implied by the true equation. The uncertainty around it also falls for larger samples. Further, if the true equation is in the accessible space of functional forms, there seems to be an optimal regularization parameter that maximizes R^2OOS bringing it to the true R^2 of the underlying DGP. The convergence appears to be the fastest when λ is set to the minimum traversal length of the true functional form. These results hold even when the noise in the data is high at true $R^2 = 5\%$ which is prevalent in economics and finance research. It is also shown, that at least in the simulated environment, it is possible to use DSR for recovering the exact structure of simple and complex functional forms behind the true DGP by tuning the proposed regularization parameter when the sample size is large enough.

By comparing the outcomes of two simulations with simple and complex DGPs, one can also infer that more complex expressions converge at a slower rate with respect to the sample size. Although intuitive, this claim requires further investigation which is beyond the scope of this study.

Empirical Results

Data

As an empirical exercise, I focus on the time-series predictability of the market equity return premium over the period from 1927 to 2021 on the monthly level and the dynamics of functional forms that can explain this predictability.

In recent years, many predictors have been proposed and tested with various success. I restrict this study to six variables used to construct key predictors in GW that are available for the entire span of the dataset: dividends, prices, lagged prices, earnings, stock variance

(svar), and book-to-market ratio (b/m). With each additional variable, the computational complexity of the DSR search process and the space of functional forms grow exponentially. This limitation of DSR restricts the number of possible predictors in the current analysis.

The data comes from an extended version of a GW dataset from Ivo Welch's website. In this dataset, the returns data is based on the S&P 500 index from 1926 to 2021 from the month-end values of the Center for Research in Security Press (CRSP) continuously compounded cum-dividend stock return data.

The risk-free rate data is the Treasury-bill rate for the period 1920 to 2005 and it is imputed by GW for the period 1871 to 1919 from Commercial paper rates for New York City that is available in the National Bureau of Economic Research (NBER) Macrohistory database. Dividends and earnings are 12-month moving sums for the S&P 500 index. The data is taken from Robert Shiller's website from 1871 to 1987. The dividends from 1988 to 2021 are provided by the S&P Corporation and the earnings data for this period are interpolated from quarterly earnings from the S&P Corporation. In the analysis, returns are defined as $R_{t+1} = \frac{P_{t+1}}{P_t}$ and the log excess return is $r_{t+1} = log\left(\frac{R_{t+1}}{R^f_{t+1}}\right)$.[11]

The stock variance is measured as a daily sum of daily returns on the S&P 500. In the book-to-market ratio, book values are from the Value Line website. The ratio is computed for the Dow Jones Industrial Average.

Measures of Predictability

A model specification that links market excess returns to lagged predictors, can be expressed as:

$$r_{t+1} = f_t(X_t) + u_{t+1} \tag{9}$$

where r_{t+1} is log market excess return at time $t+1$,[12] X_t is a matrix where column vectors are

[11] The monthly returns are computed as net dividends following GW. The yearly returns are computed as cum dividend returns so that $R_{t+1} = \frac{P_{t+1}+D_{t+1}}{P_t}$

[12] In DSR estimations, log excess market returns (instead of untransformed excess market returns) are

predictor variables as of time t, namely, prices ($X_{1,t}$), lagged prices ($X_{2,t}$), dividends ($X_{3,t}$), earnings ($X_{4,t}$), svar ($X_{5,t}$), and b/m ($X_{6,t}$). u_{t+1} is an idiosyncratic noise. In this case, $f_t(X_t)$ can be either a linear model including models estimated in GW or a model estimated by DSR.[13] For example, in the regression of returns on the log dividend-price ratio, GW estimations assume a linear relationship in the form of:

$$r_{t+1} = \alpha_t + \beta_t * log\left(\frac{D_t}{P_t}\right) + u_{t+1} \tag{10}$$

where $\hat{f}_t(D_t, P_t) = \hat{\alpha}_t + \hat{\beta}_t * log\left(\frac{D_t}{P_t}\right)$ is of a fixed functional form $f(D_t, P_t)$ and only coefficients $\hat{\alpha}_t$ and $\hat{\beta}_t$ are estimated. In DSR model estimations, functional forms themselves are assessed so that for each period $[1, t]$, DSR estimates $\hat{f}_t(X_t)$ and it can vary over time by design.

For the sample of length T[14], the specification (9) is estimated repeatedly in expanding windows $k = k_1, ..., k_{T-k_1}$ starting from an initial estimation period $t = k_1$ and where $k_{T-k_0} = T$. In-sample results are based on the full sample estimation of length $T - 1$.

In this study, I mainly focus on OOS predictions and related diagnostic statistics. This is primarily dictated by the DSR estimation method as it can always produce a functional form that will fit a given sample well enough or even perfectly. Using in-sample statistics becomes irrelevant in this case.

In order to evaluate the performance of a given model, I construct a measure of the cumulative difference between squared prediction errors of the proposed model and a null model. The difference between these errors allows to understand whether an investor would be able to time the market and gain returns above the market average. For a given estimation window k_i, I define ΔCSE_{k_i} as:

used for convenience only as it allows to directly compare the results to GW specifications. Using raw values would have no impact on DSR estimations.

[13] In a traditional sense, DSR estimates different models for each estimation window. I refer to the DSR model as a collection of best in-sample fitted equations over the expanding window for a specific cell of tuning parameters.

[14] The time frame here is defined relative to predictors so that the outcome variable r_{t+1} ranges from 2 to $T + 1$

$$\Delta CSE_{k_i} = CSE^N_{k_i} - CSE^f_{k_i} \tag{11}$$

where $CSE^f_{k_i}$ and $CSE^N_{k_i}$ are cumulative squared errors of excess return predictions. That is, for predictions $\hat{r}_{t+1} = \hat{f}_t(X_t)$ with $t \in [k_1, ..., k_i]$, CSE_{k_i} is calculated as:

$$CSE^f_{k_i} = \sum_{t=1}^{k_i} (r_{t+1} - \hat{r}_{t+1})^2 \tag{12}$$

The null model of historical mean excess return is based on average past returns defined as $\bar{r}_{t+1} = \frac{1}{t}\sum_{s=1}^{t} r_s$. Then, $CSE^N_{k_i}$ can be expressed as:

$$CSE^N_{k_i} = \sum_{t=1}^{k_i} (r_{t+1} - \bar{r}_{t+1})^2 \tag{13}$$

Prolonged periods with $\Delta CSE_{k_i} > 0$ would indicate that the investor would be able to profitably time the market had she used one of the alternative models.

Two-sided 95% confidence intervals for $CSE^f_{k_i}$ are constructed with a consistent estimate of the the asymptotic variance of the scaled mean following Diebold and Mariano (2002):

$$\hat{\Omega}_{k_i} = \frac{1}{k_i - k_1} * \sum_{s=k_1}^{k_i} \left[\Delta CSE_s - \overline{\Delta CSE_{k_i}}\right]^2 \tag{14}$$

where $\overline{\Delta CSE_{k_i}} = \frac{1}{k_i - k_1} * \sum_{s=k_1}^{k_i} \Delta CSE_s$. Similar to GW, the primary benchmark for comparing the OOS performance of a given model in this paper is the null model of past average excess market returns. I use the R^2OOS measure defined as:

$$R^2OOS = 1 - \frac{CSE^f_T}{CSE^N_T} \tag{15}$$

The R^2OOS compares the forecasting quality of the estimated model with the null model of the past mean excess return. Assuming that the past mean returns are at least some-

what informative (but not perfectly) of the future returns, these statistics can range from $-\infty$, signifying no predictability of the estimated model, to 1, indicating perfect return predictability. I test the equality of prediction errors between the conditional and null models for statistical significance with $MSE\text{-}F$ statistic advocated in McCracken (2007):

$$MSE\text{-}F = (T - k_1)\left(\frac{MSE^N - MSE^f}{MSE^f}\right) \tag{16}$$

Further, I estimate a goodness-of-fit static R^2IS for in-sample estimates over the full sample as:

$$R^2IS = 1 - \frac{\sum_{t=1}^{T}(r_{t+1} - \hat{f}_T(X_t))^2}{\sum_{t=1}^{T}(r_{t+1} - \bar{r}_T)^2} \tag{17}$$

with statistical significance levels based on a traditional F-statistic.[15]

I estimate non-overlapping 3-month horizon predictions with monthly data. The base case scenario starts with $k_1 = 240$. That is, with the first $t+1$ return data available in 1927, the first predictions are the month of the first quarter in 1947. The estimation generates 300 quarters of monthly excess return predictions.

Market Return Predictability

In the empirical part, I restrict the library of tokens in DSR estimations to $\{+, -, \times, \div, exp, log, \sqrt{}, \bullet^2, [vars], const, 1.0\}$. Compared to the dictionary in the simulation section, here I add a constant operator to account for possible fractional weights. This library also includes squaring operator $\bullet^2$ and a fixed value 1.0 that shrinks the length of traversals

[15]F-statistic and its distribution assume a linear estimation model and rely on degrees of freedom for building confidence intervals. To the best of my knowledge, at this point, there is no equivalent for F-statistics or degrees of freedom in the symbolic regression context. To make model estimates comparable, although somewhat arbitrary, I measure degrees of freedom in DSR models by the number of affine parameters. The degrees-of-freedom complexity in DSR models is also the reason why I rely on non-adjusted R^2 statistics in this paper.

for some functions.[16]

I start by comparing prediction errors of DSR models with various levels of the regularization parameter λ and using the linear model in equation (10) as a reference point.[17] Table 3 reports the IS and OOS R^2 for excess market return forecasts at a monthly frequency for 3-month non-overlapping horizons. As in the original GW paper, the linear model is not significant in IS and OOS estimations across specifications. DSR models seem to be able to identify equations with good in-sample feet when constrains are loose (large λ) based on R^2IS statistic. Columns 3 and 5 show that DSR models are able to generate meaningfully better OOS predictions with higher R^2 compared to the linear model giving the best result for the stringiest level of regularization parameter at $\lambda = 8$. Still, all DSR models appear to be under-performing relative to the null model producing negative $R2$ and low (negative) MSEF statistics. The performance seems to degrade for looser regularization parameters (larger λ). Large R^2IS together with very negative R^2OOS point to potential overfitting in DSR estimations when the regularization parameter weakens (λ large). Although these estimates generate negative R^2OOS, consistent with Kelly, Malamud, and K. Zhou (2022b), the timing Sharpe ratio is positive at 0.17 for the best-performing model. This is substantially exceeding the linear model Sharpe ratio of 0.04.

Forecasting errors can vary over time and model performance can change in different economic environments. Therefore, it is crucial to consider prediction results across the prediction window. Figure 3 plots cumulative SSE difference ΔCSE_{k_i} from 3-month predictions generated by the DSR and linear models setting the initial estimation period to 20 years. Both IS and OOS measures are presented. All DSR models experienced a substantial boost in terms of the in-sample fit during the Great Depression period. This potentially indicates that the model search is to a larger degree shaped by this crisis period.

As it can be seen from the top plot in Figure 3, for stronger levels of regularization parameter

[16] Eg. a dictionary with $\bullet^2$ operator needs two nodes to represent x^2, namely $\{\bullet^2, x\}$, while without this operator the traversal would need three nodes $\{\times, x, x\}$. Similarly, the fixed value 1.0 can be expressed as $\{\div, x, x\}$.

[17] Among four considered estimations in GW, in the current setup of 3-month predictions with monthly data, dividend-to-price ratio gives the highest $R2OOS$ compared to other specifications.

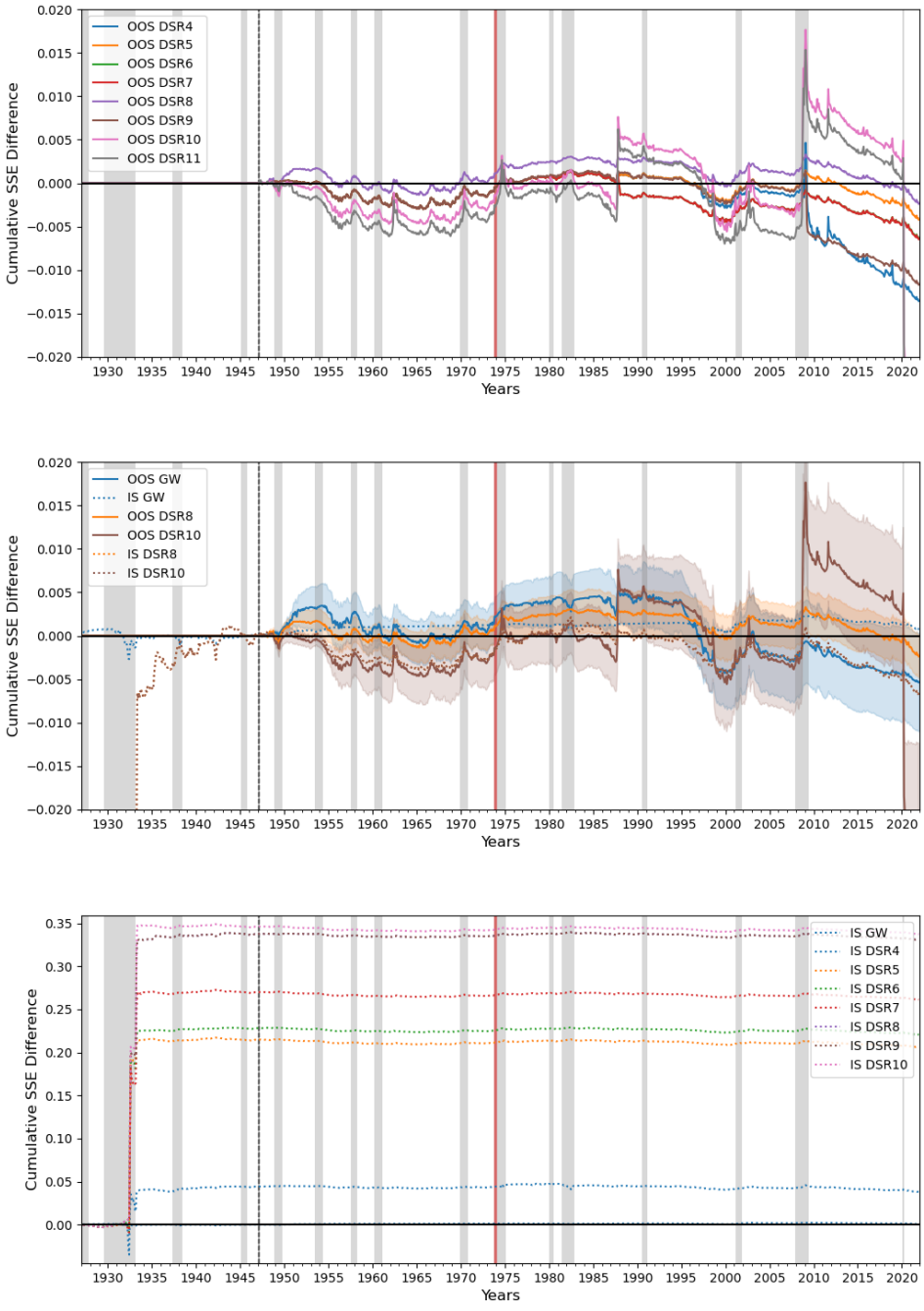
Table 3: DSR and linear models Forecasts

Table 3 presents IS and OOS R^2 for excess market return forecasts at a monthly frequency for 3-month non-overlapping horizons from GW and DSR models. Statistics for 3-month horizon predictions are based on 75-year prediction period. R^2 are in percentage terms. The IS R^2 are estimated over the full sample period. The OOS R^2 compares the forecast error of the estimated model with the forecast error of the past excess mean return. The sample runs from 1927 through 2021. Forecasts start after 20 from the beginning of the sample. Sharpe Ratio (SR) are annualized values based on monthly portfolio return estimates. Statistical significance is based on the F-statistic for IS estimates and the MSE–F statistic of McCracken (2007) for OOS estimates. ***, **, and * denote significance at the 1%, 5%, and 10% level, respectively. R^2 estimates for 3-month horizon prediction with a shorter prediction window are omitted as they are the same as with a longer prediction window.

Model	R^2IS	R^2OOS	MSEF	F-stat	SR
(1)	(2)	(3)	(4)	(5)	(6)
GW	0.01	-0.338	-3.036	0.109	0.042
DSR4	1.14	-0.856	-7.684	13.102	-0.078
DSR5	6.192	-0.265	-2.38	74.979	0.036
DSR6	6.625	-0.401	-3.603	80.603	-0.186
DSR7	7.849	-0.407	-3.656	96.766	-0.195
DSR8	9.936	-0.15	-1.35	125.318	0.171
DSR9	9.936	-0.737	-6.619	125.318	-0.193
DSR10	10.136	-1.398	-12.551	128.126	0.018
DSR11	9.937	-1.543	-13.852	125.342	0.009
DSR12	11.329	-2.103	-18.888	145.14	0.01
DSR13	11.628	-3.738	-33.568	149.469	-0.083
DSR14	11.495	-9.417	-84.567	147.548	-0.056
DSR15	12.466	-3.614	-32.45	161.776	-0.031

Figure 3: OOS Performance of GW and DSR Models

Figure 3 shows the monthly performance of DSR models relative to the linear model as specified in equation (10) with dividends and prices as predictors. The figure covers 75 years of predictions with 20 years of initial training period. The outcome in all models is the montly excess market return. Colored lines show the differences in errors in 3-month ahead predictions of estimated models relative to the null model of past mean excess returns based on ΔCSE_{k_i} in equation (11). A given model does better than the null model if the line moves up and vice versa. IS and OOS predictions from the linear model are the same across all three plots. Solid lines reflect OOS estimates while dotted lines show IS estimates. confidence intervals for OOS estimates are based on Diebold and Mariano (1995) as specified in equation (14). Top-to-bottom plots present: OOS estimates for all estimated DSR models with different constraints for λ between 4 and 11; IS and OOS estimates of two best-performing DSR models with confidence intervals and a comparison to a linear model; IS estimates for all estimated DSR models with different constraints for λ between 4 and 11. Greyed sections are NBER recessions and the red line reflects the 1974 Oil Crisis.



$\lambda \in \{4, 5\}$ that enforce relatively simple expressions, DSR is not able to identify specifications that produce positive predictions throughout most of the forecasting period compared to the linear model. While several regularization levels allow DSR to produce specifications that outperform the linear model in OOS predictions, the model with a relatively stronger regularization parameter ($\lambda = 8$) demonstrates better results up until the Great Recession. Notably, models of higher complexity ($\lambda = 10$) seem to substantially outperform during the Great Recession. This can be seen as a sign of the shared complexity of data generating process being time-varying.

Although more complex models performed particularly well during the Great Recession period, they sharply underperformed in Covid-19 times. This evidence points to the ability of DSR to capture changes in the dynamics between returns and predictors even in such a simple and restrictive setup. This also indicates that, potentially, the DSR method can effectively track changes in the average level of the economy during and following large economic disturbances, which is a difficult task as documented by Lettau and Van Nieuwerburgh (2008b). Further, the ability to forecast excess returns using models derived from DSR appears to be highly localized in time and predominantly occurs within specific, adjacent 'pockets', corroborating the results in Farmer, L. Schmidt, and Timmermann (2023). In this case, the 'pockets' of predictability seem to occur mostly, but not universally, during or just after recessions as defined by NBER.

Overall, relaxing the regularization parameter leads to DSR estimating more complex models that, although performing well in the sample, demonstrate poor out-of-sample results with occasional improbable predictions that lead to substantial jumps in ΔCSE_{k_i}.

It has been shown that random forest (RF) tends to outperform other ML techniques in prediction tasks (Mullainathan and Spiess 2017) including predictions of market returns (Gu, Kelly, and Xiu 2020). I use RF as a state-of-the-art ML benchmark method for market excess return predictability. The comparison with the tuned random forest model is shown in Figure 4. Similar to the DSR setup, the RF algorithm is given raw values of dividends and prices and it is asked to estimate the best fit of log excess market returns. 3-month OOS

predictions are then generated. RF is tuned along the tree depth and impurity parameters. Figure 4 demonstrates that even after tuning, RF substantially underperforms relative to the null hypothesis compared to the linear model or DSR8 models consistently generating inferior OOS predictions though-out the prediction horizon and over-fitting in-sample.

The model structure in DSR estimations is flexible and it can change as the estimation window increases. Therefore, examining how the equations in these models evolve over time can help us understand how they generate out-of-sample forecasts of excess returns. Table 4 and Figure 5 document equations identified by best performing DSR models based OOS predictions for corresponding prediction years for the constraints $\lambda \in \{8, 10\}$ in 3-month prediction window estimations. The table shows that although there is some variation in the model structure, the estimates overall remain stable over time with a dominant model capturing 99% and 88% of the prediction periods correspondingly. The log excess return appears to be better predicted by exponential predictions and not the log of the dividend-to-price ratio or logs of other variables as traditionally considered in the finance literature, both in-sample and out-of-sample. The dominant model is very similar in both cases with a more complex one only adding a linear term of *svar*.

In order to further explore the performance of discovered models, I generate predictions based on the top models fixing estimated parameters and compare them to the in-sample linear model.[18] As before, these models appear to predict well during periods of substantial economic distress, especially during the Great Depression and the Great Recession. More complex models exhibit higher volatility in terms of predictions fitting data well during financial crises but also underperforming over other periods. At the same time, added *svar* term causes these more complex models to miss the market movements during Covid-19. Given that the only difference between eq.1 for $\lambda = 8$ and eq.1 for $\lambda = 10$ is the linear *svar* term, the observed differences in the performance of more complex models can solely be attributed to the stock variance.

[18]It should be noted that for DSR predictions, cannot be treated as OOS estimates since the model and its parameters are taken ex-post after observing all predictions and best model from the whole sample.

Table 4: DSR Estimated Equations

Table 4 lists equations identified by best OOS performing DSR models in expanding window estimations for 3-month non-overlapping horizons from monthly estimates. The dependent variable is a monthly log excess return defined as $r^*_{t+1} = log\left(\frac{(R_{t+1}}{R^f_{t+1}}\right)$ where $R_{t+1} = \frac{P_{t+1}}{P_t}$. The set of RHS variables are: prices (X_1); lagged prices (X_2); dividends (X_3); earnings (X_4); svar (X_5); b/m (X_6).

$\lambda = 8$

#	Equation	Prediction Quarters	Total Quarters
1.	$e^{\frac{2.5-x_2\times x_3}{x_3\times x_4}}$	1947Q1-1989Q2, 1990Q2-2021Q4	297
2.	$e^{\frac{4.9-x_2\times x_3}{x_3}}$	1989Q4	1
3.	$\frac{6e^{-x_2}}{x_3-0.4}$	1990Q1	1
4.	$(181.5-242.9x_3)\times e^{-x_2}$	1990Q2	1

$\lambda = 10$

#	Equation	Prediction Quarters	Total Quarters
1.	$e^{\frac{2.5-x_2\times x_3}{x_3\times x_4}} - x_5$	1947Q1-1961Q4, 1962Q2-1964Q2, 1964Q4-1967Q4, 1968Q2-1980Q2, 1980Q4-1985Q1, 1985Q3-1988Q3, 1989Q1-Q4, 21991Q1-Q4,1992Q2-1996Q1, 1996Q3-2001Q3, 2020Q1, 2002Q4-2003Q2, 2004Q2-2007Q3, 2004Q2-2007Q3, 2008Q1-Q2, 2008Q4-2009Q1, 2009Q4-2010Q2, 2010Q4-2011Q3, 2012Q1-2014Q2, 2014Q4-2015Q4, 2016Q3-Q4, 2017Q2-2019Q2, 2019Q4-2020Q1, 2021Q2	265
2.	$2\times 10^{-4}e^{\frac{9.3-x_2}{x_3}} - x_5$	1962Q1, 1964Q3, 1980Q3, 1988Q4, 1990Q3, 1992Q1, 1996Q2, 2001Q4, 2002Q2-2002Q3, 2003Q3-Q4, 2004Q1, 2007Q4, 2008Q3, 2009Q2, 2010Q3, 2011Q4, 2014Q3, 2016Q1-Q2, 2017Q1, 2019Q3	23
3.	$2.52\times 10^{-5}x_5\times e^{\frac{11-x_2}{x_4}}$	2020Q2-2021Q1, 2021Q3-2021Q4	6
4.	$e^{\frac{5}{x_3}-2x_2} - x_5$	1968Q1, 1990Q4	2
5.	$x_5\times\left(e^{\frac{3.38}{x_4}-x_2} - 1.53\right)$	1985Q2	1
6.	$\frac{x_2}{(x_2-6.16)e^{x_2}+160.56}$	1990Q1	1
7.	$2.44*10^5\times e^{-1.70x_2-9.31x_3}$	1990Q2	1
8.	$7\times 10^{-6}e^{\frac{29.86}{x_2x_3}} - x_5$	2009Q3	1

Figure 4: OOS Performance of GW, DSR, and RF Models for Dividends and Prices

Figure 4 shows the annual performance of the DSR model relative to the OLS model in GW and a tuned Random Forest model. The outcome in both models is the annual excess market return. Lines show the differences in errors in one-year ahead prediction of past mean and corresponding models. A given model does better than the mean if the line moves up and vice versa. Top-to-bottom plots present the DSR model as a constraint for $\lambda = 8$. Greyed sections are NBER recessions and the red line reflects the 1974 Oil Crisis.

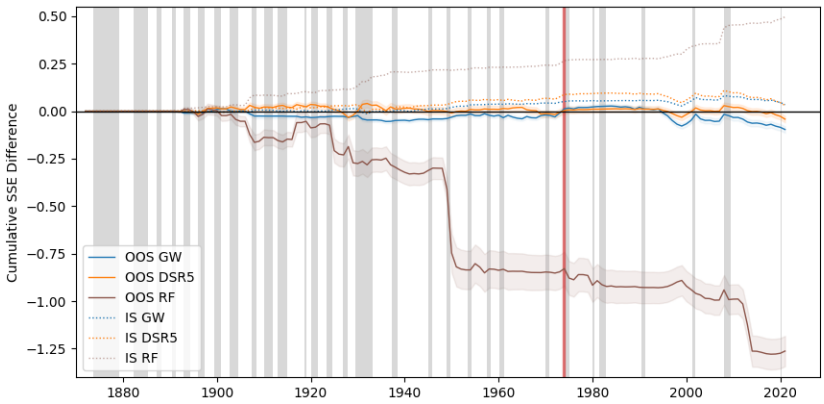


Figure 5: Prediction Periods of DSR10 Generated Models

Figure 5 show the OOS performance of the DSR models with $\lambda = 10$ and the prediction periods generated by the models in the Table 4. The figure covers 75 years of predictions with 20 years of initial training period. The outcome variable is the monthly excess market return. The orange line shows the difference in errors in 3-month ahead predictions of estimated DSR10 models relative to the null model of past mean excess returns based on ΔCSE_{k_i} in equation (11). A given model does better than the null model if the line moves up and vice versa. Light-colored sections labeled Eq1-8 reflect prediction periods generated by models estimated with DSR for $\lambda = 10$ as enumerated in Table 4. Greyed sections are NBER recessions and the red line reflects the 1974 Oil Crisis.

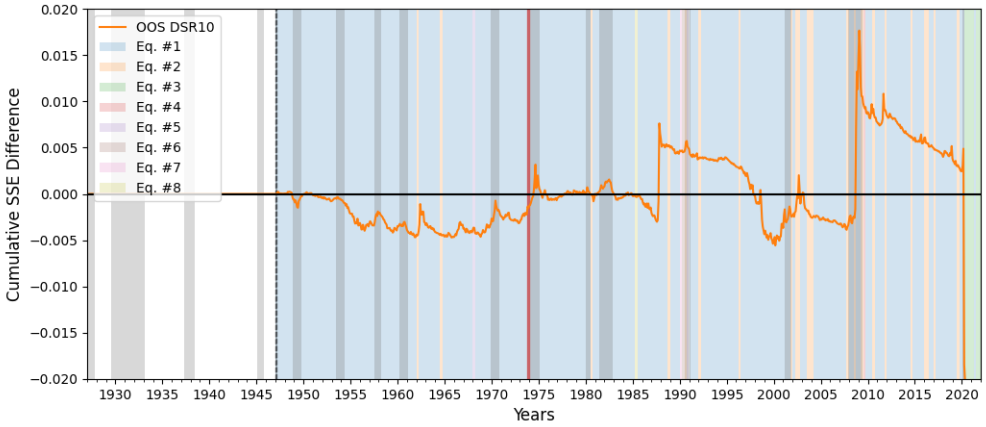
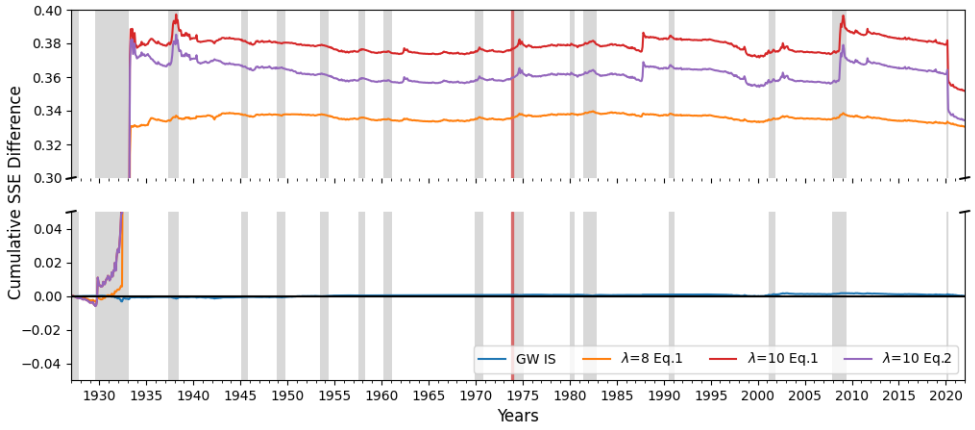


Figure 6: Preidctions of Top DSR models

Figure 6 show the performance of the most frequent DSR models with $\lambda \in [8, 10]$ relative to the in-sample GW estimate. The figure covers 75 years of predictions with 20 years of initial training period. The outcome variable is the monthly excess market return. The lines show the difference in errors in 3-month ahead predictions of estimated by DSR8 or DSR10 models relative to the null model of past mean excess returns based on ΔCSE_{k_i} in equation (11). The most frequent models are the equation #1 for $\lambda = 8$ and the equations #1 and #2 for $\lambda = 10$ in Table 4. A given model does better than the null model if the line moves up and vice versa. Greyed sections are NBER recessions and the red line reflects the 1974 Oil Crisis.



Conclusion

This paper has demonstrated the potential of symbolic regression, and DSR in particular, as a novel and powerful method for predicting market return based on a simple setting of the present value relationship. The paper has applied symbolic regression to both simulated and real data from the US stock market and compared its performance with linear regression and random forest as benchmark models.

It has been shown in a simulation exercise that symbolic regression can recover the true functional dependencies and the level of predictability, even in noisy environments. It can also handle complex data and find hidden patterns and interactions among predictors. In a real-world test, DSR outperforms linear regression and random forest in terms of out-of-sample forecasting accuracy and in-sample fit for real data from the US stock market over a sample period from 1927 to 2021. At the same time, DSR generates interpretable and parsimonious models that capture some well-known stylized facts as well as some novel nonlinearities and interactions in market return predictability.

In particular, DSR appears to be capable of generating meaningfully superior OOS predictions during large economic events like the Great Depression suggesting convoluted underlying relationships in the economy. These findings merit further investigation of the higher order approximations beyond the log-linearization of Campbell and Shiller (1988) traditionally used in the literature.

Further, a natural avenue for further research is to employ symbolic regressions in the cross-sectional asset pricing context. Unlike any other ML methods used in economics and finance literature, symbolic regressions are aimed at explaining underlying patterns in the data with succinct equations. This means that this methodology suits well the task of describing and facilitating the understanding of differences in expected returns across assets.

The paper also acknowledges some limitations of this methodology. Symbolic regression is computationally intensive and requires careful tuning of hyperparameters to produce optimal results. The space of possible functional forms that DSR searches through is predetermined by the dictionary ex-ante. If the true data-generating process is based on transformations that are not in the library, the generated equations will still be only an approximation to the true DSP. Further, given the flexibility of all possible functional forms in the span of a given dictionary, DSR tends to overfit in-sample producing poor OOS predictions unless the sample is sufficiently large. Introducing economically motivated priors to the cost function to enhance the regularization parameter by penalizing equations that produce unrealistic forecasts based on the range of the outcome and predictor variables and experimenting with the extended library to include other operations is a reasonable avenue to pursue in the future.

Appendix

Table A1: Convergence Rates of DSR estimates to True Functional Forms. (True $R^2 = 30\%$)

Table A2 reports mean rates of the exact convergence of DSR to true functional forms with simple and complex and DGPs. The exact convergence is defined as the perfect structural match of the equation estimated by DSR to the true equation behind a specific DGP. Convergence rates are on a 0 to 1 scale. All samples have 30% signal in terms of R^2. Sample sizes vary from 50 to 10,000 obs. In each specification cell, there are 101 samples for DSR model estimation. The regularization parameter can be one of the values $\lambda \in [5, 7, 8, 10, 15]$.

Panel A: Simple DGP

Panel A shows DSR convergence rates for simulations of DGP 1. The underlying true equation is $Y = X_1 + X_2 + X_1 * X_2 + \epsilon$. The minimum length of the traversal is 7.

SampleSize	5	7	8	10	15
50	0.000	0.079	0.000	0.000	0.000
100	0.000	0.198	0.020	0.000	0.000
250	0.000	0.693	0.119	0.010	0.000
1000	0.000	0.990	0.663	0.188	0.010
10000	0.000	1.000	1.000	0.931	0.257

Panel B: Complex DGP

Panel B shows DSR convergence rates for simulations of DGP 1. The underlying true equation is $Y = \log\left(\frac{X_1}{X_2}\right) + X_1^2 + \epsilon$. The minimum length of the traversal is 8.

SampleSize	5	7	8	10	15
50	0.000	0.000	0.168	0.000	0.000
100	0.000	0.000	0.238	0.020	0.000
250	0.000	0.000	0.614	0.149	0.000
500	0.000	0.000	0.733	0.287	0.000
1000	0.000	0.000	0.842	0.505	0.000
10000	0.000	0.000	1.000	0.990	0.050

Table A2: Summary Statistics of Mean R^2OOS in DSR Simulations (True $R^2 = 30\%$)

Table A2 reports summary statistics of average R^2OOS in OOS predictions by DSR models estimated over simulated data generated by simple and complex DGPs. All samples have 30% signal in terms of true R^2. Sample size varies from 50 to 10,000 obs. In each specification cell, there are 101 samples for DSR model estimation. Models estimated on one sample are then tested on the remaining 100 samples. Mean R^2OOS is obtained by averaging R^2 from test samples. The regularization parameter can be one of the values $\lambda \in [5, 7, 8, 10, 15]$. Extreme values of $R^2 < -100$ are truncated. It affects $< 0.5\%$ of obs.

Panel A: Simple DGP

Panel A reports summary statistics for simulations of DGP 1. The underlying true equation is $Y = X_1 + X_2 + X_1 * X_2 + \epsilon$.

	Max Complexity				
Sample Size	**5**	**7**	**8**	**10**	**15**
50 (n=505)	0.23	0.02	0.08	0.09	-0.39
(SD)	(0.01)	(0.10)	(0.07)	(0.05)	(0.19)
(95% CI)	[0.22,0.24]	[-0.17,0.21]	[-0.06,0.22]	[-0.01,0.19]	[-0.75,-0.02]
100 (n=505)	0.25	0.21	0.25	0.23	0.09
(SD)	(0.00)	(0.03)	(0.00)	(0.01)	(0.09)
(95% CI)	[0.24,0.26]	[0.15,0.28]	[0.24,0.26]	[0.21,0.25]	[-0.08,0.26]
250 (n=505)	0.27	0.28	0.28	0.26	0.25
(SD)	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)
(95% CI)	[0.27,0.27]	[0.28,0.29]	[0.28,0.29]	[0.22,0.30]	[0.21,0.30]
1000 (n=505)	0.27	0.30	0.29	0.29	0.26
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
(95% CI)	[0.27,0.27]	[0.30,0.30]	[0.29,0.30]	[0.29,0.29]	[0.22,0.30]
10000 (n=505)	0.27	0.30	0.30	0.30	0.30
(SD)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Panel B: Complex DGP

Panel B of reports summary statistics for simulations of DGP 2. The underlying true equation is $Y = \log\left(\frac{X_1}{X_2}\right) + X_1^2 + \epsilon$.

	Max Complexity				
Sample Size	**5**	**7**	**8**	**10**	**15**
50 (n=505)	0.09	-0.05	-0.21	-0.25	-0.56
(SE)	(0.10)	(0.13)	(0.15)	(0.15)	(0.15)
(95% CI)	[-0.10,0.29]	[-0.30,0.21]	[-0.50,0.09]	[-0.54,0.04]	[-0.86,-0.26]
100 (n=505)	0.21	0.21	0.08	0.10	0.02
(SE)	(0.01)	(0.01)	(0.08)	(0.05)	(0.07)
(95% CI)	[0.19,0.22]	[0.19,0.24]	[-0.08,0.25]	[0.00,0.21]	[-0.12,0.16]
250 (n=505)	0.25	0.28	0.28	0.27	0.26
(SE)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
(95% CI)	[0.24,0.25]	[0.28,0.29]	[0.27,0.29]	[0.26,0.28]	[0.23,0.28]
500 (n=505)	0.26	0.29	0.29	0.29	0.29
(SE)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
(95% CI)	[0.26,0.26]	[0.29,0.29]	[0.29,0.30]	[0.29,0.29]	[0.29,0.29]
1000 (n=505)	0.26	0.30	0.30	0.30	0.30
(SE)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
(95% CI)	[0.26,0.26]	[0.29,0.30]	[0.30,0.30]	[0.30,0.30]	[0.30,0.30]
10000 (n=505)	0.26	0.30	0.30	0.30	0.30
(SE)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Figure A1: R^2IS and Average R^2OOS of DSR Estimates of DGP 1 (True $R^2 = 5\%$)

Figure A1 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 1 for samples sizes of 50 and 1000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 7, 15$ that referred to as Low, True, and High correspondingly. All samples have 5% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

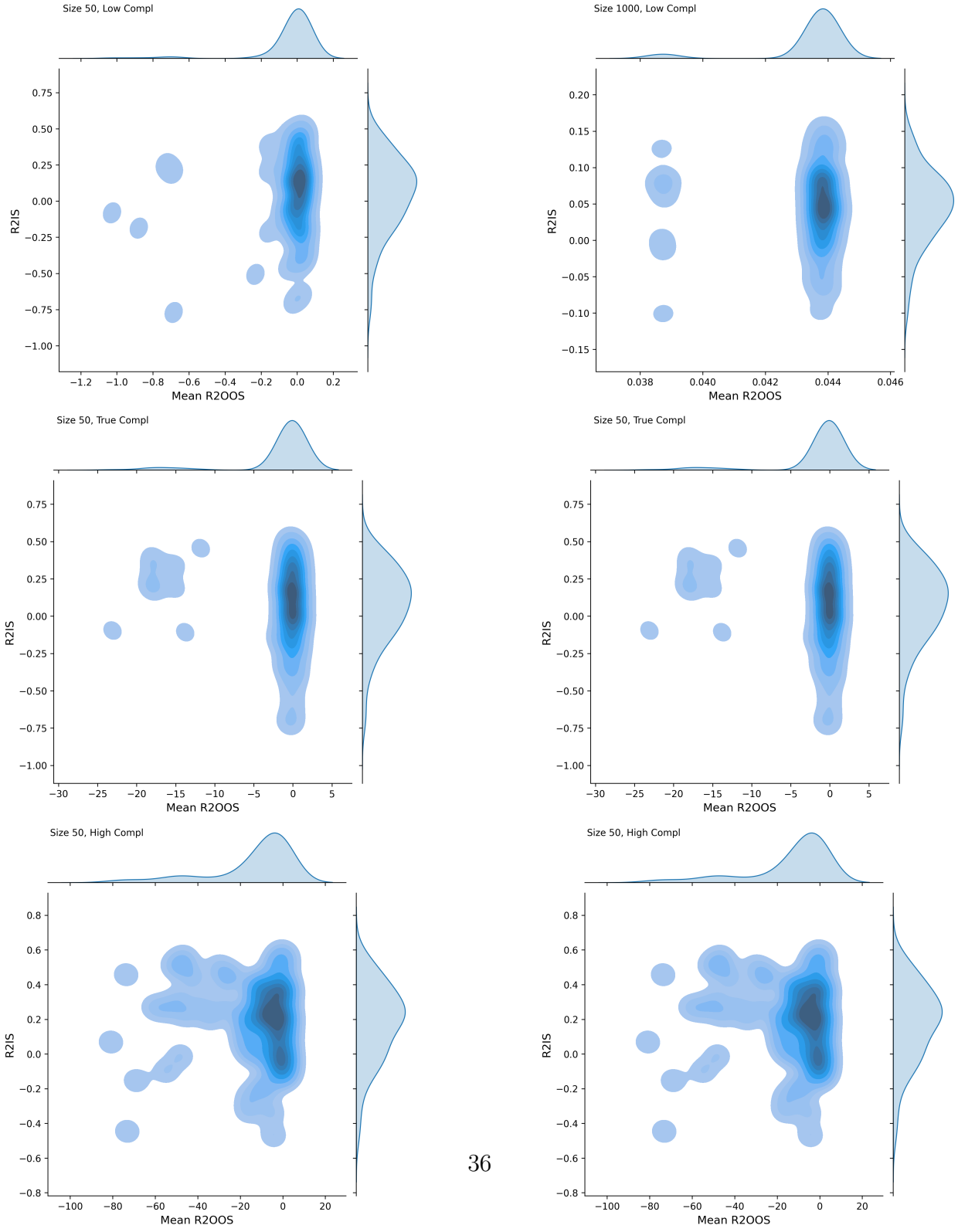


Figure A2: R^2IS and Average R^2OOS of DSR Estimates of DGP 2 (True $R^2 = 5\%$)

Figure A2 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 2 for samples sizes of 50 and 1000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 8, 15$ that referred to as Low, True, and High correspondingly. All samples have 5% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

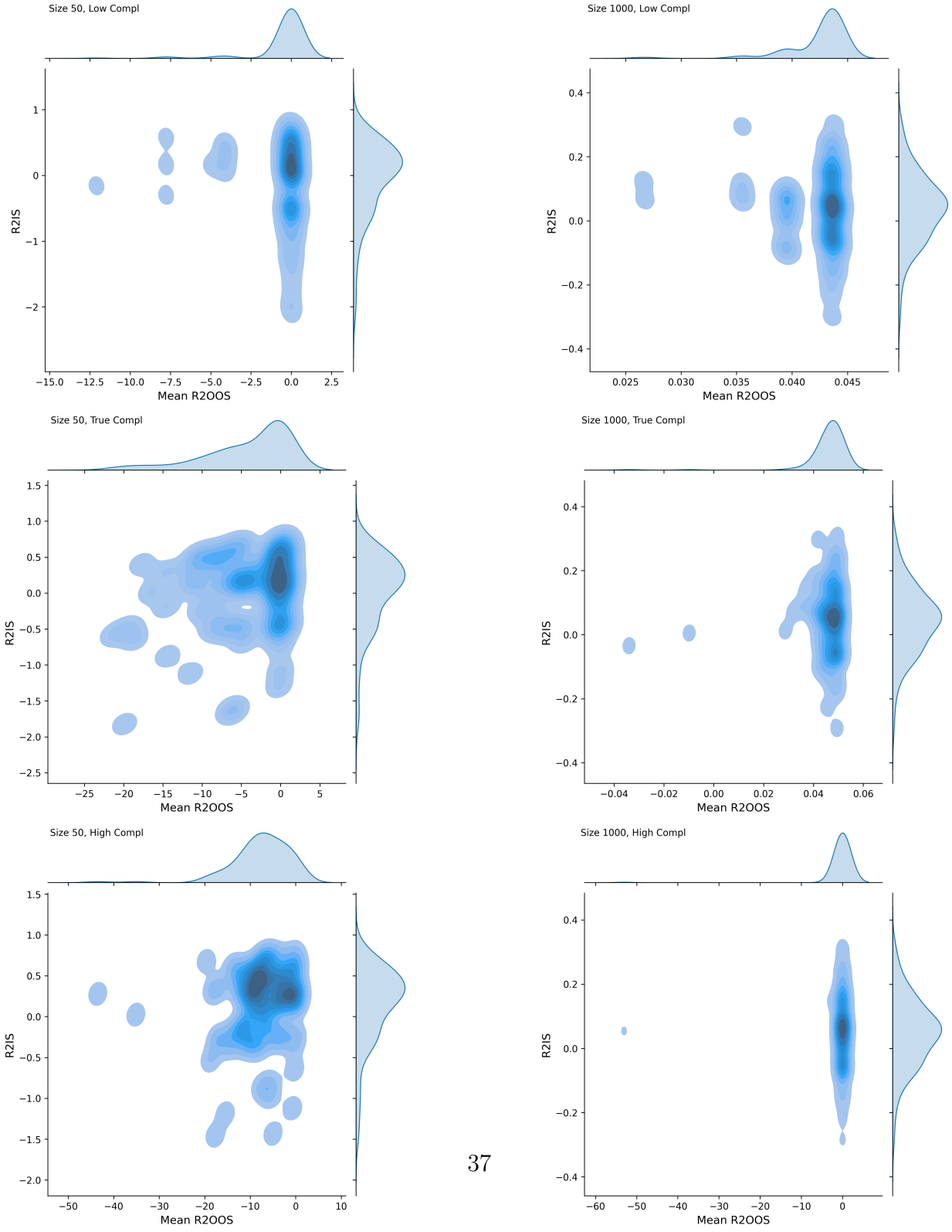


Figure A3: R^2IS and Average R^2OOS of DSR Estimates of DGP 1 (True $R^2 = 30\%$)

Figure A3 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 1 for samples sizes of 50 and 1000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 7, 15$ that referred to as Low, True, and High correspondingly. All samples have 30% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

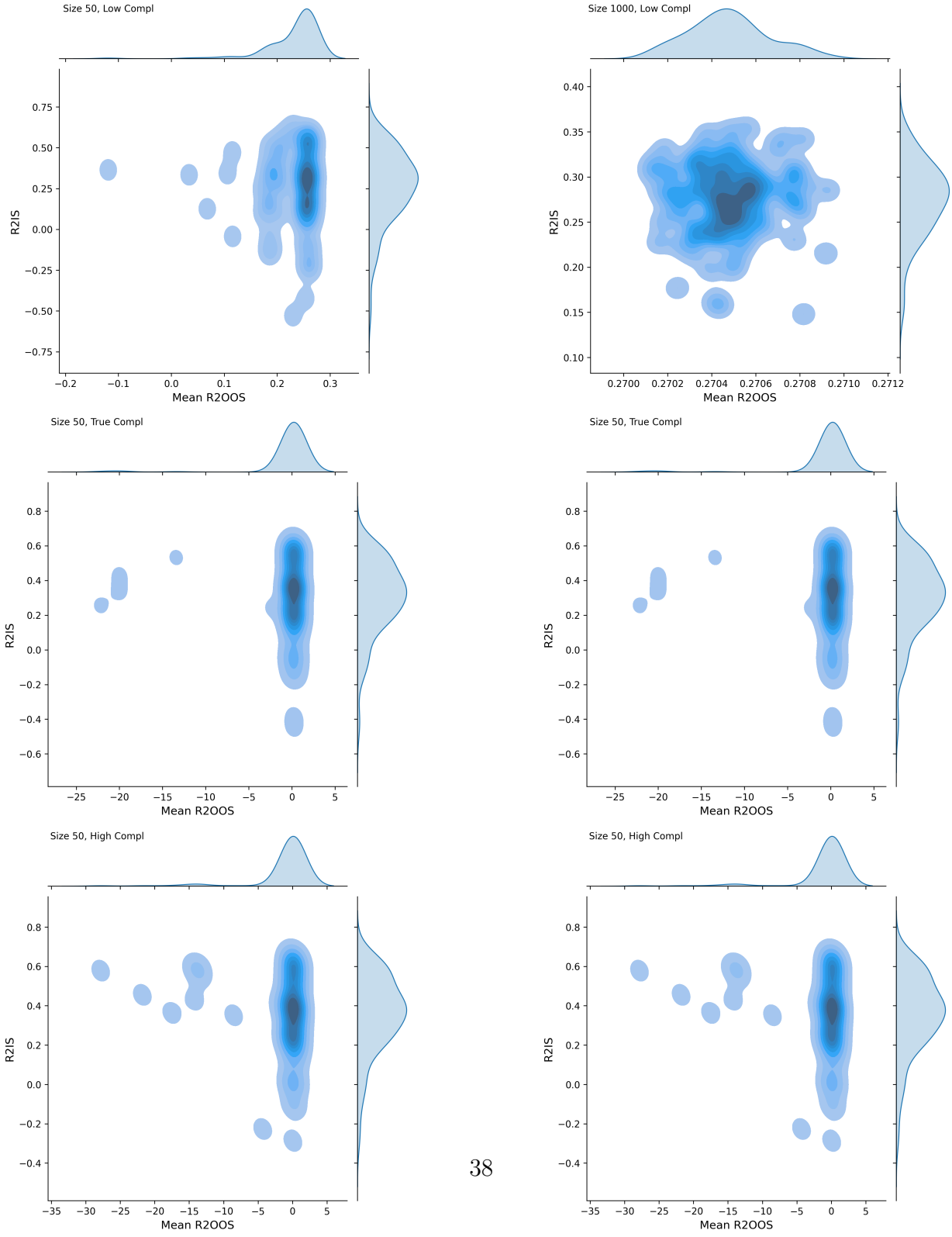


Figure A4: R^2IS and Average R^2OOS of DSR Estimates of DGP 2 (True $R^2 = 30\%$)

Figure A4 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 2 for samples sizes of 50 and 1000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 8, 15$ that referred to as Low, True, and High correspondingly. All samples have 30% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

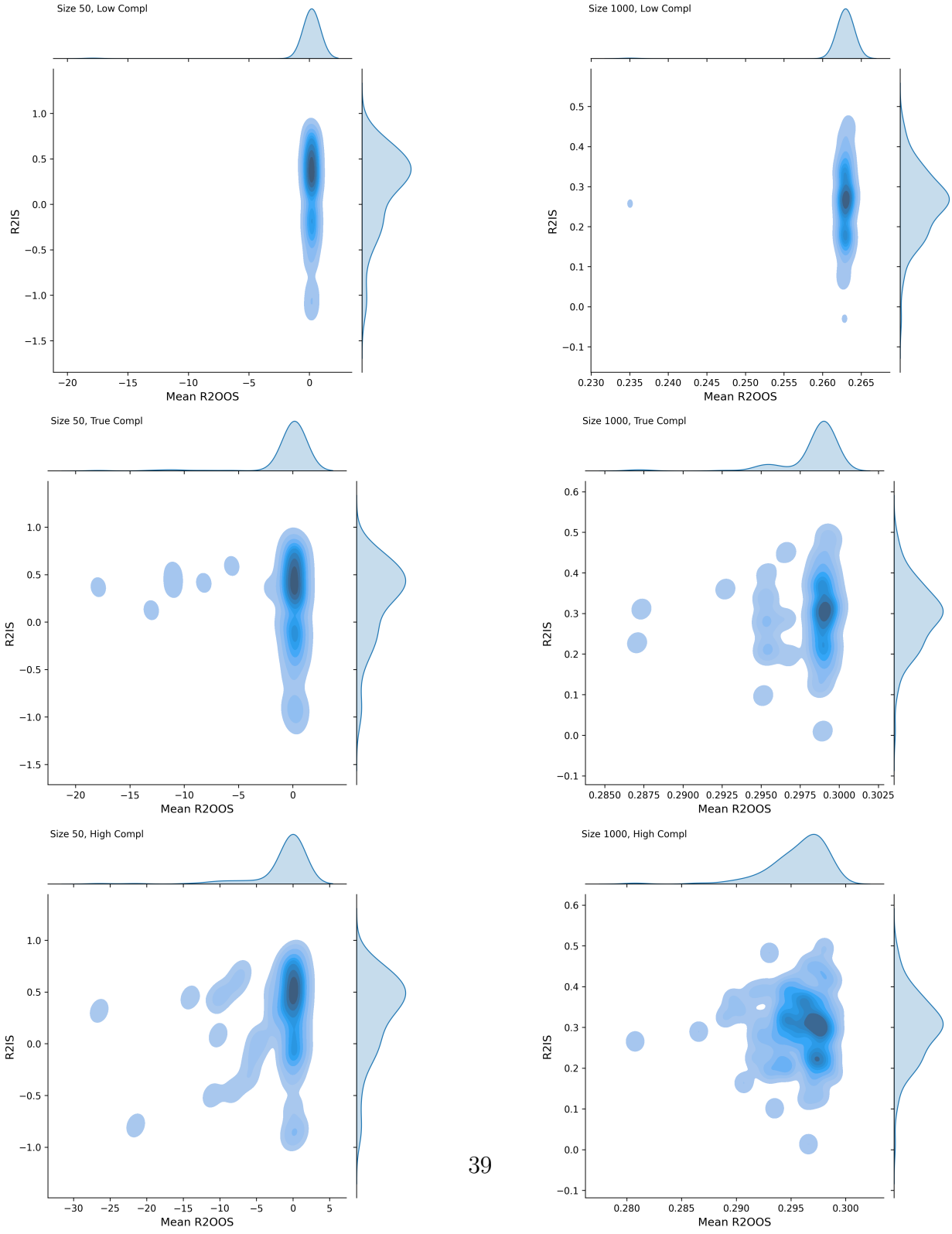


Figure A5: R^2IS and Average R^2OOS of DSR Estimates of DGP 1 (True $R^2 = 30\%$)

Figure A5 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 1 for samples sizes of 100 and 10,000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 7, 15$ that referred to as Low, True, and High correspondingly. All samples have 30% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.

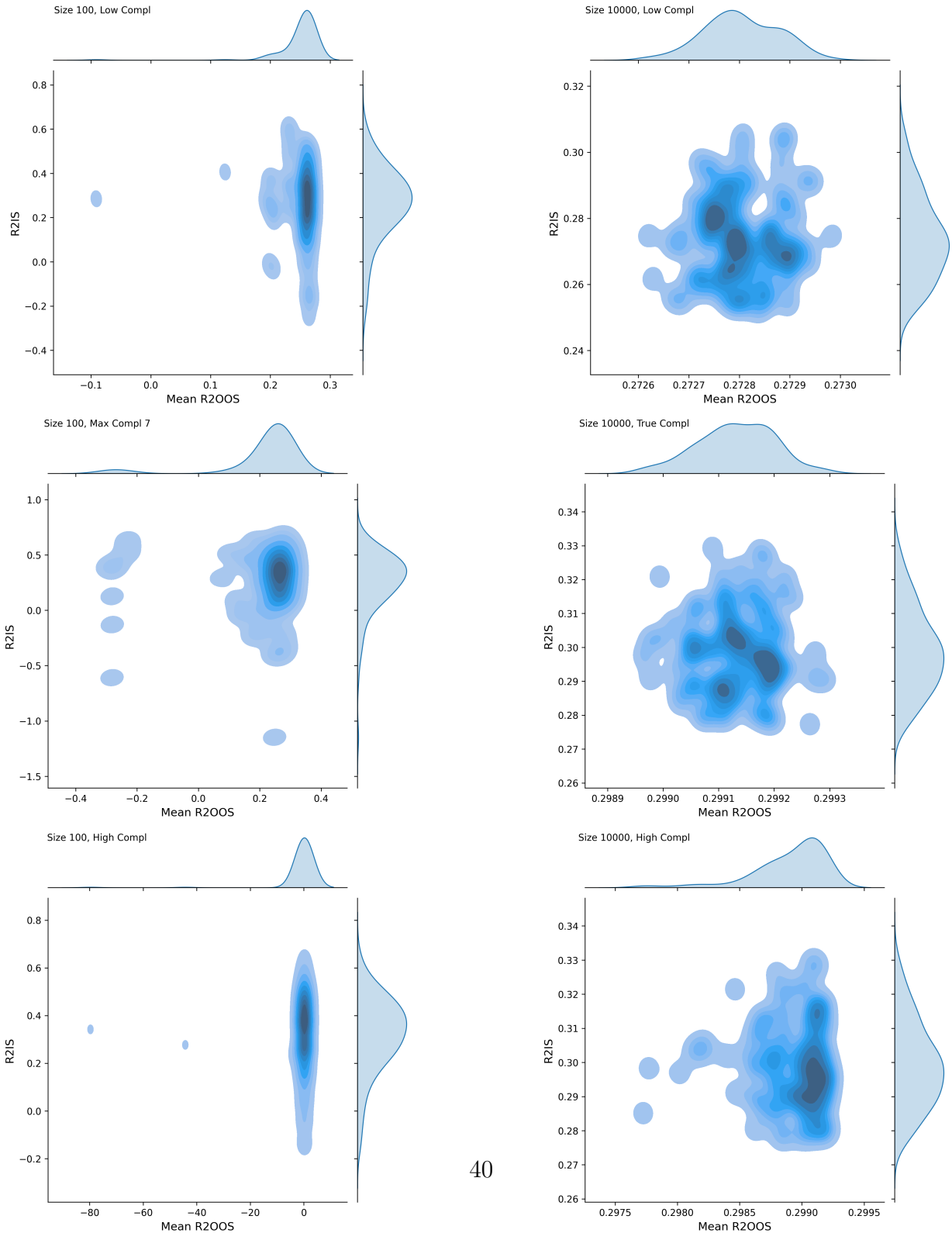
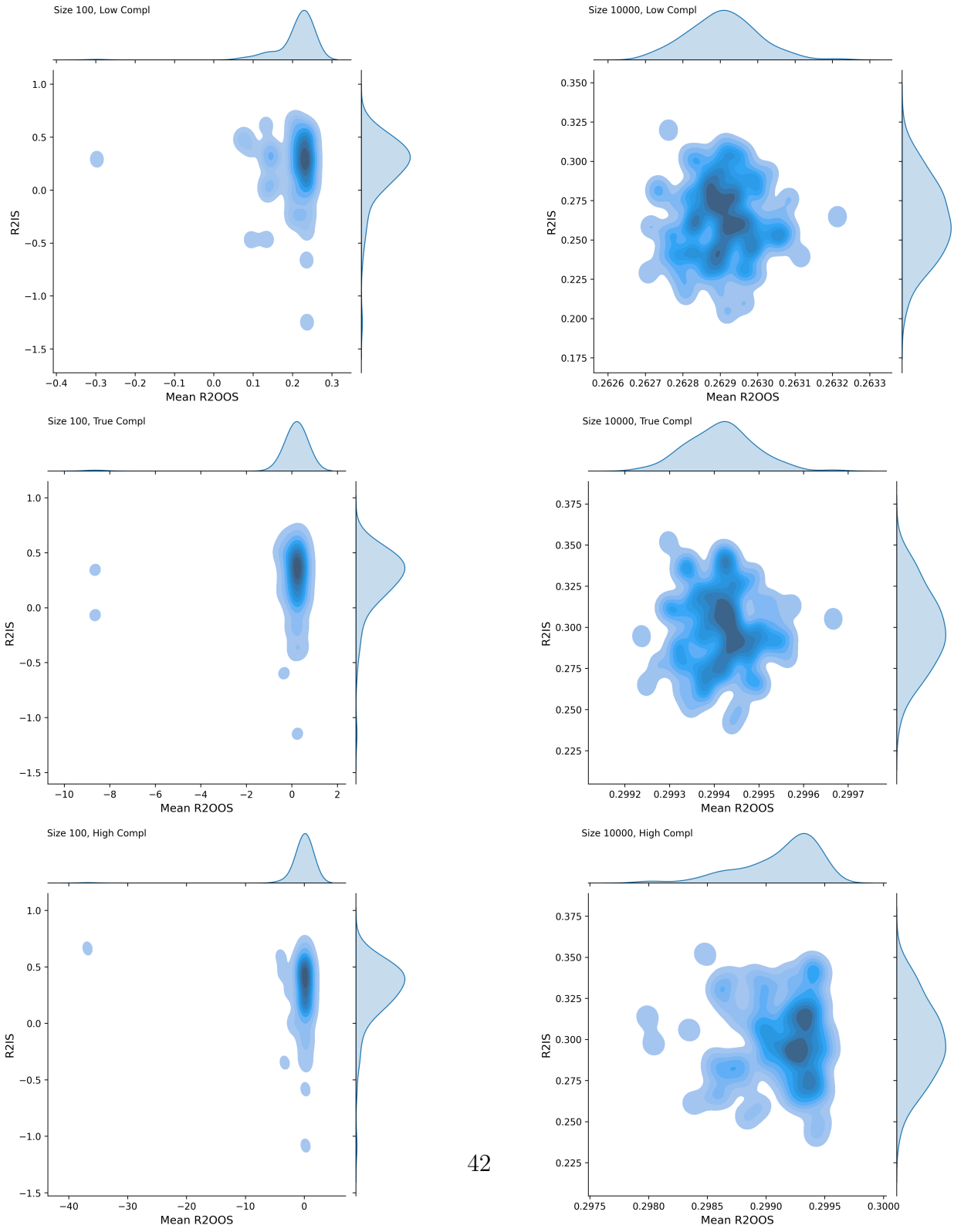


Figure A6: R^2IS and Average R^2OOS of DSR Estimates of DGP 2 (True $R^2 = 30\%$)

Figure A6 plots 2D densities of R^2IS and average R^2OOS measures for DSR model that are estimated on simulated data based on DGP 2 for samples sizes of 100 and 10,000 and 3 levels of DSR maximum complexity with $\lambda \in 5, 8, 15$ that referred to as Low, True, and High correspondingly. All samples have 30% signal in terms of true R^2. Each estimation cell in the grid of sample sizes and λ levels contains 101 random samples. For each DSR model estimated on a specific sample, mean R^2OOS is obtained by average R^2 estimates from predictions on the remaining 100 samples. Axis scales are different across plots. Extreme values of $R^2 < -100$ are winsorized. It affects $< 0.5\%$ of obs in the most demanding specifications.



References

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). "A convergence theory for deep learning via over-parameterization". *International conference on machine learning*. PMLR, 242–252.

Alvarez-Diaz, Marcos and Alberto Alvarez (Apr. 2003). "Forecasting exchange rates using genetic algorithms". *Applied Economics Letters* 10.6. Publisher: Routledge _eprint: https://doi.org/10.1080/1350485021015825O, 319–322. ISSN: 1350-4851. DOI: 10.1080/13504850210158250. URL: https://doi.org/10.1080/13504850210158250 (visited on 03/31/2023).

Bew, David et al. (Jan. 2019). "Modeling Analysts' Recommendations via Bayesian Machine Learning". en. *The Journal of Financial Data Science* 1.1. Publisher: Institutional Investor Journals Umbrella, 75–98. ISSN: 2640-3943, 2640-3951. DOI: 10.3905/jfds.2019.1.1.075. URL: https://jfds.pm-research.com/content/1/1/75 (visited on 03/30/2023).

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (Feb. 2021). "Bond Risk Premiums with Machine Learning". *The Review of Financial Studies* 34.2, 1046–1089. ISSN: 0893-9454. DOI: 10.1093/rfs/hhaa062. URL: https://doi.org/10.1093/rfs/hhaa062 (visited on 03/15/2023).

Binsbergen, Jules H. van, Xiao Han, and Alejandro Lopez-Lira (Sept. 2020). *Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases*. Working Paper. DOI: 10.3386/w27843. URL: https://www.nber.org/papers/w27843 (visited on 03/31/2023).

Campbell, John Y. and Robert J. Shiller (July 1988). "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors". *The Review of Financial Studies* 1.3, 195–228. ISSN: 0893-9454. DOI: 10.1093/rfs/1.3.195. URL: https://doi.org/10.1093/rfs/1.3.195 (visited on 03/21/2023).

Campbell, John Y. and Samuel B. Thompson (July 2008). "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *The Review of Financial Studies* 21.4, 1509–1531. ISSN: 0893-9454. DOI: 10.1093/rfs/hhm055. URL: https://doi.org/10.1093/rfs/hhm055 (visited on 03/15/2023).

Chen, Luyang, Markus Pelger, and Jason Zhu (Feb. 2023). "Deep Learning in Asset Pricing". *Management Science*. Publisher: INFORMS. ISSN: 0025-1909. DOI: 10.1287/mnsc.2023.4695. URL: https://pubsonline.informs.org/doi/full/10.1287/mnsc.2023.4695 (visited on 03/15/2023).

Claveria, Oscar, Enric Monte, and Salvador Torra (May 2017). "Assessment of the effect of the financial crisis on agents' expectations through symbolic regression". *Applied Economics Letters* 24.9. Publisher: Routledge _eprint: https://doi.org/10.1080/13504851.2016.1218419, 648–652. ISSN: 1350-4851. DOI: 10.1080/13504851.2016.1218419. URL: https://doi.org/10.1080/13504851.2016.1218419 (visited on 03/31/2023).

– (Jan. 2018). "A Data-Driven Approach to Construct Survey-Based Indicators by Means of Evolutionary Algorithms". en. *Social Indicators Research* 135.1, 1–14. ISSN: 1573-0921. DOI: 10.1007/s11205-016-1490-3. URL: https://doi.org/10.1007/s11205-016-1490-3 (visited on 03/31/2023).

– (Jan. 2022). "A Genetic Programming Approach for Economic Forecasting with Survey Expectations". en. *Applied Sciences* 12.13. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute, 6661. ISSN: 2076-3417. DOI: 10.3390/app12136661. URL: https://www.mdpi.com/2076-3417/12/13/6661 (visited on 03/31/2023).

Cochrane, John H. (2011). "Presidential Address: Discount Rates". en. *The Journal of Finance* 66.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2011.01671.x, 1047–1108. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2011.01671.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x (visited on 03/15/2023).

Cong, Lin William et al. (Dec. 2022). *Asset Pricing with Panel Tree Under Global Split Criteria*. Working Paper. DOI: 10.3386/w30805. URL: https://www.nber.org/papers/w30805 (visited on 03/30/2023).

Dabhi, Vipul K. and Sanjay K. Vij (Nov. 2011). "Empirical modeling using symbolic regression via postfix Genetic Programming". *2011 International Conference on Image Information Processing*, 1–6. DOI: 10.1109/ICIIP.2011.6108857.

Dangl, Thomas and Michael Halling (2012). "Predictive regressions with time-varying coefficients". *Journal of Financial Economics* 106.1, 157–181.

Diebold, Francis X and Robert S Mariano (Jan. 2002). "Comparing Predictive Accuracy". *Journal of Business & Economic Statistics* 20.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/073500102753410444, 134–144. ISSN: 0735-0015. DOI: 10.1198/073500102753410444. URL: https://doi.org/10.1198/073500102753410444 (visited on 03/26/2023).

Dong, Xi et al. (2022). "Anomalies and the Expected Market Return". en. *The Journal of Finance* 77.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13099, 639–681. ISSN: 1540-6261. DOI: 10.1111/jofi.13099. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13099 (visited on 03/31/2023).

Fama, Eugene F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". *The Journal of Finance* 25.2. Publisher: [American Finance Association, Wiley], 383–417. ISSN: 0022-1082. DOI: 10.2307/2325486. URL: https://www.jstor.org/stable/2325486 (visited on 03/21/2023).

Farmer, Leland E, Lawrence Schmidt, and Allan Timmermann (2023). "Pockets of predictability". *The Journal of Finance* 78.3, 1279–1341.

Feng, Guanhao, Jingyu He, and Nicholas G. Polson (Apr. 2018). *Deep Learning for Predicting Asset Returns*. arXiv:1804.09314 [cs, econ, stat]. URL: http://arxiv.org/abs/1804.09314 (visited on 03/30/2023).

Feng, Guanhao, Jingyu He, Nicholas G. Polson, and Jianeng Xu (2018). "Deep learning in characteristics-sorted factor models". *arXiv preprint arXiv:1805.01104*.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu (May 2020). "Empirical Asset Pricing via Machine Learning". *The Review of Financial Studies* 33.5, 2223–2273. ISSN: 0893-9454. DOI: 10.1093/rfs/hhaa009. URL: https://doi.org/10.1093/rfs/hhaa009 (visited on 03/15/2023).

Han, Yufeng et al. (Aug. 2022). *Expected Stock Returns and Firm Characteristics: E-ENet, Assessment, and Implications*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.3185335. URL: https://papers.ssrn.com/abstract=3185335 (visited on 03/31/2023).

Hastie, Trevor et al. (2022). "Surprises in high-dimensional ridgeless least squares interpolation". *Annals of statistics* 50.2, 949.

Henkel, Sam James, J Spencer Martin, and Federico Nardari (2011). "Time-varying short-horizon predictability". *Journal of financial economics* 99.3, 560–580.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1990). "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks". *Neural networks* 3.5, 551–560.

Huynh, Quang Nhat, Hemant Kumar Singh, and Tapabrata Ray (Dec. 2016). "Improving Symbolic Regression through a semantics-driven framework". *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. DOI: 10.1109/SSCI.2016.7849941.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". *Advances in neural information processing systems* 31.

Jin, Ying et al. (Jan. 2020). *Bayesian Symbolic Regression*. arXiv:1910.08892 [stat]. DOI: 10.48550/arXiv.1910.08892. URL: http://arxiv.org/abs/1910.08892 (visited on 03/06/2023).

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou (2022a). "The virtue of complexity everywhere". *Available at SSRN*.

– (2022b). *The virtue of complexity in return prediction*. Tech. rep. National Bureau of Economic Research.

Kelly, Bryan T and Seth Pruitt (2010). "Disaggregate Valuation Ratios and Market Expectations". en.

– (2013). "Market expectations in the cross-section of present values". *The Journal of Finance* 68.5, 1721–1756.

Korns, Michael F. (2011). "Accuracy in Symbolic Regression". en. *Genetic Programming Theory and Practice IX*. Ed. by Rick Riolo, Ekaterina Vladislavleva, and Jason H. Moore. Genetic and Evolutionary Computation. New York, NY: Springer, 129–151. ISBN: 978-1-4614-1770-5. DOI: 10.1007/978-1-4614-1770-5_8. URL: https://doi.org/10.1007/978-1-4614-1770-5_8 (visited on 03/06/2023).

Koza, John R. (June 1994). "Genetic programming as a means for programming computers by natural selection". en. *Statistics and Computing* 4.2, 87–112. ISSN: 1573-1375. DOI: 10.1007/BF00175355. URL: https://doi.org/10.1007/BF00175355 (visited on 03/06/2023).

Kusner, Matt J., Brooks Paige, and José Miguel Hernández-Lobato (July 2017). "Grammar Variational Autoencoder". en. *Proceedings of the 34th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 1945–1954. URL: https://proceedings.mlr.press/v70/kusner17a.html (visited on 03/06/2023).

Landajuela, Mikel et al. (2022). "A unified framework for deep symbolic regression". *Advances in Neural Information Processing Systems* 35, 33985–33998.

Lettau, Martin and Stijn Van Nieuwerburgh (July 2008a). "Reconciling the Return Predictability Evidence: The Review of Financial Studies: Reconciling the Return Predictability Evidence". *The Review of Financial Studies* 21.4, 1607–1652. ISSN: 0893-9454. DOI: 10.1093/rfs/hhm074. URL: https://doi.org/10.1093/rfs/hhm074 (visited on 03/21/2023).

– (July 2008b). "Reconciling the Return Predictability Evidence: The Review of Financial Studies: Reconciling the Return Predictability Evidence". *The Review of Financial Studies* 21.4, 1607–1652. ISSN: 0893-9454. DOI: 10.1093/rfs/hhm074. URL: https://doi.org/10.1093/rfs/hhm074 (visited on 04/01/2023).

Lu, Qiang, Jun Ren, and Zhiguang Wang (Jan. 2016). "Using Genetic Programming with prior formula knowledge to solve symbolic regression problem". *Computational Intelligence and Neuroscience* 2016, 1:1. ISSN: 1687-5265. DOI: 10.1155/2016/1021378. URL: https://doi.org/10.1155/2016/1021378 (visited on 03/06/2023).

McCracken, Michael W. (Oct. 2007). "Asymptotics for out of sample tests of Granger causality". en. *Journal of Econometrics* 140.2, 719–752. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2006.07.020. URL: https://www.sciencedirect.com/science/article/pii/S0304407606001539 (visited on 03/25/2023).

Mullainathan, Sendhil and Jann Spiess (May 2017). "Machine Learning: An Applied Econometric Approach". en. *Journal of Economic Perspectives* 31.2, 87–106. ISSN: 0895-3309. DOI: 10.1257/jep.31.2.87. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87 (visited on 03/15/2023).

Petersen, Brenden K. et al. (Apr. 2021). *Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients*. arXiv:1912.04871 [cs, stat]. DOI: 10.48550/arXiv.1912.04871. URL: http://arxiv.org/abs/1912.04871 (visited on 03/02/2023).

Qiu, Yue, Zhewei Song, and Zhensong Chen (Mar. 2022). "Short-term stock trends prediction based on sentiment analysis and machine learning". en. *Soft Computing* 26.5, 2209–2224. ISSN: 1433-7479. DOI: 10.1007/s00500-021-06602-7. URL: https://doi.org/10.1007/s00500-021-06602-7 (visited on 03/30/2023).

Rapach, David, Jack K. Strauss, and Guofu Zhou (2010). "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy". *The Review of Financial Studies* 23.2, 821–862.

– (2013). "International Stock Return Predictability: What Is the Role of the United States?" en. *The Journal of Finance* 68.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12041, 1633–1662. ISSN: 1540-6261. DOI: 10.1111/jofi.12041. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12041 (visited on 03/31/2023).

Rapach, David and Guofu Zhou (Jan. 2013). "Chapter 6 - Forecasting Stock Returns". en. *Handbook of Economic Forecasting*. Ed. by Graham Elliott and Allan Timmermann. Vol. 2. Handbook of Economic Forecasting. Elsevier, 328–383. DOI: 10.1016/B978-0-444-53683-9.00006-2. URL: https://www.sciencedirect.com/science/article/pii/B9780444536839000062 (visited on 03/30/2023).

Rapach, David and Guofu Zhou (2020). "Time-series and Cross-sectional Stock Return Forecasting: New Machine Learning Methods". en. *Machine Learning for Asset Management*. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119751182.ch1. John Wiley & Sons, Ltd, 1–33. ISBN: 978-1-119-75118-2. DOI: 10.1002/9781119751182.ch1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119751182.ch1 (visited on 03/31/2023).

– (Mar. 2022). *Asset Pricing: Time-Series Predictability*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.3941499. URL: https://papers.ssrn.com/abstract=3941499 (visited on 03/30/2023).

Schmidt, Michael and Hod Lipson (2009). "Distilling free-form natural laws from experimental data". *science* 324.5923, 81–85.

Trujillo, Leonardo et al. (Mar. 2016). "neat Genetic Programming: Controlling bloat naturally". en. *Information Sciences* 333, 21–43. ISSN: 0020-0255. DOI: 10.1016/j.ins.2015.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0020025515008038 (visited on 03/06/2023).

Udrescu, Silviu-Marian and Max Tegmark (Apr. 2020). "AI Feynman: A physics-inspired method for symbolic regression". *Science Advances* 6.16. Publisher: American Association for the Advancement of Science, eaay2631. DOI: 10.1126/sciadv.aay2631. URL: https://www.science.org/doi/full/10.1126/sciadv.aay2631 (visited on 03/06/2023).

Van Binsbergen, Jules H. and Ralph S. J. Koijen (2010). "Predictive Regressions: A Present-Value Approach". en. *The Journal of Finance* 65.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1
6261.2010.01575.x, 1439–1471. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2010.01575.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01575.x (visited on 03/24/2023).

Vladislavleva, Ekaterina (Katya) (Jan. 2008). "Model-Based Problem Solving through Symbolic Regression via Pareto Genetic Programming". *Tilburg University, Open Access publications from Tilburg University*.

Vladislavleva, Ekaterina J., Guido F. Smits, and Dick den Hertog (Apr. 2009). "Order of Nonlinearity as a Complexity Measure for Models Generated by Symbolic Regression via

Pareto Genetic Programming". *IEEE Transactions on Evolutionary Computation* 13.2. Conference Name: IEEE Transactions on Evolutionary Computation, 333–349. ISSN: 1941-0026. DOI: 10.1109/TEVC.2008.926486.

Welch, Ivo and Amit Goyal (July 2008). "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction". *The Review of Financial Studies* 21.4, 1455–1508. ISSN: 0893-9454. DOI: 10.1093/rfs/hhm014. URL: https://doi.org/10.1093/rfs/hhm014 (visited on 03/15/2023).