

Maximizing the Sharpe Ratio: A Genetic Programming Approach *

Yang Liu

Hunan University

Guofu Zhou

Washington University in St. Louis

Yingzi Zhu

Tsinghua University

Current version: February, 2023

*We are grateful to Michael T. Chng Allaudeen Hameed, Xuezhong (Tony) He, Chris Neely, Will Cong (discussant), Dashan Huang, Raymond Kan and the seminar participants at London Business School, Sichuan University, Washington University in St. Louis, Xi'an Jiaotong-liverpool University and Zhejiang University, and conference participants at 2020 China FinTech conference in Qingdao, 2020 Shanghai Financial Forefront Symposium, and 2021 China International Conference in Finance for very helpful comments. Liu and Zhu acknowledge the financial support from National Natural Science Foundation of China (# 71572091). Part of this work is accomplished while Liu is visiting Washington University in St. Louis.

Send correspondence to Guofu Zhou, Olin School of Business, Washington University in St. Louis, St. Louis, MO 63130; e-mail: zhou@wustl.edu; phone: 314-935-6384. Yang Liu, e-mail: liuyang9@hnu.edu.cn; Yingzi Zhu, e-mail: zhuyz@sem.tsinghua.edu.cn.

Maximizing the Sharpe Ratio: A Genetic Programming Approach

Abstract

While common machine learning algorithms focus on minimizing the mean-square errors of model fit, we show that genetic programming, GP, is well-suited to maximize an economic objective, the Sharpe ratio of investment portfolios in two contexts, the usual spread portfolio in the cross-section of expected stock returns and the standard mean-variance efficient portfolio optimization. In contrast to popular regression-based learning tools such as LASSO and the neural network, the spread portfolio of GP can double their performance in the US, and outperform them internationally. We find that, while the economic objective plays a role, GP captures nonlinearity in comparison with methods like the LASSO, and it requires smaller sample size than the neural network. We also apply GP to maximize the Sharpe ratio in the portfolio optimization with various economic constraints, and propose a regularization method based on transaction cost to control model complexity and mitigate overfitting problem. GP again substantially outperforms conventional and other machine learning-based methods by almost doubling the out-of-sample performance, especially in the case with transaction cost and short-sale constraints.

JEL Classification: G12, G14, G15

Keywords: Machine Learning, Genetic Programming, Cross-sectional Returns, Portfolio Optimization

“One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die.”

– Darwin, C., On the origin of species, 1859.

1. Introduction

Machine learning (ML) revolutionizes the research in all sciences, with its presence almost everywhere today. In finance, its applications at present appear concentrated in estimating the cross-section expected stock returns, perhaps because explaining why different assets have different returns is one of the central questions of finance. For examples, Chincó, Clark-Joseph, and Ye (2019) apply LASSO to analyze cross-firm return predictability at the one-minute horizon. Kozak, Nagel and Santosh (2020) provide a Bayesian LASSO approach to shrink factor dimensionality. Feng, Giglio, and Xiu (2020) focus on choosing factors. Gu, Kelly and Xiu (2020) apply a comprehensive set of ML tools, including generalized linear models, dimension reduction, boosted regression trees, random forests, and neural networks, to forecast individual stocks and their aggregates. Freyberger, Neuhierl, and Weber (2020) find nonlinear effects in the cross section of expected stock returns. Han, He, Rapach, and Zhou (2021) develop a new E-LASSO approach to handle a comprehensive set of firm characteristics. Filippou, Taylor, Rapach, and Zhou (2020) apply LASSO and neural network to predict foreign exchanges, and Guo, Lin, Wu, and Zhou (2022) conduct a machine learning study on corporate bonds. While all of these studies are different in their economic motivations, their solutions are based on the objective functions of the machine learning literature.

A natural question is whether we can apply or extend existing learning tools to maximize directly our economic objective function at hand. In particular, while different data sets or different methodologies are employed, existing studies, such as Lewellen (2015), Green, Hand, and Zhang (2017), and Freyberger, Neuhierl, and Weber (2020), all assess the economic significance by examining the performance of the well known spread portfolio, which longs assets in the highest estimated expected return group, and shorts those assets in the lowest. Because the GP is constructed to directly maximize the Sharpe ratio, it is likely to produce a portfolio with a higher Sharpe ratio than methods that minimize mean-squared errors.

In this paper, we show empirically that it is indeed the case with the use of the generic program-

ming (GP). While standard machine learning tools, such LASSO and neural network used in recent cross-section studies, are not readily to be adapted to maximize the Sharpe ratio directly, GP is known, at least as early as Neely, Weller and Dittmar (1997) and Allen and Karjalainen (1999) in the finance literature, as a flexible optimization method in setting the objective function. However, this flexibility in general have not receive much attention in the past, due to perhaps the extremely heavy burden in computations explained further later. In comparison with the early GP studies, we optimize the Sharpe ratio of an investment portfolio in two contexts, the usual spread portfolio in the cross-section of stock returns and the standard mean-variance efficient portfolio optimization. In fact, our paper appears the first to use the GP to maximize the Sharpe ratio, and the first to apply it for forecasting returns in the cross-section.¹ Moreover, following Zhang and Bhattacharyya (2004) and Bhowan, Johnston, Zhang and Yao (2012), we incorporate new advances into our GP set-up, allowing for multiple sets of parameters in training samples and using an ensemble method to select the ultimately mapping function.

With 15 firm characteristics that capture size, momentum and price trends over different time horizons, we find that the GP outperforms not only the leading regression-based machine learning methods, such as the ridge, LASSO, elastic net (Enet), PCR, and PLS, but also the neural network models (NN1-NN5 from one to 5 layers), the most powerful method shown by Gu, Kelly, and Xiu (2020). In the out-of-sample period from 1991 to 2019, the GP yields the greatest average return of 1.71% per month, while the second greatest return, earned by NN2, is only 1.22%. As for the Sharpe ratio, GP has the largest annualized Sharpe ratio (1.32) for the value-weighted spread portfolio. In contrast, the linear models produce a Sharpe ratio level of only about 0.70 (almost 50% lower) and the neural networks produce a level around 0.95 (almost 30% lower).

The largest differences occur during the post-2003 subperiod. Green, Hand, and Zhang (2017), noting a number of changes in firm reporting practice and government regulations, find that 2003 is a major structural break point in predicting the cross-section stock returns. Their results are replicated and confirmed with our data. Indeed, all the regression-based machine learning methods fail to generate significant average returns in their spread portfolios, though one of the five neural network models manages to get a significant average return of 0.53 per month with a t -stat of 2.19. In contrast, the GP obtains an average return of 0.72% per month with a t -stat of 3.06. Its

¹Assume knowing the true parameter, then there is a certain equivalence between finding the stochastic discount factor and maximizing the Sharpe ratio (Kozak, Nagel and Santosh (2020) and Bryzgalova, Pelger and Zhu (2021)).

Sharpe ratio is 0.77, still about 30% greater than any of the other strategies. In short, the GP makes an important economic difference when compared with existing methods. As shown later, the economic objective of maximizing the Sharpe ratio is the driving force.

What is the relation between the GP spread portfolio and the spread portfolios from other methods? We regress the expected stock return generated by GP on those generated by others in a cross-section regression, and then examine the decile portfolio sorted by the resulting residuals. Controlling for other models, GP persistently yields highly significant spread return. In contrast, controlling for GP, the spread return of other models shrinks almost to zero, indicating that GP contains substantially more information than other models by subsuming their predictability.

To understand the time-varying outperformance, we construct an idiosyncratic volatility (IVOL) index, defined as the cross-section average of the IVOL of individual stocks. The greater the index, the greater the information uncertainty. We find that the improvement of GP over other models is mainly attributed to its good performance during the high-IVOL periods. From an investment perspective, given the high noise-to-signal ratio in the stock market, it is important to predict returns with great information uncertainty. The GP appears to help exactly to do it in comparison with other methods.

We also examine the relation of the GP with various well known factor models in the literature. Following Fama and French (1993), we construct a GP factor, GPF, based on a standard 2×3 double sorting of size and the expected return generated by GP. All these machine learning methods easily produce largely unexplained alphas of about 0.90% with significant t -statistic over 3 relative to the existing factor models. However, the GPF alone can explain all the other spread portfolios. The average absolute alpha is only up to 0.11% with a negligible t -statistic of 0.43. Moreover, the p -value of the Gibbons, Ross, and Shanken (1989) (GRS) test for the GPF to price these spread portfolios is 0.88, while the p -values for other factor models are less than 10^{-4} . Moreover, adding the GPF to existing factor models improves significantly the Sharpe ratios, implying that it can serve as an additional factor based on the Sharpe ratio test of Barillas and Shanken (2017).

Our second main contribution is to apply GP to maximize the Sharpe ratio in the standard portfolio optimization problems. Our method has the following features: (1) we solve the portfolio optimization problem *without* estimating μ and Σ ; (2) we can easily incorporate various economic constraints into the optimization process; (3) we propose a new regularization method based on

transaction cost to mitigate overfitting, and the method works well in the case of large portfolios. GP again substantially outperforms conventional and other machine learning-based methods, including the maximum-Sharpe-ratio estimated and sparse regression (MAXSER) method recently proposed by Ao, Li, and Zheng (2019), by almost doubling the out-of-sample performance.

The performance of the GP is robust in a number of ways. First, once we replace the 15 firm characteristics by the 15 used by Lewellen (2015), which are primarily fundamental variables, the superior performance remains. In fact, the GP performs even better relative to existing methods, with Sharpe ratio almost more than doubling those of others. Second, GP also performs well internationally in the other G7 countries. Its spread portfolio is economically and statistically significant across the 6 markets, and its Sharpe ratio is always the greatest, consistent with the US results. It provides the evidence for *transfer learning* globally that the same model estimated by using US data can be applied to other countries.²

Thirdly, the GP is also robust to alternative setups of two parameters that determine its search for the maximum. Denote *Pop* as the individual number in each generation of the GP algorithm, and *Gen* as the maximum number of the generations. As they characterize the searching depth, it is obvious that the in-sample performance increases in either of $\langle Gen, Pop \rangle$. Indeed, even if the average Sharpe ratio for randomly generated individuals in the first generation is close to 0, the Sharpe ratio shows a strong increasing pattern as the generation increases. This evolution path suggests that GP indeed “learns” from the data and attempts to optimize the Sharpe ratio. However, increases in *Pop* is marginal compared with that of *Gen*. Moreover, we find that *Gen* also controls the volatility and convergence of the algorithm. Intuitively, simulated individuals in the earlier stage are more diversified, but as *Gen* increases, the new individual will evolve in the same direction guided by the objective. Hence, while achieving higher Sharpe ratios, they also become less-diversified and less volatile. Although the optimal $\langle Gen, Pop \rangle$ are chosen via validation, we examine a number of alternative parameters and find the results are robust.

To see the importance of setting an objective to maximize the Sharpe ratio, we provide an analysis of an alternative use of the GP algorithm with minimizing the conventional mean squared error (MSE) as the objective. The results show that our previous GP substantially outperforms

²See Jiang, Kelly and Xiu (2022) on more discussion of transfer learning and its validity in imaging analysis of financial data.

this MSE-based GP by yielding a spread portfolio with higher return and Sharpe ratio, and by subsuming its predictability.

To understand what conditions that drive the performance of the GP, we conduct two types of econometric analysis. First, we simulate data from a linear model. In this case, the GP performs similarly with other models because if the data are truly linear, learning from minimizing the MSE should learn perfectly on the data, and so the the Sharpe ratio objective makes little difference. In the second case, we simulate the data from a nonlinear model. In this case, as expected, the GP substantially outperforms the linear regression-type models with much higher Sharpe ratios. While the neural network models should capture the nonlinearity, we find that they require relative larger sample size to perform well, explaining why they performs worse than GP in the real data sets.

Our paper adds to the small literature on the applications of GP into finance. Neely, Weller and Dittmar (1997) seems of the earliest studies in finance, who apply the GP to find profitable technical rules. Allen and Karjalainen (1999) apply the GP to find profitable trading rules to beat the S&P 500 index, but unsuccessfully. Recently, Brogaard and Zareei (2018), with modified algorithms, are able to identify stronger time-series predictability of the S&P 500 index. Ready (2002) also use GP to investigate the profitability of the technical trading rules on DJIA index. In addition, Dempster and Jones (2001) and Dunis, Laws, Middleton, and Karathanasopoulos (2015) apply it to currency and commodities. All these existing studies are about using the GP for time series prediction. In contrast, our paper is about cross section prediction. As mentioned earlier, the hurdle of applying the GP is computational time, which is especially critical in our cross section context which deals with thousands of stocks. Indeed, even on a server with an Intel Xeon E7-8890 and 512 GB memory, the computation time takes days for our study. Nevertheless, with increasing computing power each year, the application of the GP in finance will surely increase drastically over time, simply due to the flexibility of the algorithm that it can be used to maximize any economic objective.³

Our paper also adds to the literature of growing applications of machine learning to finance. Chinco, Clark-Joseph, and Ye (2019), Feng, Giglio, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020), Kozak, Nagel, and Santosh (2020), Bryzgalova, Huang, and Julliard (2022), and Avramov, Cheng, Metzker, and Voigt (2022) provide various dimension reduction analysis, while

³Nordhaus (2001) shows that the computing power has increased by around 80% per year since 1980.

DeMiguel, Martin-Utrera, Nogales, and Uppal (2020), Gu, Kelly, and Xiu (2020), Chen and Velikov (2020), Chordia, Goyal, and Saretto (2020), Patton and Weller (2020), Avramov, Cheng, and Metzker (2021), Bryzgalova, Pelger, and Zhu (2021), Cong, Tang, Wang, and Zhang (2021), and Cong, Feng, He, and He (2022) focus on portfolio performance. Our paper is related more to the latter by using maximizing the Sharpe ratio as the model fitting objection. The economic goal makes the resulting spread portfolio performs the best, and this portfolio explain and understand the abnormal returns from other methods. Our paper is the first to use the powerful GP in the cross-sectional asset pricing, which appears to provide both economic insights and improved performance in general.

The rest of the paper is organized as follows. Section 2 discusses the data and the methodology of our GP model and other competing machine learning models. Section 3 presents the results in the cross-section context. Section 4 reports the results in the portfolio optimization context. Section 5 examines the robustness. Section 6 explores the explanation for GP’s good performance. Section 7 concludes.

2. Data and methodology

In this section, we first introduce the data, and then discuss the GP algorithm for maximizing the Sharpe ratio in the cross-section, along with a review of other machine learning methods for comparison.

2.1. Data

As usual, we use all domestic common stocks listed on NYSE, AMEX, and Nasdaq stock markets, and exclude close-end funds, real estate investment trusts, unit trusts, American depository receipts, and foreign stock (or stocks that do not have a CRSP share code of 10 or 11). As the literature typically does, we employ the price filter to exclude the stocks with price below \$5.

The primary set of characteristics consists of 15 variables: the market capitalization (size) and 3 past return-based signals, i.e., R_{-1} , $R_{-12,-2}$, and $R_{-60,-13}$, which correspond to the short-term reversal (*SREV*) of Lehmann (1990), Lo and MacKinlay (1990), momentum (*MOM*) of Jegadeesh and Titman (1993), and long-term reversal (*LREV*) of DeBondt and Thaler (1985),

respectively. In addition, we also include the 11 price moving average (MA) signals used in Han, Zhou and Zhu (2016), including MAs of lag lengths of 3-, 5-, 10-, 20-, 50-, 100-, 200-, 400-, 600-, 800-, and 1000-days. Following the most recent studies, we normalize each indicator in the cross-section such that it has a mean of zero and a standard deviation of one without loss of generality. We use this characteristic set because it is easy to construct, making it ideal for comparison in international markets. However, since this set relies heavily on technical signals, we also use another 15 characteristics of Lewellen (2015), which are mostly fundamental variables, and the most important 15 characteristics of Gu, Kelly, and Xiu (2020) as robustness checks.

2.2. The GP algorithm

In this subsection, we first discuss the objective function and search space, and then we introduce the optimization procedure and hyperparameter tuning.

2.2.1. Incorporating economic objective

Our economic objective is to maximize the Sharpe ratio of a portfolio based on firm characteristics, which is of importance to an investor or fund manager who would like to achieve the maximum economic gains from the information on characteristics. While we find that it is difficult to solve this problem using other existing machine learning tools, the GP appears the best to fit the purpose.

Mathematically, our objective is to find a function $G(\cdot)$ to maximize the Sharpe ratio (SR) of the usual value-weighted decile long-short spread portfolio, but here the long and short legs are determined endogenously,

$$\max_{G(\cdot) \in \mathcal{M}} SR(\text{Spread}(G(\cdot))), \quad (1)$$

where \mathcal{M} is the search space, $G(\cdot)$ is a function mapping from the stock characteristics to the expected return, and $\text{Spread}(G(\cdot))$ is the resulting spread portfolio. In particular, suppose X is a panel data of stock characteristics, in which $X_{i,t}$ is a vector of characteristics for stock i on month t . Denote the expected return for stock i in month t generated by $G(\cdot)$ as

$$ER_G^{i,t} = G(X_{i,t-1}). \quad (2)$$

Then, we can sort stocks by $ER_G^{i,t}$ in each month into decile groups and construct a value-weighted spread portfolio, so weighted as all other portfolios in the paper, and denote it as $Spread(G(\cdot))$. Put differently, we want to search for the optimal function $G(\cdot)$ to maximize the Sharpe ratio of $Spread(G(\cdot))$.

Genetic programming (GP) is a supervised machine learning method based on the principle of Darwinian natural evolution. Since its launch by Koza (1992), GP has been successfully applied in various fields, such as economics, finance, and engineering. GP randomly generates initial population of a certain number of individuals, each of which is a solution candidate to the given problem. The performances of the solution candidates are evaluated according to a problem-specific fitness function (objective function), which defines the environment for the evolution. Then, the individuals are randomly selected as parents individuals, with the selection probabilistically biased in favor of the relatively fit members. Next, the parents individuals are combined by genetic operators, such as crossover and mutation, to create offspring individuals. Afterward, successive generations are generated in the same way until the final generation.

For the optimization problem of Equation (1), the GP is ideal, as it is often used for solving optimization problems with objective functions which are non-differentiable or difficult to be expressed in other optimization approaches. In addition, as a non-parametric model, GP can discover both the model structure and model parameters, and thus are more flexible in exploring nonlinear predictability. Moreover, due to the stochastic nature, it is less likely to converge to local optima, and it is generally suitable to search for global optimum in large search space.

2.2.2. Representation and search space

In GP, the solution candidates are represented as tree structures and can be encoded as function $G(\cdot)$ mapping from characteristics X to expected returns, which is discussed in the Online Appendix in more detail. Each individual $G(\cdot)$ is built of two basic primitives, the *terminal* nodes and *function* nodes. Essentially, the terminal nodes provide the inputs to the GP program, and it includes the input characteristics (X) and some random constants. The *function* nodes come from a pre-defined function set. Panel A of Figure 1 shows an example of the tree-structure individuals. It consists of two characteristics of X_1 and X_2 , a random constant of 1, and two function operators

of MULTIPLY (\times) and ADD ($+$). It can be coded as a function $G(X) = X_1(X_2 + 1)$. In terms of its economic interpretation, this solution represents such a hypothesis about the cross-section of stock returns that stocks with greater X_1 tends to have higher future returns. In addition, it also assumes that this effect increases with X_2 by adding an interaction item of X_1 and X_2 .

The search space \mathcal{M} is spanned by a large set of functions combining an indicator set and an function set. The indicator set X includes the firm characteristics such as the 15 discussed in section 2.1.. The function set includes both commonly used linear and nonlinear operators, examples of which the linear functions are ADD, MINUS, NEGATIVE, and the nonlinear ones are MULTIPLY, DIVIDE, SIN, COS, ABS, and bool-type operator CMP. This enables GP to exploits both the linear and nonlinear predictability of the characteristics. However, though we do not assume any specific function form for $G(\cdot)$, we limit the maximum of tree depth to 30 for tractability. This still enables a sufficiently large space of millions of candidate solutions, and controls the model complexity and overfitting at the same time.

2.2.3. Optimization

It is important to examine how the GP selects the individuals to maximize the Sharpe ratio. Different from the common gradient-based method, the optimization of GP is based on the principle of Darwinian natural evolution.

Essentially, GP optimizes the given problem by iteratively producing offspring individuals based on genetic operators and then selecting strong individuals by the natural selection principle. The direction of the evolution is characterized by the fitness function, i.e, the optimization objective, which is the Sharpe ratio of the spread portfolio in our case. In particular, after initiating the random individuals in the first generation, GP will calculate their associated Sharpe ratios. Then, to produce new individuals for the next generation, the individuals are randomly selected as parent individuals, with the selection probabilistically biased in favor of the relatively fit individuals with greater Sharpe ratios. Next, the parent individuals are combined by genetic operators, such as crossover and mutation, to create new offspring individuals.

Figure 1 illustrates how the crossover and mutation operators work. As suggested by the green and red box in Panel A to D, the parent individuals in Panels A and B are combined

by the crossover operator, and the resulting offspring individuals are shown in Panels C and D. The offspring individuals can also be produced by the mutation operators. For example, the characteristics of X_2 and the constant number of 1 in the green box in Panel A can mutate to X_3 and 2 in Panel E, respectively. Also, the mutation operator can also work on the function node and the whole subtree. For example, the subtree of X_3 , shown in the red box in Panel B, can mutate to another subtree of $|\sin(X_1)|$ in Panel F.

After applying these genetic operators to produce offsprings, GP will evaluate the fitness of these offspring and parent individuals, and those with greater Sharpe ratios will survive as individuals in the next generation. Afterward, successive generations are iteratively generated in the same way, until the generation number exceeds a pre-defined max generation Gen .

Moreover, following Zhang and Bhattacharyya (2004) and Bhowan, Johnston, Zhang and Yao (2012), we adopt an ensemble approach in training our GP model to improve the model robustness and to mitigate overfitting. In particular, since GP has the advantage of parallel computing (Winschel and Krätzig, 2010, and Polachek, Das, and Thamma-Apiroam, 2015), we independently estimate GP for five times, and get $5 \times Pop$ individuals (or models) in total, as each time GP generates Pop individuals. Because of the stochastic nature of GP, this helps search for the global optima rather than being accidentally trapped by a local optimum. Finally, we take the average of the top M models with the highest training sample Sharpe ratio as the final model under the parameter $\langle Pop, Gen \rangle$. Although M is set to 5, we have also examined alternative values of 3 and 10 as robustness checks.

2.2.4. Hyperparameter tuning

There are two important hyperparameters that control the optimization process of the GP. The first is *Population* (Pop), defined as the number of individuals that GP will generate in each generation. The second is *Generation* (Gen), used to determine the maximum generation that the evolution will iterate. Clearly, the pair $\langle Pop, Gen \rangle$ characterize the searching depth for GP, and have influence on model performance. Since there is no theoretical criterion for the selection of the pair, we follow the most common approach in the literature and select the hyperparameters in a validation sample. The validation sample can be interpreted as a simulated OOS sample to learn

about model complexity and hence to mitigate overfitting.

In our paper, the parameter values for Pop are 100, 200, and 400, and those for Gen are 10, 20, and 40.⁴ Hence, there are 9 hyperparameter combinations for GP. As discussed in the previous subsection, for a given $\langle Pop, Gen \rangle$, we use the training sample to estimate the GP model, and use the average of the top M ($M=5$) model as the model, denoted as $G_{\langle Pop, Gen \rangle}$. We then evaluate the performance of the 9 models in the validation sample. The optimal model $G_{\langle Pop^*, Gen^* \rangle}^*$ is the one that earns the highest Sharpe ratio for the spread portfolio in the validation sample. Last, we use the out-of-sample subsample, which is not used for model estimating nor parameter tuning, to examine the OOS performance of the optimal GP model.

2.3. Penalized regression models

For easier comparison, we briefly introduce below other machine learning methods, i.e., those used by Gu, Kelly, and Xiu (2020).

2.3.1. Ridge

Ridge regression imposes an l_2 norm in the standard regression model,

$$\hat{\beta}_{Ridge}(\lambda) = \arg \min_{\beta} \left\{ \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (r_{i,t} - \beta_0 - \sum_{j=1}^P X_{i,t-1,j} \beta_j)^2 + \lambda \sum_{j=1}^P \beta_j^2 \right\}, \quad (3)$$

where the parameter penalization helps to prevent coefficients from becoming unduly large in magnitude.

2.3.2. Lasso

Lasso regression imposes the l_1 norm,

$$\hat{\beta}_{Lasso}(\lambda) = \arg \min_{\beta} \left\{ \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (r_{i,t} - \beta_0 - \sum_{j=1}^P X_{i,t-1,j} \beta_j)^2 + \lambda \sum_{j=1}^P |\beta_j| \right\}, \quad (4)$$

where the parameter penalization helps to force coefficients on some regressors to exactly zero, thereby selecting the most useful variables.

⁴We do not use too large parameters because the GP is computationally extensive. For example, in our applications, it takes about 24 hours to just estimate the model once under the parameter of $\langle 400, 40 \rangle$. Nevertheless, the chosen values are adequate in robustness checks.

2.3.3. Enet

The elastic net (Enet) model imposes both l_1 and l_2 norms,

$$\hat{\beta}_{Enet}(\lambda, \rho) = \arg \min_{\beta} \left\{ \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (r_{i,t} - \beta_0 - \sum_{j=1}^P X_{i,t-1,j} \beta_j)^2 + \lambda \sum_{j=1}^P (\rho \beta_j^2 + (1 - \rho) |\beta_j|) \right\}. \quad (5)$$

It is clear that $\rho = 1$ corresponds to the Ridge, and $\rho = 0$ corresponds to the Lasso. In our paper, we set $\rho = 0.5$, allowing for the associated Enet takes the advantages of both shrinkage and selection. The hyperparameter λ , in Ridge, Lasso or Enet, is determined by validation sample.

2.4. Dimension reduction models

2.4.1. PCR

Principal components regression (PCR) performs dimension reduction by zeros out coefficients on low variance components. It consists of two steps. In the first step, principal components analysis (PCA) combines the P regressors into a small set of K components ($K \leq P$), which are linear combinations that best preserve the covariance structure among the regressors. Mathematically, the k^{th} PCA component direction v_m solves:

$$\begin{aligned} & \underset{v}{\text{maximize}} && \text{Var}(Xv) \\ & \text{subject to} && \|v\| = 1, \\ & && \text{Cov}(Xv, Xv_l) = 0, \\ & && l = 1, \dots, k - 1. \end{aligned} \quad (6)$$

In the second step, regressions of stock return on the leading components are run to predict future returns.

2.4.2. PLS

Partial least square (PLS) regression performs dimension reduction by directly exploiting co-variation of regressors with the forecast target. In the optimization form, the k^{th} PLS components

solves :

$$\begin{aligned}
& \underset{v}{\text{maximize}} && \text{Cov}^2(r, Xv) \\
& \text{subject to} && \|v\| = 1, \\
& && \text{Cov}(Xv, Xv_l) = 0, , \\
& && l = 1, \dots, k - 1.
\end{aligned} \tag{7}$$

Then, a regression, similar to the PCR case, is run to determine the expected stock returns.

2.5. Neural Networks

Following Gu, Kelly, and Xiu (2020), we construct the neural networks for our study in the same way. We consider the architectures with up to five hidden layers. The shallowest neural network, denoted as NN1, has a single hidden layer of 32 neurons, NN2 has two hidden layers with 32 and 16 neurons, respectively; NN3 has three hidden layer with 32, 16, and 8 neurons, respectively; NN4 has four hidden layer with 32, 16, 8, and 4 neurons, respectively; and NN5 has four hidden layer with 32, 16, 8, 4, and 2 neurons, respectively. The nonlinear activation function is also the same rectified linear unit (ReLU) function for all nodes, defined as

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{otherwise,} \end{cases}$$

Moreover, we also employ the stochastic gradient descent (SGD) to estimate the neural network weight parameters to minimize the mean squared errors. We denote the expected return generated by NN_l ($l = 1, 2, 3, 4, 5$) for stock i in month t as $ER_{NN_l}^{i,t}$.

3. Main results

In our GP applications below, we split the full sample, from 1945:01 to 2021:12, into three subsamples. The training subsample from 1945:01 to 1980:12 is used to train the machine learning models. The validation subsample from 1981:01 to 1990:12 is used to choose the hyperparameters in these models.⁵ The out-of-sample (OOS) subsample from 1991:01 to 2021:12 is used to evaluate the models' predictive performance.

⁵Following Gu, Kelly, and Xiu (2020), we do not choose cross-validation to maintain the temporal ordering of the data for prediction.

3.1. Spread portfolios

Table 1 reports the OOS performance of the value-weighted decile spread portfolios sorted by the expected return of various models. It is interesting that there are not much differences in the linear models, whose annualized Sharpe ratios range from 0.59 to 0.69. Consistent with Gu, Kelly, and Xiu (2020), the neural networks tend to outperform the linear machine learning methods, achieving the highest annualized Sharpe ratio of 0.83. In contrast, the GP earns the best annualized Sharpe ratio up to 1.21, almost 50% greater than the next best level of 0.83 generated by NN4.

In terms of average returns, the GP also performs the best, with a monthly return of 1.53%, while the next largest average return, achieved by NN3, is only 1.14%. The linear models also have lower returns ranging from 0.84% to 1.03%. Moreover, in terms of skewness, the GP enjoys a positive skewness of 0.83, while the skewness of the linear models are lower than 0.20. However, NN2 has the largest skewness of 1.02, but it is too volatile and does not even have the highest Sharpe ratio among the neural networks models. It is also interesting to note that GP also generates the lowest maximum drawdown of only 22.56%, while the maximum drawdown is on average about 45% and 40% for the linear models and the neural networks. The results indicate that GP is resilient from recovering from the downside risk.

Table 2 reports the sub-sample results before and after 2003, a year when Green, Hand, and Zhang (2017) detect a major structural break for predicting the cross-sectional returns. Panel A shows that during the pre-2003 sub-period, the GP yields the highest spread return (2.62%) and the greatest Sharpe ratio (1.79). Interestingly, the linear models perform almost as well as the average of the neural network models. Hence, during this “easier” to predict periods as identified by Green, Hand, and Zhang (2017), existing machine learning methods, linear or nonlinear, do not seem to make much differences. This is because, though NN5 does the best, it is ex ante difficult to select NN5 out of all the models. Nevertheless, the GP still stands out and performs the best as expected.

Panel B reveals a much different pattern. In this “difficult” to predict period, all linear models fail to generate significant average returns on the spread portfolios, and even three of the 5 neural networks models fail. In contrast, the GP earns an economically and statistically significant monthly average return of 0.74%. In terms of Sharpe ratio, it has the highest of 0.71, exceeding the next

best level of 0.46, achieved by the best neural network model NN4, by 54%.

In short, empirically both in the subperiods and in the entire out-of-sample period, the GP achieves what it is designed for, to maximize the Sharpe ratio. This is one of the most important measures investors or fund managers rely upon in assessing a portfolio strategy.

3.2. Controlling for other models

Since GP and other models exploit different predictive information from the same characteristic set, it is of interest to examine which of them can provide incremental predictive power beyond the use of the other. Consider, for example, how to measure the incremental predictive power of the GP conditional on any other model. A simple approach is to regress the expected stock return generated by the GP on those generated by a given other model, and then sort the residuals into decile portfolios to see whether the new long-short spread portfolio can earn significant profits. Clearly, if the predictive power of the GP is subsumed by the given model, we should not be able to observe any profitable pattern in the resulting spread portfolio.

Panel A of Table 3 reports the results. After controlling for the expected return generated by any of the other models, the GP still produces highly significant spread returns in every single case. Conversely, we also examine the predictive power of any other model after controlling that of the GP. Panel B of the table shows that, after controlling for the GP, none of the other machine learning models can produce significant spread returns. The results clearly indicate that the GP has certain unique predictability which cannot be replaced by any of the other models, and the predictability of all the other models are subsumed by the GP.

Table 4 further controls all machine learning models simultaneously using Fama-MacBeth (1973) regression. The signal of our GP model constantly generates positive and highly significant coefficients in the full OOS period, as well as in the two subperiods. On the contrary, the results for other machine learning methods are generally insignificant or mixed across the periods. The results indicate that jointly controlling for all other machine learning models, our GP model can still provide substantial incremental predictability and economic gains.

3.3. Information uncertainty

To understand under what conditions where the GP and other methods differ, we focus on information uncertainty, and, following Zhang (2006), use idiosyncratic volatility (IVOL) to proxy for it. In particular, we construct an IVOL index, defined as the average of IVOL of individual stocks in each month, to reflect the information uncertainty at the market level. The greater the IVOL index, the greater the information uncertainty across stocks.

We carry out the following time-series predictive regression,

$$\Delta R_t = \beta_L Low_{t-1}^{Vol} + \beta_H High_{t-1}^{Vol} + \beta MKT_t + \epsilon_t, \quad (8)$$

where ΔR_t is the return of the GP spread minus that of other models, Low_{t-1}^{Vol} and $High_{t-1}^{Vol}$ are dummy variables indicating low- and high-IVOL periods of previous month, as classified based on the median level of the IVOL index. The parameters of interest are β_L and β_H , indicating either the low- or high-IVOL period or both matter for the GP outperformance.

Table 5 reports the results. The slope β_H is much greater than β_L for all of the ten models. Moreover, β_L is insignificant for eight of the ten models, while β_H is significant for all cases. On average, β_H is 1.08 with a significant t -statistic of 3.38, whereas β_L is much lower at 0.33 with a weak t -statistic of only 1.40. The results suggest that the improved performance of GP over other models is mainly attributed to the high-IVOL periods, during which the information uncertainty level is high. From an investment perspective, it is more difficult and hence more important to predict returns more accurately with greater information uncertainty. The GP appears to help exactly to do it in comparison with other methods.

3.4. GP factor

In this subsection, we consider a factor formed based on the GP and compare it with various well-known factor models in the literature.

Following Fama and French's (1993) factor formulation approach, we construct a GP factor (GPF) based on a 2×3 double sorting on size and ER_{GP} . The factors for the comparison are: the CAPM, the Fama and French (1993) 3-factor model (FF-3), Fama and French (2015) 5-factor model (FF-5), Hou, Xue, and Zhang (2015) 4-factor model (HXZ-4), Stambaugh and Yuan (2016)

mispricing-factor model (SY-4), and Daniel, Hirshleifer, and Sun (2020) behavioral-factor model (DHS-3), with data from their websites.

Table 6 reports the results.⁶ The GPF earns the greatest monthly average return of 1.16%, which is 45% greater than the next best factor of the market factor (0.80%). Its annualized Sharpe ratio, 1.63, is also the maximum, almost doubling the next best too, 0.88. It has the large skewness of 0.91, indicating a desirable positive return pattern. Panel B provides the correlation matrix of the factors. It shows that the GPF has low correlation with the well known factors.

Although the GPF other factors and has little correlation with each one of them, it does not rule out the hypothesis that a portfolio of other factors can replicate the performance of the GPF. To test this hypothesis, we carry out six spanning tests: Wald test under conditional homoskedasticity, Wald test under independent and identically distributed (IID) elliptical distribution, Wald test under conditional heteroskedasticity, Bekerart-Urias spanning test with errors-in-variables (EIV) adjustment, Bekerart-Urias spanning test without the EIV adjustment and DeSantis spanning test (see Kan and Zhou, 2012).

Panel A of Table 7 provides the results for the spanning tests. The spanning hypothesis is strongly rejected, indicating that the GPF can add substantial investment value to existing factor models. Barillas and Shanken (2017) show that investment value is related to model comparison. If a new factor can add substantial Sharpe ratio to an existing factor model, an extended model by adding the factor must outperform the existing model in explaining asset returns, irrespective of the test assets. Along this line, we conduct the Sharpe ratio test to compare the Sharpe ratios (Sh^2) of the various models with and without the GPF.

Panel B of Table 7 reports the results. It is apparent that adding the GP factor substantially improves the Sh^2 for all of other models. For example, the Sh^2 for CAPM increases significantly from 0.03 to 0.23, where the significance level is computed based on a studentized bootstrap procedure due to Ledoit and Wolf (2008). The virtually zero p -values cross the models suggest that the GP factor can improve the pricing ability of existing models substantially.

⁶Due to the factor data availability, the sample for SY-4 ends in 2016:12, and the sample for DHS-3 ends in 2018:12.

3.5. Risk-adjusted performances

Table 8 reports the alphas of the spread portfolios of the machine learning methods under different factor models. The first 6 rows show that all of the 11 spread portfolios earn highly significant alphas with respect to all the well known existing factor models: the CAPM, FF-3, FF-5, XHZ-4, SY-4, and DHS-3, indicating that existing factor models cannot explain the predicted returns of the machine learning methods. In fact, the magnitude of the alphas are much larger than almost all of those classic anomalies in the literature (see, e.g., Hou, Xue, and Zhang (2015)).

In contrast, as shown by the last row, all the alphas become insignificant relative to the extended CAPM with the GPF as the added factor. Indeed, the largest alpha is now only 30 basis points, while the average alphas exceed 1% previously. The results are similar if the GPF is added to any other factor models, suggesting that the GP factor improves substantially the pricing ability of existing models.

4. Maximize SR for portfolio optimization with economic constraints

In the previous section, we model $G(\cdot)$ as a function mapping from the stock characteristics to the expected return to maximize the Sharpe ratio (SR) of the resulting spread portfolio. In this section, we extend our GP model to maximize the Sharpe ratio under a different setting, that is, portfolio optimization with economic constraints. The intuition is to use GP to estimate a weighting function $G(\cdot)$, mapping from historical data of individual risky assets to a weight vector, to maximize the Sharpe ratio of the resulting portfolio.

4.1. A GP model for maximize portfolio SR

Consider the standard portfolio choice problem of achieving mean-variance efficiency. Denote the *excess* returns for the N risky assets by $\mathbf{r} = (r_1, r_2, \dots, r_N)'$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the expected mean and covariance matrix of \mathbf{r} . The mean-variance efficient portfolio of Markowitz (1952) has an explicit expression that depends only on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. However, it faces the following challenges: (1) $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are widely accepted to be very difficult to be estimated accurately. The commonly used

“plug-in” portfolio based on the sample mean and sample covariance is shown to generate poor out-of-sample (OOS) performance (e.g., Michaud, 1989); (2) it assumes no market friction and a zero transaction cost; (3) the OOS performance worsens as the asset number increases (e.g., Best and Grauer, 1991; Britten-Jones, 1999; Kan and Zhou, 2007).⁷

Here, we use GP to solve the portfolio optimization problem. Our method has the following features: (1) we solve the portfolio optimization problem *without* estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$; (2) we can incorporate various economic constraints, such as transaction cost, into the optimization process; (3) we propose a new regularization method based on transaction cost to mitigate overfitting, and the method works well in the case of large portfolios.

4.1.1. Baseline problem representation

The intuition for our method is that we interpret a portfolio rule as a weighting function of historical data (not limited to return data) and take advantage of the powerful searching ability of machine learning to estimate the function. Mathematically, denote the historical data at the end of time t as Φ_t . Φ_t can include the historical return data, other characteristics for individual stocks, and macro- or market-level indicators. Our goal is to estimate a weighting function $G(\cdot)$, mapping from historical data of individual risky assets ($\Phi_{i,t}$) to (unscaled) weights on the risky assets. Denote

$$w_{i,t+1} = \frac{G(\Phi_{i,t})}{\sum_i |G(\Phi_{i,t})|}, \quad (9)$$

where $w_{i,t+1}$ is the weight on risky asset i at the beginning of period $t+1$. Note that $\sum_i |w_{i,t+1}| = 1$, and this indicates that we require a 100% margin on long and short positions.⁸ Since $w_{i,t+1}$ is essentially determined by $G(\cdot)$, we define the scaling function $\dot{G}(\cdot)$ and re-write Eq. (9) as

$$w_{i,t+1} = \dot{G}(G(\Phi_t)) \equiv \frac{G(\Phi_{i,t})}{\sum_i |G(\Phi_{i,t})|}. \quad (10)$$

For a given $G(\cdot)$, denote the portfolio return in period $t+1$ as R_{t+1} :

$$R_{t+1} = \sum_i w_{i,t+1} r_{i,t+1} = \sum_i \dot{G}(G(\Phi_t)) r_{i,t+1}. \quad (11)$$

⁷Some new methods, such as MAXSER proposed by Ao, Li, and Zheng (2019), partially solve the above problems, and we use them as the benchmark for our new method.

⁸The leverage level will not affect the portfolio Sharpe ratio.

Denote $\tilde{r}_{t+1} = \{r_{1,t+1}, r_{2,t+1}, \dots, r_{N,t+1}\}$, and then we can define function $F(\cdot)$ and re-write Eq. (11) as

$$F(G(\cdot), \Phi_t, \tilde{r}_{t+1}) \equiv \sum_i \dot{G}(G(\Phi_t)) r_{i,t+1} = R_{t+1}. \quad (12)$$

The objective is to maximize the Sharpe ratio (SR) of the portfolio return $\{R_t\}$, and hence the optimization problem can be represented as:

$$\max_{G(\cdot) \in \mathcal{M}} SR\{F(G(\cdot), \Phi_t, \tilde{r}_{t+1})|_{t=0}^{t=T}\}, \quad (13)$$

where \mathcal{M} is the search function space, and $G(\cdot)$ is a function mapping from historical data to (unscaled) weights on the risky assets.

4.1.2. Incorporating economic constraints

We first incorporate the short-sale constraint and transaction cost into the baseline representation, separately. Then, we discuss the combination of various economic constraints.

Case 1. Short-sale constraint:

To model short-sale constraint, paralleling to Eq. (10), define

$$w_{i,t+1} = |\dot{G}(G(\Phi_t))| \equiv \left| \frac{G(\Phi_{i,t})}{\sum_i |G(\Phi_{i,t})|} \right|. \quad (14)$$

Paralleling to Eq. (12), we can define $\dot{F}(\cdot)$ and denote the portfolio return in period $t+1$ as

$$\dot{F}(G(\cdot), \Phi_t, \tilde{r}_{t+1}) \equiv \sum_i |\dot{G}(G(\Phi_t))| r_{i,t+1} = R_{t+1}. \quad (15)$$

Hence, the portfolio optimization problem with short-sale constraint can be represented as:

$$\max_{G(\cdot) \in \mathcal{M}} SR\{\dot{F}(G(\cdot), \Phi_t, \tilde{r}_{t+1})|_{t=0}^{t=T}\}. \quad (16)$$

Case 2. Transaction cost:

Following Robert, Robert, and Jeffrey (2012), we define the portfolio return net of transaction costs as follows:

$$R_{net,t+1} = \left\{ 1 - \sum_i c_{i,t+1} |w_{i,t+2} - w_{i,(t+1)+}| \right\} (1 + R_{t+1}) - 1, \quad (17)$$

where $c_{i,t+1}$ is a cost level that measures the transaction cost per dollar traded for trading asset i ,⁹ $w_{i,t+2}$ is the weight on asset i at the beginning of period $t+2$, and $w_{i,(t+1)+}$ is the weight on asset i at the end of period $t+1$, and R_{t+1} is the portfolio return without transaction cost. The portfolio turnover in the period $t+1$ is accordingly defined as:

$$Turn_{t+1} = \sum_i |w_i(t+1) - w_i(t+)|. \quad (18)$$

Given Eq. (10) and Eq. (12), We can re-write Eq. (17) as

$$R_{net,t+1} = \{1 - \sum_i c_{i,t+1} |\dot{G}(G(\Phi_{i,t+1})) - w_{i,(t+1)+}| \} (1 + F(G(\cdot), \Phi_t, \tilde{r}_{t+1})) - 1. \quad (19)$$

Denote $\tilde{c}_{t+1} = \{c_{i,t+1}, \dots, c_{N,t+1}\}$. Notice that $w_{i,(t+1)+}$ is determined by $w_{i,t+1}$, namely $\dot{G}(G(\Phi_{i,t}))$, and \tilde{r}_{t+1} . Hence, paralleling to Eq. (12), we can define $\ddot{F}(\cdot)$ and re-write Eq. (19) as

$$\ddot{F}(G(\cdot), \Phi_t, \Phi_{t+1}, \tilde{r}_{t+1}, \tilde{c}_{t+1}) \equiv R_{net,t+1}. \quad (20)$$

Then, the portfolio optimization problem with transaction cost $\{\tilde{c}_{t+1}\}$ can be represented as:

$$\max_{G(\cdot) \in \mathcal{M}} SR\{\ddot{F}(G(\cdot), \Phi_t, \Phi_{t+1}, \tilde{r}_{t+1}, \tilde{c}_{t+1})|_{t=0}^{t=T}\}. \quad (21)$$

The optimization problem in (13) is a special case of (21) with a zero transaction cost ($c_{i,t} \equiv 0$).

Case 3. Other constraints:

It is convenient to consider the short-sale constraint and the transaction cost simultaneously. Also, the above representation can be easily extended to other economic constraints. For example, consider an equity fund manager with a preference for a certain factor style, say value stocks. We can add the constraint that the portfolio weight on value stocks should be greater than a pre-defined threshold.

4.2. Benchmark models

Table 9 lists the benchmark portfolios and their abbreviations. In particular, Ao, Li, and Zheng (2019) propose a new approach named maximum-Sharpe-ratio estimated and sparse regression (MAXSER) to construct optimal mean-variance portfolios.

⁹A potential extension is to apply higher transaction cost for short-sale trading.

We consider the economic constraints of certain transaction costs and short-sale constraints. For the GP model:

- “GP-NoCost” stand for the GP model without short-sale constraint and without transaction cost;
- “GP-WithCost” stand for the GP model without short-sale constraint and with transaction cost;
- “GP-NoCost-SS” stand for the GP model with short-sale constraint and without transaction cost;
- “GP-WithCost-SS” stand for the GP model with short-sale constraint and with transaction cost.

4.3. Empirical results

We apply GP to solve the optimization problem of (13), (16), and (21) for its advantage of flexibility. The historical data Φ_t , used to estimate the weighting function $G(\cdot)$, consists of the same 15 variables introduced in Section 2.1.: market capitalization and 3 past return-based signals, i.e., R_{-1} , $R_{-12,-2}$, and $R_{-60,-13}$, and the 11 price moving average (MA) signals used in Han, Zhou, and Zhu (2016). For our GP strategies to be compared with the MAXSER of Ao, Li, and Zheng (2019), we follow them to use the stock universe of S&P 500 index constituents to implement GP optimization. All portfolios are constructed using the monthly excess return.¹⁰

The sample splitting is the same as in the previous section. The training sample for GP is from 1945:01 to 1980:12, the validation sample is from 1981:01 to 1990:12, and the whole OOS sample is from 1991:01 to 2021:12. For now, we report results on the OOS period from 1997:01 to 2016:12 to compare the results with Ao, Li, and Zheng (2019).

4.3.1. Baseline results

Table 10 reports the OOS performance without transaction cost.¹¹ Our GP substantially outperforms other models by doubling their Sharpe ratios. When there is no short-sale constraint, Panel A and Panel B show that our GP portfolio (GP-NoCost) generates the greatest annualized Sharpe ratio up to 1.089, almost doubling that of the second-highest annualized Sharpe ratio of 0.591 produced by MAXSER. Incorporating the short-sale constraint, Panel C and Panel D show that our GP portfolio (GP-NoCost-SS) again yields the greatest annualized Sharpe ratio of 0.637, doubling that of the second-highest annualized Sharpe ratio of 0.316 produced by MV-P-SSCV.

¹⁰The data for risk-free rate comes from Kenneth French Data Library.

¹¹The results of MAXSER and other models come from Table 5 and Table 6 of Ao, Li, and Zheng (2019).

Table 11 reports the OOS performance with a transaction cost of 0.1%. Our GP portfolios still significantly outperform other models by doubling their Sharpe ratios. When there is no short-sale constraint, Panel A and Panel B show that our GP portfolio (GP-WithCost) generates the greatest annualized Sharpe ratio up to 0.794, much higher than the second-highest annualized Sharpe ratio of 0.483 produced by MAXSER. In terms of the case with short-sale constraint, Panel C and Panel D show that our GP portfolio (GP-WithCost-SS) again yields the greatest annualized Sharpe ratio of 0.594, tripling that of the second-highest annualized Sharpe ratio of 0.174 produced by MV-P-SSCV.

4.3.2. Penalty for overfitting based on transaction cost

An advantage of our method is that we can incorporate the transaction cost into the optimization objective. In the previous results, we treat the transaction cost as a *fixed* parameter. For example, consider the case of a transaction cost of 0.1% in Table 11, the GP is estimated with a transaction cost of 0.1%, and we examine its OOS performance with the same transaction cost. In this section, we show that the transaction cost can work as a *hyper-parameter* to control model complexity and mitigate overfitting. Assuming the true transaction cost is C_{True} , we show that a transaction cost which is a little bit higher than C_{True} , rather than C_{True} , should be used in (21) to *train* the GP model.¹² This enlargement in transaction cost essentially works as a penalty to control the portfolio turnover and model complexity, which shrinks the in-sample performance but facilitates the OOS performance.

Consider an intuitive example. For any portfolio optimizer M , suppose the turnover for the optimal portfolio under the transaction cost of C is $Turn_{M,C}$. Consider two cost levels $C_1 < C_2$. Then, a favourable optimizer M is expected to generate portfolios with $Turn_{M,C_1} > Turn_{M,C_2}$. In other words, an intelligent portfolio optimizer should learn to decrease its portfolio turnover in response to increasing transaction costs. However, $Turn_{C_1} = Turn_{C_2}$ for MAXSER and other benchmark models, since the transaction cost does not come into the optimization process. On the contrary, the transaction cost \tilde{c}_{t+1} plays a role in our optimization objective (21). Consistently, Table 12 shows that GP can optimize the portfolio according to various transaction cost levels.

Panel A of Table 12 reports the results with short-sale constraints. We estimate the GP model

¹²We should use C_{True} for the OOS test.

with and without transaction costs, and then conduct the OOS backtesting with and without transaction costs, respectively.¹³ Hence, we have four portfolios. There are two interesting findings. First, consistent with our expectation, GP generates portfolios with lower turnover as the transaction cost grows. For example, the turnover for GP-NoCost-SS is 1.810, while the turnover for GP-WithCost-SS is lower at 1.648. Although the difference is relatively small, note that the results are based on a relatively small transaction cost of 0.1%, given that our stock universe is S&P 500 constituents. If we apply a higher transaction cost in smaller-cap and less liquid stocks, the effect should be more significant.

Second, the inclusion of transaction cost in GP estimation helps to improve OOS performance. To see this, notice that the model with transaction cost (GP-WithCost-SS) outperforms the model without transaction cost (GP-NoCost-SS) in both backtesting with and without transaction cost. It might be counter-intuitive at the first glance that GP-WithCost-SS produces a higher annualized Sharpe ratio (0.680) than GP-NoCost-SS (0.637) even when the OOS backtesting is conducted without transaction cost, which is consistent with the objective of GP-NoCost-SS rather than GP-WithCost-SS. One potential explanation is the transaction cost essentially serves as a penalty for overfitting. Intuitively, a model with greater complexity is prone to have higher turnover and is also more likely to overfit. Hence, while the transaction cost decreases the portfolio turnover, it naturally controls model complexity and helps to mitigate overfitting, thus leading to a stronger OOS performance.

The above findings indicate that we can use transaction cost as a *hyper-parameter* to control model complexity and mitigate overfitting. Suppose the true transaction cost is C_{True} , say $C_{True} = 0.10\%$, we can set the transaction cost to $\{0.10\%, 0.12\%, 0.14\%, 0.16\%, 0.18\%, 0.20\%\}$, respectively, to estimate GP model using the *training* sample. Then, we evaluate the associated 6 models in the validation sample using the C_{True} of 0.10% and select the optimal one as the final model to apply in OOS. The final model is not necessarily the one estimated with C_{True} , and is expected to be the one estimated with a cost greater than C_{True} . This enlargement in transaction cost essentially works as a penalty to control the portfolio turnover and model complexity, which shrinks the in-sample performance but facilitates the OOS performance.

¹³In Table 10, GP model in-sample training and OOS backtesting are both conducted without transaction costs. In Table 11, GP model in-sample training and OOS backtesting are both conducted with transaction costs.

5. Robustness

5.1. *Alternative characteristics*

In this subsection, we examine the performance of the machine learning methods when applied to two alternative characteristic set. The first is the typical one of 15 characteristics used by Lewellen (2015). Different to the characteristic set used in the main results which relies heavily on technical indicators, this new data set are mainly fundamental variables: size, book-to-market ratio, the growth in split-adjusted shares outstanding from month -36 to month -1, accrual, ROA, annual growth of total asset, dividend yield, the growth in split-adjusted shares outstanding from month -12 to month -1, market beta, the return from month -12 to month -2, the return from month -36 to month -13, return volatility, turnover, debt-to-price ratio, and sales-to-price ratio.¹⁴ Since this characteristic set uses the accounting data from the Compustat, the sample period is much shorter and starts in January 1976. Hence, we set below the training sample from 1976:01 to 1995:12, the validation sample from 1996:01 to 2000:12, and the OOS sample from 2001:01 to 2021:12. In addition, we use the same price filter as before to exclude stocks with price below \$5.

Table 13 reports the OOS performance under the characteristics in Lewellen (2015). It is important to note that none of the linear models can generate significant returns, although they still have positive returns and still outperform substantially the OLS model (unreported). The neural network models, however, do yield significant gains in 3 out of 5 cases. In contrast, GP still performs the best, earning the greatest significant return of 0.76%, improving the next best one of NN5 (90.5%) by about 35%. Its Sharpe ratio is the largest, 0.63, as expected, which improves the next best one of NN5 (0.41) by about 55%. In contrast to the previous set of characteristics, the new one has less predictability on the cross-section of the stock returns. In this case, the GP outperforms other methods even more in terms of percentage improvement.

The second characteristic set we examined is the top 15 characteristics used in Gu, Kelly, and Xiu (2020). These characteristics are selected based on NN5, which is the best performing model based on the first 30 years in Gu, Kelly, and Xiu (2020), and hence there is so looking ahead bias in selecting the top variables. Note that to ensure comparability, we follow Gu, Kelly, and Xiu (2020)

¹⁴The detailed constructions of these variables are provided by Lewellen (2015) and are also available in the Online Appendix.

to use the full sample stocks to conduct the analysis without applying the price filter used in our main test.

Table 14 reports the OOS performance during the same period from 1987 to 2016 as in Gu, Kelly, and Xiu (2020). Consistent with their results, we also find that neural networks perform considerably better than the linear models. The mean return and the Sharpe ratio of the neural networks are on average about 50% greater than those of the linear models. However, it is important to note that our GP further dominates the neural networks by earning the highest return of 1.68% and the greatest Sharpe ratio of 1.53. On the other hand, the best performance for other machine learning method is lower at 1.66% and 1.10 for NN3. Interestingly, GP’s Sharpe ratio (1.53) is even higher than that of the best one (1.35 for NN4) shown in Table 7 in Gu, Kelly, and Xiu (2020). However, their results are based on a far more complicated characteristics set of 920 variables (including variable interaction), while ours are based on only 15 variables. The results indicate that GP does a good job in uncovering the the non-linear interaction predictability among the variables.

In short, GP consistently outperforms other machine learning models, especially in terms of the Sharpe ratio, under the alternative characteristic sets.

5.2. *International markets*

In this subsection, we examine the performance of the GP in the major international stock markets. As emphasized by Schwert (2003), the use of alternative data sets is one way to mitigate the concern of data-snooping. For brevity, we focus on other G7 countries: the UK, Canada, Japan, Italy, France, and Germany.

There is one unique feature in our applications to the international markets. Instead of re-estimating the machine learning models in each market, we directly apply all of them estimated in the US directly to other markets. Since the data in other markets are not used for neither model estimation nor parameter tuning, they offer a perfect setting to examine the OOS performance.

Table 15 reports the results. There are two notable patterns. First, the GP substantially outperforms other machine learning methods, achieving the largest Sharpe ratio in all the 6 markets. For example, in UK, it has a Sharpe ratio of 1.07 with an average monthly return of 1.78%. Although

the Sharpe ratio from NN2 is high, but the average across the other methods is about 30% lower than the GP. The result is echoed by the average cross the markets, reported in Panel G. The second pattern is that linear models perform well in the international markets relative to the nonlinear neural networks. This differs from the US market where the latter dominates the former. The pattern is interesting and puzzling, and is a subject of future research.

In short, GP performs well not only in the US, but also internationally in other G7 markets, even with the same model estimated in the US. The strong performance of the GP internationally indicates that the method captures salient features of the market and is robust to alternative data sets.

5.3. *Alternative parameters*

For the main results, the GP model is estimated under the hyperparameters $\langle Pop, Gen \rangle = \langle 200, 40 \rangle$, which is determined by the validation sample. We now further examine the robustness under alternative parameters.

5.3.1. *In-sample performance evolution*

Consider alternative parameters for *Pop*: 100, 200, and 400, and that for *Gen*: 10, 20, and 40. There are a total of 9 sets of the hyperparameters. For a given $\langle Pop, Gen \rangle$, we independently estimate GP for 5 times, and get $5 \times Pop$ models (individuals) in total. We use the average of the top M models with the highest Sharpe ratios in the training sample as the final model. Note that $M = 5$ in our main results, and here we also consider alternative values of 3 and 10.

Table 16 reports the results with the alternative $\langle Pop, Gen \rangle$'s and M 's. There are a few interesting facts. First, since $\langle Pop, Gen \rangle$ characterizes the searching depths for GP, the Sharpe ratio in the training sample increases with $\langle Pop, Gen \rangle$. For example, for $Pop=100$ and $M = 5$, the annualized Sharpe ratio grows from 2.27 to 2.87 as *Gen* increases from 10 to 40. Second, while the training sample Sharpe ratio generally increases with *Pop*, the effect is weaker. For example, for $Gen=20$ and $M = 5$, the Sharpe ratio increases from 2.35 to only 2.45 as *Pop* increases from 100 to 400. In general, the in-sample performance increases with $\langle Pop, Gen \rangle$, but is more sensitive with respect to *Gen*.

Third, by comparing the validation Sharpe ratios of various parameters in Panel A, we find that the parameter $\langle Pop, Gen \rangle$ of $\langle 200, 40 \rangle$ achieves the largest validation sample Sharpe ratio of 2.56, supporting our earlier parameter choice. This choice also achieves the best OOS performance: the spread portfolios earns the largest annualized Sharpe ratio of 1.21, as shown earlier in Table 1. Forth, although the objective of GP is to maximize the spread portfolio's Sharpe ratio, we also report the average return of the spread portfolios for other parameters. In general, the spread return exhibits similar patterns as the Sharpe ratios. For example, the training sample return generally increases with $\langle Pop, Gen \rangle$. The largest validation return is also achieved at $\langle Pop, Gen \rangle = \langle 200, 40 \rangle$. Fifth, Panel B and Panel C show similar patterns to Panel A, indicating that the performance is robust to M . Overall, the results are economically not too far apart even though the parameter values are substantially different.

5.3.2. Sharpe ratio evolution

To understand further the performances under the alternative parameters, we examine now how the Sharpe ratio changes in the GP algorithm as the population grows.

Figure 2 presents the plots of the Sharpe ratios. Consider Figure A. Since the Pop is 100, the max number in the X-axis is 100 for the individual. The blue curve plots the training sample Sharpe ratios averaged over the 5 ($M=5$) estimations. The green curve and the red line one are those for the validation sample and OOS sample, respectively. Since we sort the individuals (models) by their training sample Sharpe ratio, the blue curve shows a monotonic increasing pattern. It is clear that OOS performance is weakened in comparison with in-sample and validation. However, it does share the same pattern, indicating that greater in-sample Sharpe ratios tend to generate stronger predictability in the OOS sample.

For a given Pop , as Gen increases, the green and red line become less volatile. For example, as Gen grows from 10 in Figures B to 40 in Figure H, the green and red line becomes increasingly flatter, suggesting that the solution converges to a stable OOS performance. In the Online Appendix, we provide detailed results to show that the model performance volatility decreases with Gen .

6. What drives GP’s performance?

In this section, we explore the reasons why the GP can outperform the other machine learning methods.

6.1. Objective function

An obvious question is whether the objective function plays a role in the performance. To examine this, instead of maximizing the Sharpe ratio as we did before, we now consider the objective of minimizing the conventional mean squared error (MSE) of the predicted returns. We denote this model as GP_{MSE} , and denote the previous GP model of maximizing the Sharpe ratio as GP_{SR} .

Table 17 compares the performance of the two models. The spread portfolio of GP_{MSE} yields an average monthly return of 1.44% with an annualized Sharpe ratio of 1.01. However, it is important to note that GP_{MSE} is dominated by GP_{SR} . Since the Sharpe ratio is a comprehensive metric which considers the trade-off between return and risk, the results show that it does do better in terms of both return and volatility. Indeed, compared with GP_{MSE} , GP_{SR} not only earns a higher spread return of 1.53%, but also produces a lower volatility of 4.38%. As a result, GP_{SR} yields a greater Sharpe ratio of 1.21, about 20% larger than that of GP_{MSE} . In addition, GP_{SR} also earns a higher positive skewness and lower maximum drawdown than GP_{MSE} . This suggests that GP_{SR} does a good job in extreme scenarios, and illustrates the gains of directly considering the economic return-risk trade-off in the optimization.

As an alternative way to compare GP_{MSE} with GP_{SR} , we regress the expected returns generated by the two GP models on each other, and then examine the performance of the resulting spread portfolio sorted by the residuals. The right panel of Table 17 reports the results. Controlling for GP_{SR} , GP_{MSE}^ω generates a negligible spread return of 0.01% with a weak t -statistic of only 0.05, indicating that the predictability of GP_{MSE} is subsumed by GP_{SR} . In contrast, controlling for GP_{MSE} , GP_{GP}^ω still earns a persistent spread return of 0.72% with a significant t -statistic of 4.31, suggesting that GP_{SR} contains additional predictability beyond GP_{MSE} .

In short, compared with the conventional MSE-based models, the reason for the economic gains of using our proposed GP model arises from maximizing the spread portfolio’s Sharpe ratio directly.

By considering both return and risk, the metric produces much higher Sharpe ratio and outperforms the MSE-based models of the GP and other machine learning methods.

6.2. Linearity vs nonlinearity

It is well known that the standard MSE estimator of the parameters is efficient if the data are normally and independent and identically distributed. In this case, there is likely little difference between MSE minimization and Sharpe ratio maximization. However, when the true data have nonlinearity (see, e.g., Freyberger, Neuhierl, and Weber, 2020), the difference will likely be large. We show that this is indeed the case via simulations.

Consider the linear case first. Following Freyberger, Neuhierl, and Weber (2020), we simulate data from a linear model with a set of fixed predictors:

1. Assume the “true” predictor set Z consists of *Size*, *SREV*, *MOM*, and *LREV*.
2. Regress the stock return R on the assumed predictor set Z in a panel regression, pooled over the entire sample from 1945 to 2019. Then, decompose R into the fitted part ($\hat{R}_{i,t}$) and the residual ($\epsilon_{i,t}$).
3. Generate returns according to $\tilde{R}_{i,t} = \hat{R}_{i,t} + \tilde{\epsilon}_{i,t}$, where $\tilde{\epsilon}_{i,t}$ is resampled with replacement from the empirical residuals in step 2. To generate the residuals in a particular month t , we first draw a random time period, say month s , from which we sample the residuals. Moreover, to ensure we sample from the distribution with zero means, we re-center the original residuals each month.
4. Based on the simulated return $\tilde{R}_{i,t}$ from step 3 and the predictor set for investment use, Q , consisting of *Size*, *SREV*, *MOM*, and *LREV*, we estimate the GP and other benchmark models, and examine their OOS performance.

Note that Q is the same as Z from step 1, and this is equivalent to assuming the true predictors are known to investors.

5. Redo steps 3-4 for 500 times.

For the nonlinear case, the simulation procedure is similar, except that we add the interaction terms in the true predictor set. That is, Z now consists of 10 variables: $Size$, $SREV$, MOM , and $LREV$, as well as 6 pairwise interaction terms of these four variables. Suppose the true data process is generated by this new predictor set. We then estimate the coefficients in a panel regression. In particular, we scale the slopes on the interaction terms to make them comparable to those of the four original predictors.¹⁵

Note that in the linear simulation, the true predictor set Z is the same as the indicator set Q , which is the input data for training the models. In this case, the linear model is the true model and hence is expected to perform well. In the nonlinear simulation, however, the true predictor set Z include the nonlinear interaction effects, while we still use the same indicator set Q for forecasting. Since GP captures nonlinearity, we expect that GP will show its strength in the nonlinear simulation.

Table 18 reports the average OOS statistics in the linear and nonlinear simulations. In particular, we also consider a special benchmark model, the fitted return $\hat{R}_{i,t}$ from step 2 in the simulation procedure. $\hat{R}_{i,t}$, by construction, contains *all* the predictability in the simulated return $\tilde{R}_{i,t}$, and hence, it can be interpreted as the optimal model.

The left panel reports the Sharpe ratios. In the linear simulation, the model of $\hat{R}_{i,t}$ produce a Sharpe ratio of 1.33. GP earns a Sharpe ratio of 1.24, which is only slightly less than the optimal model. Meanwhile, consistent with our prediction, all the linear models performs well in the linear simulation, and yield Sharpe ratios around 1.30, very close to that of $\hat{R}_{i,t}$. It is not surprising to see the good performance of the linear models, because, in the linear simulation, these models are the true models and hence are expected to achieve similar performance with $\hat{R}_{i,t}$.

In the nonlinear simulation, as we include the nonlinear interaction terms, which increases the overall predictability, the Sharpe ratio of the optimal model $\hat{R}_{i,t}$ increases substantially by 153%, from 1.33 to 3.36. It is important to note that the Sharpe ratio of GP also grow significantly by 148%, from 1.24 to 3.08, which is close to that of the $\hat{R}_{i,t}$. On the contrary, although linear models also produce better performance in the nonlinear case, the resulting Sharpe ratio of around 2.05 is much lower than that for GP and $\hat{R}_{i,t}$, and the rowth rate of 50% is also much smaller than that

¹⁵We multiply the slopes on the interaction terms by 8, and also get qualitatively robust results under alternative values.

for GP.

The right panel reports the mean returns. In the linear case, the return of the linear models are very close to that of $\hat{R}_{i,t}$. Since the objective of our GP is to maximize the Sharpe ratio rather than the return, GP earns a little bit lower OOS returns than other linear models in the linear case. But in the nonlinear case, GP yields higher return than other linear models.

In short, consistent with our prediction, while linear models perform well in the linear simulation, GP outperforms linear models in the nonlinear simulation. This evidence suggests that the ability to exploit nonlinear predictability is another source for GP's good performance.

6.3. Bootstrap with different sample size

In this subsection, we carry out bootstrap analysis to compare the performance of GP and NN under different sample sizes.

We choose two different sample size. In the first simulation, in each month t , we resample stocks so that the stock number in the simulated data is only *half* of the actual stock number. In the second simulation, we do the same but double the stock number in the simulation each month. Then, based on the simulated data, we estimate GP and NN models, and examine their OOS performance.

Table 19 reports the average OOS statistics for the spread portfolios of GP and NN.¹⁶ The performance of GP is robust to the sample size. For example, in both sample sizes, GP earns an annualized Sharpe ratio of about 1.00 and a mean return of 1.60%. In contrast, the performance of NN is sensitive to the sample size. In particular, when we reduce the sample size to half, the OOS performances of NN become substantially worse. As the sample size increases, they become better. In short, GP has stable performances under reasonable sample sizes, while NN is more sensitive to it. This provides another reason (besides objective functions) why, although both are nonlinear models, the GP has better performances previously than the NNs.

¹⁶The time for estimating GP model increases with sample size. To save time, the parameter $\langle Pop, Gen \rangle$ for GP in this analysis is $\langle 100, 10 \rangle$. We repeat the simulation for 10 times, and the table reports the average statistics.

7. Conclusion

In this paper, we propose to maximize the Sharpe ratio of a portfolio via genetic programming (GP), one of the machine learning tools applied here the first time for the study of the cross-section of stock returns. Our approach directly optimizes the Sharpe ratio by searching a function that maps from the stock characteristics to the expected stock returns in a large functional space. We find that the performance of the GP spread portfolio in the cross-section outperforms substantially the usual MSE-based models, such as ridge, lasso, Enet, PCR, and PLS. It also outperforms significantly the more powerful neural networks by subsuming their predictability. While existing factor models fail to explain the performance of the MSE-based machines learning methods, a single factor based on the GP fully captures all their spread portfolios. The performance of the GP is robust to alternative parameters, different characteristics, and international data sets. We find further that the good performance of the GP is due to its economic objection optimization, and it is less sensitive to sample size than the neural networks.

Our second contribution is to apply GP to maximize the Sharpe ratio in the standard portfolio optimization problem, and propose a regularization method based on transaction cost to control model complexity and mitigate overfitting problem. We solve the portfolio optimization *without* estimating μ and Σ and can easily incorporate various economic constraints into the optimization process. GP again substantially outperforms conventional methods and other machine learning-based methods, including the maximum-Sharpe-ratio estimated and sparse regression (MAXSER) recently proposed by Ao, Li, and Zheng (2019), by almost doubling the out-of-sample performance, especially in the case with transaction cost and short-sale constraints.

Our empirical evidence suggests that it is important to apply machine learning tools to maximize economic objectives, beyond the scope of the traditional model fitting. Since the Sharpe ratio is one of the most important performance measure of a trading strategy, the present framework can be applied in many areas to maximize the Sharpe ratio. It will not only be useful for a fund managers to improve investment performance in various asset classes, but also be useful for researchers to identify potentially the largest anomalies in currencies, corporate bonds or commodities. These are interesting issues for future research.

References

- Allen, F., Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics* 51, 245-271.
- Ao, M., Yingying, L., Zheng, X., 2019. Approaching mean-variance efficiency for large portfolios. *Review of Financial Studies* 32, 2890-2919.
- Avramov, D., Cheng, S., Metzker, L., Voigt, S., 2022. Integrating factor models. *Journal of Finance*, forthcoming.
- Avramov, D., Cheng, S., Metzker, L., 2021. Machine learning versus economic restrictions: Evidence from stock return predictability. *Management Science*, forthcoming.
- Barillas, F., Shanken, J., 2017. Which alpha?. *Review of Financial Studies* 30, 1316-1338.
- Best, M.J., Grauer, R.R., 1991. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies* 4, 315-342.
- Bhowan, U., Johnston, M., Zhang, M., Yao, X., 2012. Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*, 17, 368-386.
- Britten-Jones, M., 1999, The sampling error in estimates of mean-variance efficient portfolio weights, *Journal of Finance* 54, 655-671.
- Brogaard, J., Zareei A., 2018. Machine learning and the stock market. Workingpaper.
- Bryzgalova, S., Huang, J., Julliard, C., 2022. Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance*, forthcoming.
- Bryzgalova, S., Pelger, M., Zhu, J., 2021. Forest through the trees: Building cross-sections of stock returns. Available at SSRN 3493458.
- Chen, A. Y., Velikov, M., 2020. Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, forthcoming

- Chinco, A., Clark-Joseph, A.D., Ye, M., 2019. Sparse signals in the cross-section of returns. *Journal of Finance* 74, 449-492.
- Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. *Review of Financial Studies* 33, 2134–2179
- Cong, L. W., Feng, G., He, J., He, X., 2022. Asset pricing with panel tree under global split criteria. Working paper
- Cong, L. W., Tang, K., Wang, J., Zhang, Y., 2021. Alphaportfolio: Direct construction through reinforcement learning and interpretable AI. Working paper
- Daniel, K., Hirshleifer, D., Sun, L., 2020. Short-and long-horizon behavioral factors. *Review of Financial Studies* 33, 1673-1736
- Darwin, C., 2004. *On the origin of species*, 1859. Routledge.
- DeBondt, W.F.M., Thaler, R., 1985. Does the stock market overreact? *Journal of Finance* 40, 783-805.
- DeMiguel, V., Martin-Utrera, A., Nogales, F. J., Uppal, R., 2020. A portfolio perspective on the multitude of firm characteristics. *Review of Financial Studies* 33, 2180–2222
- Dempster, M.A. and Jones, C.M., 2001. A real-time adaptive trading system using genetic programming. *Quantitative Finance*, 1, 397-413.
- Dunis, C.L., Laws, J., Middleton, P.W., Karathanasopoulos, A., 2015. Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models. *The European Journal of Finance*, 21, 352-375.
- Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33, 3-56.
- Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. *Journal of financial economics* 116, 1-22.
- Fama, E. F., MacBeth, J. D. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3), 607-636 .

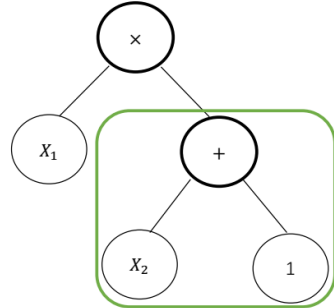
- Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. *Journal of Finance*, 75, 1327-1370.
- Filippou, I., Rapach, D., Taylor, M.P., Zhou, G., 2020. Exchange Rate Prediction with Machine Learning and a Smart Carry Portfolio. Available at SSRN 3455713.
- Freyberger, J., Neuhierl A., Weber, M., 2020. Dissecting characteristics nonparametrically, *Review of Financial Studies*, 33, 2326-2377.
- Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, 30, 4389-4436.
- Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223-2273.
- Guo, X., Lin, H., Wu, C., Zhou, G., 2022. Investor Sentiment and the Cross-Section of Corporate Bond Returns. Available at SSRN 3223846.
- Han, Y., He, A., Rapach, D. and Zhou, G., 2021. Expected stock returns and firm characteristics: E-LASSO, assessment, and implications. Available at SSRN 3185335.
- Han, Y., Zhou, G., Zhu, Y. 2016. A trend factor: Any economic gains from using information over investment horizons?. *Journal of Financial Economics* 1222, 352-375.
- Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28, 650-705.
- Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48, 65-91.
- Jiang, J., Kelly, B. T., Xiu, D, 2020. (Re-)Imag(in)ing price trends. Available at SSRN 3756587.
- Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis* 42, 621-656.
- Kan, R., Zhou, G., 2012. Tests of mean-variance spanning. *Annals of Economics and Finance* 13, 139-187.

- Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge
- Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. *Journal of Financial Economics*, 135, 271-292.
- LeBaron, B., Arthur, W.B., Palmer, R., 1999. Time series properties of an artificial stock market. *Journal of Economic Dynamics and control*, 23, 1487-1516.
- Ledoit, O., Wolf, M., 2008. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance* 15, 850-859.
- Lehmann, B.N., 1990. Fads, martingales and market efficiency. *Quarterly Journal of Economics* 105, 1-28
- Lewellen, J. 2015. The Cross-section of Expected Stock Returns. *Critical Finance Review* 4, 1-44.
- Lo, A.W., MacKinlay, A.C., 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3, 175-205.
- Markowitz, H., 1952, Portfolio selection, *Journal of Finance* 7, 77–91.
- Michaud, R.O., 1989. The Markowitz optimization enigma: Is ‘optimized’ optimal?. *Financial analysts Journal* 45, 31-42.
- Neely, C., Weller, P., Dittmar, R., 1997. Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of financial and Quantitative Analysis*, 32, 405-426.
- Newey, W.K., West, K. D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703-708
- Nordhaus, W.D., 2001. The progress of computing. Available at SSRN 285167.
- Patton, A. J., Weller, B., 2020. What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics* 137, 515–549

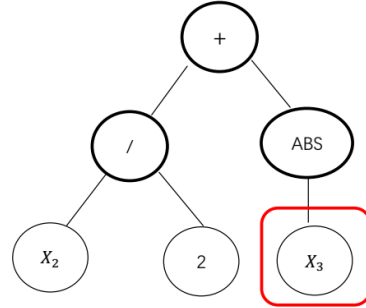
- Polachek, S.W., Das, T., Thamma-Apiroam, R., 2015. Micro-and macroeconomic implications of heterogeneity in the production of human capital. *Journal of Political Economy*, 123, 1410-1455.
- Ready, M.J., 2002. Profits from technical trading rules. *Financial Management*, 43-61.
- Robert, E., Robert, F., Jeffrey, R., 2012. Measuring and modeling execution cost and risk. *Journal of Portfolio Management* 38, 14-28.
- Schwert, G.W., 2003. Anomalies and market efficiency. In: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), *Handbook of the Economics of Finance*, 1. Elsevier, Amsterdam, Netherlands, pp. 939-974. chap. 15.
- Stambaugh, R.F., Yuan, Y., 2016. Mispricing factors. *The Review of Financial Studies* 30, 1270-1315.
- Winschel, V., Kräitzig, M., 2010. Solving, estimating, and selecting nonlinear dynamic models without the curse of dimensionality. *Econometrica*, 78, 803-821.
- Zhang, X.F., 2006. Information uncertainty and stock returns. *Journal of Finance*, 61, 105-137.
- Zhang, Y., Bhattacharyya, S., 2004. Genetic programming in classifying large-scale data: an ensemble method. *Information Sciences*, 163, 85-101.

Figure 1: **Tree-structured representation and genetic operators**

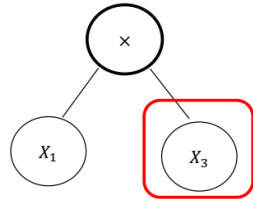
This figure illustrates the tree-structured individuals and the genetic operators of crossover and mutation. The parents individuals in Panel A and B are combined by the crossover operator, and the resulting offspring individuals are shown in Panel C and D. The offspring individual in Panel E (F) is produced by the mutation operator from the individual in Panel A (B).



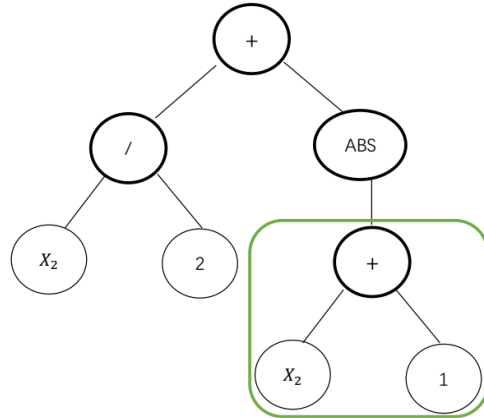
(A) $G(X) = X_1 * (X_2 + 1)$



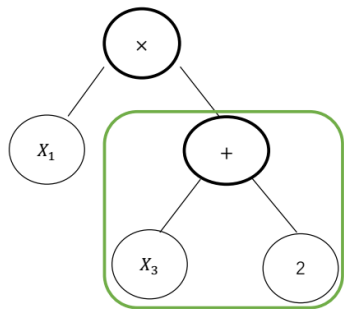
(B) $G(X) = 0.5X_2 + |X_3|$



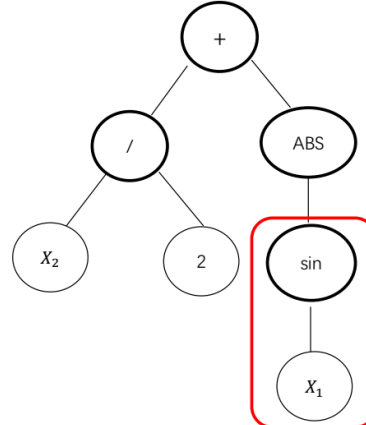
(C) $G(X) = X_1 * X_3$



(D) $G(X) = 0.5X_2 + |X_2 + 1|$



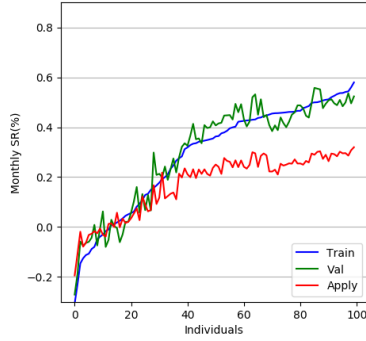
(E) $G(X) = X_1 * (X_3 + 2)$



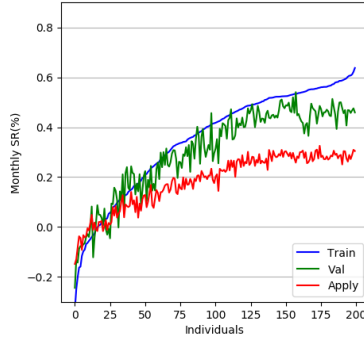
(F) $G(X) = 0.5X_2 + |\sin(X_1)|$

Figure 2: **GP's performance under various hyperparameters**

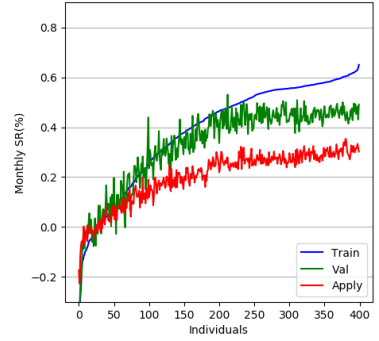
This figure shows the monthly Sharpe ratio of the spread portfolios generated by GP under various parameters. For a given set of the parameter $\langle Pop, Gen \rangle$, we independently estimate GP model using training sample for five times and get Pop individuals each time. We sort the individuals within each time by their associated Sharpe ratio in the training sample. The blue (green, or red) lines show the Sharpe ratios of the individuals average over the five estimations in the training (validation, or OOS) sample.



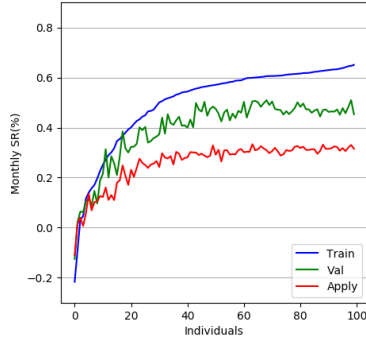
(A) $\langle Pop, Gen \rangle: \langle 100, 10 \rangle$



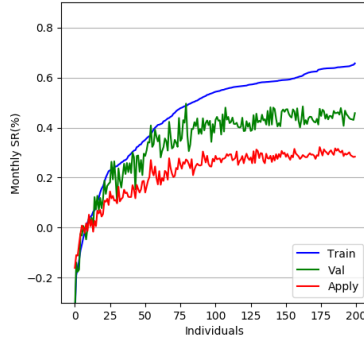
(B) $\langle Pop, Gen \rangle: \langle 200, 10 \rangle$



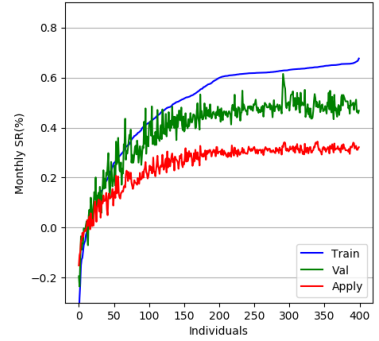
(C) $\langle Pop, Gen \rangle: \langle 400, 10 \rangle$



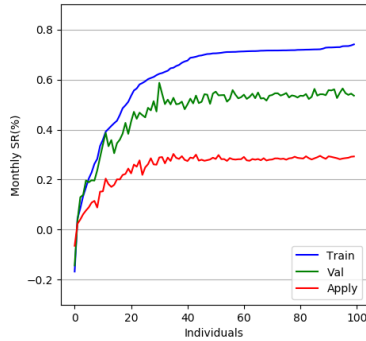
(D) $\langle Pop, Gen \rangle: \langle 100, 20 \rangle$



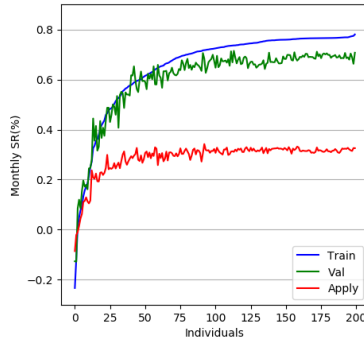
(E) $\langle Pop, Gen \rangle: \langle 200, 20 \rangle$



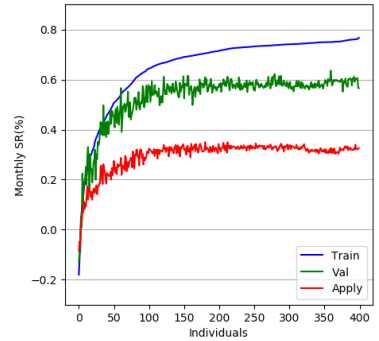
(F) $\langle Pop, Gen \rangle: \langle 400, 20 \rangle$



(G) $\langle Pop, Gen \rangle: \langle 100, 40 \rangle$



(H) $\langle Pop, Gen \rangle: \langle 200, 40 \rangle$



(I) $\langle Pop, Gen \rangle: \langle 400, 40 \rangle$

Table 1

Spread portfolios

The table reports the summary statistics for the decile spread portfolios generated by the GP and other models. For each model, we report the average monthly return in percentage points, the Newey-west (1987) robust t -statistic, standard deviation in percentage points, the annualized Sharpe ratio (*Sharpe*), the skewness (*Skew*), and the maximum drawdown (MDD) in percentage points. The sample period is from 1991:01 to 2021:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Low	0.29	0.75	0.72	0.70	0.79	0.75	0.86	0.83	0.79	0.57	0.46
2	0.60	0.78	0.78	0.86	0.81	0.79	0.88	0.80	0.97	0.73	0.77
3	0.67	0.94	0.96	0.88	0.87	0.93	0.90	0.99	0.92	0.84	0.78
4	0.59	0.95	1.01	1.01	1.01	0.97	0.97	0.99	1.04	0.90	0.93
5	0.81	1.18	1.15	1.14	1.20	1.17	1.19	1.12	1.18	0.92	0.98
6	1.10	1.14	1.14	1.16	1.18	1.14	1.13	1.30	1.06	0.97	0.98
7	1.16	1.30	1.43	1.37	1.24	1.30	1.12	1.31	1.19	1.19	1.01
8	1.53	1.43	1.38	1.40	1.52	1.42	1.16	1.40	1.40	1.23	1.10
9	1.56	1.56	1.47	1.53	1.42	1.58	1.30	1.41	1.46	1.40	1.39
High	1.82	1.71	1.75	1.68	1.64	1.69	1.67	1.94	1.93	1.70	1.49
H-L	1.53***	0.95***	1.03***	0.98***	0.84***	0.93***	0.80***	1.10***	1.14***	1.12***	1.02***
t-stat	(6.00)	(3.14)	(3.26)	(3.29)	(2.71)	(3.07)	(3.24)	(3.39)	(4.17)	(4.63)	(4.00)
Std. dev.	4.38	5.18	5.15	5.11	4.98	5.20	4.61	4.86	5.37	4.68	4.30
Sharpe	1.21	0.64	0.69	0.67	0.59	0.62	0.61	0.79	0.74	0.83	0.82
Skew	0.83	0.08	-0.11	-0.06	0.15	0.06	-0.36	1.02	0.82	0.21	0.15
MDD	22.56	46.19	43.95	43.25	43.26	46.25	39.29	56.05	46.45	31.04	30.74

Table 2

Subperiod performance

The table reports the summary statistics for the decile spread portfolios generated by the GP and other models over two subperiods. For each model, we report the average monthly return in percentage points, the Newey-west (1987) robust t -statistic, standard deviation in percentage points, the annualized Sharpe ratio (*Sharpe*), the skewness (*Skew*), and the maximum drawdown (MDD) in percentage points. The sample period in Panel A is from 1991:01 to 2003:12, and in Panel B is from 2004:01 to 2021:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Panel A: 1991:01-2003:12											
H-L	2.62***	2.03***	2.09***	1.92***	1.75***	2.01***	1.60***	2.27***	1.88***	1.85***	2.02***
t-stat	(6.85)	(4.64)	(4.76)	(4.40)	(3.52)	(4.56)	(4.51)	(5.07)	(4.45)	(5.47)	(5.49)
Std. dev.	5.07	5.52	5.29	5.35	5.07	5.51	4.86	5.12	5.93	4.82	4.41
Sharpe	1.79	1.27	1.38	1.25	1.20	1.27	1.15	1.54	1.10	1.34	1.59
Skew	0.89	0.50	0.25	0.35	0.42	0.55	0.26	1.70	1.17	0.13	0.43
MDD	15.11	21.57	24.09	24.23	29.96	22.23	15.01	22.05	34.25	26.03	12.87
Panel B: 2004:01-2021:12											
H-L	0.74***	0.17	0.25	0.29	0.19	0.15	0.22	0.26	0.60**	0.59**	0.29
t-stat	(3.90)	(0.55)	(0.73)	(0.90)	(0.58)	(0.49)	(0.83)	(0.76)	(1.98)	(2.11)	(1.35)
Std. dev.	3.61	4.78	4.91	4.82	4.81	4.81	4.34	4.47	4.86	4.50	4.06
Sharpe	0.71	0.13	0.18	0.22	0.14	0.11	0.18	0.20	0.43	0.46	0.25
Skew	0.13	-0.54	-0.52	-0.57	-0.12	-0.62	-1.10	0.23	0.21	0.23	-0.20
MDD	22.56	46.19	43.95	43.25	43.26	46.25	39.29	56.05	46.45	31.04	30.74

Table 3

Spread portfolios controlling for other models

This table reports the summary statistics for the decile spread portfolios of each model controlling for one of the other models. Panel A provides the results for the GP controlling for one of the other models, and Panel B provides the results for other models controlling for the GP. The sample period is from 1991:01 to 2021:12.

	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Panel A: GP, controlling for other models										
Low	0.73	0.73	0.72	0.68	0.73	0.63	0.71	0.74	0.68	0.65
2	1.00	1.01	1.04	0.97	0.99	0.87	0.91	0.91	0.96	0.89
3	1.11	1.12	1.14	1.12	1.10	1.07	1.06	0.97	0.89	0.98
4	1.12	1.11	1.01	1.34	1.18	0.98	1.17	1.14	1.03	1.11
5	1.10	1.28	1.34	1.08	1.11	1.26	1.08	1.11	1.19	1.20
6	1.22	1.20	1.23	1.14	1.15	1.26	1.23	1.21	1.23	1.25
7	1.55	1.25	1.52	1.49	1.53	1.05	1.15	1.42	1.16	1.25
8	1.24	1.57	1.46	1.17	1.23	1.35	1.41	1.30	1.36	1.28
9	1.50	1.52	1.40	1.54	1.48	1.33	1.31	1.25	1.46	1.34
High	1.36	1.34	1.34	1.38	1.38	1.34	1.30	1.34	1.40	1.42
H-L	0.63***	0.60***	0.62***	0.70***	0.64***	0.70***	0.58***	0.60***	0.71***	0.76***
t-stat	(4.08)	(3.97)	(3.81)	(4.27)	(4.14)	(4.10)	(3.38)	(4.15)	(4.68)	(4.37)
Panel B: Other models, controlling for GP										
Low	1.14	1.18	1.24	1.15	1.14	1.07	1.00	0.74	1.06	1.23
2	1.34	1.55	1.22	1.55	1.34	1.21	1.13	1.05	1.26	1.41
3	1.49	1.37	1.47	1.39	1.52	1.05	1.00	1.29	1.42	1.31
4	1.32	1.27	1.41	1.33	1.23	1.39	1.00	1.17	1.26	1.48
5	1.18	1.39	1.30	1.22	1.20	1.12	1.10	1.27	1.31	1.35
6	1.30	1.38	1.19	1.18	1.33	1.12	1.03	1.38	1.22	1.27
7	1.15	1.07	1.15	1.16	1.16	1.22	1.10	1.15	1.09	1.18
8	1.11	1.11	1.08	1.14	1.10	1.12	0.98	1.06	1.18	1.21
9	0.92	0.95	0.95	0.97	0.92	1.12	1.01	1.08	1.01	1.02
High	1.00	0.99	0.99	1.00	1.00	0.94	1.00	1.01	0.99	0.98
H-L	-0.14	-0.18	-0.25	-0.14	-0.13	-0.13	-0.00	0.26	-0.06	-0.24
t-stat	(-0.58)	(-0.69)	(-0.95)	(-0.54)	(-0.57)	(-0.69)	(-0.02)	(1.24)	(-0.16)	(-0.77)

Table 4

Fama-MacBeth regressions

This table reports results for the Fama-MacBeth regressions. The dependent variable is the individual stock returns. The independent variables are the signals generated by various machine learning methods. Panel A reports the results during the full OOS period from 1991:01 to 2021:12. Panel B and Panel C reports the results for the subperiod from 1991:01 to 2003:12 and the subperiod from 2004:01 to 2021:12, respectively.

	Panel A: 1991:01-2021:12			Panel B: 1991:01-2003:12			Panel C: 2004:01-2021:12		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GP	0.0056*** (7.42)	0.0039*** (6.46)	0.0036*** (6.16)	0.0091*** (9.08)	0.0055*** (5.90)	0.0053*** (5.52)	0.0030*** (5.91)	0.0027*** (4.12)	0.0023*** (4.25)
Ridge		-1.1041 (-0.35)	-1.7151 (-0.52)		4.2526 (0.92)	2.7283 (0.55)		-4.9729 (-1.26)	-4.9242 (-1.16)
LASSO		-3.8861** (-2.39)	-2.9744* (-1.92)		-2.3544 (-0.82)	-1.0420 (-0.38)		-4.9924*** (-2.68)	-4.3700** (-2.50)
ENET		4.0598** (2.05)	3.1125* (1.69)		2.4384 (0.74)	1.2867 (0.41)		5.2308** (2.17)	4.4312** (2.01)
PCR		0.4953* (1.73)	0.3545 (1.12)		-0.2384 (-0.54)	-0.4674 (-0.96)		1.0253*** (3.18)	0.9482*** (2.65)
PLS		0.7304 (0.23)	1.0978 (0.33)		-3.5306 (-0.74)	-2.5990 (-0.52)		3.8079 (0.96)	3.7677 (0.89)
NN1			0.0654 (1.48)			0.0976 (1.59)			0.0421 (0.68)
NN2			0.3039*** (4.62)			0.4899*** (5.50)			0.1696** (2.13)
NN3			0.1390 (1.23)			0.3423 (1.42)			-0.0078 (-0.11)
NN4			0.1956* (1.91)			0.2481 (1.55)			0.1576 (1.15)
NN5			-0.2221 (-1.34)			-0.6287** (-2.17)			0.0714 (0.45)
Intercept	0.0119*** (4.82)	0.0081*** (3.23)	0.0072*** (3.25)	0.0149*** (4.77)	0.0075** (2.39)	0.0084** (2.39)	0.0099*** (2.75)	0.0086** (2.32)	0.0063** (2.17)

Table 5

Performance under information uncertainty

This table reports the β_L and β_H and their t -stats for the regression:

$$\Delta R_t = \beta_L Low_{t-1}^{Vol} + \beta_H High_{t-1}^{Vol} + \beta MKT_t + \epsilon_t,$$

where ΔR_t is the spread portfolio return of the GP minus the spread of one of the other models, and Low_{t-1}^{Vol} and $High_{t-1}^{Vol}$ are dummy variables indicating high- and low-IVOL periods, as classified based on the median level of the IVOL index, which is defined as the cross-sectional mean of the IVOL of individual stocks. The last row ‘‘Average’’ reports the statistics average over the 10 models. The sample period is from 1991:01 to 2021:12.

	β_L	t-stat	β_H	t-stat
Ridge	0.33	(1.47)	1.26***	(4.26)
LASSO	0.26	(1.05)	1.18***	(3.74)
ENET	0.19	(0.80)	1.34***	(3.99)
PCR	0.28	(1.29)	1.46***	(4.77)
PLS	0.35	(1.48)	1.28***	(4.32)
NN1	0.10	(0.36)	1.25***	(3.57)
NN2	0.38	(1.56)	0.78**	(2.47)
NN3	0.57**	(2.23)	0.80**	(2.04)
NN4	0.37	(1.62)	0.72**	(2.16)
NN5	0.47**	(2.16)	0.68**	(2.48)
Average	0.33	1.40	1.08	3.38

Table 6

Comparison with existing factors

This table provides the summary statistics of the GP factor (GPF) and the well known factors. Panel A reports the average monthly return (*Mean*) (%), the Newey-west (1987) robust *t*-statistics, the standard deviation (*Std.dev.*) (%) the annual Sharpe ratio (*Sharpe*), the skewness (*Skew*), and kurtosis (*Kurt*). Panel B reports the correlation matrix. The sample period is from 1991:01 to 2021:12.

	GPF	MKT	SMB	HML	RMW	CMA	ME	IA	ROE	PEAD	FIN
Panel A: Summary Statistics											
Mean	1.16***	0.80***	0.18	0.13	0.35**	0.19	0.24	0.18	0.42***	0.51***	0.50*
t-stat	(6.59)	(3.33)	(1.19)	(0.64)	(2.16)	(1.54)	(1.55)	(1.60)	(2.77)	(4.32)	(1.96)
Std. dev.	2.47	4.31	3.07	3.16	2.66	2.02	3.14	2.06	2.85	2.04	4.32
Sharpe	1.63	0.65	0.21	0.15	0.47	0.33	0.27	0.31	0.52	0.88	0.40
Skew	0.91	-0.63	0.38	0.08	-0.38	0.62	0.75	0.43	-0.86	0.32	0.01
Panel B: Correlation matrix											
GPF	1.00	0.16	0.14	-0.11	-0.17	-0.13	0.16	-0.14	-0.13	0.10	-0.18
MKT	0.16	1.00	0.23	-0.10	-0.35	-0.34	0.24	-0.29	-0.43	-0.12	-0.53
SMB	0.14	0.23	1.00	-0.02	-0.46	-0.02	0.97	-0.08	-0.46	0.06	-0.46
GML	-0.11	-0.10	-0.02	1.00	0.36	0.62	0.02	0.65	0.07	-0.26	0.63
RMW	-0.17	-0.35	-0.46	0.36	1.00	0.24	-0.44	0.29	0.70	-0.08	0.77
CMA	-0.13	-0.34	-0.02	0.62	0.24	1.00	0.01	0.91	0.13	-0.12	0.57
ME	0.16	0.24	0.97	0.02	-0.44	0.01	1.00	-0.06	-0.36	0.07	-0.44
IA	-0.14	-0.29	-0.08	0.65	0.29	0.91	-0.06	1.00	0.15	-0.18	0.65
ROE	-0.13	-0.43	-0.46	0.07	0.70	0.13	-0.36	0.15	1.00	0.20	0.55
PEAD	0.10	-0.12	0.06	-0.26	-0.08	-0.12	0.07	-0.18	0.20	1.00	-0.11
FIN	-0.18	-0.53	-0.46	0.63	0.77	0.57	-0.44	0.65	0.55	-0.11	1.00
MGMT	-0.12	-0.45	-0.32	0.68	0.50	0.74	-0.29	0.76	0.34	-0.08	0.81
PERF	0.04	-0.45	-0.14	-0.23	0.41	0.04	-0.12	0.00	0.66	0.43	0.24

Table 7

Spanning test and Sharpe ratio test

Panel A reports six spanning tests of whether the GP factor can be spanned by various factor models: W , the Wald test under conditional homoskedasticity; W_e , the Wald test under the IID elliptical; W_a the Wald test under the conditional heteroskedasticity; J_1 , the Bekaert-Urias test with the Errors-in-Variables (EIV) adjustment; J_2 is the Bekaert-Urias test without the EIV adjustment, and J_3 , the DeSantis test. The p -values are in brackets. Panel B reports the results of the Sharpe ratio test. “*Original*” reports the squared monthly Sharpe ratios (Sh^2) of a model. “*With GPF*” reports the squared monthly Sharpe ratios for a model plus the GP factor. “ $\Delta(Sh^2)$ ” reports the Sh^2 difference of the two models. The bootstrap p -value, for the null hypothesis of no difference, is reported in brackets, following Ledoit and Wolf (2008) with a repetition number of 4999. The sample period is from 1991:01 to 2021:12.

Panel A: Spanning test						
	W	W_e	W_a	J_1	J_2	J_3
CAPM	964.8*** [0.00]	451.0*** [0.00]	647.8*** [0.00]	72.5*** [0.00]	72.2*** [0.00]	445.4*** [0.00]
FF-3	260.7*** [0.00]	171.9*** [0.00]	166.1*** [0.00]	74.6*** [0.00]	83.1*** [0.00]	145.9*** [0.00]
FF-5	117.1*** [0.00]	67.3*** [0.00]	88.4*** [0.00]	72.1*** [0.00]	81.1*** [0.00]	113.6*** [0.00]
HXZ-4	115.5*** [0.00]	67.5*** [0.00]	89.6*** [0.00]	68.2*** [0.00]	76.0*** [0.00]	93.2*** [0.00]
DHS-3	102.4*** [0.00]	61.5*** [0.00]	71.8*** [0.00]	56.2*** [0.00]	63.2*** [0.00]	84.6*** [0.00]
SY-4	77.0*** [0.00]	54.0*** [0.00]	62.1*** [0.00]	46.5*** [0.00]	54.3*** [0.00]	61.7*** [0.00]
Panel B: Sh^2 in the Sharpe ratio test						
	Original	With GPF	$\Delta(Sh^2)$	p -value		
CAPM	0.03	0.23	0.19***	[0.00]		
FF-3	0.04	0.25	0.20***	[0.00]		
FF-5	0.13	0.36	0.22***	[0.00]		
HXZ-4	0.14	0.36	0.22***	[0.00]		
DHS-3	0.19	0.44	0.24***	[0.00]		
SY-4	0.21	0.45	0.24***	[0.00]		

Table 8

Risk-adjusted returns

The table reports the risk-adjusted returns of the spread portfolios generated by the GP and other methods. Newey-west (1987) robust t -statistics are reported in parentheses. The sample period is from 1991:01 to 2021:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
CAPM	1.32*** (5.86)	0.60** (2.13)	0.68** (2.32)	0.64** (2.27)	0.56* (1.94)	0.58** (2.06)	0.73*** (2.79)	0.84*** (2.94)	0.68*** (2.63)	0.84*** (3.32)	0.80*** (3.16)
FF-3	1.33*** (5.76)	0.54** (2.41)	0.62*** (2.59)	0.57** (2.53)	0.47** (2.17)	0.52** (2.29)	0.66*** (2.91)	0.76*** (3.26)	0.65*** (2.92)	0.78*** (3.31)	0.75*** (3.21)
FF-5	1.49*** (5.51)	0.59*** (2.63)	0.69*** (2.86)	0.64*** (2.80)	0.49** (2.21)	0.58** (2.51)	0.62*** (2.68)	0.78*** (3.14)	0.83*** (3.62)	0.68*** (2.87)	0.72*** (3.25)
HXZ-4	1.41*** (5.72)	0.53** (2.30)	0.58** (2.35)	0.53** (2.25)	0.41* (1.78)	0.51** (2.16)	0.33 (1.26)	0.68*** (2.86)	0.75*** (3.34)	0.66** (2.51)	0.62** (2.52)
DHS-3	1.43*** (6.42)	1.22*** (4.72)	1.25*** (4.61)	1.24*** (4.82)	1.04*** (3.83)	1.19*** (4.59)	0.91*** (3.13)	1.29*** (4.77)	1.31*** (5.39)	1.10*** (3.96)	1.00*** (3.85)
SY-4	1.43*** (5.83)	0.70*** (2.87)	0.79*** (2.90)	0.71*** (2.86)	0.55** (2.21)	0.68*** (2.76)	0.41 (1.51)	0.84*** (3.53)	0.94*** (4.02)	0.88*** (3.23)	0.66*** (2.73)
GPF	-0.10 (-0.75)	-0.28 (-0.96)	-0.18 (-0.60)	-0.19 (-0.65)	-0.25 (-0.86)	-0.30 (-0.99)	-0.04 (-0.16)	-0.04 (-0.16)	-0.07 (-0.25)	0.02 (0.08)	-0.09 (-0.46)

Table 9

List of benchmark portfolios and their abbreviations

The table lists the benchmark portfolios and their abbreviations. “GP” stands for genetic programming, “MAXSER” stands for maximum-Sharpe-ratio estimated and sparse regression, “MV” stands for mean-variance portfolio, and “GMV” stands for global minimum variance portfolio.

Portfolio	Abbreviation
<i>Panel A: Baseline models</i>	
Plug-in MV on Fama and French (1993) 3-factors	Factor
Three-fund portfolio by Kan and Zhou (2007)	KZ
MAXSER by Ao, Li, and Zheng (2019)	MASER
GP with no transaction cost	GP-NoCost
GP with transaction cost	GP-WithCost
<i>Panel B: MV/GMV with different covariance matrix estimates</i>	
MV with sample cov	MV-P
MV with linear shrinkage cov	MV-LS
MV with nonlinear shrinkage cov	MV-NLS
MV with nonlinear shrinkage cov adjusted for factor models	MV-NLSF
GMV with linear shrinkage cov	GMV-LS
GMV with linear nonshrinkage cov	GMV-NLS
<i>Panel C: Models with short-sale constraints</i>	
MV with sample cov and short-sale and cross-validation	MV-P-SSCV
MV with linear shrinkage cov and short-sale and cross-validation	MV-LS-SSCV
MV with nonlinear shrinkage cov and short-sale and cross-validation	MV-NLS-SSCV
GP with no transaction cost and short-sale	GP-NoCost-SS
GP with transaction cost and short-sale	GP-WithCost-SS
<i>Panel D: MV with L1-norm constraint and cross-validation</i>	
MV with sample cov and L1-CV	MV-P-L1CV
MV with linear shrinkage cov and L1-CV	MV-LS-L1CV
MV with nonlinear shrinkage cov and L1-CV	MV-NLS-L1CV

Table 10

Portfolio performance without transaction costs

This table reports the OOS performances without transaction costs for the portfolios based on S&P 500 constituents. We report the annualized Sharpe ratio, monthly Sharpe ratio, monthly return and monthly volatility. The OOS period is from 1997:01 to 2016:12.

	Annual SR	Mon SR	Mon Rt (%)	Mon Vol (%)
<i>Panel A: Baseline models</i>				
Index	0.302	0.087	0.384	4.400
EW	0.344	0.099	0.487	4.900
Factor	0.409	0.118	0.531	4.500
KZ	0.331	0.096	0.831	8.700
MAXSER	0.591	0.171	0.904	5.300
GP-NoCost	1.089	0.314	0.420	1.338
<i>Panel B: MV/GMV with different covariance matrix estimates</i>				
MV-P	0.257	0.074	2.723	36.700
MV-LS	0.093	0.027	0.209	7.800
MV-NLS	0.091	0.026	0.168	6.400
MV-NLSF	0.141	0.041	0.232	5.700
GMV-LS	0.436	0.126	0.315	2.500
GMV-NLS	0.467	0.135	0.364	2.700
<i>Panel C: Models with short-sale constraint</i>				
MV-P-SSCV	0.316	0.091	0.611	6.700
MV-LS-SSCV	0.157	0.045	0.263	5.800
MV-NLS-SSCV	0.139	0.040	0.233	5.800
GP-NoCost-SS	0.637	0.184	1.298	7.059
<i>Panel D: MV with L1-norm constraint and cross-validation</i>				
MV-P-L1CV	0.128	0.037	4.700	6.700
MV-LS-L1CV	-0.053	-0.015	4.800	5.800
MV-NLS-L1CV	-0.059	-0.017	4.800	5.800

Table 11

Portfolio performance with transaction costs

This table reports the OOS performances with transaction costs of 0.1% for the portfolios based on S&P 500 constituents. We report the annualized Sharpe ratio, monthly Sharpe ratio, monthly return and monthly volatility. The OOS period is from 1997:01 to 2016:12.

	Annual SR	Mon SR	Mon Rt (%)	Mon Vol (%)
<i>Panel A: Baseline models</i>				
Index	0.302	0.087	0.384	4.400
EW	0.330	0.095	0.467	4.900
Factor	0.330	0.095	0.429	4.500
KZ	0.160	0.046	0.402	8.700
MAXSER	0.483	0.139	0.739	5.300
GP-WithCost	0.794	0.229	0.309	1.347
<i>Panel B: MV/GMV with different covariance matrix estimates</i>				
MV-P	-0.018	-0.005	-0.186	35.700
MV-LS	0.011	0.003	0.025	7.800
MV-NLS	0.022	0.006	0.041	6.400
MV-NLSF	0.054	0.016	0.089	5.700
GMV-LS	0.350	0.101	0.253	2.500
GMV-NLS	0.398	0.115	0.310	2.700
<i>Panel C: Models with short-sale constraint</i>				
MV-P-SSCV	0.174	0.050	1.183	6.800
MV-LS-SSCV	0.083	0.024	0.490	5.900
MV-NLS-SSCV	0.075	0.022	0.435	5.800
GP-WithCost-SS	0.594	0.171	1.137	6.630
<i>Panel D: MV with L1-norm constraint and cross-validation</i>				
MV-P-L1CV	-0.013	-0.004	-0.061	4.700
MV-LS-L1CV	-0.112	-0.032	-0.538	4.800
MV-NLS-L1CV	-0.110	-0.032	-0.528	4.800

Table 12

The effects of transaction cost on GP portfolios

This table reports the OOS performances of GP portfolios estimated with and without transaction costs. Panel A reports the results with short-sale constraints, while Panel B reports the results without short-sale constraints. The models of GP-NoCost and GP-NoCost-SS are estimated with no transaction cost. The models of GP-WithCost and GP-WithCost-SS are estimated with a transaction cost of 0.1%. We then calculate their OOS performances with and without a transaction cost of 0.1%, respectively. The OOS period is from 1997:01 to 2016:12.

<i>Panel A: Short-sale constraints</i>				
	Without cost		With cost	
	GP-NoCost-SS	GP-WithCost-SS	GP-NoCost-SS	GP-WithCost-SS
Annual SR	0.637	0.680	0.548	0.594
Mon SR	0.184	0.196	0.158	0.171
Mon Rt (%)	1.298	1.301	1.117	1.137
Mon Vol (%)	7.059	6.628	7.060	6.630
Mon Turn	1.810	1.648	1.810	1.648
Mon MDD	0.715	0.638	0.726	0.651

<i>Panel B: Short-sale allowed</i>				
	Without cost		With cost	
	GP-NoCost	GP-WithCost	GP-NoCost	GP-WithCost
Annual SR	1.089	1.117	0.751	0.794
Mon SR	0.314	0.322	0.217	0.229
Mon Rt (%)	0.420	0.434	0.290	0.309
Mon Vol (%)	1.338	1.347	1.338	1.347
Mon Turn	1.300	1.254	1.300	1.254
Mon MDD	0.060	0.058	0.078	0.064

Table 13

Performance under the 15 characteristics of Lewellen (2015)

This table reports the performance of the decile spread portfolios based on the alternative characteristic set of 15 variables used in Lewellen (2015) . For each spread portfolio, we report the average monthly return in percentage points, the Newey-west (1987) robust t -statistic, the annualized Sharpe ratio (*Sharpe*), the skewness (*Skew*), and the maximum drawdown (*MDD*) in percentage. The sample period is from 2001:01 to 2021:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Low	0.26	0.53	0.31	0.41	0.39	0.31	0.26	0.39	0.42	0.39	0.30
2	0.64	0.31	0.43	0.37	0.31	0.32	0.64	0.42	0.57	0.42	0.70
3	0.61	0.81	0.64	0.73	0.80	0.80	0.62	0.80	0.77	0.73	0.70
4	0.75	0.69	0.81	0.78	0.75	0.72	0.82	0.80	0.82	0.83	0.70
5	0.91	0.82	0.85	0.83	0.80	0.83	0.88	0.75	0.79	0.80	0.77
6	0.96	0.91	0.79	0.91	0.84	0.65	0.76	0.92	0.67	0.71	0.67
7	0.88	0.76	0.90	0.80	0.86	0.99	0.74	0.68	0.89	0.67	0.94
8	0.92	0.82	0.68	0.76	0.76	0.86	0.89	0.78	0.83	0.66	1.05
9	0.97	0.73	0.84	0.78	0.85	0.80	0.79	0.87	0.66	1.01	0.76
High	1.02	0.66	0.71	0.67	0.66	0.73	0.74	0.88	0.81	0.91	0.88
H-L	0.76***	0.13	0.39	0.25	0.26	0.41	0.47	0.48*	0.39	0.51*	0.57*
t-stat	(2.69)	(0.35)	(0.98)	(0.66)	(0.71)	(1.23)	(1.38)	(1.68)	(1.22)	(1.68)	(1.81)
Std. dev.	4.18	5.79	6.10	6.04	5.79	5.79	5.08	4.57	4.82	4.77	4.88
Sharpe	0.63	0.08	0.23	0.15	0.16	0.25	0.32	0.37	0.28	0.38	0.41
Skew	0.31	-0.51	-0.28	-0.39	-0.28	-0.68	0.47	-0.44	0.25	-0.05	-0.50
MDD	26.53	66.22	56.02	64.87	60.34	47.44	40.61	38.96	41.14	38.34	43.41

Table 14

Performance under the top 15 characteristics of Gu, Kelly, and Xiu (2020)

This table reports the performance of the decile spread portfolios based on the top 15 characteristics used in Gu, Kelly, and Xiu (2020). The top 15 characteristics are selected based on NN5, which is the best performing model based on the first 30 years in Gu, Kelly, and Xiu (2020). For each spread portfolio, we report the average monthly return in percentage points, the Newey-west (1987) robust t -statistic, the annualized Sharpe ratio (*Sharpe*), the skewness (*Skew*), and the maximum drawdown (*MDD*) in percentage. The sample period is from 1987:01 to 2016:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Low	-0.46	0.22	0.28	0.27	0.39	0.19	-0.09	-0.11	-0.19	-0.20	-0.04
2	0.17	0.79	0.81	0.80	0.58	0.64	0.40	0.27	0.48	0.40	0.60
3	0.57	0.90	0.85	0.88	0.80	0.91	0.59	0.60	0.58	0.61	0.77
4	0.91	0.94	0.96	0.95	0.84	0.80	0.89	0.79	0.89	0.69	0.74
5	0.95	1.00	0.96	0.98	0.86	0.88	1.01	0.95	0.87	0.86	0.86
6	0.89	1.04	1.12	1.09	1.10	0.96	1.06	1.00	0.93	0.96	1.00
7	0.87	1.22	1.14	1.14	1.17	1.13	0.94	1.02	1.04	1.01	0.94
8	1.05	1.11	1.11	1.16	1.12	1.12	0.98	0.94	0.99	1.03	0.88
9	1.07	1.16	1.18	1.16	1.22	1.14	1.02	1.01	1.05	1.20	0.95
High	1.22	1.15	1.18	1.17	1.41	1.33	1.23	1.40	1.48	1.26	1.11
H-L	1.67***	0.93***	0.89***	0.90***	1.02***	1.13***	1.31***	1.51***	1.66***	1.45***	1.15***
t-stat	(6.03)	(3.39)	(3.32)	(3.26)	(3.76)	(3.98)	(3.78)	(3.77)	(4.89)	(3.88)	(4.63)
p-value	3.80	4.70	4.82	4.79	4.92	4.75	5.20	5.42	5.23	4.81	4.49
Sharpe	1.53	0.69	0.64	0.65	0.72	0.83	0.88	0.97	1.10	1.05	0.89
Skew	-0.79	0.20	0.25	0.22	0.38	0.26	-0.01	-0.16	0.50	-0.29	-0.16
MDD	28.40	42.60	46.35	46.36	36.23	37.14	54.94	56.99	31.39	57.08	28.44

Table 15

International evidence

The table reports the performance of the decile spread portfolios in other G7 markets. For each spread portfolio, we report the average monthly return in percentage points, the Newey-west (1987) robust t -statistic, the annualized Sharpe ratio (*Sharpe*). Panel A to F report the statistics for each of the six markets, whereas Panel G reports the average over the six markets. The sample period is from 1991:01 to 2021:12.

	GP	Ridge	LASSO	ENET	PCR	PLS	NN1	NN2	NN3	NN4	NN5
Panel A: UK											
Mean	1.78	1.36	1.35	1.47	1.23	1.36	1.53	1.79	1.29	1.02	1.27
t-stat	6.76	4.67	4.83	5.49	4.98	4.80	5.09	5.68	4.09	3.30	4.07
Sharpe	1.07	0.75	0.74	0.83	0.64	0.76	0.83	1.04	0.72	0.52	0.70
Panel B: Canada											
Mean	2.31	2.16	2.12	2.12	1.46	2.14	1.95	1.79	1.14	0.99	1.04
t-stat	5.28	4.44	4.36	4.32	2.99	4.27	3.32	3.80	2.13	1.87	2.12
Sharpe	0.96	0.74	0.74	0.72	0.51	0.73	0.67	0.70	0.42	0.35	0.36
Panel C: Germany											
Mean	2.11	1.30	1.48	1.30	1.35	1.30	0.88	0.71	0.65	0.85	1.41
t-stat	6.76	3.55	3.58	3.15	3.28	3.57	2.53	2.09	1.71	2.00	4.25
Sharpe	1.22	0.62	0.68	0.60	0.64	0.62	0.48	0.36	0.31	0.41	0.74
Panel D: Japan											
Mean	1.46	1.36	1.48	1.40	1.52	1.34	0.89	1.14	1.11	1.14	1.37
t-stat	5.57	4.81	5.26	4.90	4.78	4.73	4.41	4.99	5.10	5.27	5.19
Sharpe	1.22	0.97	1.06	1.00	1.03	0.95	0.75	0.97	0.92	0.99	1.06
Panel E: Italy											
Mean	1.29	1.40	1.24	1.29	1.24	1.31	1.27	0.99	1.31	1.40	0.98
t-stat	3.53	4.07	3.74	3.86	4.35	3.74	3.78	3.07	3.68	4.65	2.57
Sharpe	0.74	0.68	0.63	0.65	0.64	0.64	0.67	0.52	0.65	0.74	0.49
Panel F: France											
Mean	1.86	1.56	1.60	1.56	1.60	1.55	0.98	1.23	1.48	1.31	1.19
t-stat	4.77	4.91	5.84	5.27	4.55	4.78	2.21	3.67	4.22	4.35	2.89
Sharpe	1.13	0.84	0.90	0.86	0.82	0.83	0.51	0.69	0.79	0.72	0.66
Panel G: Average statistics over the markets											
Mean	1.80	1.52	1.55	1.52	1.40	1.50	1.25	1.28	1.16	1.12	1.21
t-stat	5.44	4.41	4.60	4.50	4.15	4.31	3.56	3.88	3.49	3.57	3.51
Sharpe	1.06	0.77	0.79	0.78	0.71	0.76	0.65	0.71	0.64	0.62	0.67

Table 16

Performance under alternative parameters

The table reports the annualized Sharpe ratio and average return of the spread portfolios generated by GP under alternative hyperparameters $\langle Pop, Gen \rangle$ and M . Panel A, B, and C reports the results for $M = 5, 3,$ and $10,$ respectively. The training sample is from 1945:01 to 1980:12. The validation sample is from 1981:01 to 1990:12. The OOS sample is from 1991:01 to 2021:12.

Gen\Pop	Sharpe									Return								
	Train			Validation			OOS			Train			Validation			OOS		
	100	200	400	100	200	400	100	200	400	100	200	400	100	200	400	100	200	400
Panel A: Average of top 5 models																		
10	2.27	2.53	2.38	2.22	2.08	1.62	1.12	0.87	1.06	2.07	2.34	2.31	1.97	2.02	1.71	1.62	1.35	1.62
20	2.35	2.44	2.45	1.82	1.55	1.68	1.13	0.92	1.11	1.96	2.23	2.25	1.61	1.52	1.63	1.40	1.35	1.66
40	2.87	3.12	2.90	1.95	2.56	2.05	1.04	1.21	0.88	2.32	2.21	2.42	1.77	2.00	1.96	1.51	1.53	1.26
Panel B: Average of top 3 models																		
10	2.11	2.58	2.43	1.75	2.24	1.68	1.15	0.92	1.02	2.03	2.31	2.29	1.65	2.10	1.69	1.58	1.43	1.62
20	2.28	2.44	2.47	1.85	1.54	1.63	1.15	0.90	1.16	1.95	2.21	2.26	1.66	1.48	1.57	1.42	1.33	1.72
40	2.92	3.15	2.90	1.69	2.60	2.04	1.07	1.20	0.90	2.34	2.22	2.43	1.57	2.02	1.93	1.56	1.56	1.26
Panel C: Average of top 10 models																		
10	2.30	2.36	2.28	2.11	1.80	1.54	1.11	0.99	1.02	2.05	2.29	2.24	1.89	1.85	1.69	1.64	1.55	1.51
20	2.32	2.33	2.43	1.66	1.59	1.86	1.10	0.99	1.10	2.24	2.19	2.29	1.61	1.61	1.87	1.47	1.46	1.65
40	2.85	3.16	2.90	2.39	2.53	2.16	1.12	1.12	1.03	2.35	2.37	2.44	2.04	2.16	2.08	1.58	1.56	1.45

Table 17

Comparison with different objective functions

This table reports the summary statistics for the decile portfolios generated by GP under two objectives, i.e., to maximize the resulting spread portfolio's Sharpe ratio (GP_{SR}) and to minimize the conventional mean squared error (GP_{MSE}). We also report the results for the two methods controlling for each other. GP_{SR}^ω reports the results for GP_{SR} controlling for GP_{MSE} . Each month, the expected return under GP_{SR} is regressed in a cross-section regression on that under GP_{MSE} . Stocks are then sorted by the resulting residual into ten decile portfolios. GP_{MSE}^ω reports the results for GP_{MSE} controlling for GP_{SR} . The sample period is from 1991:01 to 2021:12.

	Original		Controlling each other	
	GP_{SR}	GP_{MSE}	GP_{SR}^ω	GP_{MSE}^ω
Low	0.29	0.36	0.77	0.95
2	0.60	0.65	0.90	1.41
3	0.67	0.76	0.99	1.41
4	0.59	1.03	1.26	1.40
5	0.81	0.99	1.24	1.25
6	1.10	1.16	1.16	1.25
7	1.16	1.21	1.41	1.12
8	1.53	1.18	1.51	1.12
9	1.56	1.54	1.54	0.90
High	1.82	1.81	1.49	0.97
H-L	1.53***	1.44***	0.72***	0.01
t-stat	(6.00)	(4.99)	(4.31)	(0.05)
Std. dev.	4.38	4.95	2.88	5.51
Sharpe	1.21	1.01	0.87	0.01
Skew	0.83	0.62	0.60	-0.17
MDD	22.56	32.35	17.37	85.13

Table 18

Simulation: Linear vs nonlinear

This table reports the OOS performances of various models in the linear and nonlinear simulations. $\hat{R}_{i,t}$ is the fitted return from step 2 in the simulation procedure. The simulation procedure is discussed in section 6.2.

	Annual SR		Mean Rt	
	Linear	Nonlinear	Linear	Nonlinear
$\hat{R}_{i,t}$	1.33	3.36	1.26	3.49
GP	1.24	3.08	1.08	2.50
Ridge	1.30	2.03	1.24	1.97
Lasso	1.32	2.05	1.27	2.03
Enet	1.32	2.07	1.25	2.01
PCR	1.30	2.03	1.24	1.97
PLS	1.30	2.03	1.25	1.97

Table 19

Bootstrap with various sample size

This table reports the OOS performances of various models in bootstrap with different sample size. For “Half”, in each month t , we resample stocks so that the stock number in the simulated data is only *half* of the actual stock number. For “Double”, we do the same but double the stock number in the simulated data each month. Then, based on the simulated data, we estimate GP and NN models, and examine the spread portfolio in the OOS sample.

	Annualied SR		Mean Rt	
	Half	Double	Half	Double
GP	1.01	1.07	1.65	1.61
NN1	0.25	0.74	0.30	1.10
NN2	0.47	0.86	0.58	1.27
NN3	0.46	0.72	0.58	1.04
NN4	0.49	0.62	0.56	0.86
NN5	0.20	0.90	0.20	1.32