# Industry Return Predictability:
# *A Machine Learning Approach*

**David E. Rapach, Jack K. Strauss, Jun Tu, and Guofu Zho**

**David E. Rapach**

is a professor of economics and the John Simon Endowed Chair in Economics at the Saint Louis University Chaifetz School of Business in St. Louis, MO.
david.rapach@slu.edu

**Jack K. Strauss**

is a professor of economics and the Miller Chair of Applied Economics at the University of Denver Daniels School of Business in Denver, CO.
jack.strauss@du.edu

**Jun Tu**

is an associate professor of finance at the Singapore Management University Lee Kong Chian School of Business in Singapore.
tujun@smu.edu.sg

**Guofu Zhou**

is the Frederick Bierman and James E. Spears Professor of Finance at Washington University in the St. Louis Olin Business School in St. Louis, MO.
zhou@wustl.edu

**ABSTRACT:** *In this article, the authors use machine learning tools to analyze industry return predictability based on the information in lagged industry returns. Controlling for post-selection inference and multiple testing, they find significant in-sample evidence of industry return predictability. Lagged returns for the financial sector and commodity- and material-producing industries exhibit widespread predictive ability, consistent with the gradual diffusion of information across economically linked industries. Out-of-sample industry return forecasts that incorporate the information in lagged industry returns are economically valuable: Controlling for systematic risk using leading multifactor models from the literature, an industry-rotation portfolio that goes long (short) industries with the highest (lowest) forecasted returns delivers an annualized alpha of over 8%. The industry-rotation portfolio also generates substantial gains during economic downturns, including the Great Recession.*

**TOPICS:** *Big data/machine learning, analysis of individual factors/risk premia, portfolio construction, performance measurement**

A voluminous literature investigates aggregate stock market return predictability.[1] In contrast, a relatively limited number of studies examine stock return predictability along industry lines, even though analyst reports and asset allocations are often industry based. Studies that analyze industry return predictability typically rely on popular predictor variables from the literature on aggregate market return predictability, such as the aggregate dividend yield, nominal yields, and yield spreads (e.g., Ferson and Harvey 1991, 1999; Ferson and Korajczyk 1995; Avramov 2004). In this article, we examine industry return predictability using a different information set—namely, lagged industry returns from across the entire economy.

Our study is the first to directly analyze industry return predictability based on lagged industry returns from across the economy. The lack of previous studies investigating this topic is perhaps due to the statistical challenges inherent in estimating regression models with a large number of predictors.[2] Our use of lagged industry returns to forecast individual industry returns is motivated by the theoretical model of Hong, Torous, and Valkanov (2007), who introduce information frictions into an economy with multiple

---

[1] See Rapach and Zhou (2013) for a survey.

[2] Moskowitz and Grinblatt (1999) analyze industry momentum, and Cohen and Frazzini (2008) and Menzly and Ozbas (2010) investigate customer-supplier links. In contrast, our study focuses on whether lagged industry returns have predictive power for future industry returns and on how returns across all industries are related to one another.

linked industries. Due to the industry links, cash flow shocks originating in one industry can affect expected cash flows in related industries. In a frictionless, rational-expectations equilibrium, investors readily recognize all of the interindustry implications of a cash flow shock in a particular industry. Consequently, equity prices across all relevant industries immediately adjust to fully impound the interindustry effects of the cash flow shock, and lagged industry returns do not have predictive power. However, incorporating insights from Merton (1987) and Hong and Stein (1999), Hong, Torous, and Valkanov (2007) posit that investors with limited information-processing capabilities specialize in specific market segments. In this environment, when a cash flow shock arises in a particular industry, information-processing limitations prevent investors specializing in related industries from quickly working out the full implications of the shock. Information thus diffuses gradually across industries, and the resulting delayed adjustment in equity prices gives rise to industry return predictability on the basis of lagged industry returns.

We analyze the predictive ability of lagged industry returns in a general predictive regression framework that allows each industry's return to respond to the lagged returns for all industries, thereby accommodating a wide array of industry links, both direct and indirect. Given the plethora of potential predictors in the predictive regression models, conventional ordinary least squares (OLS) estimation has potential drawbacks. First, if all lagged returns are used, the abundance of predictors makes OLS estimation susceptible to overfitting. Second, if only a few lagged returns are chosen, it is difficult to know a priori which are the most important. Therefore, we use a machine learning approach to guard against overfitting the data in our high-dimensional setting and to select the most relevant predictors. Along this line, Rapach, Strauss, and Zhou (2013); Gu, Kelly, and Xiu (2018); Chinco, Clark-Joseph, and Ye (2019); Han et al. (2019); and Freyberger, Neuhierl, and Weber (forthcoming) recently use machine learning to predict aggregate and/or individual stock returns. Although we focus on industry return predictability in the present article, we share with these studies the use of machine learning to deal with the challenges posed by high-dimensional data in finance.

We use the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), a powerful and popular technique in machine learning. Similarly to ridge regression (Hoerl and Kennard 1970), the LASSO induces shrinkage in the estimated coefficients by including a convex penalty term in the objective function for fitting a model. In contrast to the $\ell_2$ penalty term in ridge regression, the LASSO relies on an $\ell_1$ penalty term, so that it permits shrinkage to exactly zero for some coefficients. It thus performs variable selection, which typically yields a sparse model. Sparsity has two key advantages. First, it helps to avoid overfitting the data by setting unimportant coefficients to zero. Second, it facilitates interpretation of the estimated model by identifying the most relevant predictor variables.

Although the LASSO's $\ell_1$ penalty term mitigates overfitting the data via sparsity, it also tends to over-shrink the coefficients for the selected variables. This tendency can lead to substantial downward biases (in magnitude) in the estimated coefficients (Fan and Li 2001). Recent studies propose OLS post-LASSO estimation to reduce the biases (e.g., Efron et al. 2004; Meinshausen 2007; Belloni and Chernozhukov 2011, 2013). The idea is to first use the LASSO to reduce the dimension of the model; to lessen the biases in the LASSO coefficient estimates, we then re-estimate the coefficients for the selected predictor variables using OLS. We use OLS post-LASSO to estimate predictive regression models for each industry, where the set of candidate predictors includes the lagged returns for all 30 industries that we consider. OLS post-LASSO estimation allows us to identify the most relevant set of lagged industry returns for predicting a given industry's return, while generating more accurate estimates of the coefficients for the relevant lagged industry returns.

We use both in-sample and out-of-sample tests to analyze the ability of lagged industry returns to predict individual industry returns. With respect to the in-sample tests, we address the important issues of *post-selection inference* and *multiple testing*. In the context of OLS post-LASSO estimation, post-selection inference recognizes that the predictor variables that appear in the regression model are first selected from a large group of candidate predictors, so that conventional inferences for the selected variables' coefficients are not necessarily reliable—we peeked at the data twice. We incorporate insights from Berk et al. (2013); Leeb, Pötscher, and Ewald (2015); and Zhao, Shojaie, and Witten (2017) to make accurate post-selection inferences concerning industry return predictability.

Multiple testing relates to the fact that we test a large number of individual null hypotheses. In this context, conventional *p*-values—which implicitly assume that a researcher tests a given null hypothesis in isolation—can present a misleading picture of statistical significance. To account for multiple testing, we use the Benjamini and Hochberg (2000) adaptive version of the Benjamini and Hochberg (1995) linear step-up procedure to control the false discovery rate (FDR) when testing hypotheses concerning individual coefficients for the entire set of LASSO-selected lagged industry returns.[3]

For the in-sample analysis, we estimate predictive regression models via OLS post-LASSO using monthly return data spanning 1960 to 2016 for 30 industry portfolios from Kenneth French's Data Library. The LASSO selects at least one lagged industry return as a predictor for 29 of the individual industries, while multiple lagged industry returns are selected for 22 of the individual industries. Furthermore, the OLS post-LASSO estimation results indicate that the lagged industry returns selected by the LASSO are often statistically significant predictors of industry returns, even after controlling for post-selection inference and multiple testing. Although monthly equity returns inherently contain a relatively small predictable component, the Campbell and Thompson (2008) metric indicates that the degree of return predictability is economically meaningful for nearly all industries.

Lagged returns for the financial sector and commodity- and material-producing industries are selected as return predictors for numerous individual industries. The prominent role of these sectors is economically intuitive from the perspective of the theoretical model in Hong, Torous, and Valkanov (2007). Firms in many industries rely on financial intermediaries for financing. When the financial sector experiences a positive return shock, financial firms have larger capital buffers and thus become more willing to provide credit on favorable terms to industries across the economy; borrowers benefit directly from the better terms, and their customers benefit indirectly. In the presence of information

frictions, we thus expect lagged financial sector returns to positively affect future returns in many industries, which is what we find. Furthermore, commodity price shocks raise product prices and returns for sectors located in earlier production stages, whereas they squeeze profit margins and lower returns for sectors located in later production stages. With information frictions, we thus expect returns for lagged commodity- and material-producing industries to negatively affect future returns for industries located in later stages of the production chain, which is again what we find.

The out-of-sample analysis simulates the situation of an investor in real time and measures the economic value of industry return predictability. We first compute out-of-sample forecasts of monthly industry returns based on predictive regressions estimated via OLS post-LASSO. We then sort the 30 industries according to their forecasted returns over the next month and construct a zero-investment portfolio that goes long (short) the top (bottom) quintile of sorted industries. For the January 1970 to December 2016 out-of-sample period, the industry-rotation portfolio generates an average return of 7.33% per annum, annualized Sharpe (Sortino) ratio of 0.65 (1.16), and Goetzmann et al. (2007) manipulation-proof performance measure (MPPM) of 4.84% for a relative risk aversion coefficient of four. The industry-rotation portfolio performs especially well during business-cycle recessions, thereby providing a hedge against bad times for the macroeconomy. The portfolio also exhibits negative exposures to the broad equity market and Fama and French (1993) value factors, and it generates a substantial annualized alpha of more than 8% in the context of both the Carhart (1997) four-factor and Hou, Xue, and Zhang (2015) *q*-factor models.

## PREDICTIVE REGRESSION FRAMEWORK

Our basic framework is the following general predictive regression model specification:

$$\mathbf{r}_i = a_i^* \iota_T + \mathbf{X}\mathbf{b}_i^* + \varepsilon_i \quad \text{for} \quad i = 1, \dots, N, \qquad (1)$$

where $\mathbf{r}_i$ is the *T*-vector of monthly observations for the return on industry portfolio *i* in excess of the one-month Treasury bill return (with typical element $r_{i,t}$), *T* is the usable number of monthly observations, $\mathbf{X}$ is a *T*-by-*N* data matrix of lagged industry excess return observations, *N* is the number of industry portfolios, $\mathbf{b}_i^* = [b_{i,1}^* \ \cdots \ b_{i,N}^*]'$

---

[3] Harvey, Liu, and Zhu (2016); Green, Hand, and Zhang (2017); Linnainmaa and Roberts (2018); and Hou, Xue, and Zhang (forthcoming) highlight the relevance of multiple testing for identifying significant firm characteristics in the cross section of stock returns. Arnott, Harvey, and Markowitz (2019) stress the importance of accounting for multiple testing when using machine learning techniques.

is an $N$-vector of regression coefficients, and $\varepsilon_i$ is a $T$-vector of zero-mean disturbance terms. Equation 1 allows for lagged returns for all industries across the economy to affect a given industry's excess return, thereby accommodating general industry links. Because of the high-dimensional nature ($N = 30$) of Equation 1, conventional OLS estimation runs the risk of overfitting. To deal with the challenges posed by the high dimensionality of the predictive regression models, we use the LASSO from machine learning.

Tibshirani (1996) introduces the LASSO as a regularization device for performing shrinkage in regressions with a large number of candidate predictor variables. For Equation 1, the LASSO objective function can be expressed as

$$\underset{a_i \in \mathbb{R}, \mathbf{b}_i \in \mathbb{R}^N}{\arg\min} \left( \frac{1}{2T} \left\| \mathbf{r}_i - a_i \iota_T - \mathbf{X}\mathbf{b}_i \right\|_2^2 + \lambda_i \left\| \mathbf{b}_i \right\|_1 \right), \qquad (2)$$

where $\lambda_i \geq 0$ is a regularization parameter. The first component in parentheses in Equation 2 is the familiar sum of squared residuals; the objective function thus reduces to that for OLS when $\lambda_i = 0$. The second component in parentheses is an $\ell_1$ penalty term that shrinks the slope coefficient estimates to prevent overfitting. Unlike ridge regression—which relies on an $\ell_2$ penalty—the $\ell_1$ penalty in Equation 2 allows for shrinkage to zero (for sufficiently large $\lambda_i$), so that it performs variable selection. By reducing the dimension of the model, variable selection helps to guard against overfitting and facilitates interpretation of the estimated model. Powerful algorithms are available for computing the solution to Equation 2, such as cyclical coordinate descent (Friedman, Hastie, and Tibshirani 2010), making LASSO estimation feasible even when $N$ is large.

The LASSO generally performs well in selecting the most relevant predictor variables (e.g., Zhang and Huang 2008; Bickel, Ritov, and Tsybakov 2009; Meinshausen and Yu 2009). However, LASSO estimates of the coefficients for the selected predictors suffer from downward biases in magnitude (Fan and Li 2001). Intuitively, the LASSO penalty term overshrinks the coefficients for the selected predictors. To alleviate the biases in the LASSO estimates, Efron et al. (2004), Meinshausen (2007), and Belloni and Chernozhukov (2011, 2013) propose re-estimating the coefficients for the LASSO-selected predictors via OLS (OLS post-LASSO estimation).

Because the predictor variables are first selected from a set of candidate predictors, statistical inference based on OLS post-LASSO estimation constitutes post-selection inference, an active area of recent research (e.g., Wasserman and Roeder 2009; Meinshausen, Meier, and Bühlmann 2009; Berk et al. 2013; Leeb, Pötscher, and Ewald 2015; Lee et al. 2016; Tibshirani et al. 2016; Zhao, Shojaie, and Witten 2017).[4] When making post-selection inferences, we first need to identify the inferential target. The traditional target is the set of true regression coefficients for the full model that includes all of the potential predictors (Leeb, Pötscher, and Ewald 2015), which, in our context, corresponds to $\mathbf{b}_i^*$ in Equation 1. Alternatively, Berk et al. (2013) advocate targeting the true regression coefficients for a submodel:

$$\mathbf{r}_i = a_i^{(\mathcal{M})} \iota_T + \mathbf{X}_{\mathcal{M}} \mathbf{b}_i^{(\mathcal{M})} + \varepsilon_i^{(\mathcal{M})}, \qquad (3)$$

where $\mathcal{M} \subseteq \{1, \ldots, N\}$ is the index set of predictors for the submodel, and $\mathbf{X}_{\mathcal{M}}$ is composed of the columns of $\mathbf{X}$ indexed by $\mathcal{M}$. Recognizing that the actual data-generating process is essentially unknowable, a useful submodel provides a parsimonious representation of the data that focuses on the most relevant predictors.[5] Because they measure different things, the true regression coefficients for the full model and submodel are generally different: $b_{i,j}^*$ measures the response to a given predictor, conditional on the remaining $N - 1$ predictors in the full model; $b_{i,j}^{(\mathcal{M})}$ measures the response to the predictor, conditional on the other $|\mathcal{M}| - 1$ predictors in the submodel, where $|\mathcal{M}|$ is the cardinality of $\mathcal{M}$.

Let $\text{CI}_{i,j}^{(\widehat{\mathcal{M}})}$ for any $j \in \widehat{\mathcal{M}}$ denote the conventional OLS confidence interval for $b_{i,j}^{(\widehat{\mathcal{M}})}$ in the model selected by the LASSO, where $\widehat{\mathcal{M}}$ is the index set of predictors for the LASSO-selected model. Leeb, Pötscher, and Ewald (2015) use simulations to investigate the coverage properties of $\text{CI}_{i,j}^{(\widehat{\mathcal{M}})}$. For the case in which the true regression coefficients for a submodel are targeted, they find that $\text{CI}_{i,j}^{(\widehat{\mathcal{M}})}$ has approximately correct coverage. This finding is surprising, since the naïve OLS post-LASSO confidence interval does not explicitly account for the fact that the data are used twice—first

---

[4] See Dezeure et al. (2015) and Taylor and Tibshirani (2015) for recent surveys of the post-selection inference literature.

[5] Targeting the submodel is in the spirit of George Box's famous dictum, "All models are wrong, but some are useful" (Box 1979, p. 202).

to select the predictor variables from the complete set of potential predictors via the LASSO and then to estimate the regression coefficients for the LASSO-selected predictors via OLS. Zhao, Shojaie, and Witten (2017) provide a theoretical explanation for this finding. Under reasonably weak assumptions, they show that the set of LASSO-selected predictors is deterministic with high probability, so that the conventional OLS post-LASSO confidence intervals and $t$-statistics are asymptotically valid—asymptotically, it is as if the data are only used once. We target the true regression coefficients for the LASSO-selected submodel. In line with the results of Leeb, Pötscher, and Ewald (2015) and Zhao, Shojaie, and Witten (2017), we assess statistical significance using conventional OLS post-LASSO $t$-statistics.

The results of Leeb, Pötscher, and Ewald (2015) and Zhao, Shojaie, and Witten (2017) demonstrate that the $p$-value for the conventional OLS post-LASSO $t$-statistic is valid for making post-selection inferences for an individual coefficient $b_{i,j}^{(\widehat{\mathcal{M}})}$. However, when analyzing the significance of a large number of individual coefficient estimates, the issue of multiple testing remains. When testing many individual null hypotheses, conventional $p$-values can present a misleading picture of statistical significance.

Suppose that we test $m$ individual null hypotheses ($m = 167$ in our application in the next section), each at significance level $q$. One approach for addressing multiple testing is to control the family-wise error rate (FWER): FWER = $\Pr(V \geq 1) \leq q$, where $V$ is the number of wrongly rejected null hypotheses. The familiar Bonferroni procedure controls the FWER by multiplying each of the unadjusted $p$-values by $m$. However, FWER control tends to be extremely conservative, with little power to detect false null hypotheses.

Instead of the FWER, Benjamini and Hochberg (1995) propose to control the FDR: FDR = $E(V/R) \leq q$, where $R$ is the number of rejections. They develop a popular linear step-up procedure based on adjusted $p$-values to control the FDR. The Benjamini and Hochberg (1995) procedure orders the unadjusted $p$-values and their corresponding null hypotheses ($p_{(1)}, \ldots, p_{(m)}$ and $H_{(1)}, \ldots, H_{(m)}$, respectively) and rejects the first $k$ null hypotheses ($H_{(1)}, \ldots, H_{(k)}$), where

$$k = \max_{j}\{1 \leq j \leq m : p_{(j)}^{\text{BH}} \leq q\}, \tag{4}$$

and

$$p_{(j)}^{\text{BH}} = \frac{m}{j} p_{(j)} \tag{5}$$

is the adjusted $p$-value for $H_{(j)}$. Benjamini and Hochberg (1995) show that the linear step-up procedure controls the FDR when $p_{(1)}, \ldots, p_{(m)}$ are independent, while Benjamini and Yekutieli (2001) demonstrate that the procedure controls the FDR for a particular type of dependence (positive regression dependence) in the $p$-values.[6]

When the number of true null hypotheses $m_0$ is less than $m$, the Benjamini and Hochberg (1995) procedure controls the FDR at too low a level: FDR $\leq \pi_0 q$, where $\pi_0 = m_0/m$. To improve its performance, a number of studies (e.g., Benjamini and Hochberg 2000; Storey, Taylor, and Siegmund 2004; Benjamini, Krieger, and Yekutieli 2006; Blanchard and Roquain 2009) adapt the Benjamini and Hochberg (1995) procedure to the data by replacing $p_{(j)}^{\text{BH}}$ in Equation 5 with

$$p_{(j)}^{\text{adapt}} = \hat{\pi}_0 \frac{m}{j} p_{(j)}, \tag{6}$$

where $\hat{\pi}_0$ is a conservative estimate of $\pi_0$. We control for multiple testing via the FDR using the Benjamini and Hochberg (2000) adaptive procedure, which employs the Hochberg and Benjamini (1990) estimator for $\pi_0$. For values of $m$ relevant to our application, simulations by Benjamini, Krieger, and Yekutieli (2006) indicate that the Benjamini and Hochberg (2000) adaptive procedure delivers good overall performance in terms of FDR control and power, including for dependent $p$-values.[7]

Finally, to implement OLS post-LASSO estimation, we need to choose the regularization parameter $\lambda_i$ in Equation 2. We choose $\lambda_i$ using the Hurvich and

---

[6] Benjamini and Yekutieli (2001) modify the Benjamini and Hochberg (1995) procedure so that it controls the FDR under arbitrarily dependent $p$-values. However, the modified procedure is typically very conservative (at times more so than the Bonferroni procedure), so that it has limited ability to identify false null hypotheses.

[7] The adaptive procedures can be shown to control the FDR when $p_{(1)}, \ldots, p_{(m)}$ are independent. Theoretical results concerning FDR control for adaptive procedures under dependent $p$-values are more limited (e.g., Farcomeni 2007; Blanchard and Roquain 2009), and simulations are typically used to analyze FDR control and power.

**Summary Statistics, Industry Portfolio Excess Returns, December 1959 to December 2016**

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Industry Portfolio | Ann. Mean (%) | Ann. Vol (%) | Min. (%) | Max. (%) | Ann. Sharpe Ratio | Industry Portfolio | Ann. Mean (%) | Ann. Vol. (%) | Min. (%) | Max. (%) | Ann. Sharpe Ratio |
| Food | 8.29 | 15.03 | −18.15 | 19.89 | 0.55 | Carry | 8.62 | 21.80 | −31.10 | 23.39 | 0.40 |
| Beer | 8.53 | 17.63 | −20.19 | 25.51 | 0.48 | Mines | 6.58 | 25.82 | −34.55 | 35.15 | 0.26 |
| Smoke | 11.79 | 21.00 | −25.32 | 32.38 | 0.56 | Coal | 8.83 | 35.31 | −38.09 | 45.55 | 0.25 |
| Games | 8.42 | 24.88 | −33.40 | 34.52 | 0.34 | Oil | 7.81 | 18.53 | −18.96 | 23.70 | 0.42 |
| Books | 6.34 | 20.12 | −26.56 | 33.13 | 0.32 | Util | 5.80 | 13.83 | −12.94 | 18.26 | 0.42 |
| Hshld | 6.65 | 16.49 | −22.24 | 18.22 | 0.40 | Telcm | 6.26 | 16.04 | −16.44 | 21.22 | 0.39 |
| Clths | 8.03 | 22.12 | −31.50 | 31.79 | 0.36 | Servs | 8.20 | 22.61 | −28.67 | 23.38 | 0.36 |
| Hlth | 7.81 | 17.07 | −21.06 | 29.01 | 0.46 | BusEq | 6.97 | 23.34 | −32.07 | 24.66 | 0.30 |
| Chems | 6.24 | 19.12 | −28.60 | 21.68 | 0.33 | Paper | 6.12 | 17.51 | −27.74 | 21.00 | 0.35 |
| Txtls | 8.01 | 24.33 | −33.11 | 59.03 | 0.33 | Trans | 7.02 | 19.90 | −28.50 | 18.50 | 0.35 |
| Cnstr | 6.28 | 20.75 | −28.74 | 25.02 | 0.30 | Whlsl | 7.46 | 19.42 | −29.25 | 17.53 | 0.38 |
| Steel | 3.49 | 25.26 | −32.99 | 30.30 | 0.14 | Rtail | 7.97 | 18.53 | −29.74 | 26.49 | 0.43 |
| FabPr | 6.75 | 21.18 | −31.63 | 22.91 | 0.32 | Meals | 8.47 | 21.15 | −31.89 | 27.38 | 0.40 |
| ElcEq | 8.45 | 21.52 | −32.80 | 23.21 | 0.39 | Fin | 7.31 | 18.75 | −22.53 | 20.59 | 0.39 |
| Autos | 5.45 | 23.17 | −36.49 | 49.56 | 0.24 | Other | 4.64 | 20.09 | −28.11 | 19.96 | 0.23 |

*Notes: The exhibit reports summary statistics for excess returns for 30 value-weighted industry portfolios from Kenneth French's Data Library. Excess returns are computed relative to the one-month Treasury bill return.*

*Autos = Automobiles and Trucks; Beer = Beer and Liquor; Books = Printing and Publishing; BusEq = Business Equipment; Carry = Aircraft, Ships, and Railroad Equipment; Chems = Chemicals; Clths = Apparel; Cnstr = Construction and Construction Materials; Coal = Coal; ElcEq = Electrical Equipment; FabPr = Fabricated Products and Machinery; Fin = Banking, Insurance, Real Estate, and Trading; Food = Food Products; Games = Recreation; Hlth = Healthcare, Medical Equipment, and Pharmaceutical Products; Hshld = Consumer Goods; Meals = Restaurants, Hotels, and Motels; Mines = Precious Metals, Non-Metallic, and Industrial Metal Mining; Oil = Petroleum and Natural Gas; Other = Everything Else; Paper = Business Supplies and Shipping Containers; Rtail = Retail; Servs = Personal and Business Services; Smoke = Tobacco Products; Steel = Steel Works, Etc.; Telcm = Communication; Trans = Transportation; Txtls = Textiles; Util = Utilities; Whlsl = Wholesale.*

Tsai (1989) corrected version of the Akaike information criterion (AIC; Akaike 1973). In the realistic case in which the true model is not included among the candidate models, Flynn, Hurvich, and Simonoff (2013) show that the corrected AIC ($AIC_c$) is asymptotically efficient for choosing $\lambda_i$ for OLS post-LASSO estimation. This means that the $AIC_c$ asymptotically selects the best-performing model from among the candidate models in terms of the $\ell_2$ loss criterion for predictive performance. In addition, Flynn, Hurvich, and Simonoff (2013) find that the $AIC_c$ performs well for selecting the best-performing model in finite-sample simulations.

A popular alternative strategy for selecting $\lambda_i$ in Equation 2 is $K$-fold cross-validation. We use the $AIC_c$ for simplicity, because $K$-fold cross-validation involves the choice of the number of folds $K$, as well as how the folds are defined. Nevertheless, the use of a typical five-fold cross-validation yields results similar to those based on the $AIC_c$.

## IN-SAMPLE RESULTS

We estimate predictive regressions via OLS post-LASSO using monthly excess return data for 30 value-weighted industry portfolios from Kenneth French's Data Library,[8] in which the industries are defined based on the Standard Industrial Classification system. Exhibit 1 reports summary statistics for the industry portfolio excess returns for December 1959 to December 2016. We refer to the industries by their Data Library abbreviations, which are given in the notes to Exhibit 1.

---

[8] Available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Along with the fact that the industry portfolios are value weighted, starting the sample in December 1959 mitigates illiquidity and thin-trading concerns. We start the sample in December 1959 to account for the lagged predictors when estimating the predictive regressions, so that the available estimation sample covers January 1960 to December 2016 (684 observations). Smoke (Tobacco Products) displays the highest annualized average excess return and Sharpe ratio at 11.79% and 0.56, respectively, in Exhibit 1, whereas Steel (Steel Works, Etc.) has the lowest annualized average excess return and Sharpe ratio at 3.49% and 0.14, respectively.

Exhibit 2 reports the OLS post-LASSO coefficient estimates for each industry.[9] As discussed previously, we target the true regression coefficients for the LASSO-selected submodel. To conserve space, we use a bold (italicized bold) entry to indicate that a coefficient estimate is significant at the 10% (5%) level based on the conventional OLS post-LASSO $t$-statistic.

Overall, the OLS post-LASSO estimates in Exhibit 2 highlight the importance of lagged industry returns for predicting individual industry returns. The LASSO selects 167 lagged industry returns (out of a possible 900) as individual industry return predictors. At least one lagged industry return is selected as a return predictor by the LASSO for 29 of the 30 individual industries, and multiple lagged industry returns are selected for 22 individual industries. According to the conventional OLS post-LASSO $t$-statistics, 82 (53) of the 167 LASSO-selected lagged industry returns are significant predictors at the 10% (5%) level.[10] In general, autocorrelation plays a limited role in Exhibit 2: An industry's own lagged return is only selected by the LASSO for seven industries.

Exhibit 3 controls for multiple testing across all 167 of the OLS post-LASSO coefficient estimates in Exhibit 2. The exhibit depicts the first 85 elements of the sequence of sorted unadjusted $p$-values for the conventional OLS post-LASSO $t$-statistics, along with corre-sponding adjusted $p$-values based on the Benjamini and Hochberg (2000) adaptive procedure. In line with the bold (italicized bold) entries in Exhibit 2, there are 82 (53) rejections of $H_0 : b_{i,j}^{(\widehat{\mathcal{M}})} = 0$ at the 10% (5%) significance level based on the unadjusted $p$-values. According to the Benjamini and Hochberg (2000) adjusted $p$-values, 72 (21) of these rejections remain when we control for multiple testing via the FDR. In conjunction with the out-of-sample results in the next section, the numerous significant coefficient estimates in Exhibit 2 that survive multiple-testing control suggest that our evidence of industry return predictability is not simply an artifact of data mining.

Many of the coefficient estimates in Exhibit 2 appear economically plausible from the perspective of gradual information diffusion across economically related industries (Hong, Torous, and Valkanov 2007). For example, lagged Fin (Banking, Insurance, Real Estate, and Trading) returns are selected by the LASSO for 19 of the 30 individual industries, and 11 (7) of the coefficient estimates are significant at the 10% (5%) level according to the conventional OLS post-LASSO $t$-statistics. Furthermore, all of the coefficient estimates for lagged Fin returns are positive. This is economically reasonable, as firms in many industries rely extensively on financial intermediaries for financing. A positive return shock in the financial sector increases financial firms' capital buffers, so that financial firms become more willing to make credit available to firms throughout the economy. In contrast, adverse shocks to the financial sector curtail intermediaries' capacity to lend, thereby driving up borrowing costs and driving down returns for many sectors. Financial sector shocks have direct effects for firms that borrow from financial intermediaries, as well as indirect effects for the customers of the borrowing firms. The coefficient estimates for lagged Fin returns are typically sizable in Exhibit 2, reaching as high as 0.18 (Txtls, Textiles).

Another noteworthy pattern in Exhibit 2 involves industries located in different stages of the production process. Lagged returns for commodity- and material-producing industries located in earlier stages of the production chain—such as Coal (Coal) and Oil (Petroleum and Natural Gas)—are often negatively related to returns for industries located in later stages of the production chain—such as Smoke, Books (Printing and Publishing), Txtls, Paper (Business Supplies and Shipping Containers), Whlsl (Wholesale), and Meals (Restaurants, Hotels, and Motels).

---

[9] All computations are performed using R (R Core Team 2017). We use the glmnet package (Friedman et al. 2017) to compute the LASSO solution. Following convention, the predictor variables are standardized before solving Equation 2. The estimated coefficients in Exhibit 2 correspond to the scales of the original predictor variables.

[10] As shown in Exhibit S1 of the online supplement, inferences are similar for $t$-statistics based on White (1980) heteroskedasticity-robust standard errors.

**OLS Post-LASSO Predictive Regression Estimation Results, January 1960 to December 2016**

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Regressand | | | | | |
| Regressor | Food | Beer | Smoke | Games | Books | Hshld | Clths | Hlth | Chems | Txtls |
| Food | – | *0.12* | – | – | – | – | – | – | – | – |
| Beer | – | – | – | – | – | – | – | – | – | – |
| Smoke | – | – | – | – | – | – | – | – | – | – |
| Games | – | – | – | – | 0.03 | – | – | – | – | – |
| Books | – | – | – | *0.18* | 0.04 | – | 0.06 | *0.10* | – | – |
| Hshld | – | – | – | – | – | – | – | – | – | – |
| Clths | 0.04 | 0.05 | – | 0.05 | – | *0.10* | 0.07 | – | *0.08* | 0.09 |
| Hlth | – | – | – | – | – | – | – | – | – | – |
| Chems | – | – | – | – | – | – | **0.15** | – | – | – |
| Txtls | – | – | 0.06 | – | – | – | – | – | – | – |
| Cnstr | – | – | – | – | – | – | – | – | – | – |
| Steel | – | – | – | – | – | – | –0.08 | – | – | – |
| FabPr | – | – | – | – | – | – | – | – | – | – |
| ElcEq | – | – | – | – | – | – | *–0.27* | – | – | – |
| Autos | – | – | – | – | – | – | – | – | – | **0.11** |
| Carry | – | – | *0.17* | – | – | – | 0.05 | – | – | – |
| Mines | – | – | –0.02 | – | – | – | – | –0.06 | – | – |
| Coal | *–0.06* | *–0.06* | –0.03 | *–0.07* | –0.04 | *–0.05* | –0.05 | –0.05 | – | *–0.07* |
| Oil | – | – | **–0.10** | – | –0.17 | – | –0.15 | – | – | *–0.15* |
| Util | **0.09** | – | **0.27** | – | **0.13** | – | **0.16** | **0.11** | – | – |
| Telcm | – | – | –0.11 | – | – | – | **–0.14** | – | – | – |
| Servs | – | – | *–0.15* | – | 0.05 | – | 0.12 | – | – | – |
| BusEq | – | – | – | – | 0.06 | – | **0.13** | – | – | – |
| Paper | – | – | *–0.19* | – | – | – | – | – | – | – |
| Trans | – | – | 0.10 | – | – | – | – | – | – | – |
| Whlsl | – | – | – | – | – | – | – | – | – | – |
| Rtail | 0.02 | – | – | – | 0.03 | 0.05 | 0.06 | – | – | 0.07 |
| Meals | – | – | – | – | – | – | – | – | – | – |
| Fin | – | – | 0.11 | 0.10 | 0.08 | – | 0.08 | – | – | *0.18* |
| Other | – | – | – | – | – | – | – | – | – | – |
| $R^2$ | 2.24% | 2.52% | 6.54% | 5.05% | 6.30% | 2.97% | 7.93% | 2.68% | 0.78% | 7.91% |
| | (0.93) | (1.35) | (2.73) | (5.62) | (8.20) | (2.29) | (7.93) | (1.61) | (0.89) | (9.60) |

| | | | | | Regressand | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Regressor | Cnstr | Steel | FabPr | ElcEq | Autos | Carry | Mines | Coal | Oil | Util |
| Food | – | – | – | – | – | – | – | – | – | 0.10 |
| Beer | – | – | – | – | – | – | – | *–0.27* | –0.08 | *–0.10* |
| Smoke | – | – | – | – | – | – | – | –0.09 | – | 0.02 |
| Games | – | – | – | – | – | – | – | – | – | – |
| Books | – | – | – | – | – | – | – | 0.13 | – | – |
| Hshld | – | – | – | – | *–0.30* | – | – | – | – | –0.08 |
| Clths | 0.04 | – | – | – | 0.04 | – | – | – | – | – |
| Hlth | – | – | – | – | – | – | – | – | *–0.13* | –0.08 |
| Chems | – | – | – | – | – | – | – | – | – | – |
| Txtls | – | – | – | – | – | – | – | – | – | – |

*(continued)*

**OLS Post-LASSO Predictive Regression Estimation Results, January 1960 to December 2016**

| | Regressand | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Regressor** | **Cnstr** | **Steel** | **FabPr** | **ElcEq** | **Autos** | **Carry** | **Mines** | **Coal** | **Oil** | **Util** |
| Cnstr | – | – | – | – | – | – | – | – | – | *−0.18* |
| Steel | – | – | – | – | – | – | – | – | – | – |
| FabPr | – | – | – | – | – | – | – | – | – | **0.12** |
| ElcEq | – | – | – | – | – | – | – | – | – | – |
| Autos | – | – | – | – | – | – | – | −0.002 | – | – |
| Carry | – | – | – | – | – | – | – | – | *0.17* | **0.08** |
| Mines | – | – | – | – | – | – | – | – | – | −0.04 |
| Coal | *−0.06* | – | – | – | −0.05 | – | – | **0.08** | – | – |
| Oil | *−0.14* | – | – | – | **−0.13** | – | – | *−0.20* | – | *−0.08* |
| Util | *0.15* | – | – | – | *0.17* | – | – | – | – | **0.09** |
| Telcm | – | – | – | – | – | – | – | – | – | 0.07 |
| Servs | – | – | – | – | – | – | – | – | – | – |
| BusEq | – | – | – | – | *0.12* | – | – | – | – | 0.03 |
| Paper | – | – | – | – | – | – | – | 0.19 | – | – |
| Trans | 0.06 | – | 0.06 | – | – | *0.16* | – | – | – | – |
| Whlsl | – | – | – | – | – | – | – | – | – | *−0.14* |
| Rtail | 0.001 | – | – | – | **0.18** | – | – | 0.10 | – | – |
| Meals | – | – | – | – | – | – | – | – | – | – |
| Fin | **0.15** | *0.15* | 0.09 | *0.10* | 0.14 | – | – | – | – | *0.13* |
| Other | – | – | – | – | – | – | – | – | – | **0.10** |
| $R^2$ | 5.13% | 1.29% | 1.56% | 0.80% | 6.13% | 2.27% | – | 2.84% | 2.52% | 7.88% |
| | (7.14) | (8.21) | (1.89) | (0.64) | (14.25) | (1.80) | – | (5.65) | (1.77) | (5.93) |

| | Regressand | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Regressor** | **Telcm** | **Servs** | **BusEq** | **Paper** | **Trans** | **Whlsl** | **Rtail** | **Meals** | **Fin** | **Other** |
| Food | – | – | – | – | – | −0.10 | – | – | – | – |
| Beer | −0.06 | – | – | – | – | −0.05 | – | – | – | – |
| Smoke | −0.03 | **−0.09** | *−0.14* | – | – | −0.05 | – | −0.06 | – | – |
| Games | – | – | – | – | – | – | – | – | – | – |
| Books | 0.09 | 0.10 | *0.12* | – | – | *0.14* | – | 0.06 | – | – |
| Hshld | −0.07 | – | – | – | – | – | – | – | – | – |
| Clths | – | – | – | 0.06 | – | – | – | 0.10 | – | 0.08 |
| Hlth | – | – | – | – | – | −0.06 | – | – | – | – |
| Chems | – | – | – | – | – | – | – | – | – | – |
| Txtls | – | – | – | – | – | – | – | – | – | – |
| Cnstr | −0.07 | – | – | – | – | – | – | – | – | – |
| Steel | – | −0.08 | – | – | – | – | – | *−0.12* | – | – |
| FabPr | – | – | – | – | – | – | – | – | – | – |
| ElcEq | – | – | – | – | – | – | – | – | – | – |
| Autos | −0.04 | – | – | – | – | – | – | – | – | – |
| Carry | −0.04 | – | – | – | – | 0.06 | – | 0.04 | – | – |
| Mines | −0.01 | – | – | – | – | – | – | – | – | – |
| Coal | −0.02 | – | – | **−0.04** | – | −0.04 | – | **−0.05** | – | – |
| Oil | **−0.09** | *−0.12* | – | *−0.11* | – | *−0.15* | – | **−0.15** | – | – |
| Util | *0.16* | 0.12 | *0.18* | – | – | *0.25* | – | 0.12 | – | – |
| Telcm | – | – | – | – | – | **−0.12** | – | – | – | – |
| Servs | 0.02 | – | – | – | – | 0.05 | – | 0.07 | – | – |

*(continued)*

**OLS Post-LASSO Predictive Regression Estimation Results, January 1960 to December 2016**

| | Regressand | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Regressor** | **Telcm** | **Servs** | **BusEq** | **Paper** | **Trans** | **Whlsl** | **Rtail** | **Meals** | **Fin** | **Other** |
| BusEq | 0.05 | – | – | – | – | 0.03 | – | 0.06 | – | – |
| Paper | – | – | – | – | – | – | – | – | – | – |
| Trans | – | – | – | – | – | – | – | – | – | – |
| Whlsl | – | – | – | – | – | – | – | – | – | – |
| Rtail | 0.10 | – | – | 0.03 | – | – | *0.13* | – | – | – |
| Meals | **−0.10** | – | – | – | – | – | – | 0.05 | – | – |
| Fin | *0.16* | **0.16** | – | **0.10** | *0.12* | 0.11 | – | 0.05 | *0.13* | **0.10** |
| Other | – | – | – | – | – | 0.05 | – | – | – | – |
| $R^2$ | 5.18% | 2.88% | 2.75% | 3.24% | 1.29% | 7.46% | 1.61% | 7.91% | 1.70% | 2.69% |
| | (4.36) | (2.74) | (3.84) | (3.33) | (1.28) | (6.64) | (1.08) | (6.52) | (1.38) | (6.26) |

*Notes: The exhibit reports OLS slope coefficient estimates and the $R^2$ statistic for the predictive regression model selected by the LASSO. The regressand is the excess return for the industry portfolio in the column heading. The regressors are selected from the complete set of lagged industry excess returns in the first column. Each predictive regression model includes an intercept term. Bold (italicized bold) indicates significance at the 10% (5%) level according to conventional OLS post-LASSO t-statistics; − indicates that the lagged industry excess return was not selected by the LASSO. Parentheses report the Campbell and Thompson (2008) measure of the proportional increase in average excess return for a mean-variance investor who uses return predictability when allocating between a given industry portfolio and risk-free bills.*

Lagged Coal and Oil returns are selected by the LASSO for 16 and 13 of the individual industries, respectively, in Exhibit 2. Eleven and 13 (7 and 9) of the coefficient estimates for lagged Coal and Oil returns, respectively, are significant at the 10% (5%) level based on the conventional OLS post-LASSO *t*-statistics. The estimated coefficients for lagged Coal and Oil returns are all negative in Exhibit 2, with the exception of the autoregression coefficient for Coal. These negative relationships presumably stem from supply shocks that raise product prices and returns for sectors located in earlier production stages but squeeze profit margins and lower returns for sectors located in later production stages. The significant coefficient estimates are again sizable in magnitude.

Although additional predictive relationships in Exhibit 2 readily accord with gradual information diffusion across related industries, there are other relationships that are more challenging to explain; for example, it is not obvious what economic channel links lagged Beer (Beer and Liquor) to future Coal returns. It is well known that machine learning is an effective means for discovering new relationships in the data, and we uncover some unusual relationships in Exhibit 2. Future research is needed to better understand their economic underpinnings.

For the industries with at least one lagged return selected by the LASSO, the $R^2$ statistics at the bottom of Exhibit 2 range from 0.78% (Chems, Chemicals) to 7.93% (Clths, Apparel). Because monthly stock returns inherently contain a sizable unpredictable component, the degree of monthly stock return predictability will necessarily be limited. To assess the economic importance of the $R^2$ statistics, we use the convenient metric suggested by Campbell and Thompson (2008):

$$CT_i = \left( \frac{R_i^2}{1 - R_i^2} \right)\left( \frac{1 + S_i^2}{S_i^2} \right), \qquad (7)$$

where $R_i^2$ is the $R^2$ statistic for the predictive regression when $r_{i,t}$ is the regressand, and $S_i$ is the unconditional Sharpe ratio for $r_{i,t}$. Equation 7 measures the proportional increase in average excess return for a mean–variance investor who allocates between the industry $i$ equity portfolio and risk-free bills when the investor uses return predictability relative to when the investor ignores return predictability.

The parentheses below the $R^2$ statistics in Exhibit 2 report the $CT_i$ metric for each industry based on its monthly Sharpe ratio in Exhibit 1 and $R^2$ statistic in Exhibit 2.[11] According to the $CT_i$ metrics at the bottom of Exhibit 2, industry return predictability based on

---

[11] We divide the annualized Sharpe ratio in Exhibit 1 by $\sqrt{12}$ to arrive at the monthly Sharpe ratio.

**Ordered Unadjusted and Adjusted *p*-Values**
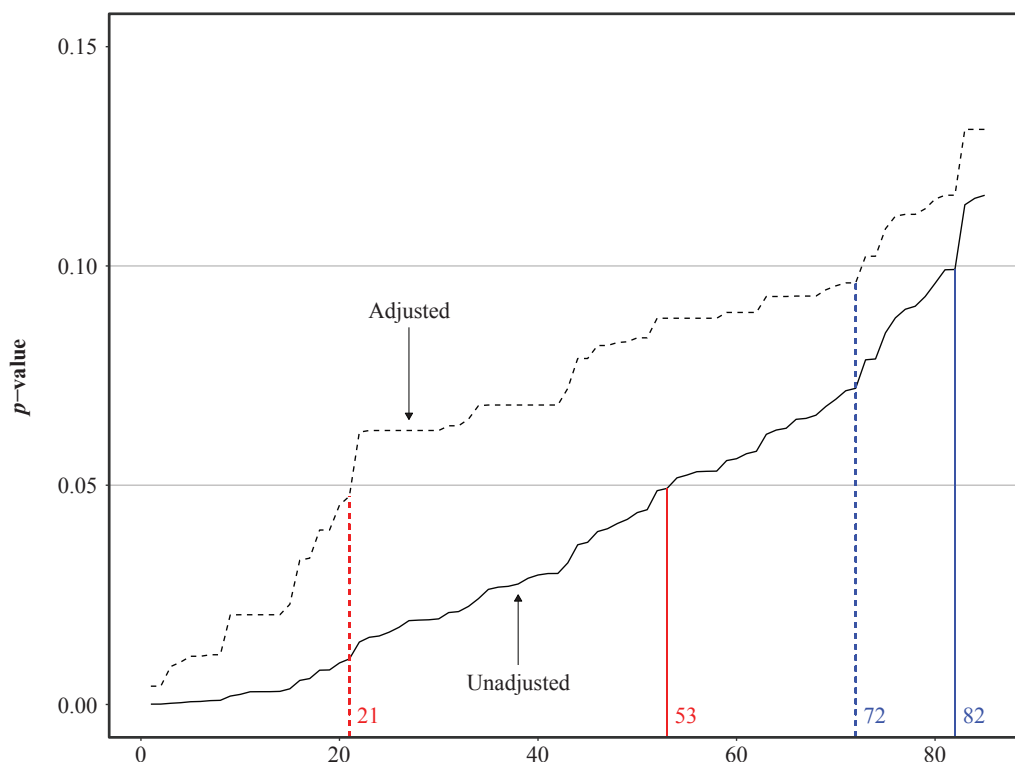


*Notes: The exhibit depicts the first 85 elements for the sequence of ordered p-values for testing the significance of the OLS post-LASSO coefficient estimates in Exhibit 2. The solid line delineates unadjusted p-values. The dashed line delineates adjusted p-values based on the Benjamini and Hochberg (2000) adaptive procedure. The second (first) vertical solid and second (first) vertical dashed lines indicate the number of rejections according to the unadjusted and adjusted p-values, respectively, at the 10% (5%) level.*

lagged industry returns increases the average excess return for a mean–variance investor by proportional factors ranging from 0.64 (ElcEq, Electrical Equipment) to 14.25 (Autos, Automobiles and Trucks). The vast majority of the $CT_i$ measures are above one, so that the average excess return for a mean–variance investor more than doubles. The substantial proportional increases in average excess return indicate that the $R^2$ statistics represent an economically meaningful degree of return predictability for nearly all of the industries.

Compared to OLS estimation of the full model, the LASSO sets numerous unimportant coefficients to zero, thereby reducing the noise in the data to better identify the predictive signal in lagged industry returns. As previously discussed, however, the LASSO tends to over-shrink the full-model OLS coefficient estimates for the predictors selected by the LASSO. Indeed, on average,

the LASSO coefficient estimates are less than half as large (in magnitude) as the corresponding full-model OLS estimates for the lagged industry returns selected by the LASSO.[12] The degree of shrinkage induced by OLS post-LASSO estimation is more limited: On average, the OLS post-LASSO estimates are 15% smaller (in magnitude) than the corresponding full-model OLS estimates for the LASSO-selected predictors. Of course, the $R^2$ statistics for the full models estimated via OLS are larger than the corresponding statistics at the bottom of Exhibit 2, since OLS estimation of the full model maximizes the $R^2$ statistic by construction. However, the $R^2$ statistics for the full models likely overstate the actual degree of return predictability, as OLS estimation fails to

---

[12] The LASSO and full-model OLS coefficient estimates are reported in Exhibits S2 and S3, respectively, of the online supplement.

adequately filter the noise in the data. In summary, OLS post–LASSO estimation more accurately measures the predictive signal in lagged industry returns by avoiding both the overfitting associated with full-model OLS estimation and overshrinking of coefficient estimates engendered by the LASSO for the LASSO-selected predictors.

We check the robustness of the results along a number of dimensions. First, to examine the robustness of the post-selection inferences, we compute confidence intervals using the approach of Lee et al. (2016). Demonstrating that the LASSO selection event can be characterized as a union of polyhedra, they construct valid confidence intervals conditional on the event $\mathcal{M} = \widehat{\mathcal{M}}$. Many of the coefficient estimates that are significant in Exhibit 2 are also significant based on Lee et al. (2016) confidence intervals.[13]

Next, we check whether the return predictability in Exhibit 2 reflects time variation in risk premiums by augmenting Equation 1 with four lagged predictor variables similar to those used by Ferson and Harvey (1991, 1999), Ferson and Korajczyk (1995), and Avramov (2004): the S&P 500 dividend yield, three-month Treasury bill yield, difference between yields on a 10-year Treasury bond and a three-month Treasury bill (term spread), and difference between yields on BAA- and AAA-rated corporate bonds (credit spread).[14] These variables represent popular return predictors from the literature, and they are often viewed as capturing time-varying risk premiums. When the lagged economic variables are included in Equation 1, the lagged industry returns selected by the LASSO and corresponding OLS post–LASSO coefficient estimates are typically quite similar to those reported in Exhibit 2.[15] Popular measures of time-varying risk premiums thus do not readily account for the predictive power of lagged industry returns.

Furthermore, because the LASSO tends to select one predictor variable from among a group of correlated predictors while dropping the other predictors, we select the lagged industry returns in Equation 1 using the elastic net (ENet; Zou and Hastie 2005). The penalty term for the ENet is a convex combination of $\ell_1$ (LASSO) and $\ell_2$ (ridge) components. The ENet has a stronger tendency than the LASSO to select correlated predictor variables as a group. The ENet and LASSO typically select similar sets of predictor variables for the individual industries. Indeed, the two procedures select the same set of lagged industry returns as predictors for many individual industries.[16] Practically speaking, the LASSO's simpler penalty term appears sufficient for identifying the relevant lagged industry returns for predicting individual industry returns.

## OUT-OF-SAMPLE RESULTS

This section reports out-of-sample measures of the economic value of industry return predictability in the context of a monthly long-short industry-rotation portfolio. We construct a long-short industry-rotation portfolio for January 1970 to December 2016 using out-of-sample forecasts of monthly industry excess returns based on OLS post–LASSO estimation of predictive regressions for each industry.

We construct the long-short industry-rotation portfolio as follows. We first use data from the beginning of the sample through December 1969 to estimate Equation 1 for each industry via OLS post–LASSO and generate a set of 30 out-of-sample industry excess return forecasts for January 1970. We sort the industries in ascending order according to the excess return forecasts and form equal-weighted quintile portfolios; we then create a zero-investment portfolio that goes long (short) the top (bottom) quintile portfolio. Next, we use data through January 1970 to compute an updated set of out-of-sample industry excess return forecasts for February 1970 based on OLS post–LASSO estimation of Equation 1, sort the industries according to the forecasts, and form equal-weighted quintile portfolios; the zero-investment portfolio again goes long (short) the top (bottom) quintile portfolio. Continuing in this fashion, we construct a monthly long-short industry-rotation portfolio based on the OLS post–LASSO industry excess return forecasts for the January 1970 to December 2016 out-of-sample period (564 months). In generating the out-of-sample excess return forecasts, observe that we only use data available at the time of forecast formation when selecting

---

[13] The complete results are reported in Exhibit S4 of the online supplement. We compute Lee et al. (2016) confidence intervals using the selectiveInference package (Tibshirani et al. 2017).

[14] Data for these variables are from Global Financial Data at https://www.globalfinancialdata.com.

[15] Complete results for the augmented predictive regressions are reported in Exhibit S5 of the online supplement.

[16] The complete ENet results are reported in Exhibit S6 of the online supplement.

# Exhibit 4
**Industry-Rotation Portfolio Performance, January 1970 to December 2016**

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|
| Forecasts | Ann. Mean (%) | Ann. Vol. (%) | Ann. Sharpe Ratio | Ann. Downside Risk (%) | Ann. Sortino Ratio | Max. Drawdown (%) | Ann. MPPM (%) |
| Prevailing Mean | −2.22 | 10.98 | −0.20 | 8.34 | −0.27 | 73.97 | −4.73 |
| OLS | 5.52 | 11.64 | 0.47 | 7.17 | 0.77 | 29.39 | 2.84 |
| OLS Post-LASSO | 7.33 | 11.29 | 0.65 | 6.31 | 1.16 | 25.65 | 4.84 |

*Notes: The exhibit reports performance measures for long–short industry-rotation portfolios. At the end of each month, we sort 30 industry portfolios according to out-of-sample forecasts of their excess returns for the subsequent month. The out-of-sample industry excess return forecasts are based on predictive regression models estimated via OLS or OLS post-LASSO, as well as prevailing mean forecasts, as indicated in the first column. We then form equal-weighted quintile portfolios based on the sorts, and each long–short industry-rotation portfolio is a zero-investment portfolio that goes long (short) the top (bottom) quintile portfolio. Downside risk is measured as the semi-standard deviation. MPPM is the Goetzmann et al. (2007) manipulation-proof performance measure for a relative risk aversion coefficient of four.*

the predictor variables via the LASSO and estimating the predictive regressions.

For the purpose of comparison, we construct a pair of benchmark long–short industry-rotation portfolios. The first benchmark is constructed in the same manner as the portfolio described previously, except that we use out-of-sample industry excess return forecasts based on the prevailing mean. The prevailing mean forecast corresponds to the constant expected excess return model— Equation 1 with $b_{i,j}^{*} = 0$ for $j = 1, \ldots, N$—and assumes that individual industry excess returns do not depend on lagged industry returns. The forecast is simply the mean industry excess return based on data from the beginning of the sample through the month of forecast formation. The prevailing mean forecast is a popular benchmark in out-of-sample tests of return predictability (e.g., Goyal and Welch 2008). We use this benchmark to assess portfolio performance under the assumption of no industry excess return predictability (apart from the mean excess return).

The second benchmark portfolio is again constructed in the same manner, but it relies on out-of-sample industry excess return forecasts based on OLS estimation of Equation 1. Comparing the performance of the benchmark portfolio based on the OLS forecasts to that of the portfolio based on the OLS post-LASSO forecasts provides a sense of the ability of machine learning to increase the economic value of out-of-sample industry return forecasts.

Exhibit 4 reports performance measures for the industry-rotation portfolio based on the OLS post-LASSO forecasts and the two benchmark portfolios.

In addition to the annualized mean, volatility, and Sharpe ratio, the exhibit reports the annualized downside risk and Sortino ratio, maximum drawdown, and annualized Goetzmann et al. (2007) MPPM for a relative risk aversion coefficient of four. MPPM measures the continuously compounded certainty equivalent return in excess of the risk–free return for a power-utility investor.

The benchmark portfolio based on the prevailing mean forecasts generates an annualized average return of −2.22%, so that simply going long (short) industries with the highest (lowest) historical returns is not a profitable strategy over the out-of-sample period. Because of the negative average return, the Sharpe and Sortino ratios are both negative. The prevailing mean benchmark portfolio also has a sizable annualized downside risk (8.34%), large maximum drawdown (73.97%), and negative annualized MPPM (−4.73%).

In contrast to the prevailing mean benchmark, the benchmark portfolio based on the OLS forecasts produces a positive annualized average return of 5.52%. Although it has somewhat higher volatility than the prevailing mean benchmark, the OLS benchmark evinces a smaller downside risk (7.17%), much lower maximum drawdown (29.39%), positive Sharpe and Sortino ratios (0.47 and 0.77, respectively), and positive MPPM (2.84%). Overall, incorporating the information in lagged industry returns, as captured by the OLS forecasts, improves portfolio performance relative to ignoring such information.

Although the OLS benchmark portfolio generally outperforms the prevailing mean benchmark in Exhibit 4,

Exhibit 5
**Log Cumulative Returns for Industry-Rotation Portfolios, January 1970 to December 2016**
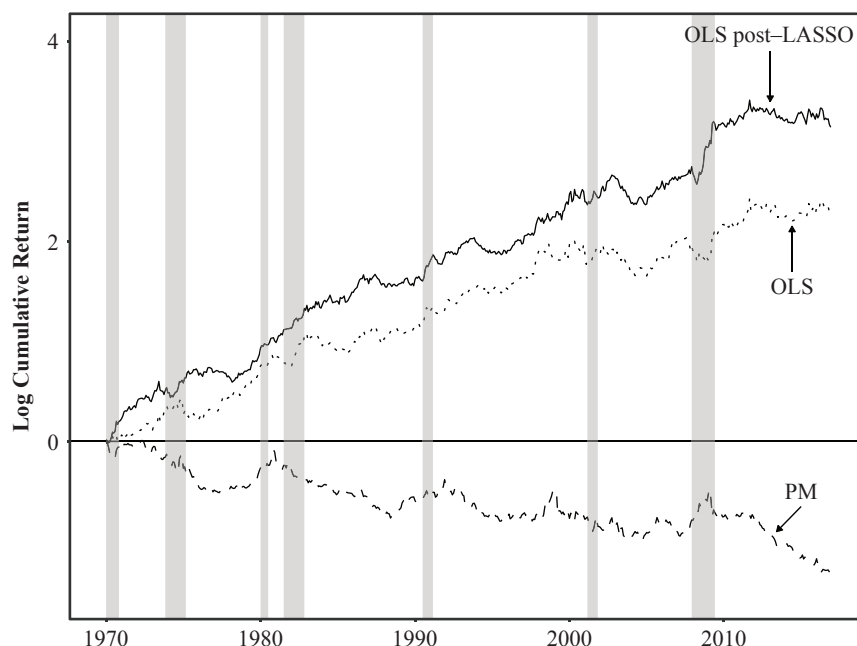


*Notes: The exhibit shows log cumulative returns for long-short industry-rotation portfolios that go long (short) the six industries with the highest (lowest) out-of-sample excess return forecasts using the forecasts indicated in the exhibit (PM is the prevailing mean). Vertical bars delineate NBER-dated recessions.*

OLS estimation of the predictive regressions is vulnerable to overfitting, which potentially detracts from the usefulness of the OLS forecasts as inputs for constructing the industry-rotation portfolio. The performance measures for the industry-rotation portfolio based on the OLS post–LASSO forecasts indicate that our machine learning approach increases the economic value of industry return forecasts, as the OLS post–LASSO forecasts improve portfolio performance vis-à-vis the OLS forecasts across all measures in Exhibit 4. The annualized average return and volatility of 7.33% and 11.29%, respectively, for the OLS post–LASSO portfolio translate into a substantive Sharpe ratio of 0.65. The downside risk and maximum drawdown (6.31% and 25.65%, respectively) are smaller for the OLS post–LASSO portfolio vis-à-vis the OLS portfolio, whereas the Sortino ratio (1.16) is considerably larger. Furthermore, the annualized MPPM increases by 200 bps (to 4.84%) when we allocate across industries using the OLS post–LASSO forecasts in lieu of the OLS forecasts.

Exhibit 5 depicts log cumulative returns for the long-short industry–rotation portfolios. In line with its negative average return in Exhibit 4, the prevailing mean benchmark portfolio (dashed line) experiences prolonged periods of poor performance in Exhibit 5. The OLS benchmark (dotted line) produces gains on a relatively consistent basis, and it typically performs well during business–cycle recessions as dated by the National Bureau of Economic Research (NBER). The industry-rotation portfolio based on the OLS post–LASSO forecasts (solid line) also delivers sizable gains on a consistent basis and appears to perform even better than the OLS benchmark during recessions, especially the recent Great Recession.

Exhibit 6 examines portfolio performance over the business cycle in greater detail. The exhibit reports annualized average portfolio returns during good and bad macroeconomic states, measured as months when the economy is in an NBER-dated expansion or recession, respectively. As a robustness check, we also measure the state of the economy using the Chicago Fed national activity index (CFNAI), in which months in the bottom quintile (top 80%) of CFNAI observations are deemed a bad (good) state.[17]

---

[17] Monthly indicator variables for NBER recessions and CFNAI observations are from Federal Reserve Economic Data at https://fred.stlouisfed.org.

## Exhibit 6
### Annualized Average Returns during Good and Bad States, January 1970 to December 2016

| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| | NBER Business-Cycle Phases | | | CFNAI Observations | | |
| Forecasts | Expansion | Recession | Difference | Upper 80% | Lower 20% | Difference |
| Prevailing | –2.54% | –0.57% | 1.97% | –3.49% | 2.72% | 6.22% |
| mean | [–1.45] | [–0.14] | [0.45] | [–1.95]* | [0.77] | [1.57] |
| OLS | 4.20% | 12.54% | 8.34% | 5.56% | 5.34% | –0.22% |
| | [2.27]** | [2.94]*** | [1.79]* | [2.92]*** | [1.42] | [–0.05] |
| OLS | 4.62% | 21.75% | 17.13% | 5.07% | 16.14% | 11.08% |
| post-LASSO | [2.61]*** | [5.31]*** | [3.84]*** | [2.76]*** | [4.45]*** | [2.72]*** |

*Notes: The exhibit reports annualized average returns for long-short industry-rotation portfolios during good and bad macroeconomic states. At the end of each month, we sort 30 industry portfolios according to out-of-sample forecasts of their excess returns for the subsequent month. The out-of-sample industry excess return forecasts are based on predictive regression models estimated via OLS or OLS post-LASSO, as well as prevailing mean forecasts, as indicated in the first column. We then form equal-weighted quintile portfolios based on the sorts, and each long-short industry-rotation portfolio is a zero-investment portfolio that goes long (short) the top (bottom) quintile portfolio. Good (bad) states are months when the economy is in an expansion (recession) according to NBER-dated cyclical turning points or months in the upper 80% (lower 20%) of CFNAI observations. Difference is the difference in average returns during bad versus good states. Brackets report t-statistics.*

*\*, \*\*, \*\*\* indicate significance at the 10%, 5%, and 1% levels, respectively.*

The average returns for the benchmark portfolio based on the prevailing mean forecasts are negative during both expansions and recessions (although neither is significant), and the difference in average returns during recessions vis-à-vis expansions is insignificant. Confirming the visual impression in Exhibit 5, the benchmark portfolio based on the OLS forecasts generates annualized average returns of 4.20% and 12.54% during expansions and recessions, respectively, both of which are significant. The difference in average returns is a sizable 8.34%, which is significant at the 10% level. The portfolio based on the OLS post-LASSO forecasts produces a significant annualized average return of 4.62% during expansions and a striking (and significant) average return of 21.75% during recessions. The difference (17.13%) is significant at the 1% level. When we measure the macroeconomic state using the CFNAI, the OLS post-LASSO portfolio provides significant average returns of 5.07% and 16.14% during good and bad states, respectively, and the difference of 11.08% is again significant at the 1% level. In contrast, the average returns across good and bad states (5.56% and 5.34%, respectively) become much closer for the OLS benchmark portfolio, and the difference (–0.22%) is no longer significant.[18]

The performance of the industry-rotation portfolio based on the OLS post-LASSO forecasts during cyclical downturns is difficult to explain with a rational macroeconomic risk–based story. By paying out handsomely in bad states when marginal utility is likely to be high, the long-short portfolio provides a hedge against bad times for the macroeconomy. From a rational asset pricing perspective, such a portfolio should have a high price (i.e., low expected return); however, the portfolio generates a sizable annualized average return (7.33%) in Exhibit 4.

Next, we test whether exposures to equity risk factors in leading multifactor models from the literature can account for the behavior of the long-short industry-rotation portfolios. Specifically, we estimate a portfolio's alpha in the context of the popular Carhart (1997) four-factor and Hou, Xue, and Zhang (2015) q-factor models. The former includes the three Fama and French (1993) factors—market, size, and value—and a momentum factor (Jegadeesh and Titman 1993). Motivated by the q theory of investment, the Hou, Xue, and Zhang (2015) model contains market, size, investment, and profitability factors.[19]

---

[18] The inferences in Exhibit 6 are similar for t-statistics based on White (1980) heteroskedasticity-robust standard errors; see Exhibit S7 of the online supplement.

[19] Factor data for the Carhart model are from Kenneth French's Data Library. We thank Kewei Hou for providing us with factor data for the q-factor model.

**Multifactor Model Estimation Results, January 1970 to December 2016**

| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| | PM Forecasts | | OLS Forecasts | | OLS Post-LASSO Forecasts | |
| Factor | Carhart Four-Factor Model | HXZ $q$-Factor Model | Carhart Four-Factor Model | HXZ $q$-Factor Model | Carhart Four-Factor Model | HXZ $q$-Factor Model |
| Ann. α | −1.84% | −2.91% | 6.64% | 5.57% | 8.78% | 8.04% |
| | [−1.29] | [−1.84]* | [3.78]*** | [3.04]*** | [5.19]*** | [4.55]*** |
| Market | −0.03 | −0.02 | −0.05 | −0.02 | −0.13 | −0.10 |
| | [−1.17] | [−0.63] | [−1.2] | [−0.56] | [−3.87]*** | [−3.09]*** |
| Size | −0.12 | −0.04 | −0.12 | −0.13 | −0.07 | −0.07 |
| | [−3.19]*** | [−0.85] | [−2.60]*** | [−2.56]** | [−1.61] | [−1.54] |
| Value | −0.38 | | −0.10 | | −0.12 | |
| | [−9.05]*** | | [−1.87]* | | [−2.35]** | |
| Momentum | 0.24 | | −0.01 | | 0.004 | |
| | [8.98]*** | | [−0.37] | | [0.12] | |
| Investment | | −0.37 | | 0.13 | | 0.002 |
| | | [−5.37]*** | | [1.64] | | [0.02] |
| Profitability | | 0.45 | | −0.04 | | 0.02 |
| | | [8.98]*** | | [−0.61] | | [0.38] |
| $R^2$ | 27.80% | 18.19% | 2.22% | 2.30% | 4.06% | 3.28% |

*Notes: The exhibit reports Carhart (1997) four-factor and Hou, Xue, and Zhang (2015; HXZ) q-factor model estimation results for long-short industry-rotation portfolios. At the end of each month, we sort 30 industry portfolios according to out-of-sample forecasts of their excess returns for the subsequent month. The out-of-sample industry excess return forecasts are based on predictive regression models estimated via OLS or OLS post-LASSO, as well as prevailing mean (PM) forecasts, as indicated at the top of the exhibit. We then form equal-weighted quintile portfolios based on the sorts, and each long-short industry-rotation portfolio is a zero-investment portfolio that goes long (short) the top (bottom) quintile portfolio. The regressand for the multiple regression model is the return for the portfolio indicated at the top of the exhibit. Brackets report t-statistics.*

*\*, \*\*, \*\*\* indicate significance at the 10%, 5%, and 1% levels, respectively.*

The benchmark portfolio based on the prevailing mean forecasts evinces significant negative (positive) exposures to the size and value (momentum) factors in the Carhart model. The annualized alpha is −1.84%, which is insignificant at conventional levels. The portfolio's annualized alpha is also negative (−2.91%, significant at the 10% level) in the $q$-factor model, and it exhibits significant negative (positive) exposure to the investment (profitability) factor. The factors collectively explain 27.80% and 18.19% of the variation in portfolio return in the Carhart and $q$-factor models, respectively.

Turning to the benchmark portfolio based on the OLS forecasts, the portfolio displays significant negative exposures to the size and value factors (the latter at the 10% level) in the Carhart model. The portfolio also generates an economically sizable annualized alpha of 6.64% (significant at the 1% level). For the $q$-factor model, the portfolio again exhibits significant negative

exposure to the size factor and produces a sizable annualized alpha (5.57%, significant at the 1% level). The $R^2$ statistics are below 2.50% for both multifactor models. Comparing the multifactor model estimation results for the prevailing mean and OLS benchmark portfolios in Exhibit 7, as in Exhibit 4, we see that incorporating the information in lagged industry returns improves portfolio performance.

For the Carhart model, the industry-rotation portfolio based on the OLS post-LASSO forecasts delivers an annualized alpha of 8.78% (significant at the 1% level); the portfolio thus generates a substantial risk-adjusted average return. The risk-adjusted average return is nearly 150 bps higher than its unadjusted counterpart in Exhibit 4, due to the portfolio's significant negative exposures to the market and value factors (−0.13 and −0.12, respectively) and insignificant exposures to the size and

momentum factors in Exhibit 7.[20] The portfolio's annualized alpha is again above 8% (8.04%, significant at the 1% level) in the $q$-factor model, with significant negative exposure to the market factor and insignificant exposures to the remaining factors.

The annualized alphas for the OLS post-LASSO portfolio are over 200 bps higher than those for the OLS benchmark in Exhibit 7, providing further evidence that the OLS post-LASSO forecasts add value by reducing the noise in the OLS forecasts. The $R^2$ statistics for the Carhart and $q$-factor models are relatively small (4.06% and 3.28%, respectively), indicating that the equity risk factors have limited explanatory power for the OLS post-LASSO portfolio return.[21] Results from Frazzini, Israel, and Moskowitz (2015) suggest that, at least for a large institutional investor, transaction costs will have relatively little impact on the substantial average return and alpha generated by the industry-rotation portfolio based on the OLS post-LASSO forecasts.

## CONCLUSION

We use machine learning tools to analyze industry return predictability based on the information in lagged industry returns from across the economy. We begin by specifying a general predictive regression model for an individual industry's return that includes lagged returns for all 30 of the industries that we consider as regressors. Because conventional OLS estimation of predictive regressions with such a plethora of predictor variables runs the risk of overfitting, we use the LASSO from machine learning to fit sparse models. To circumvent

the downward biases (in magnitude) in the LASSO coefficient estimates themselves, we re-estimate the coefficients for the LASSO-selected predictor variables via OLS (OLS post-LASSO estimation). Controlling for post-selection inference and multiple testing, in-sample results provide extensive evidence of industry return predictability, pointing to the existence of industry-related information frictions in the equity market.

We also compute out-of-sample forecasts of industry returns based on OLS post-LASSO estimation of the predictive regressions to construct a zero-investment industry-rotation portfolio that goes long (short) industries with the highest (lowest) forecasted returns. The long-short industry-rotation portfolio earns a significant average return, performs well during cyclical downturns, and delivers a substantial annualized alpha of over 8% in the context of leading multifactor models. The information in lagged industry returns thus appears quite valuable for generating risk-adjusted average returns.

Our article shows that the historical information in industry returns contains statistically and economically significant predictive power, thereby providing machine learning-based evidence against the weak form of market efficiency. With additional information, such as industry characteristics, return predictability may be even stronger. Along this line, the encompassing LASSO approach developed by Han et al. (2019), which currently provides the most accurate out-of-sample forecasts of cross-sectional expected stock returns according to the Lewellen (2015) test, holds promise for further improving predictive performance. In summary, we expect machine learning techniques to continue to enhance our understanding of industry return predictability in future research.

---

[20] We also constructed a cross-sectional industry-momentum portfolio along the lines of Moskowitz and Grinblatt (1999). Specifically, each month we sort the 30 industries according to their cumulative excess returns over the previous 12 months and go long (short) the top (bottom) quintile of sorted industries. Consistent with Moskowitz and Grinblatt (1999), the cross-sectional industry-momentum portfolio evinces large exposure (0.87, significant at the 1% level) to the momentum factor in the Carhart model and generates insignificant alpha.

[21] As shown in Exhibit S8 of the online supplement, the inferences in Exhibit 7 are similar for $t$-statistics based on White (1980) heteroskedasticity-robust standard errors. The portfolio based on the OLS post-LASSO forecasts also generates an annualized alpha of 8.12% (significant at the 1% level) in the Fama and French (2015) five-factor model and an annualized alpha of 8.17% (significant at the 1% level) in a six-factor model composed of the five Fama and French (2015) factors and the momentum factor from the Carhart model.

for insightful comments that substantively improved the article. The authors are also grateful to Frank Fabozzi (the editor) for very helpful comments.

## REFERENCES

Akaike, H. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the 2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 267–281. Budapest: Akadémiai Kiadó, 1973.

Arnott, R., C. R. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science* 1 (1): 64–74.

Avramov, D. 2004. "Stock Return Predictability and Asset Pricing Models." *The Review of Financial Studies* 17 (3): 699–738.

Belloni, A., and V. Chernozhukov. "High Dimensional Sparse Econometric Models: An Introduction." In *Inverse Problems and High-Dimensional Estimation*, edited by P. Alquier, E. Gautier, and G. Stoltz, pp. 121–156. Berlin, Heidelberg: Springer, 2011.

——. 2013. "Least Squares after Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19 (2): 521–547.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

——. 2000. "On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics." *Journal of Educational and Behavioral Statistics* 25 (1): 60–83.

Benjamini, Y., A. M. Krieger, and D. Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika* 93 (3): 491–507.

Benjamini, Y., and D. Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29 (4): 1165–1188.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao. 2013. "Valid Post-Selection Inference." *Annals of Statistics* 41 (2): 802–837.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics* 37 (4): 1705–1732.

Blanchard, G., and E. Roquain. 2009. "Adaptive False Discovery Rate Control under Independence and Dependence." *Journal of Machine Learning Research* 10: 2837–2871.

Box, G. E. P. "Robustness in the Strategy of Scientific Model Building." In *Robustness in Statistics: Proceedings of a Workshop*, edited by R. L. Launer and G. N. Wilkinson, pp. 201–236. New York: Academic Press, 1979.

Campbell, J. Y., and S. B. Thompson. 2008. "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *The Review of Financial Studies* 21 (4): 1509–1531.

Carhart, M. M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Chinco, A., A. D. Clark-Joseph, and M. Ye. 2019. "Sparse Signals in the Cross-Section of Returns." *The Journal of Finance* 74 (1): 449–492.

Cohen, L., and A. Frazzini. 2008. "Economic Links and Predictable Returns." *The Journal of Finance* 63 (4): 1977–2011.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen. 2015. "High-Dimensional Inference: Confidence Intervals, $p$-Values and R-Software hdi." *Statistical Science* 30 (4): 533–558.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2): 407–451.

Fama, E. F., and K. R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

——. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

Fan, J., and R. Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96 (456): 1348–1360.

Farcomeni, A. 2007. "Some Results on the Control of the False Discovery Rate under Dependence." *Scandinavian Journal of Statistics* 34 (2): 275–297.

Ferson, W. E., and C. R. Harvey. 1991. "The Variation of Economic Risk Premiums." *Journal of Political Economy* 99 (2): 385–415.

——. 1999. "Conditioning Variables and the Cross Section of Stock Returns." *The Journal of Finance* 54 (4): 1325–1360.

Ferson, W. E., and R. A. Korajczyk. 1995. "Do Arbitrage Pricing Models Explain the Predictability of Stock Returns?" *The Journal of Business* 68 (3): 309–349.

Flynn, C. J., C. M. Hurvich, and J. S. Simonoff. 2013. "Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models." *Journal of the American Statistical Association* 108 (503): 1031–1043.

Frazzini, A., R. Israel, and T. J. Moskowitz. "Trading Costs of Asset Pricing Anomalies." Working paper, 2015.

Freyberger, J., A. Neuhierl, and M. Weber. "Dissecting Characteristics Nonparametrically." *The Review of Financial Studies* (forthcoming).

Friedman, J., T. Hastie, N. Simon, J. Qian, and R. Tibshirani. "glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models." R Package Version 2.0-13, 2017.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22.

Goetzmann, W., J. Ingersoll, M. Spiegel, and I. Welch. 2007. "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures." *The Review of Financial Studies* 20 (5): 1503–1546.

Goyal, A., and I. Welch. 2008. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *The Review of Financial Studies* 21 (4): 1455–1508.

Green, J., J. R. M. Hand, and X. F. Zhang. 2017. "The Characteristics that Provide Independent Information about Average US Monthly Stock Returns." *The Review of Financial Studies* 30 (12): 4389–4436.

Gu, S., B. Kelly, and D. Xiu. "Empirical Asset Pricing via Machine Learning." Working paper, 2018.

Han, Y., A. He, D. E. Rapach, and G. Zhou. "What Firm Characteristics Drive the Cross Section of Expected Stock Returns?" Working paper, 2019.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. "…and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

Hochberg, Y., and Y. Benjamini. 1990. "More Powerful Procedures for Multiple Significance Testing." *Statistics in Medicine* 9 (7): 811–818.

Hoerl, A. E., and R. W. Kennard. 1970. "Ridge Regression: Applications to Nonorthogonal Problems." *Technometrics* 12 (1): 69–82.

Hong, H., and J. C. Stein. 1999. "A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets." *The Journal of Finance* 54 (6): 2143–2184.

Hong, H., W. Torous, and R. Valkanov. 2007. "Do Industries Lead Stock Markets?" *Journal of Financial Economics* 83 (2): 367–396.

Hou, K., C. Xue, and L. Zhang. 2015. "Digesting Anomalies: An Investment Approach." *Review of Financial Studies* 28 (3): 650–703.

———. "Replicating Anomalies." *The Review of Financial Studies* (forthcoming).

Hurvich, C. M., and C. L. Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76 (2): 297–307.

Jegadeesh, N., and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* 48 (1): 65–91.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. "Exact Post-Selection Inference, with Application to the LASSO." *Annals of Statistics* 44 (3): 907–927.

Leeb, H., B. M. Pötscher, and K. Ewald. 2015. "On Various Confidence Intervals Post-Model-Selection." *Statistical Science* 30 (2): 216–227.

Lewellen, J. 2015. "The Cross-Section of Expected Stock Returns." *Critical Finance Review* 4 (1): 1–44.

Linnainmaa, J. T., and M. R. Roberts. 2018. "The History of the Cross Section of Stock Returns." *The Review of Financial Studies* 31 (7): 2606–2649.

Meinshausen, N. 2007. "Relaxed Lasso." *Computational Statistics and Data Analysis* 52 (1): 374–393.

Meinshausen, N., L. Meier, and P. Bühlmann. 2009. "*p*-Values for High-Dimensional Regression." *Journal of the American Statistical Association* 104 (488): 1671–1681.

Meinshausen, N., and B. Yu. 2009. "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data." *Annals of Statistics* 37 (1): 246–270.

Menzly, L., and O. Ozbas. 2010. "Market Segmentation and the Cross-Predictability of Returns." *The Journal of Finance* 65 (4): 1555–1580.

Merton, R. C. 1987. "A Simple Model of Capital Market Equilibrium with Incomplete Information." *The Journal of Finance* 42 (3): 483–510.

Moskowitz, T. J., and M. Grinblatt. 1999. "Do Industries Explain Momentum?" *The Journal of Finance* 54 (4): 1249–1290.

R Core Team. *R Version 3.4.2: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2017.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2013. "International Stock Return Predictability: What Is the Role of the United States?" *The Journal of Finance* 68 (4): 1633–1662.

Rapach, D. E., and G. Zhou. "Forecasting Stock Returns." In *Handbook of Economic Forecasting*, edited by G. Elliott and A. Timmermann, pp. 328–383. Amsterdam: Elsevier, 2013.

Storey, J. D., J. E. Taylor, and D. Siegmund. 2004. "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66 (1): 187–205.

Taylor, J., and R. Tibshirani. 2015. "Statistical Learning and Selective Inference." *Proceedings of the National Academy of Sciences* 112 (25): 7629–7634.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani. 2016. "Exact Post-Selection Inference for Sequential Regression Procedures." *Journal of the American Statistical Association* 111 (514): 600–620.

Tibshirani, R. J., R. Tibshirani, J. Taylor, J. Loftus, and S. Reid. "selectiveInference: Tools for Post-Selection Inference." R Package Version 1.2.4, 2017.

Wasserman, L., and K. Roeder. 2009. "High-Dimensional Variable Selection." *Annals of Statistics* 37 (5A): 2178–2201.

White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity." *Econometrica* 48 (4): 817–838.

Zhang, C. H., and J. Huang. 2008. "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression." *Annals of Statistics* 36 (4): 1567–1594.

Zhao, S., A. Shojaie, and D. Witten. "In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference." Working paper, 2017.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–320.

*To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.*

## ADDITIONAL READING

### A Backtesting Protocol in the Era of Machine Learning
ROB ARNOTT, CAMPBELL R. HARVEY, AND HARRY MARKOWITZ
*The Journal of Financial Data Science*
https://jfds.pm-research.com/content/1/1/64

**ABSTRACT:** *Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning's early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, who may be influenced by the actions of others and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. Although investors need to be cautious—indeed, more cautious than in past applications of quantitative methods—these new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.*