Optimal Peers*

Florian S. Peters†

This version: 17 July 2024

Abstract

This paper provides a theoretical foundation for constructing optimal benchmarks via machine learning (ML). For a broad class of models, the optimal benchmark is given by an appropriately weighted portfolio of peers. While Ordinary Least Squares (OLS) provides the theoretically optimal weights in the population, ML methods, notably the lasso, can provide a robust, implementable solution. In an application to a large sample of U.S. public firms, ML-based benchmarks strongly outperform traditional industry benchmarks in out-of-sample explanatory power. This suggests that ML-based benchmarks can substantially improve outcomes in a wide range of applications, such as incentive contracts or relative performance evaluation.

Keywords: Peer groups; benchmarking; machine learning; lasso; relative performance evaluation; industry classification.

JEL Classification: D2, D4, D8, G1, G3, L1, L2.

*I especially thank Yenan Wang for countless helpful discussions. I also thank Spyros Terovitis and Simon Rottke for many helpful comments. All errors are mine.

†University of Amsterdam, email: f.s.peters@uva.nl, web: florianpeters.org.

1 Introduction

Benchmarking is a common practice in virtually all areas of business, finance, and economics. For example, corporate and investment managers are often evaluated or paid based on performance relative to an appropriate benchmark; investment projects are commonly evaluated relative to other comparable projects; company valuations are frequently based on the valuations of comparable firms.

Benchmarking serves to decompose the overall outcome of a project into a component that can be attributed to factors outside the manager's control and a component that is specific to the manager. The specific component provides a more precise signal of the manager's unique contribution to the project, thereby helping the principal to more effectively learn about the manager's ability or more effectively incentivize managerial effort.

But what benchmark should one choose? From a theoretical perspective, the properties of an optimal benchmark are straightforward: a benchmark should measure, as precisely as possible, all exogenous components of performance. In practice however, the choice – or construction – of a benchmark is often ambiguous. Consider, for example, a portfolio manager whose performance is to be benchmarked against "the market." What is the market portfolio? Should it include all asset classes or just equities? How should the constituents of the market portfolio be weighted? Or consider a corporate manager who is to be benchmarked against her "competitors." How should one select those competitors and how should each of them be weighted in a benchmark portfolio?

This paper presents a theoretically founded and empirically implementable solution to the problem of constructing optimal benchmarks. I start by restating the properties of optimal benchmarks based on Blackwell's (1951) Informativeness Theorem and Holmström's (1979, 1982) Informativeness Principle and sufficient statistics results. When the factors outside the manager's control are unobservable risks that are shared by other firms in the economy, the optimal benchmark is an appropriately weighted average of the outcomes of *all* other firms in the economy, i.e. the optimal benchmark is an optimally

weighted peer portfolio.

I show that, under mild assumptions and in the population, there is a simple general solution to the optimal peer portfolio for a given focal firm: the Ordinary Least Squares (OLS) estimand from a regression of the focal firm outcome on *all* potential peer outcomes gives the optimal weights for each peer firm in the benchmark portfolio.

This result suggests that optimal benchmarks can conveniently be constructed via OLS regression. However, OLS is not a feasible method when the number of regressors is large relative to the available sample size. Considering *all* other firms in the economy as potential peers means that such a regression needs to include thousands of regressors. But in practice, the available samples are at least an order of magnitude smaller.

For practical purposes, machine learning (ML) techniques can provide a feasible and robust solution to the problem. I implement this approach on a large sample of U.S. corporations. The key result is summarized in Figure 1. The red line plots the out-of-sample explanatory power of optimal peer groups. The other lines represent peer groups based on existing industry classifications (SIC, NAICS, GICS), an alternative peer classification (TNIC) and the market portfolio.

Figure 1 shows that optimal benchmarks perform markedly better than all other peer groups except in the very early part of the sample. Their outperformance shows a continuous increase over the past 30 years, and the gap is substantial in recent years. These results suggest that ML-based, optimal benchmarks have the potential to greatly improve incentive contracts and any decision that is driven by inference of managerial skill, such as promotions, demotions or dismissals.

Historically, benchmarks for firm performance have been constructed using a weighted average of the focal firm's industry peers. But industry classifications are often ambiguous and to some extent arbitrary. This is evident from the many different industry classifications that exist.[1] Also, traditional industry classifications do not offer any guidance on how to weight industry peers in a peer portfolio for a given focal firm. Finally, benchmarking a

[1] SIC, NAICS, GICS, BICS, TNIC, etc.

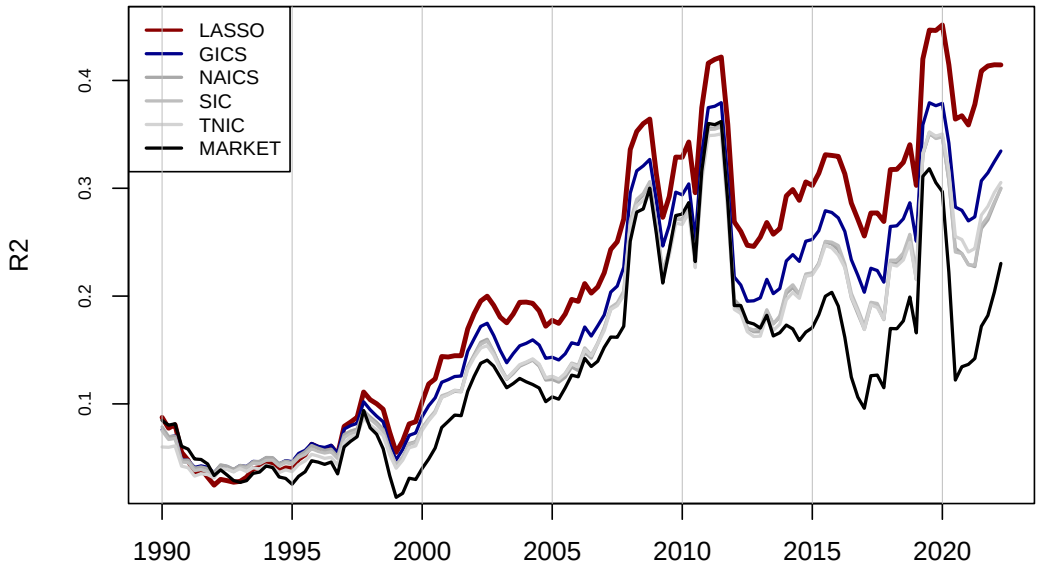


Figure 1: Average out-of-sample R^2 over time.

firm to other firms in the same industry category seems less appropriate in world where industry boundaries appear more and more blurred. While in the past century, an automobile company could reasonably be benchmarked against other automobile companies, it is much less clear if today, a company like Tesla that, besides electric vehicles, also develops self-driving software, energy storage solutions, and solar panels, should be compared to "industry peers" such as General Motors or Ford. More generally, the use of modern technologies such as Artificial Intelligence or Big Data has transcended traditional industry boundaries, thereby rendering traditional industry peers less appropriate for benchmarking.

2 Theory

2.1 Model classes

There are n firms. Each firm, indexed by i, generates monetary outcome x_i according to

$$x_i = \gamma_i + \delta_i' z + \varepsilon_i, \quad i = 1, ..., n. \tag{1}$$

I will refer to the entities indexed by i as firms, but they may represent any entity that generates measurable outcomes: an individual agent, a firm, an investment project, an investment fund, etc. I will refer to x_i as the outcome or performance of firm i. In most applications of interest, x_i represents a monetary outcome such as the profits, sales, or stock return of a firm. But the framework is, in principle, more general. It can be applied to any measurable characteristic that is determined according to eq. (1). I will refer to firm i as the *focal firm* and to all other firms $j \neq i$ as *potential peers*.

In eq.(1), z is a vector of exogenous, common factors affecting the performance of firm i; δ_i is the vector of "exposures" of firm i to these factors. The factors may represent technology or demand shocks that are particular to the industry or product market the focal firm operates in; they may represent currency or commodity price shocks that the firm is exposed to, etc. The factors are 'common' in the sense that they do not only affect firm i's performance, but all firms in the economy, to possibly different extents. The ε_i's are idiosyncratic, independent risks. Let $z \sim N(0, \sigma_z^2)$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_i^2)$.

Coefficient γ_i is the performance component that represents the manager's unique contribution to the firm. The contribution can be the effort the manager exerts or the skill s/he has. Performance x_i is observable, but factors z, exposures δ_i and contributions γ_i are not. Because common shocks are unobserved, an optimally weighted portfolio of peer firms will turn out to be best proxy for the common component of firm i's performance, $\delta_i' z$.

2.2 Optimal peers

I analyze the optimal choice of peers for two different classes of models. The first is the standard moral hazard model of Holmström (1982): The principle's objective is to maximize output x_i less wage costs, and γ_i represents the unobservable equilibrium action of an agent. Hence, the problem for the principal is one of inducing the agent to provide optimal amounts of inputs given that actions cannot be contracted for directly.

If, as in eq. (1), firms' outcomes are related via exposure to common risks, the optimal incentive contract depends on x_i and an aggregate measure of

peer performance, $\overline{x}_{-i} = \sum_{j \neq i} w_j x_j$. Holmström (1982) proves that there exists a weighting scheme $\{w_j\}$ for peer portfolio $\overline{x}_{-i}$ such that the pair $(x_i, \overline{x}_{-i})$ is a sufficient statistic with respect to the agent's equilibrium action. As a consequence, optimal incentive contracts depend on x_i and $\overline{x}_{-i}$ alone. The portfolio weights are given by $w_j = \tau_j / \overline{\tau}$, where $\tau_j = \delta_j^2 / \sigma_j^2$ and $\overline{\tau} = \sum_j \tau_j$ (Theorem 8).

The solution is intuitive. The optimal portfolio weights are proportional to the precision of each x_j as a signal for common risk z. A peer whose idiosyncratic risk is low should receive a larger weight in the peer portfolio as its performance is more informative about z. For the same reason, a peer whose exposure to the common shock is greater should receive a larger weight.

The second is a class of models where γ_i represents some fixed characteristic of the agent, say ability. Here, the problem of the principal is one of learning about the agent's ability and making optimal decisions based on that inference, such as promoting or replacing her (see, e.g. Hermalin and Weisbach, 1998).

For the second class of models, Blackwell's Informativeness Theorem (Blackwell, 1951) provides a similar optimality criterion. Blackwell's Theorem states that if there are several informative signals about the relevant outcomes, and these signals differ in the amount of noise they contain, then a Bayesian agent will prefer the least noisy signal for any possible decision problem.

The following proposition proves that (i) there is a unique solution to the optimal peer portfolio for both classes of models, and (ii) the optimal portfolio is given by the solution to a minimum mean squared error problem.

Proposition 1. *In the linear-normal performance model of eq.(1), the optimal weighting scheme for peer portfolio* $\overline{x}_{-i} = \sum_{j \neq i} w_j x_j$ *is given by the solution to the minimum mean squared error (MMSE) problem*

$$\min_b \mathbb{E}[(x_i - b_0 - \sum_{j \neq i}^{n} b_j x_j)^2]. \tag{2}$$

In other words, the optimal weights are given by the population estimand of an Ordinary Least Squares (OLS) regression of focal firm outcome on the outcomes of all potential peers:

$$x_i = b_0 + \sum_{j \neq i}^{n} b_j x_j + \nu_i. \tag{3}$$

The proper, optimal portfolio weights are scaled OLS coefficients:

$$w_j^* = \frac{b_j^*}{\sum_{j \neq i} b_j^*}.$$

The optimal peer portfolio is the most precise proxy for the common shock, i.e.,

$$w^* = \arg\min_{w} \mathbb{V}\text{ar}\left[\sum_{j \neq i} w_j x_j\right] \tag{4}$$

Proof. In the Appendix. □

It may not be surprising that the solutions of the two model classes coincide. The subtle difference is that, in contracting models, no inference occurs. The optimal contract is specified such that optimal effort is provided. In the second class of models, inference occurs. The solutions coincide, because the optimal contract depends on the same statistic that enables optimal inference about γ_i.

2.3 Optimal implementable peers

Proposition 1 suggests that one can find the optimal peer portfolio for a focal firm by simply regressing focal firm performance on the performance outcomes of all potential peers. *In the population*, the coefficients provide optimal portfolio weights as shown in Section 2.2.

However, if the weights are estimated from a sample, estimation error will drive a wedge between the theoretical, optimal weights and their empirical counterparts. If the sample size is large relative to the number of potential

peers (i.e. regressors), OLS coefficients yield accurate estimates of the optimal weights. In practice, the number of potential peers is typically large relative to the number of observations. For example, if we wanted to estimate optimal weights using four years of daily stock return data, the sample would have about $4 \times 250 = 1,000$ observations. If we define the universe of potential peers as all U.S. stocks, the number of regressors is several times greater than the number of observations. (There are about 5,000 tradable, publicly listed stocks in the U.S.) With less frequently observed data such as quarterly accounting earnings, the ratio of regressors to observations is even greater.

We define model *complexity*, C, as the ratio of the number of regressors to the number of observations, $C = P/T$. In general, for a fixed sample size T, OLS estimates become less precise as the number of regressors P increases. This leads to a trade-off: on the one hand, more regressors can lead to a better empirical approximation of the (unknown) true model. On the other hand, increasing estimation error deteriorates the empirical approximation. In sample, the MSE mechanically declines with the number of regressors since there is more flexibility to fit the data. But as the number of regressors gets large, OLS increasingly fits noise – or coincidental, non-structural associations – rather than signal to the data.

Out-of-sample, the MSE has a hockey-stick shape: as model complexity increases, the OOS MSE first declines but then rises. As P approaches T, the OOS MSE explodes. When $P = T$, the OLS system of equations is exactly determined. The estimation sample is fit perfectly, but the OOS fit is worst. When $P > T$, OLS cannot be applied as it does not yield a unique solution to the system of equations. The system is underdetermined. There are more parameters than equations, and infinitely many combinations of parameters perfectly fit the sample.

Machine learning (ML) addresses the problems of OLS in settings of high model complexity. ML algorithms are designed to make optimal out-of-sample predictions, i.e. to minimize MSE out of sample rather than in sample.

For the linear, additive regression model of eq. (3) a number of ma-

chine learning algorithms have been developed. These methods fall into three classes: subset selection, dimension reduction, and shrinkage.

Subset selection. Subset selection methods fit separate least squares regressions to subsets of the P predictors and select the best fitting model. These methods are conceptually appealing but too computationally demanding when the number of potential peers is large: there are 2^P models that involve P potential peers. So, even with just 100 potential peers, more than $1.26 \cdot 10^{30}$ regressions would need to be run – an impossible task. There are computationally more efficient alternatives such as forward and backward stepwise selection. But these are not guaranteed to find the best possible model.

Dimension reduction. Dimension reduction methods transform the original regressors to only a few linear combinations. In a second step, a linear regression is fit to a model featuring only the transformed regressors. One way to construct transformed regressors is via principal component analysis. Dimension reduction methods are not ideally suited for the problem of constructing optimal peer portfolios for two reasons: Dimension reduction methods are not feature selection methods. Since each principal component is a linear combination of *all* regressors, these methods do not reduce the number of peers that end up in the peer portfolio. Every potential peer is part of the peer portfolio albeit with negligible weight. Second, the transformed regressors are statistical constructs and are thus difficult to interpret. They do not correspond to particular peers.

Shrinkage. For the problem at hand, regularization or shrinkage methods seem to be best suited. Shrinkage methods tackle the problem of estimation error by shrinking coefficients towards zero. The two most popular such methods are ridge regression and the lasso (an acronym for *least absolute shrinkage and selection operator*, see Tibshirani (1996)). The lasso achieves shrinkage by adding a penalty term for large absolute coefficients to the standard OLS objective:

$$\min_b \sum_{t=1}^{T} \left(x_{it} - b_0 - \sum_{j \neq i}^{n} b_j x_{jt} \right)^2 + \lambda \sum_{j \neq i}^{n} |b_j|. \tag{5}$$

Shrinkage methods improve the OOS fit by optimizing the bias-variance tradeoff in the OOS MSE. For any predictor $\hat{f}(x)$, the mean squared prediction error on a test sample can be expressed as

$$\mathbb{E}[(y - \hat{f}(x))^2] = \mathbb{V}\text{ar}[\hat{f}(x)] + (Bias[\hat{f}(x)])^2 + \mathbb{V}\text{ar}[\varepsilon].$$

From this expression, we can see why OLS does not make optimal OOS forecasts when complexity is high. While OLS is unbiased for the linear model of eq.(3), the variance of the coefficient estimates, and hence the variance of predictions, increases with model complexity. Said differently, when complexity is high, OLS typically produces some very large coefficients which fit non-structural associations in the specific training sample. These will produce large prediction errors in test data. By penalizing large coefficients, shrinkage methods optimize the bias-variance tradeoff and produce better OOS predictions.

Lasso vs. ridge regression. Due to its particular penalty term, an ℓ_1 norm, the lasso sets some coefficients exactly to zero. This makes it also a feature selection method. In the applications we consider in Section 3 below, the lasso will typically set about 98% of potential peer coefficients to zero, i.e. it will select only about 60 out of 3,000 potential peers. The feature selection property makes lasso-optimal peer portfolios easy to interpret. They are also close in size to traditional industry groups such as 4-digit SIC or 6-digit GICS groups. The lasso is therefore our preferred method for the problem at hand.

The ridge regression does not produce sparse solutions. All coefficient are shrunk towards zero, but none is set to exactly zero. Similar to dimension reduction methods, ridge results are hard to interpret, because all potential peers end up in the optimal peer portfolio.

Tuning. The extent of shrinkage is determined by the tuning parameter, λ. When $\lambda = 0$, the objective collapses to OLS, and OOS predictions are poor due to high variance. When λ approaches infinity, all coefficient are shrunk to zero, and OOS predictions are poor because of large bias. The optimal tuning parameter minimizes the OOS MSE. The most popular method for choosing the tuning parameter is by cross-validation. This method splits the

training sample into K equal sized parts or *folds*. For each fold k, the model is fit on the other $K-1$ parts of the data, and the prediction error of the fitted model is calculated on the kth fold. The average prediction error from all K folds forms an estimate of the expected OOS MSE. The optimal tuning parameter is chosen as the value of λ that minimizes the estimate of the OOS MSE over a grid of λ's.

3 Application

3.1 Data and implementation

To assess its empirical performance, I apply the lasso, i.e. objective (5), to a large sample of U.S. public companies. For each calendar quarter between 1990Q1 and 2022Q4, I construct an initial sample consisting of all public U.S. firms that are present in CRSP and Compustat with non-missing daily stock return data over the four years before, and the year following, the end of the quarter. I exclude firms with non-positive sales and assets of less than \$1 million. A typical quarter contains between 2,500 and 3,500 firms that meet these requirements.

I define the outcome variable x_{it} as the daily stock return of firm i net of the risk free rate on date t: $x_{it} = r_{it} - r_{ft}$. I use daily stock returns as the outcome variable, because they provide a large training sample relative to other candidate outcome variables like sales or earnings.[2]

I solve the lasso objective (eq. 5) for each quarter q and stock i using all other stocks $j \neq i$ as potential peers, i.e. using the returns of all other stocks as regressors. The estimation sample comprises the four years up until the end of quarter q. This means that each optimization uses about 1,000 observations to estimate between 2,500 and 3,500 parameters. To select the optimal tuning parameter, I follow the standard procedure of K-fold cross-validation with $K = 10$ folds (Hastie et al., 2009). The year following quarter

[2] The risk free rate is from Ken French's data library. Excess stock returns are winsorized at -2 and +4 times the contemporaneous market return (Welch, 2022). Results are unchanged with a symmetric top/bottom 0.5% winsorization for each trading day.

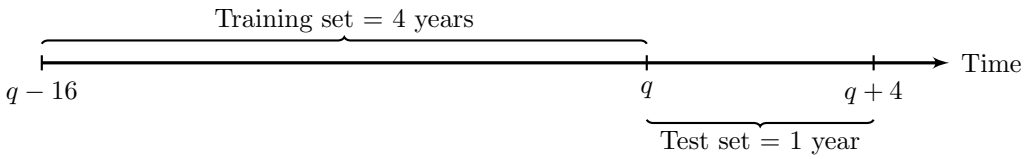


Figure 2: Training and test set windows.

q is the test set, the sample used to assess out-of-sample performance. See Figure 2 for an illustration of the timing.

3.2 Identification

The lasso identifies relevant peers from contemporaneous stock return correlation between the focal firm and potential peers. If the model of eq. (1) is a reasonable approximation of reality, the lasso will select peers that have large exposures to the common shocks that most affect the focal firm. For example, if the focal firm's stock price strongly co-moves with the oil price, then the lasso will pick up peers that have similarly large exposures to the oil price.

An interesting feature of peer identification based on the lasso objective (5) is that the exposures of focal and potential peer to a particular common shock have to be large in *absolute value* for a potential peer to be selected. If the focal firm (e.g. an oil and gas corporation) has large positive exposure to the oil price, the lasso may select peer firms that have large positive (e.g. a competing oil and gas corporation) as well as large *negative* exposure (e.g. a cosmetics company) to the oil price. The sign of the exposure is irrelevant. In the first case, the lasso will estimate a positive weight; in the second case, a negative weight. A positive weight is equivalent to the peer firm being a "long position" in the peer portfolio, a negative weight implies a "short position". Both long and short positions are helpful for measuring the common shock and therefore for benchmarking and prediction.

It is worth noting that, in the application considered here, the common shocks do not necessarily need to be 'fundamental'. Because we use stock return as the outcome variable, the lasso picks up co-movement between focal

and peer firms that is caused by any shock that drives stock prices. Most of these shocks will likely be fundamental, e.g. commodity price, currency, technology, demand, or interest rate risks. But the lasso may also pick up co-movement that is driven by sentiment shocks. If, for instance, two stocks are correlated, because they are heavily traded by a coordinated group of speculative investors, the lasso may classify them as peers even if their fundamentals are not related (cf. the speculative period in AMC and Gamestop stocks in 2021).

Whether one wants non-fundamental shocks to affect peer selection depends on the purpose of the peer group. For the purpose of relative performance evaluation in its purest form, all factors outside the manager's control should be filtered out. Thus, firms that are correlated with the focal firm via similar exposure to sentiment should be part of the peer group. After all, would a board want to reward its CEO for riding a sentiment wave? If one does not want sentiment shocks to affect peer selection, the methodology is flexible enough to accommodate that: one can include a variable that measures sentiment as an additional regressor to the lasso objective (5).[3] With such a specification, only components of performance that are orthogonal to sentiment will determine peer selection.[4]

Stock prices incorporate shocks to current earnings as well as expectations of future earnings. This implies that the lasso will tend to select peers whose current profits or future expected profits are generated in similar businesses as those of the focal firm. These businesses are not necessarily those that currently command the most capital investment, fixed assets or the largest workforce. For example, if Amazon Inc. generates most of its profits from its cloud computing services and no profits from its online book store, the lasso will tend to select peers that are in the cloud computing business rather than book stores.

Similarly, for firms whose stock price is mainly driven by earnings that are expected to materialize years in the future, lasso peers may be very different

[3] E.g., one may think of the price of Bitcoin as a good proxy for non-fundamental, speculative sentiment.

[4] Modern implementations of the lasso allow one to force the inclusion of particular regressors.

from firms that have similar assets in place as the focal firm. For example, if investors regard Tesla Inc. as an A.I. company, its stock price may move similarly to that of Google, Microsoft or Nvidia – hence, the lasso will select companies such as these – even though current earnings mostly come from their automobile manufacturing business.

3.3 Examples

It is instructive to study a few examples to gain intuition for how the lasso identifies peers. I construct optimal peer portfolios for three companies that are categorized as part of the automotive industry by all traditional industry classifications: Ford Motor Inc., General Motors Company and Tesla Inc.[5] The training sample is 01/01/2018-12/31/2021. The details of the implementation are as described in Section 3.1. For the three companies, Table 1 reports the ten most highly weighted lasso peers (column 2), their respective portfolio weight (column 2), the NAICS 3-digit code (column 3) and the NAICS industry description (column 4).

General Motors Company. The selected peers for GM are predominantly from the transportation equipment and other manufacturing industries. By far the most important peer of GM is Ford. Lear, Magna and Aptiv are all automotive suppliers. This makes sense. What seems puzzling at first sight, is that three financial companies also feature among the top ten peers. This is explained by the importance of GM's financing arm *GM Financial*. For example, in fiscal year 2021, about 10% of GM's revenue and about one third of its earnings came from its financing division.

Ford Motor Company. A mirror image emerges for the peers of Ford: (i) its most important peer is GM; (ii) several automotive suppliers feature among its top ten peers as well as (iii) financial companies. In fact, financial companies make up an even larger fraction of Ford's peer portfolio. The aggregate weight is 16%, whereas it is 13% for GM. This is strikingly consistent with the fact that an even greater fraction of Ford's profits comes from its

[5] All three companies have primary SIC 3711 (Motor vehicles and car bodies), primary NAICS 336110 (Automobile and Light Duty Motor Vehicle Manufacturing), and GICS 251020 (Automobiles),

financing unit, *Ford Credit*. In 2021, 42% of Ford's earnings were attributed to its financing unit.

Tesla Inc. For Tesla, the picture looks very different. No automobile manufacturer is selected as a peer; neither GM nor Ford are not seen as Tesla's peers by the lasso. Only one automotive supplier features among the top ten peers. Instead, computer software and hardware companies dominate the peer group. This is a reflection of the fact that Tesla investors regard the company rather as a software or computer company than a car company. It is also noteworthy that Tesla's peer portfolio is much less concentrated than that of GM or Ford. The lasso does not identify any dominant peer of Tesla. Apparently, the stock market does not view Tesla as having close competitors.

3.4 Out-of-sample performance

In this section I document the out-of-sample performance of optimal peers and compare it with alternative peer portfolio choices. It is common in academia and practice to use existing industry groups as peer groups. In academia, the most commonly used industry classifications are SIC and NAICS. More recently, Hoberg and Phillips' TNIC (Hoberg and Phillips, 2016) have also become popular. In business and investment practice, Standard & Poor's GICS, Bloomberg's BICS or FactSet's RBICS are widely used. I compare optimal peer portfolios with peer groups based on SIC, NAICS, TNIC and GICS.

For each calendar quarter q between 1990Q1 and 2022Q4 and each focal firm i, I solve the lasso objective (eq. 5) using excess daily stock returns over the preceding four years as the outcome variable. The estimation yields a set of coefficients $(\hat{b}_0, \hat{b}_j, j \neq i)$ for each quarter and firm. I rescale the estimated lasso coefficients to proper portfolio weights: $\hat{w}_j \coloneqq \hat{b}_j / \sum_{j \neq i} \hat{b}_j$, and define $\hat{c}_0^{Lasso} \coloneqq \hat{b}_0$, $\hat{c}_1^{Lasso} \coloneqq \sum_{j \neq i} \hat{b}_j$, and $r_{it}^{Lasso} \coloneqq \sum_{j \neq i}^{n} \hat{w}_j r_{jt}$.

The model

$$r_{it} = \hat{c}_0^{Lasso} + \hat{c}_1^{Lasso} r_{it}^{Lasso} + \nu_{it} \tag{6}$$

generates the same predictions as the unscaled, original model $r_{it} = \hat{b}_0 + \sum_{j\neq i}^{n} \hat{b}_j r_{jt} + \nu_{it}$, but is more easily interpreted as a simple regression with intercept $\hat{c}_0^{Lasso}$ and exposure $\hat{c}_1^{Lasso}$ to the the weighted peer portfolio return $\sum_{j\neq i}^{n} \hat{w}_j x_{jt}$ as the only independent variable.

To obtain the training set coefficients for traditional industry groups, I construct daily industry peer returns, r_{it}^{bm}, as the equally and value-weighted returns of firms in the same SIC, NAICS, TNIC, or GICS group as the focal firm. I also add the CRSP equally and value-weighted indices as broader benchmarks. I then run the regression $r_{it} = c_0^{bm} + c_1^{bm} r_{it}^{bm} + e_{it}$, on the training set where $bm \in \{SIC, NAICS, TNIC, GICS, MKT\}$. This yields a set of coefficients $(\hat{c}_0^{bm}, \hat{c}_1^{bm})$ for each quarter and firm – the counterparts of the lasso-derived coefficients $(\hat{c}_0^{Lasso}, \hat{c}_1^{Lasso})$.

I measure the out-of-sample fit by the R^2 of the regression

$$r_{it} = \hat{c}_0^{bm} + \hat{c}_1^{bm} r_{it}^{bm} + e_{it}, \tag{7}$$

using daily excess stock returns over the year following focal quarter q, and keeping fixed the training set coefficients $(\hat{c}_0^{bm}, \hat{c}_1^{bm})$.

Figure 1 depicts the out-of-sample R^2 for each benchmark over time. Each point on a line represents the average R^2 from regression (7) over all focal firms $i = 1, ..., n$ for a given quarter q. For benckmarks SIC, NAICS, TNIC, GICS, and MKT, the figure shows the performance of value weighted portfolios. Equally weighted portfolios perform unequivocally worse than value weighted ones.

Figure 1 shows that optimal peers dominate traditional peer groups in terms of out-of-sample R^2 in almost every period except at the very beginning of the sample period. The advantage that lasso peers display grows over time and is very substantial at the of the sample period. A likely reason for the increasing lasso advantage is that stock prices have become more informative over time. Because the lasso relies on stock prices accurately reflecting the relevant shocks, it becomes better over time at selecting peers.

Several other aspects of the Figure 1 stand out: First, the best fitting traditional peer group is based on GICS. This is likely due to the capital

market oriented approach that S&P takes to determine GICS. Second, three industry classifications that are popular in academia all show very similar out-of-sample performance: the lines for SIC, NAICS and TNIC peers are almost on top of each other. Third, lasso, SIC, NAICS, TNIC, GICS all perform better out-of-sample than the market portfolio. Fourth, there seems to be a long-term, possibly secular, trend in the explanatory power of peer groups for individual stocks. A higher R^2 from regression (7) is equivalent to low idiosyncratic risk. Thus, Figure 1 shows a secular decline in idiosyncratic risk. Importantly though, this trend is strongest for optimal peers, and only for optimal peers it seems to continue after the year 2010. Understanding the economic reasons for these patterns seems to be a fruitful avenue of research.

Table 2 contains detailed statistics of out-of-sample fit. The entries in columns 1-3 at the top of the table are means of out-of-sample R^2 over different time periods and samples. Columns 1 is based on the largest sample of all Compustat firms in the period 1990Q1-2022Q4. Column 3 is based on the same time period but the smaller cross section of S&P 500 firms. Column 2 restricts to Execucomp firms where the sample starts in 1992Q1. We can see that out-of-sample fit increases for all methods as we go from the broad Compustat universe to smaller samples of larger firms. This is not surprising: there is less idiosyncratic risk in larger, more transparent, more broadly owned stocks.

The table also reports R^2 statistics for equally weighted peer portfolios for the SIC, NAICS, TNIC, GICS, and MKT benchmarks. Equally weighted peer portfolios feature unambiguously worse out-of-sample performance than value weights.

3.5 Peer group size

Given the strong out-of-sample performance of optimal peers, it is instructive to examine with how many peers this performance is achieved. Does the lasso construct peer groups that are similar in size to standard industry groups?

Table 3 reports descriptive statistics on size of lasso peer groups and traditional industry groups. It turns out that lasso peer groups are very

similar in size to existing industry groups on average. All methods roughly produce peer groups of an average size 60 to 80. One difference that stands out, however, is that lasso peer group size shows much less variation than traditional industry groups; in particular, they have a much fewer outliers in the right tail of the distribution.

This is also visible in the histograms of Figure 3. The top panel shows the peer group size histograms of lasso peers and GICS-8 industry groups for all focal firms with sufficient data in CRSP and Compustat. The bottom panel shows the same histogram for focal firms in the S&P500 index.

Lasso peer groups have a bell-shaped size distribution centered around 60-80 (with the exception of being censored at zero). GICS have a very different size distribution with many small and a few very large industry groups.[6]

In the full Compustat sample, the lasso does not find any peer for about 8 percent of firms (see the leftmost red bar in the top panel). These are mainly cases of very illiquid, infrequently traded, microcap stocks whose daily stock prices contain little or no information. Once we move to the larger S&P500 stocks, the lasso always finds a reasonably sized peer group.

3.6 Stability of peer weights over time

One may wonder how stable lasso peer groups are over time. This question is of particular interest because SIC, NAICS, and GICS classifications are sometimes criticized for being too static. These industry definitions are updated infrequently.[7] As a result, firms entering new industries might not be recognized as such by these classifications. Additionally, there appears to be a certain inertia among the entities assigning industry codes (e.g., government agencies, data vendors) in updating these codes over time.

[6] For ease of exposition, Figure 3 only compares lasso peer groups with GICS-8. The comparisons with SIC-3, NAICS-4, and TNIC-3 are similar and are reported in the Internet Appendix.

[7] SIC industry definition updates have been discontinued since 1987. NAICS definitions, which were intended to replace the SIC system, are only updated every five years (1997, 2002, 2007, 2012, 2017, 2022). GICS industry definitions underwent significant updates only in 2016, 2018, and 2023 (since its inception in 1999).

The machine learning-based approach proposed in this paper does not rely on official industry definitions nor is it subject to human inertia. When a firm enters a new industry, lasso weights reflect this as soon as stock prices capture the opportunities and risks of the new industry for all firms operating within it. Consequently, the lasso is potentially more flexible and dynamic than traditional classification systems.

Table 4 reports statistics on the stability of peer groups over time. The numbers in the first row show the average absolute change in lasso weights over three different time horizons: one quarter, four quarters, and 16 quarters. The number in the first column and first row indicates that the weight of a lasso peer changes by two percentage points from one quarter to the next, on average. As the horizon increases, the weight changes also increase. At a horizon of 16 quarters, the training sets (which span four years) are non-overlapping, so there is no mechanical stickiness of the peer weights. Nevertheless, the lasso weights are reasonably stable, with an average change of three percentage points.

Figure 4 illustrates the average weight changes over time. Each point on the line represents the cross-sectional average of the absolute lasso weight change over the respective horizon at that point in time. The lasso weights were the most dynamic in the mid-1990s and have become more stable ever since. The peak in weight dynamics coincides with the trough in explanatory power seen in Figure 1. This is not surprising: when the lasso struggles to find peers with good out-of-sample explanatory power, there is little structural similarity between firms, as revealed by stock prices. Therefore, the lasso changes the peer weights more dynamically as the training data changes.

The middle section of Table 4 compares the probabilities of dropping out of the peer group in quarter $q + h$, conditional on being a peer in quarter q. For SIC, NAICS, and GICS, the one-year drop-out probability is 7-8 percent. The four-year drop-out probability is slightly less than twice as high for SIC and NAICS and about 2.5 times as high for GICS. Apparently, GICS industry groups are somewhat more dynamic than SIC and NAICS. Hoberg and Phillips' TNIC groups are significantly more dynamic than SIC, NAICS, and GICS. For example, the one-year drop-out rate is 19 percent,

which is two to three times higher than for SIC and NAICS.

The last row in the middle section of Table 4 reports the drop-out probabilities for lasso peers. These are substantially higher than those of all existing industry classifications. For instance, there is a 56 percent chance for a current lasso peer to drop out within one year, and a 66 percent chance to drop out within four years. However, as the bottom two rows of the table show, the high drop-out rates are mostly driven by the peers with the smallest weights. Drop-out rates of the most important peers are much lower. For example, the probability of a current lasso peer with a top-5 weight to drop out within one year is only three percent. This indicates that the lasso method successfully identifies structurally similar firms over time.

4 Conclusions

This paper proposes a theoretically founded, implementable method to construct benchmarks via machine learning. The benchmarks are theoretically optimal in a broad range of economic models and applications.

An application to a large sample of U.S. public firms reveals that optimal benchmarks show superior out-of-sample performance relative to the most commonly used benchmarks. Their advantage over traditional benchmarks sharply increases with market capitalization, trading volume, (negative) bid-ask spreads, and other measures of stock price informativeness. As a consequence, their advantage over traditional benchmarks has steadily increased over time.

References

Blackwell, D., 1951. Comparison of experiments, in: Neyman, J. (Ed.), Second Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. volume 1, pp. 93–102.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference and prediction. 2 ed., Springer. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Hermalin, B., Weisbach, M., 1998. Endogenously chosen boards of directors and their monitoring of the CEO. American Economic Review 88, 96–118.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. The Journal of Political Economy 124, 74—-91.
Holmström, B., 1982. Moral hazard in teams. The Bell Journal of Economics 13, 324–340.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.
Welch, I., 2022. Simply Better Market Betas. Critical Finance Review 11, 37–64.

Proof of Proposition 1

With slight abuse of notation, normalize performance $x_j = x_j/\delta_j$, so that $\gamma_j = \gamma_j/\delta_j$, $\varepsilon_j = \varepsilon_j/\delta_j$, and $\sigma_j^2 = \sigma_j^2/\delta_j^2$. Then

$$
\begin{aligned}
&\arg\min_{b_0,b} \mathbb{E}\left[(x_i - b_0 - \sum_j b_j x_j)^2\right] \\
=&\arg\min_{b_0,b} \mathbb{E}\left[[\gamma_i + z + \varepsilon_i - b_0 - \sum_j b_j(\gamma_j + z + \varepsilon_j)]^2\right] \\
=&\arg\min_{b_0,b} \mathbb{E}\left[[z(1 - \sum_j b_j) + (\varepsilon_i - \sum_j b_j\varepsilon_j) + (\gamma_i - \sum_j b_j\gamma_j) - b_0]^2\right] \\
=&\arg\min_{b_0,b} \mathbb{E}\left[z^2(1 - \sum_j b_j)^2 + (\varepsilon_i - \sum_j b_j\varepsilon_j)^2 + (\gamma_i - \sum_j b_j\gamma_j - b_0)^2\right] \\
=&\arg\min_{b_0,b} \mathbb{E}\left[z^2(1 - \sum_j b_j)^2 + (\varepsilon_i^2 - 2\varepsilon_i\sum_j b_j\varepsilon_j + [\sum_j b_j\varepsilon_j]^2) + (\gamma_i - \sum_j b_j\gamma_j - b_0)^2\right] \\
=&\arg\min_{b_0,b} \mathbb{E}\left[z^2(1 - \sum_j b_j)^2\right] + \mathbb{E}\left[\varepsilon_i^2\right] - \mathbb{E}\left[2\varepsilon_i\sum_j b_j\varepsilon_j\right] + \mathbb{E}\left[[\sum_j b_j\varepsilon_j]^2\right] + (\gamma_i - \sum_j b_j\gamma_j - b_0)^2 \\
=&\arg\min_{b_0,b}\ \sigma_z^2(1 - \sum_j b_j)^2 + \sigma_i^2 + \sum_j b_j^2\sigma_j^2 + (\gamma_i - \sum_j b_j\gamma_j - b_0)^2 \\
=&\arg\min_{b_0,b}\ \sigma_z^2(1 - \sum_j b_j)^2 + \sum_j b_j^2\sigma_j^2 + (\gamma_i - \sum_j b_j\gamma_j - b_0)^2
\end{aligned}
$$

Rescaling the coefficients as $\hat{b}_j = b_j/\bar{b}$ where $\bar{b} = \sum_j b_j$, we get the constrained optimization problem

$$
\begin{aligned}
&\arg\min_{b_0,\hat{b},\bar{b}}\ \sigma_z^2(1 - \bar{b})^2 + \bar{b}^2\sum_j \hat{b}_j^2\sigma_j^2 + (\gamma_i - \bar{b}\sum_j \hat{b}_j\gamma_j - b_0)^2 \qquad (8) \\
&s.t.\ \sum_j \hat{b}_j = 1.
\end{aligned}
$$

The Lagrangian of this problem is given by

$$
\mathcal{L}(b_0, \hat{b}, \bar{b}, \lambda) = \sigma_z^2(1 - \bar{b})^2 + \bar{b}^2\sum_j \hat{b}_j^2\sigma_j^2 + (\gamma_i - \bar{b}\sum_j \hat{b}_j\gamma_j - b_0)^2 + \lambda(\sum_j \hat{b}_j - 1).
$$

The FOCs w.r.t. b_j, b_0, and $\bar{b}$ are

$$\begin{aligned} 2\bar{b}^2\hat{b}_j\sigma_j^2 - 2(\gamma_i - \bar{b}\sum \hat{b}_j\gamma_j - b_0)\bar{b}\gamma_j + \lambda = 0 \quad \forall j \\ \gamma_i - \bar{b}\sum_j \hat{b}_j\gamma_j - b_0 = 0 \\ -2\sigma_z^2(1-\bar{b}) + 2\bar{b}\sum_j \hat{b}_j^2\sigma_j^2 - 2(\gamma_i - \bar{b}\sum_j \hat{b}_j\gamma_j - b_0)(\sum_j \hat{b}_j\gamma_j) = 0 \end{aligned}$$

The second FOC gives

$$b_0 = \gamma_i - \bar{b}\sum_j \hat{b}_j\gamma_j,$$

and simplifies the first FOC to

$$2\bar{b}^2\hat{b}_j\sigma_j^2 + \lambda = 0 \quad \forall j$$

or

$$\hat{b}_j = \frac{-\lambda}{2\bar{b}^2}\frac{1}{\sigma_j^2} \quad \forall j.$$

Summing up the $\hat{b}_j$'s and using the constraint gives

$$\sum_j \hat{b}_j = \frac{-\lambda}{2\bar{b}^2}\sum_j \frac{1}{\sigma_j^2} = 1$$

Hence, $-\frac{\lambda}{2\bar{b}^2} = (\sum_j \frac{1}{\sigma_j^2})^{-1}$, and

$$\hat{b}_j = \frac{\tau_j}{\bar{\tau}}.$$

Use the third FOC to solve for $\bar{b}$. Again, the second FOC simplifies the third to

$$-2\sigma_z^2(1-\bar{b}) + 2\bar{b}\sum_j \hat{b}_j^2\sigma_j^2 = 0$$
$$\sigma_z^2(1-\bar{b}) = \bar{b}\sum_j \hat{b}_j^2\sigma_j^2$$

which gives

$$\bar{b} = \frac{\sigma_z^2}{\sum_j \hat{b}_j^2\sigma_j^2 + \sigma_z^2}$$

One way to understand Proposition 1 is by viewing the peer portfolio as a hedge portfolio for x_i, and the term $x_i - b_0 - \sum_j b_j x_j$ as the hedged outcome of firm i. By minimizing the mean squared hedged outcome, OLS constructs the optimal, i.e. the minimum variance, hedge portfolio. We can see from the constrained minimization problem of eq. (8) that OLS achieves three things simultaneously:

(i) it chooses the intercept b_0 such that level differences between the focal firm and peer firms do not affect the prediction error (the optimal intercept sets the third term to zero);

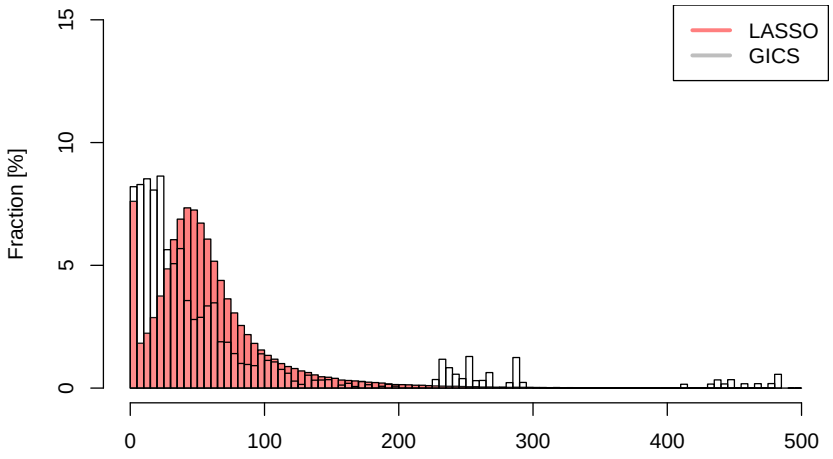
(ii) it chooses $\bar{b}$ to find the optimal balance between exposure to the common risk and the idiosyncratic risks contained in the peer portfolio. Since the peer portfolio contains non-zero risk, $\sum_j \hat{b}_j^2\sigma_j^2 > 0$, it is not generally optimal to have full exposure ($\bar{b} = 1$) to the peer portfolio and zero exposure to the common risk.

(iii) it chooses the portfolio weights of the peer firms, the $\hat{b}_j$'s, such that the peer portfolio variance is minimized, i.e.

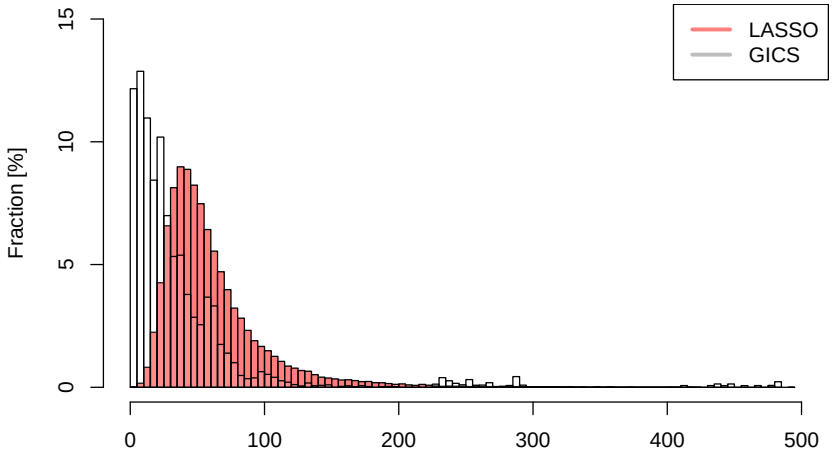
$$\hat{b}^* = \arg\min_{\hat{b}} \mathbb{V}\text{ar}\left[\sum_{j\neq i} \hat{b}_j x_j\right] \tag{9}$$

This implies that the solution to the MMSE problem of eq. (2) gives the most precise proxy for the common risk.

Figures and Tables



(a) Compustat firms.



(b) S&P firms.

Figure 3: Histograms of peer group size.

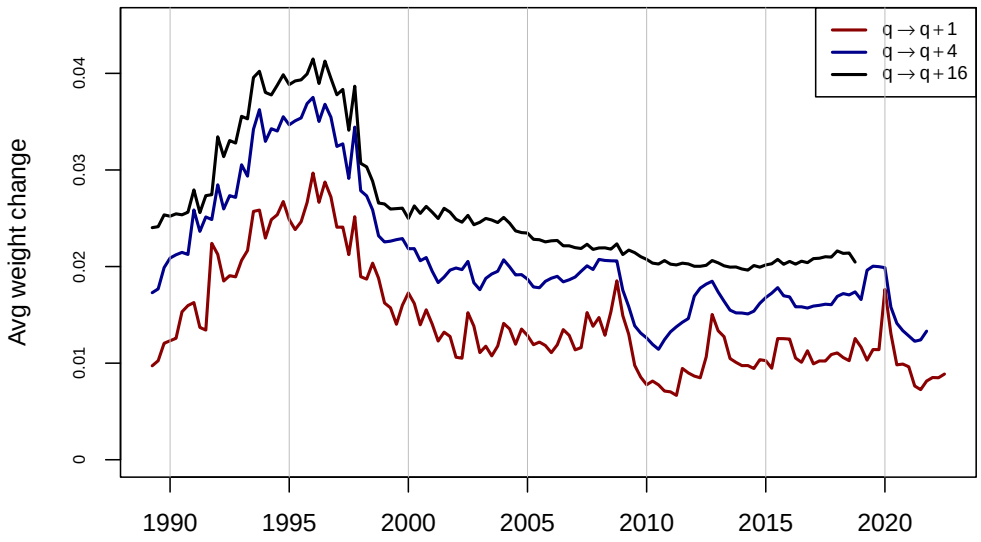


Figure 4: Average absolute change of peer weight over time.

Panel A: General Motors Company			
Weight	Peer firm	NAICS	NAICS Description
0.40	FORD MOTOR CO DEL	336	Transportation Equipment Manufacturing
0.11	LEAR CORP	336	Transportation Equipment Manufacturing
0.10	MAGNA INTERNATIONAL INC	336	Transportation Equipment Manufacturing
0.08	APTIV PLC	336	Transportation Equipment Manufacturing
0.06	POPULAR INC	522	Credit Intermediation and Related Activities
0.05	CITIZENS FINANCIAL GROUP INC	522	Credit Intermediation and Related Activities
0.05	AUTOLIV INC	336	Transportation Equipment Manufacturing
0.04	PARKER HANNIFIN CORP	332	Fabricated Metal Product Manufacturing
0.03	CELANESE CORP DEL	325	Chemical Manufacturing
0.02	MAIN STREET CAPITAL CORP	523	Securities, Commodity Contracts, and Other
0.94	Sum of top ten weights		
Panel B: Ford Motor Company			
Weight	Peer firm	NAICS	NAICS description
0.57	GENERAL MOTORS CO	336	Transportation Equipment Manufacturing
0.08	BORGWARNER INC	336	Transportation Equipment Manufacturing
0.07	PRINCIPAL FINANCIAL GROUP INC	523	Securities, Commodity Contracts, and Other
0.05	PRUDENTIAL FINANCIAL INC	524	Insurance Carriers and Rel. Activities
0.04	B O K FINANCIAL CORP	522	Credit Intermediation and Rel. Activities
0.04	JETBLUE AIRWAYS CORP	481	Air Transportation
0.02	GAP INC	458	Clothing, Accessories, and Shoe Retailers
0.02	AMERICAN AXLE & MFG HLGDS INC	336	Transportation Equipment Manufacturing
0.02	MAGNA INTERNATIONAL INC	336	Transportation Equipment Manufacturing
0.02	HONDA MOTOR LTD	336	Transportation Equipment Manufacturing
0.93	Sum of top ten weights		
Panel C: Tesla Inc.			
Weight	Peer firm	NAICS	NAICS Description
0.12	APPLE INC	334	Computer and Electronic Product Manufacturing
0.09	SYNOPSYS INC	511	Software Publishers
0.07	NETEASE INC	519	Web Search Portals, Libraries, Archives
0.07	ALLOT LTD	334	Computer and Electronic Product Manufacturing
0.07	PERFICIENT INC	541	Professional, Scientific, and Technical Svcs
0.06	APTIV PLC	336	Transportation Equipment Manufacturing
0.05	BALLARD POWER SYSTEMS INC	335	Electrical Eqmt, Appliance, and Component Mfg
0.05	BLOCK INC	518	Computing Infra Provs, Data Prcs, Web Hosting
0.04	A C M RESEARCH INC	333	Machinery Manufacturing
0.04	SALESFORCE INC	511	Software Publishers
0.66	Sum of top ten weights		

Table 1: Optimal peer groups of General Motors, Ford, and Tesla.

Peer group classification and weighting scheme	Potential peer firm universe → / Focal firm universe →	Compustat North America: Compustat NA (1)	Execucomp (2)	S&P 500 (3)
	Lasso			
	- optimal weights	**0.20**	**0.36**	**0.43**
	Hoberg-Phillips			
	- equally-weighted	0.16	0.27	0.28
	- value-weighted	0.15	0.27	0.33
	SIC			
	- equally-weighted	0.15	0.25	0.26
	- value-weighted	0.15	0.27	0.34
	NAICS			
	- equally-weighted	0.15	0.25	0.26
	- value-weighted	0.15	0.27	0.35
	GICS			
	- equally-weighted	0.17	0.28	0.29
	- value-weighted	0.18	0.31	0.39
	Market portfolio			
	- equally-weighted	0.12	0.21	0.20
	- value-weighted	0.12	0.22	0.25
	Improvement of lasso R2 over alternative			
	- Hoberg-Phillips	29%	32%	30%
	- SIC	30%	31%	25%
	- NAICS	31%	30%	25%
	- GICS	10%	15%	12%
	- Market	63%	62%	71%

Table 2: This table reports the mean out-of-sample R^2 of regressions of focal firm stock return on peer portfolio return. Peer groups are based on various classifications and peer portfolios are constructed using different weighting schemes as indicated in the leftmost column. For comparability, the mean values in column (2) are based on the common sample of focal firms with non-missing data for the out-of-sample R2 using TNIC-3, SIC-3, NAICS-4 and GICS-8. Out-of-sample R2 are winsorized at 0.5% and 99.5% levels. In column (1) and (3), the sample period is 1990Q1-2022Q4. In column (2), the sample period is 1992Q1-2022Q4 since Execucomp coverage begins in 1992.

Peer group classification	Mean	Std	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Lasso	58	45	0	0	12	32	50	73	109	142	230
TNIC	75	121	0	0	0	3	17	79	296	379	467
SIC	62	80	0	0	2	8	21	89	199	249	300
NAICS	83	120	0	0	3	9	27	100	296	401	459
GICS	58	84	0	0	0	11	29	63	135	253	442

Table 3: Peer group size statistics. The table reports statistics of peer group size for the peer group classification methods listed in the leftmost column. The sample period is 1990Q1-2022Q4. The sample comprises all firms in CRSP/Compustat with non-missing daily stock returns in the four-year training period and the one year test period around the focal quarter as described in Section 3.1.

Horizon h (in quarters) $\rightarrow$	$h=1$	$h=4$	$h=16$
$\lvert\Delta$ weight$\rvert$ $q \rightarrow q+h$			
Mean	0.02	0.03	0.03
Median	0.01	0.01	0.01
Std	0.87	0.70	0.80
N	9,317,115	8,657,464	6,769,666
Prob[SIC industry peer in $q \rightarrow$ no peer in $q+h$]	-	0.07	0.13
Prob[NAICS industry peer in $q \rightarrow$ no peer in $q+h$]	-	0.07	0.12
Prob[GICS industry peer in $q \rightarrow$ no peer in $q+h$]	-	0.08	0.19
Prob[TNIC peer in $q \rightarrow$ no peer in $q+h$]	-	0.19	0.28
Prob[lasso peer in $q \rightarrow$ no lasso peer in $q+h$]	0.36	0.56	0.66
Prob[top-5 lasso peer in $q \rightarrow$ no lasso peer in $q+h$]	0.01	0.03	0.05
Prob[top-10 lasso peer in $q \rightarrow$ no lasso peer in $q+h$]	0.02	0.06	0.12

Table 4: Stability of peer weights over time.