KING'S
BUSINESS
SCHOOL

Data Analytics
for Finance
& Macro
Research Centre

KING'S
College
LONDON

Interpretable Machine Learning Modelling for Asset Pricing

G. Kapetanios, F. Kempf

Working paper No. 2022/1 | February 2022 | ISSN 2516-5933

Interpretable Machine Learning Modelling for Asset Pricing[*]

WORKING PAPER

George Kapetanios
King's College London

Felix Kempf
King's College London

February 15, 2022

Abstract

We use deep neural networks to estimate time-varying equity risk premia. The key innovations are the nonlinear and non-parametric generalisation of Fama-Macbeth regressions through partial derivatives of an arbitrary estimator function with respect to its input and the introduction of Jacobian regularisation in the objective function to empirical asset pricing. Our methodology outperforms all linear benchmarks out-of-sample. Moreover, we introduce the concept of sensitivity-sorted portfolios. Most importantly, we move elements of interpretability and explainability to the foreground.

Keywords: Conditional asset pricing model, nonlinear factor model, cross-section of expected returns, machine learning, deep learning, Jacobian regularisation, interpretability, explainability

1 Introduction

Asset pricing centres around the question of how to explain variations in expected returns across assets. While being a notoriously challenging task empirically, the answer to this question from an asset pricing theory point of view is clear. In the absence of arbitrage, variations in expected returns should be reflected by an asset's exposure to a stochastic discount factor (SDF) m_{t+1} for any excess return $r_{i,t+1}$. Borrowing notation from Kelly et al. (2019), it follows that

$$\mathbb{E}_t[m_{t+1} r_{i,t+1}] = 0 \iff \mathbb{E}_t[r_{i,t+1}] = \underbrace{\frac{\mathrm{Cov}_t(m_{t+1}, r_{i,t+1})}{\mathrm{Var}_t(m_{t+1})}}_{\beta'_{i,t}} \underbrace{\left(-\frac{\mathrm{Var}_t(m_{t+1})}{\mathbb{E}_t[m_{t+1}]}\right)}_{\lambda_t}, \tag{1}$$

[*] We are grateful for the helpful comments and suggestions from Markus Pelger (Stanford University), Harald Lohre (Invesco), Maximilian Stroh (Quoniam), and the seminar participants of the 2019 and 2021 Bank of England Conference *Modelling with Big Data and Machine Learning: Interpretability and Model Uncertainty*, where we first presented the idea, https://www.bankofengland.co.uk/events/2019/november/modelling-with-big-data-and-machine-learning.

where $\beta_{i,t}$ denotes the exposures to systematic risk factors of assets $i = 1, ..., N_{t+1}$ and λ_t are risk prices which is defined as compensation an investor can expect to receive for bearing systematic risk. Without loss of generality, if the SDF is linear in factors f_{t+1} (i.e. $m_{t+1} = a + b' f_{t+1}$; see Cochrane (2009)), the conditional mean estimator function of equation (1) is equivalent to a beta pricing model (e.g. see Ross (1976) or Hansen and Richard (1987)), such that

$$r_{i,t+1} = \alpha_{i,t} + \beta'_{i,t} f_{t+1} + \epsilon_{i,t+1}, \tag{2}$$

where $\mathbb{E}_t[\epsilon_{i,t+1}] = \mathbb{E}_t[f_{t+1}\epsilon_{i,t+1}] = 0$ and $\mathbb{E}_t[f_{t+1}] = \lambda_t$, with most asset pricing tests typically focussing on $\mathbb{E}_t[\alpha_{i,t}] = 0$. The conditional expectation of equation (1) and the beta representation of equation (2) are the starting point of two of the most widely used approaches in empirical asset pricing: cross-sectional regressions on firm characteristics and (one or) two-pass regressions such as the original Fama-Macbeth (FM hereafter; Fama and MacBeth (1973)) regressions on (conditional) portfolio sorts, respectively predefined and common risk factors. Thus, in the former case, the betas are assumed to be observable (e.g. Lewellen (2015)), while in the latter case, the risk factors are assumed to be observable (e.g. see Jagannathan and Wang (1998)). What unifies both approaches is the estimation and testing of the marginal importance of each characteristic, respectively risk factor. We show that our proposed methodology can be applied universally in both settings.

Current literature has four key challenges regarding the empirical estimation and interpretation of risk exposures and prices. First, in practice, neither factors nor risk exposures are directly observable and could theoretically depend on all observable information, yielding a potentially high-dimensional estimation problem. The empirical hunt for new risk factors or firm characteristics that can explain variations in expected returns has produced hundreds of potential candidates in recent years, as documented by Harvey et al. (2016), Mclean and Pontiff (2016) or Hou et al. (2020). Consequently, modern asset pricing models should be capable of handling high-dimensional data. It is in this environment that Cochrane (2011) points to alternative methods to navigate this ever-growing *factor zoo* effectively.

Second, there is little intuitive justification and strong empirical evidence against the linearity assumption in equation (2), as discussed, for example, by Campbell and Cochrane (1999), Bansal and Yaron (2004) or He and Krishnamurthy (2013). Therefore, an asset pricing model should (at least partly) capture intrinsic nonlinearities in the return data. Non-parametric modelling, such as machine learning, offers the opportunity to impose no restrictions on the estimator's functional form. This data-driven approach allows for nonlinear estimations and yields statistical factors and exposures instead of pre-specified (and potentially misspecified) factors and exposures.

Third, despite an abundance of dynamic models, the overwhelming majority of empirical research is concerned with estimating risk premia unconditionally, meaning that the compensation an investor can expect to receive for bearing systematic risk is assumed to remain constant over time. Notable counter-examples include Ferson and Harvey (1991), Ilmanen et al. (2019) or Umlandt (2020) who study time-varying risk premia and risk exposures. Still, little is known about the time-variation of risk exposures and premia and their non-parametric estimation. In addition, the ever-changing nature of financial markets, technological advancements and even newly introduced regulation make it seem unlikely that risk premia should remain constant over time; e.g. see Oh and Patton (2018). The poor performance of traditional value investing (e.g. Agrawal (2020)) in recent years or changing cross-asset correlation structures during times of crises support this hypothesis from an asset manager's or regulator's point of view.

Fourth, researchers, asset managers, or regulators have the fiduciary duty to understand their models' inner structure and communicate it clearly with all associated risks to their clients or stakeholders. Hence, it is desirable to engineer models with powerful out-of-sample performance (which is laborious enough already) and move elements of interpretability and explainability to the foreground. In this context, the terms *interpretability* and *explainability* are only loosely defined in the machine learning literature. Miller (2019), for example, states that interpretability can be summarised as the degree to which a human can understand the cause of a model's output, with Kim et al. (2016) or Murdoch et al. (2019) argue in the same vein and point out that interpretability is a useful umbrella term that captures the identification of relevant knowledge extracted from a machine learning model. Rudin (2019) differentiates between the two terms and argues that interpretable models are intrinsically understandable, while explainable models are black-box models whose outputs become interpretable post-hoc. In addition, Doshi-Velez and Kim (2017) point out that single performance measures, such as the cross-sectional mean R^2, predictive R^2 or even out-of-sample mean-squared errors, are incomplete descriptions for a model's suitability and elements of interpretability and explainability should be part of the model selection process.

Interpretable or explainable machine learning in empirical asset pricing is fundamentally different from other disciplines. For example, consider the domain of image recognition, where the data is characterised by extraordinarily high signal-to-noise ratios, combined with an abundant training data availability. Rudin (2019)'s arguments of introducing interpretable layers into machine learning models, such as neural networks, follows common sense: if, for example, a model detects the head of a bird and is capable of comparing the bird's head to an extensive library of other birds' heads and, thus, bases its classification on the most probable commonality, this type of model mechanic is highly interpretable for humans. On the other hand, datasets in economics and finance are fundamentally different due to their meagre signal-to-noise ratio. The extent to which a human can fully understand where the data's signal is extracted from is, thus, diametrically different compared to other disciplines, such as image recognition. Therefore, direct comparability of model interpretability across disciplines is limited or at least difficult.

We build on existing research, such as Chen et al. (2019), Gu et al. (2020a) and Freyberger et al. (2020), that presents compelling evidence in favour of modelling the conditional mean of equation (1) non-parametrically and nonlinearly. To estimate risk premia, we conflate the seminal cross-sectional regressions of Fama and MacBeth (1973) with machine learning. Machine learning certainly is already an integral component of modern asset pricing literature. However, to the best of our knowledge, there exists little insight into the direct implementation of FM regressions using machine learning. The main reason for this is that previous research has often deemed periodic refitting as computationally too expensive. However, thanks to easy access to increased and on-demand computing power and significantly reduced computing costs, those limitations are slowly but surely reduced. Despite being computationally more expensive, machine learning helps mitigate one of the central and well-documented problems present in cross-sectional regressions: in-sample overfitting. By drawing on tools such as sample-splitting, hyper-parameter tuning or regularisation, machine learning specifically attempts to avoid in-sample overfitting and seeks to generalise well out-of-sample[1]. This shift in focus helps to improve statistical inference about risk premia.

With this paper, we aim to make four key contributions. First, under the condition of differentiability, we propose the use of partial derivatives of the conditional mean estimator function with respect to its inputs (i.e. the input gradients) as a means to estimate nonlinear and time-varying risk premia (e.g. see Dixon and Polson (2019)). The nonlinearity of the thereby estimated risk premia directly results from the estimator's inherent nonlinearity (e.g. see Kapetanios (2007)). Refitting periodically in the style of Fama-Macbeth regressions induces a time variation in our estimation. Partial derivatives are a natural way to conceptualise risk premia and distinguish themselves through easy and fast implementation. We show that our proposed methodology is a generalisation that nests the Fama-Macbeth estimator as a special case assuming that the conditional mean estimator function is linear. Most importantly, the combination of nonlinear modelling and input gradients allows for more general pricing factors. Most cross-sectional research fundamentally draws on the seminal works by Black et al. (1972), Fama and MacBeth (1973) and Gibbons et al. (1989) and perform statistical inference on the time-series averages of risk premia estimates, typically with some form of adjustment for multiple testing as proposed by Newey and West (1987) or Benjamini and Yekutieli (2001). Using averages is an intuitive and well-established procedure. However, evaluating an overall average fails to recognise that risk premia may change over time. We explicitly allow risk premia to vary over time and transition in and out of empirical importance by constructing tolerance bands. This transition is particularly beneficial when analysing times of economic stress. Moreover, we show that nonlinear risk premia facilitate much richer insights as we can conduct analyses, for example, by company size or industry, without refitting the model on the subset of interest.

[1] For an introduction to machine learning in general or machine learning in finance, we refer readers to Bishop (1995), Hastie et al. (2009), Goodfellow et al. (2016) or De Prado (2018).

Second, we introduce Jacobian regularisation as part of the objective function to empirical asset pricing. The objective function generalises variable selection and shrinkage for nonlinear models analogously to LASSO, Ridge or Elastic Net (e.g. see Tibshirani (1996), Zou and Hastie (2005) or Hastie et al. (2009)), and thus increases model interpretability. The objective function minimises the residual sum of squares subject to the regularisation of the gradient norm[2]. While the concept of Jacobian regularisation is not new in the general machine learning literature, and in particular in image recognition (e.g. see Drucker and Cun (1992), Sokolić et al. (2017), Varga et al. (2018) or Hoffman et al. (2019)), we introduce it to asset pricing. Due to the constrained loss function's nature, certain risk premia are set exactly to (or shrunk towards) zero. Financial machine learning fundamentally differs from machine learning applications in other disciplines, such as image recognition or natural language processing, where machine learning thrives. Due to limited data availability, the dynamic nature of financial markets and a low signal-to-noise ratio, out-of-the-box machine learning algorithms show a high failure rate in practice. De Prado (2018) argues that one of the critical factors to make financial machine learning applications more successful is to move away from a purely data-driven approach to a *quantamental* strategy. Quantamental, in this context, describes the combination of data-driven algorithms with economic and financial theory. Our proposed objective function makes exactly that possible. Economic theory suggests that the true but unknown asset pricing model is approximately low-dimensional as recognised, for example, by Kozak et al. (2018) or Barillas and Shanken (2018). By setting specific risk premia to exactly zero, the objective function performs model selection. The resulting model is, consequently, more easily interpretable due to its reduced dimensionality. Additionally, from a practitioners' point of view, it is desirable to achieve competitive out-of-sample performance while manually imposing restrictions on the influence of specific risk factors or firm characteristics in the model. A significant amount of academic research in financial machine learning finds that the most influential factors can frequently be attributed to market frictions or momentum and regularly attests short-term-reversal to be the source of the most significant signal (e.g. Gu et al. (2020b), or Lewellen (2015)). However, those factors can be difficult to trade on in real-life, due to high portfolio turnovers and trading cost considerations (e.g. Leung et al. (2021)). Therefore, there is great demand for performant machine learning models such as deep neural networks and the possibility of simultaneously imposing individual restrictions on the learning algorithm. While many practical applications are imaginable, such as ESG (economic, social and governance) restrictions, we merely focus on the objective function's theoretical properties and implementation methods in this paper.

Third, we explicitly do not wish to shift existing out-of-sample performance frontiers with our empirical analysis. Instead, the collectivity of all innovations mentioned above intends to move interpretability and explainability elements to the foreground. Those aspects have become increasingly important in

[2] The flexible form of the objective function also allows for simultaneous weight penalisation.

recent years, not least because of asset managers, regulators, and researchers' fiduciary duty to communicate all associated risks of their models to all relevant stakeholders. For example, we show how the nonlinear interaction of model sensitivities to changes in the input helps to understand complex inner model mechanics better. Moreover, our proposed methodology allows for model analysis on the asset level. Such granular model inspection is still underrepresented in current literature to the best of our knowledge but offers valuable insights. Examples include detecting individual assets that the model is not handling well or extreme sensitivity outliers. An inspection on the asset level also helps with software debugging or the discovery of unwanted model biases. It further allows for computationally inexpensive and detailed analyses on subgroups of assets, such as by industry or size class, defined by an asset's market capitalisation.

Finally, we introduce the concept of double-sorted portfolios based on model sensitivities. Portfolios sorts are ubiquitous in the empirical asset pricing literature and are used to test fundamental asset pricing theories, establish pricing anomalies or identify profitable investment strategies (e.g. see Cattaneo et al. (2020)). Portfolio sorts are most fundamentally based on the idea of sorting firm characteristics into baskets on which, for example, equal or value-weighted portfolios are constructed and has been informally recognised as a non-parametric alternative to imposing linearity on the relationship between the returns of assets and firm characteristics (e.g. see Fama and French (2008), or Cochrane (2011)). We develop a framework by casting double-sorted out-of-sample portfolios as sorts that are based on firm characteristics and the expected volatility in expected returns due to changes in firm characteristics. By incorporating expected out-of-sample return sensitivities to (potentially unexpected) changes in firm characteristics, we offer a tool to manage a double-sorted portfolio's expected risk. The critical difference to existing approaches in current literature is that our methodology fundamentally relies on an out-of-sample estimation that is assumed to generalise well to unseen data. This out-of-sample approach, thus, is an empirically more difficult task. As a consequence, we do not claim to discover previously unknown anomalies. Instead, we show that incorporating model sensitivities into the portfolio construction exercise can help to manage out-of-sample risk.

Our empirical study follows the standard procedure in the empirical asset pricing literature. We source monthly stock returns from CRSP for firms listed on NYSE, AMEX, and NASDAQ and construct 103 firm characteristics based on fundamental data sourced from Compustat, I/B/E/S, FRED and BLS. The list of firm characteristics includes current literature's most relevant candidates. Further, we consider an alternative scenario in which only 49 of the most commonly used firm characteristics are included. We refer to these 49 firm characteristics as *core characteristics*. The data sample spans over 492 months from January 1980 to December 2020, yielding over four decades of data. We investigate ten different neural networks, each with a separate objective function. In particular, we consider objective functions with no model parameter penalisation, model weights regularisation and various forms of Jacobian (input gradient) regularisation.

Moreover, we benchmark the neural networks' model performances against five standard linear benchmarks, including ordinary least squares (OLS), weighted least squares (WLS), ridge, elastic net and least absolute shrinkage and selection operator (LASSO). All ten neural networks under consideration outperform the linear benchmarks out-of-sample . We compare the out-of-sample performances using the two key metrics, cross-sectional mean R^2 and predictive R^2. The best performing models in the core characteristics only case are the neural networks with L_1 norm weight, respectively column-wise L_1 norm Jacobian penalisation as part of their objective function, with 17.20%, respectively 15.83% cross-sectional mean R^2 and 0.12%, respectively 0.09% predictive R^2. This paper primarily focuses on the newly introduced methodological contributions, which is why we do not claim to shift existing state-of-the-art prediction performance frontiers. The competitive empirical performances of neural networks whose objective function includes Jacobian regularisation are robust across many sample-splitting and data pre-processing regimes, input dimensionalities, asset size classes (defined by their market capitalisation) and specific microcap considerations (such as excluding microcaps entirely from the investment universe).

The empirical findings are six-fold. First, we confirm the empirical insights presented by Gu et al. (2020b) or Chen et al. (2019) that deep neural networks can explain intrinsic return structures due to their non-parametric and nonlinear form. However, performances can be subpar when using out-of-the-box neural network training regimes, including objective functions with no model parameter regularisation. The crucial innovation is the inclusion of the Jacobian or input gradient penalty, as it not only generalises the linear model selection techniques, such as the LASSO or Ridge- to neural networks. It also introduces an economically interpretably penalty term.

Second, we confirm the theoretical properties of the Jacobian regularisation term and show that input gradient penalisation does indeed yield non-parametric model selection. In addition, the empirical asset pricing literature has long argued that a parsimonious model representation is desirable. Thus, the marginal benefit of including the Jacobian penalty term in the objective function is relatively more significant in high-dimensional settings compared to lower dimensionalities. For example, the cross-sectional mean R^2 in the case of using all 103 firm characteristics is nearly 18 times higher for neural networks that are trained with an objective function that includes an L_1 norm column-wise penalty term of the Jacobian compared to a neural network with no model parameter regularisation as part of the objective function (16.12%, compared to 0.92%). However, in the case of manually reducing the firm characteristic universe from 103 to the most relevant 49 firm characteristics, this marginal benefit is significantly reduced, as the cross-sectional mean R^2 of neural networks with L_1 norm column-wise Jacobian penalisation in their objective function is merely 1.1 times higher compared to neural networks with no model parameter penalisation in the objective function. In addition, the most critical variables selected by the best performing neural networks are significantly less correlated than, for example, the linear OLS benchmark. The advantageous correlation structure of the most influential firm characteristics suggests that the best performing neural networks are better capable of extracting signals from firm characteristics, even in the presence of mild multicollinearity, making them an empirically robust model choice.

Third, our applied study provides strong empirical evidence suggesting a time-varying nature of risk premia. In particular, we show that the unconditional and constant risk premia, which are typically reported in the empirical asset pricing literature, are far less informative compared to the time-varying risk premia with tolerance bands that we report in this paper. In our empirical study, we show, for example, that the estimated risk premium associated with being exposed to the systematic risk factor return-on-assets spikes during the financial crisis but levels out in the following years. Moreover, the risk premia estimates are robust across different asset size classes defined by their market capitalisation.

Fourth, similarly to Chen et al. (2019) we show that nonlinear firm characteristic interactions matter, which is also discussed by Gu et al. (2020b), or Bryzgalova et al. (2019). Moreover, pairwise locally weighted regressions help to understand the nonlinear interactions between return prediction sensitivities to changes in input firm characteristics. These nonlinear interactions provide valuable model insights to all stakeholders. They help explain the expected model prediction sensitivities to changes in one firm characteristic, given the model prediction sensitivity to changes in another firm characteristic for a particular asset.

Fifth, the input gradients provide valuable model insights on the asset level. In particular, we show that they are beneficial for evaluating the general functioning of the objective function of choice. Further, they can help with software debugging and the detection of unwanted biases in the model estimation. In an empirical study, we also discuss the concept of *prediction stability*, which is closely related to model sensitivity outliers. While individual assets typically do not negatively influence the overall model performance, model stakeholders may still be interested in stable return predictions for individual assets. In this context, we stress that the objective function choice should become an integral part of the model design exercise, such that the general functioning of the objective function adequately fulfils the required features.

Sixth, double-sorted portfolios on firm characteristics and sensitivities can offer a practical tool that can help to manage expected out-of-sample portfolio volatilities. We show that for neural networks with column-wise L_1 norm Jacobian regularisation, double-sorted value-weighted portfolios sorted on, for example, return-on-assets and low expected model prediction sensitivities yield a Sharpe ratio that is nearly 38% higher than the Sharpe ratio of portfolios that are sorted on high model sensitivities. A similar pattern emerges for equal-weighted portfolios where the portfolio sorted on low sensitivities is nearly 28% compared to the portfolio constructed on high sensitivities. We do not claim that such patterns are universally applicable to all firm characteristics and models. A possible explanation for this pattern – which is also closely related to the concept of prediction stability – is that given an estimator function g that generalises well out-of-sample, assets with low return prediction sensitivities to changes in a specific characteristic are expected to be less volatile compared to assets with high sensitivities.

We intend to extend several different strands of the empirical asset pricing and econometrics literature. This paper seamlessly blends into the extensive asset pricing literature investigating the risk factor identification and econometrics literature estimating factor models. However, our paper most closely relates to the newly established literature utilising non-parametric and nonlinear machine learning to shrink the high-dimensional cross-section in asset pricing.

Starting with the capital asset pricing model (CAPM) of Sharpe (1964), Lintner (1965) and Mossin (1966), which is building on the mean-variance portfolio optimisation proposed by Markowitz (1952)), empirical asset pricing has attempted to capture the implications of the SDF: the empirical estimation of risk exposures β and prices of risk λ[3]. The most important theme of this early seminal research are the testable implications (e.g. see Jensen (1968), Black (1972) or Gibbons et al. (1989)) of the null hypothesis that the intercept (alpha) and time-series average of risk prices are zero. Subsequently, the single-factor approach of the CAPM[4] has been continuously extended. Fama and French (1993) propose a three-factor model, while Hou et al. (2015) present a four-factor model. Fama and French (2015) move on to a five-factor model and Barillas and Shanken (2018) suggest a six-factor model. Similarly to Lewellen (2015), we aim to shed some light on the navigation of this ever-growing factor zoo.

Methodologically, the seminal contribution by Fama and MacBeth (1973) is the infamous two-step procedure that combines time-series with cross-sectional regressions in order to estimate β's and λ's in equation (2). FM regressions are still heavily used today. In a first step, excess returns are regressed on previously constructed risk factor portfolio returns[5] using rolling time-series regressions such as for months $t-60$ to $t-1$. After estimating the betas, cross-sectional regressions of excess returns on the betas are utilised to estimate prices of risk, where the null hypothesis is that the mean risk premium is zero. We translate this fundamental principle into a nonlinear and non-parametric setting where we allow risk exposures and risk premia to vary over time. Fama and French (2020) show that observable firm characteristics are equivalent to time-varying risk exposures in cross-sectional regressions.

We also build on extensive econometrics research investigating time-varying models, in contrast to, for example, static-beta models. Ferson and Harvey (1991), Ghysels (1998), Chaieb et al. (2018), Ilmanen et al. (2019) or Umlandt (2020) study time-varying risk premia. Similarly, Jagannathan and Wang

[3] Other pioneering work includes Black (1972), Merton (1973), Ross (1976), Banz (1981), Basu (1983) or Rosenberg et al. (1985)

[4] Empirical evidence against the static single-factor CAPM is presented, for example, by Banz (1981), Reinganum (1981), Gibbons (1982), Basu (1983), Chan et al. (1985), Shanken (1985) or Bhandari (1988)

[5] Typically, they are constructed as long-short portfolio returns based on firm characteristics. For example, at time t all firms are ranked based on their characteristic $c_{t,ik}$ where an investor would go long the top decile and short the bottom decile.

(1996a) allow risk exposures and risk premia to vary over time. We extend this approach by discussing the usefulness of confidence, respectively tolerance bands for time-varying risk premia estimations. In particular, we discuss theoretical methodologies such as bootstrapping and put them into practical context by considering their computational cost. To keep computational costs feasible, we provide an empirical approach that yields empirical tolerance bands, such that risk premia can transition in and out of empirical importance, where we define risk premia as *empirically important* if the tolerance bands do not include zero.

In times of unprecedented data and computing power availability machine learning has emerged as a powerful alternative in modern asset pricing to study the cross-section of stock returns[6]. Gu et al. (2020b) conduct a large-scale comparison of various different machine learning methods and show the benefits of non-parametric and nonlinear modelling. In particular, they show the competitive performance of deep neural networks, not least because of their universal approximation capabilities; see, for example, Cybenko (1989) or Hornik et al. (1989). Similarly, Messmer (2017a), Feng et al. (2018) and Gu et al. (2020a) predict stock returns using deep neural networks. Chen et al. (2019) propose to further include macroeconomic time-series in their neural network training. Feng et al. (2020) utilise deep neural networks in connection with an economic objective function that minimises pricing errors which shows a significant improvement in the efficient portfolio[7]. We build on this strand of literature and also apply deep neural networks in our empirical application. However, our proposed methodology generalises to a large number of models under the condition of differentiability.

In addition to deep neural networks, we also build on the newly established strand of literature studying the dimensionality reduction of the high-dimensional cross-section of returns. Rapach et al. (2013) apply LASSO to predict returns using lagged information. Similarly, Messmer (2017b), Feng and Giglio (2017) or Han et al. (2019) propose the LASSO for model selection in order to navigate through the factor zoo[8]. Freyberger et al. (2020) and Kozak et al. (2020) use shrinkage and selection methods to estimate the SDF. Kelly et al. (2019) propose the instrumented principal components that extends typical PCA with time-varying loadings. Lettau and Pelger (2020b) study the estimation of latent factors using a generalisation of PCA. Gu et al. (2020a) propose non-linear PCA using autoencoders.

Last but not least, our paper also relates to Dixon and Polson (2019) as they also draw on partial derivatives of deep neural networks with respect to their inputs. However, our approach differs from Dixon and Polson (2019) in three ways: First, we apply partial derivatives in the context of asset pricing models and, therefore, estimate not only systematic risk factors but also risk compensation. Second, we directly incorporate the partial derivatives as part of the objective function through Jacobian

[6] The rise of machine learning in asset pricing has arguably been fuelled even more by spillover effects from successes in other disciplines such as such as computer vision (e.g. Goodfellow et al. (2014)), natural language processing (e.g. Kumar et al. (2016)) or complex gaming (eg. Silver et al. (2016)).

[7] Other notable papers include Heaton et al. (2017), Messmer (2017a), Bryzgalova et al. (2019), Imajo et al. (2020), Harvey and Liu (2014).

[8] Other examples include Chernozhukov et al. (2018).

penalisation. Third, we use partial derivatives as an instrument to shed light on complex inner model structures, which are often described as *black boxes*. Moreover, our empirical application more closely resembles current state-of-the-art data as we replicate the factor zoo of Green et al. (2017).

The rest of this paper is structured as follows. Section 2 lays out our proposed methodology. Section 3 discusses the empirical estimation strategy. In section 4 we introduce the considered investment universe and the empirical results. Section 5 presents concluding remarks. We relegate all additional empirical results and discussion to the appendix.

2 Methodology

In the following, we do not impose any restrictions on the functional form of returns and generalise the beta representation of excess returns from equation (2) as an additive prediction error model such that

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \epsilon_{i,t+1}, \tag{3}$$

with

$$\mathbb{E}_t[r_{i,t+1}] = g_t(\mathbf{x}_{i,t};\ \mathbf{W}_t, \theta_t), \tag{4}$$

where $\mathbb{E}_t[r_{i,t+1}] = \mathbb{E}[r_{i,t+1}|\mathcal{I}_t]$ denotes the conditional expectation of returns given all available information $\mathcal{I}_t$ that is observable by an investor up until time $t = 1, ..., T$, and $\epsilon \sim$ i.i.d.$(0, \sigma_\epsilon^2)$ are random errors. In its most general form, g_t is arbitrary, unknown and only depends on the K-dimensional input vector $\mathbf{x}_{i,t}$. Potential remaining model (hyper-)parameters are captured by $\mathbf{W}_t$ and θ_t. More specifically, we apply common terminology and refer to the model parameters captured by $\mathbf{W}_t$ as *learnable* model parameters, as their *optimal*[9] values are not set manually. Instead they are found as part of the model fitting process, such as gradient-based optimisation. In contrast, the model parameters captured by θ_t cannot be *learned* through optimisation, but need to be set manually. Their *optimal* values are typically found through k-fold cross-validation[10]. Individual stocks in equations (3) and (4) are indexed by $i = 1, ..., N_t$, where N_t describes the total number of stocks at time t. The time index emphasises that financial data is generally unbalanced and that the number of stocks can vary over time.

2.1 Generalisation of Fama-Macbeth Two-Pass Procedure

In this section, we review the standard two-step FM regression procedure (see Fama and MacBeth (1973) or Cochrane (2009)) that remains popular in asset pricing literature to this day due to its simplicity. Our approach can be seen as a generalisation of the static FM asset pricing approach. What unites the classical FM and our generalised approach is estimating the risk exposures and the prices of risk in two

[9] Where the term optimal does not necessarily refer to a global optimum.

[10] We refer readers to Hastie et al. (2009) for a more comprehensive discussion.

separate steps. Equation (4) raises the central question of how to approximate and estimate the unknown estimator function g_t. The conditional asset pricing model in equation (1) and the beta representation in equation (2) form the basis of FM regressions as they assume a linear relationship between the risk exposures (β) and risk prices (λ) and, thus, defines g_t as

$$\mathbb{E}_t[r_{i,t+1}] = g_t(\boldsymbol{\beta}_{i,t}, \boldsymbol{\lambda}) = \boldsymbol{\beta}_{i,t}\boldsymbol{\lambda}_t, \tag{5}$$

where $\boldsymbol{\beta}_{i,t} = (\beta_{i,t}^{(1)}, \cdots, \beta_{i,t}^{(k)}, \cdots, \beta_{i,t}^{(K)})$ denotes a K-vector of risk exposures for asset $i = 1, ..., N_{t+1}$ and $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \cdots, \lambda_{k,t}, \cdots, \lambda_{K,t})'$ is a K-vector of risk prices. The fundamental problem (and the starting point of the first step in FM regressions) is that neither the risk exposures nor the risk prices are directly observable and, therefore, must be estimated first. The standard procedure assumes (e.g. see Bai and Zhou (2015)) that returns are governed by a K-factor model, which – expressed in vector and matrix notation – yields

$$\boldsymbol{r}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t^{(1)} f_{1,t+1} + \cdots + \boldsymbol{\beta}_t^{(K)} f_{K,t+1} + \boldsymbol{\epsilon}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{B}_t \boldsymbol{f}_{t+1} + \boldsymbol{\epsilon}_{t+1} \tag{6}$$

for the cross-section of returns, with $\boldsymbol{r}_{t+1} = (r_{1,t+1}, \cdots, r_{N_{t+1},t+1})'$ is an N_{t+1}-vector of excess returns, $\boldsymbol{\beta}_t^{(1)}, \cdots, \boldsymbol{\beta}_t^{(K)}$ are N_{t+1}-vectors of the multiple-regression betas, $\boldsymbol{B}_t = (\boldsymbol{\beta}_t^{(1)}, \cdots, \boldsymbol{\beta}_t^{(K)})$ is an $[N_{t+1} \times K]$ matrix, and $\boldsymbol{f}_{t+1} = (f_{1,t+1}, \cdots, f_{k,t+1}, \cdots, f_{K,t+1})'$, $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \cdots, \alpha_{N_{t+1},t})'$, with $\mathbb{E}_t[\boldsymbol{\alpha}_t] = 0$, and let $\boldsymbol{\epsilon}_{t+1} = (\epsilon_{1,t+1}, \cdots, \epsilon_{N_{t+1},t+1})'$ be an N_{t+1}-vector of errors.

In the first step of the classical FM procedure, we find the beta estimates through linear time-series regressions for each asset $i = 1, ..., N$, where Fama and MacBeth (1973) suggest a rolling regression approach to induce a time-variation in the betas (see also Chen et al. (1986), Ferson and Harvey (1991) or Petkova and Zhang (2005))), following a fixed-window look-back period of window size W, such that

$$\boldsymbol{r}_{i,t+1} = h_t^{(i)}(\boldsymbol{F}_{t+1}) = \boldsymbol{\alpha}_{i,t} + \boldsymbol{F}_{t+1}\boldsymbol{\beta}'_{i,t} + \boldsymbol{\epsilon_{i,t+1}}, \; \forall i = 1, ..., N, \; \forall t = W, ..., T \tag{7}$$

where N denotes the total number of assets in the sample, $\boldsymbol{r}_{i,t+1}$ denotes a W-vector of excess returns for asset i, $\boldsymbol{\beta}_{i,t} = (\beta_{i,t}^{(1)}, \cdots, \beta_{i,t}^{(k)}, \cdots, \beta_{i,t}^{(K)})$ is a K-dimensional vector, and $\boldsymbol{F}_{t+1}$ is a $[W \times K]$ matrix of observable factors. Consequently, running rolling time-series regressions for each asset across time yields

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{i,t}|\boldsymbol{F}_{t+1}] = \hat{\boldsymbol{\beta}}_{i,t} = (\boldsymbol{F}'_{t+1}\boldsymbol{F}_{t+1})^{-1}\boldsymbol{F}'_{t+1}\boldsymbol{r}_{i,t+1}, \; \forall i = 1, ..., N, \; \forall t = W, ..., T \tag{8}$$

While equation (8) can also be estimated over the entire sample, there is compelling empirical evidence in favour of time-varying betas (e.g. see Jagannathan and Wang (1996b) or Adrian et al. (2015)) which is why we concentrate on a rolling estimation. Most importantly, in the standard Fama-Macbeth procedure, equations (6)-(7) assume that the the factors $\boldsymbol{f}_{t+1}$ are observable and may or may not be tradable. The nuance on tradeability is subtle but important: as discussed by Giglio and Xiu (2016), if the factors are themselves portfolios (i.e. they are tradable), the risk premium may be directly estimated as the time-series average of the excess return of the factor (see Cochrane (2009) for a detailed discussion). However, investors may be concerned about non-tradable risks (i.e. risks that are not themselves portfolios, such as inflation, consumption or liquidity), requiring the second pass in the FM procedure.

Once the betas are estimated, they are considered observable and substituted into equation (6). Subsequently, the second step consists of T cross-sectional regressions such that the conditional asset pricing model takes the form

$$\mathbb{E}_t[\boldsymbol{r}_{t+1}] = g_t(\hat{\boldsymbol{\beta}}_t) = \hat{\boldsymbol{\beta}}_t \boldsymbol{\lambda}_t \tag{9}$$

and the prices of risk are estimated through T cross-sectional regressions

$$\hat{\boldsymbol{\lambda}}_t = (\hat{\boldsymbol{\beta}}_t' \hat{\boldsymbol{\beta}}_t)^{-1} \hat{\boldsymbol{\beta}}_t' \boldsymbol{r}_{t+1}, \ \ \forall t = 1, ..., T \tag{10}$$

where $\hat{\boldsymbol{\lambda}}_t = (\hat{\lambda}_{1,t}, \cdots, \hat{\lambda}_{K,t})'$ is a K-vector of estimated risk prices at t, and $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_t^{(1)}, \cdots, \boldsymbol{\beta}_t^{(K)})$. For simplicity, note that equation (9) does not include a constant (e.g. see Cochrane (2009)), such that the intercepts are pricing errors. Typically, the unconditional risk price is estimated as the overall time-series average (e.g. see Green et al. (2017))

$$\hat{\lambda}_k = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_{k,t}. \tag{11}$$

We do not intend to discuss all econometric properties – such as standard errors — of the estimators in their entirety as they are well-documented and refer readers to Shanken (1992), Jagannathan and Wang (1998), Cochrane (2011) and Bai and Zhou (2015).

In the following, we build on Ullah (1988), Dixon and Polson (2019) and Farrell et al. (2021) to generalise the first and second pass of the classical FM procedure and assume that $\hat{h}^{(i)}$, respectively $\hat{h}_t^{(i)}$ and $\hat{g}_t$, are consistent estimators (which are collective denoted by $\hat{f}$) of the true but unknown function f_*, that is continuous and differentiable everywhere. Moreover, we require $\hat{f}$ to be Lipschitz continuous, meaning that there is a positive real constant A such that $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^K$, $\|\hat{f}(\boldsymbol{x}_1) - \hat{f}(\boldsymbol{x}_2)\| \leq A\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. Then, for the first pass of the two-pass procedure, the partial derivative of the estimator function with respect to its inputs is bounded and a consistent estimator of the risk exposures such that

$$\hat{\beta}_{i,t}^{(k)}(f_{k,t+1}) = \frac{\partial \hat{h}^{(i)}(\boldsymbol{F}_{t+1})}{\partial f_{k,t+1}} = \nabla_k \hat{h}^{(i)}, \ \forall \ i = 1, ..., N; \ \forall \ k = 1, ..., K \tag{12}$$

respectively

$$\hat{\beta}_{i,t}^{(k)}(\boldsymbol{f}_{t+1}^{(k)}) = \bar{\boldsymbol{\beta}}_{i,t}^{(k)}(\boldsymbol{f}_{t+1}^{(k)}), \tag{13}$$

with

$$\boldsymbol{\beta}_{i,t}^{(k)}(\boldsymbol{f}_{t+1}^{(k)}) = \frac{\partial \hat{h}_t^{(i)}(\boldsymbol{F}_{t+1})}{\boldsymbol{f}_{t+1}^{(k)}} = \nabla_k \hat{h}_t^{(i)},\ \forall\, i = 1, ..., N; \forall\, k = 1, ..., K; \forall\, t = W, ..., T. \tag{14}$$

Equation (12) describes the simplest case of estimating h for asset i once over the entire sample, whereas equations (13) to (14) generalise the rolling window estimation from equation (8). Thus, in equation (14), $\hat{h}_t^{(i)}$ is a rolling window estimate over a fixed-window look-back period os size W for each asset $i = 1, ..., N$, where $\boldsymbol{F}_{t+1}$ denotes a $[W \times K]$ matrix of observable factors[11], and $\boldsymbol{f}_{t+1}^{(k)}$ is a W-dimensional vector of the k-th factor. In equation (13), $\hat{\beta}_{i,t}^{(k)}$ denotes the fixed response estimate, which requires an aggregation of the W-vector of partial derivatives summarised by $\boldsymbol{\beta}_{i,t}^{(k)}$. The aggregation into a fixed response estimate is further highlighted by the bar notation in equation (13). We discuss various aggregation forms in section 2.3, but examples include the expected value (e.g. see Ullah (1988) or the median.

Analogously for the second pass, the partial derivatives of $\hat{g}_t$ with respect to the risk exposures yield consistent and bounded estimates for the prices of risk, where

$$\hat{\lambda}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)}) = \bar{\boldsymbol{\lambda}}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)}), \tag{15}$$

with

$$\boldsymbol{\lambda}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)}) = \frac{\partial \hat{g}_t(\hat{\boldsymbol{\beta}}_t)}{\partial \hat{\boldsymbol{\beta}}_t^{(k)}} = \nabla_k \hat{g}_t,\ \forall\, k = 1, ..., K,\ \forall\, t = 1, ..., T, \tag{16}$$

where $\hat{g}_t$ was estimated cross-sectionally at each t, and $\bar{\boldsymbol{\lambda}}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)})$ is a fixed response estimate.

From equations (12)-(16) it is easy to see that if h and g are linear in factors, respectively betas, the partial derivatives yield the classical Fama-Macbeth regressions as a special case. However, there is little intuitive justification for the strong assumption of linearity. In contrast, there exists strong empirical evidence in favour of non-linear return dynamics (e.g. see Campbell and Cochrane (1999), He and Krishnamurthy (2013), Pohl et al. (2018) or Gu et al. (2020a)). Consequently, we do not intend to impose any restrictions on the functional forms of h and g, allowing for (potentially strong) nonlinearities, which offers advantageous properties, especially in the context of non-parametric estimation, which we discuss in section 2.3.

2.2 Special Case: Observable Risk Exposures – Firm Characteristics

The previous section discussed the generalisation of the Fama-MacBeth two-pass procedure. In particular, we showed that the starting point of the first pass is that neither the betas nor the prices of risk are observable. Consequently, the betas must be estimated first through time-series regressions. However, Fama and French (2020) use the insight from Fama (1976), that the coefficients from cross-sectional regressions of excess returns on firm characteristics are equivalent to zero-investment portfolio returns, and

[11] Note that in equation (12), $\boldsymbol{F}_{t+1}$ denotes a $[T \times K]$ matrix of observable factors.

show that if the cross-sectional regressions are stacked over t, the setup becomes an asset pricing model. In other words, firm characteristics are treated as observable and time-varying betas. There exists a myriad of research that investigates the explanatory power of firm characteristics in the cross-section of returns (e.g. see Fama and French (1992, 1996), Daniel and Titman (1997), Avramov and Chordia (2006), Chordia et al. (2015), Lewellen (2015), Kelly and Pruitt (2015) or Freyberger et al. (2020)). Prominent examples include firm characteristics such book-to-market equity, 1-month momentum, size or investments, but modern asset pricing literature lists over 300 different firm characteristics candidates (e.g. see Mclean and Pontiff (2016), Harvey et al. (2016) or Hou et al. (2020)).

When we treat firm characteristics as observable and time-varying betas – analogously to Fama and French (2020) – we skip the first step of the FM two-step procedure and directly run T cross-sectional regressions of excess returns on firm characteristics in order to estimate the prices of risk. Consequently, we substitute $\hat{\boldsymbol{\beta}}_t = \boldsymbol{c}_t$ into equations (15) and (16), with $\boldsymbol{c}_t = (\boldsymbol{c}_t^{(1)}, \cdots, \boldsymbol{c}_t^{(k)}, \cdots, \boldsymbol{c}_t^{(K)})$, where $\boldsymbol{c}_t^{(k)}$ denotes an N_{t+1} dimensional vector of the k-th observable firm characteristic for assets $i = 1, ..., N_{t+1}$. It follows that the k-th time-varying price of risk is

$$\hat{\lambda}_{k,t}(\boldsymbol{c}_t^{(k)}) = \bar{\boldsymbol{\lambda}}_{k,t}(\boldsymbol{c}_t^{(k)}), \tag{17}$$

with

$$\boldsymbol{\lambda}_{k,t}(\boldsymbol{c}_t^{(k)}) = \frac{\partial \hat{g}_t(\boldsymbol{c}_t)}{\partial \boldsymbol{c}_t^{(k)}} = \nabla_k \hat{g}_t, \ \forall\, k = 1, ..., K, \ \forall\, t = 1, ..., T \tag{18}$$

where $\hat{g}_t$ was estimated cross-sectionally at each t, and $\bar{\boldsymbol{\lambda}}_{k,t}(\boldsymbol{c}_t^{(k)})$ is a fixed response estimate which can, for example, be defined as $\bar{\boldsymbol{\lambda}}_{k,t}(\boldsymbol{c}_t^{(k)}) = \mathbb{E}_t[\boldsymbol{\lambda}_{k,t}(\boldsymbol{c}_t^{(k)})]$ (e.g. see Ullah (1988)), or see 2.3 for a more comprehensive discussion. Note that the firm characteristics in equation (20) and (18) are treated as observable to an investor at t.

The general model setup in equation (6) introduces a return forecasting approach using lagged firm characteristics that is common in the empirical asset pricing literature. Our approach differs from a pure forecasting exercise in that we are not only interested in estimating g_t consistently, but also deriving prices of risks as well as deeper model insights. The aspect of model interpretability and communication of inner model mechanics in particular is underrepresented in current literature. We, therefore, intend to shift the focus from a pure forecasting perspective and introduce elements of model interpretability. Moreover, our approach yields an asset pricing model analogously to equation (1).

One of the key problems in current literature is that there is an abundance of cross-sectional predictors (firm characteristics) to chose from. Moreover, a large number of proposed predictors are highly correlated, which calls for the need of a parsimonious model representation to effectively navigate the current *factor zoo*. In a first step, equations (20) and (18) do not introduce the possibility for model selection as they merely focus on a generalised way to estimate prices of risk as the partial derivatives of $\hat{g}_t$ with respect to its inputs (firm characteristics). Equally, equations (20) and (18) do not limit the number of firm characteristics that can be used for return estimations, making the setup very flexible. However, we introduce nonlinear and non-parametric model selection that is conform with the above introduced methodology as part of the objective function in section 3.3.

Goyal (2012) points to the problem that in a high-dimensional setting, the amount of independent information in firm characteristics is unclear. Backed by the empirical analyses conducted by Messmer (2017a) or Gu et al. (2020a), we follow a non-parametric approach analogously to the second step discussed in section 2.1, and allow for nonlinear interactions between firm characteristics through the flexible functional form of g_t.

Due to its popularity in the empirical asset pricing literature, we adopt the "one-step" procedure introduced in this section, which treats firm characteristics as observable betas, in an empirical study in section 4, which also makes our results directly comparable to Green et al. (2017). Consequently, we merely concentrate on the estimation and statistical inference of risk prices assuming we observe the betas through firm characteristics in the following. However, all subsequently introduced methodologies are analogously applicable to the beta estimates.

2.3 Economic Interpretation and Discussion

In the following, we discuss the benefits and economic interpretation of the proposed methodology, starting with the first pass of the two-pass procedure. First, nonlinear estimator functions yield time-varying estimates. Equation (12) shows that even if we estimate h only once over the entire sample, the resulting risk exposure estimates are still time-varying since they are a function of the time-varying factors $\boldsymbol{f}_{t+1}$ (e.g. see Kapetanios (2007)). The same also applies for the estimation of risk prices. In the context of non-parametric modelling, a single estimation is particularly beneficial as it dramatically reduces computational costs. Single estimations, rather than periodic model re-fitting, are not uncommon in the empirical asset pricing literature using machine learning (e.g. see Gu et al. (2020b)). Moreover, there exists a long-standing debate about the time-varying nature of betas and the prices of risk (e.g. see Fama and French (1989), Harvey (1989), Chen (1991), Ferson and Harvey (1991, 1993), Ferson and Korajczyk (1995), Ghysels (1998), Gagliardini et al. (2016) or Umlandt (2020)), which is not fully explored yet. On top of existing empirical evidence indicating time-variations in the risk exposures and prices of risk, we argue that due to the ever-changing nature of financial markets, the introduction of novel regulation and disruptive technologies, allowing for the time-varying estimation of risk exposures and prices of risk is reasonable. Therefore, the proposed methodology is aimed to contribute to the on-going debate about the time-varying nature of risk exposures and prices of risk by offering an alternative way of estimation. However, to make our proposed methodology directly comparable to traditional FM regressions, we also allow for periodic re-fitting. While this also applies for the estimation of risk exposures (which, for example, could involve a rolling or expanding window approach, see equation (14)), we specifically emphasise this approach for the cross-sectional risk price estimations as it directly corresponds to the original FM regressions (see equation (16)).

Equation 16 raises some questions about the economic interpretation of the partial derivatives with respect to the betas (or firm characteristics in equation (18)) – the N_t-dimensional vector[12] $\boldsymbol{\lambda}_{k,t}$. In the linear regression case, Fama (1976) and Fama and French (2020) show that the regression coefficients from cross-sectional regressions $\forall t$ are equivalent to the month t zero-investment portfolio with weights for the assets that set the month $t-1$ portfolio value of that variable to one and zero for all other, such that the portfolio weights are summarised by $\boldsymbol{W}_t = (\boldsymbol{\beta}'_t \boldsymbol{\beta}_t)^{-1} \boldsymbol{\beta}'_t$. Fama (1976) point out that OLS implies a zero-investment portfolio, since $\sum_{i+1}^{N_{t+1}} w_{ik,t} = 0$. In other words, ordinary least squares yields the portfolio weights for forming the k-th risk factor portfolio, with all constraint discussed in Fama (1976). From this perspective, the partial derivative can be seen as a *marginal* return. To see why this is, let us assume returns are linear in firm characteristics as summarised in equations (9) and (10). Thus, if the conditional estimator function is linear in firm characteristics, it follows that

$$\hat{\lambda}_{k,t} = \hat{\lambda}_{ik,t} = \frac{\partial \hat{g}_t(C_t)}{\partial c_{ik,t}} = (C_t C'_t)^{-1} C'_t R_{t+1}, \quad \forall i \tag{19}$$

where C denoted the $[N_{t+1} \times K]$ matrix of firm characteristics, and R is an N_{t+1}-vector of returns. Most importantly, however, equation (19) states that the partial derivative conceptually coincides with a portfolio return. The reason why we refer to this portfolio return as marginal is, that we can estimate the partial derivative on the asset-level: the partial derivative is also a sensitivity, which quantifies the expected change in expected return for a particular asset, given a unit change in the k-th characteristic. In equation (19), however, if g is linear, the marginal portfolio return is invariant across assets.

However, if g is nonlinear, the partial derivatives (and thus the marginal returns) vary across assets and time, yielding a distribution of marginal portfolio returns, requiring an aggregation into a single risk portfolio return (e.g. Kapetanios (2007)). Ullah (1988), for example, proposes an expected value approach such that $\mathbb{E}_t[\boldsymbol{\lambda}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)})]$ denotes an aggregated estimate of the price of risk. This approach is also followed by Dixon and Polson (2019)[13]. In the case of the expected value, and analogously to the weight matrix in case of linear regressions, the portfolio weight for each asset is assumed to be $1/N_{t+1}$. However, the functional form of the aggregated portfolio remains unknown, which is why the aggregation is not limited to the expected value approach. In this paper, for example, we propose the median as a form of aggregation, which is a more robust estimate in the case of skewed partial derivative distributions.

Therefore, the economic interpretation of the aggregated derivative or price of risk is the expected

[12] Note that for notational simplicity we do not apply the commonly used *hat* notation to indicate an estimate.

[13] While Dixon and Polson (2019) do not explicitly mention the usage of the expected value in their paper, they seem to apply it in their published code which can be found on `https://github.com/mfrdixon/Deep_Fundamental_Factors/blob/master/DNNs_vs_OLS.ipynb`.

change in return given a change in risk exposure across all assets. The reference to the risk price's conformity to all assets is subtle but important as it aligns the partial derivative approach with classical asset pricing theory, where we assume that the price of risk must be the same for all assets – merely the asset-specific risk exposure (or beta) causes cross-sectional variations in expected returns.

Empirically, however, we are confronted with a dilemma as the estimated risk prices naturally depend on the investment universe choice. This choice induces an unavoidable arbitrariness, as it is up to the researcher to decide which assets to include or what time horizon to consider in an empirical study. For example, the empirical US-only investment universe considered by Gu et al. (2020b) includes almost 30,000 assets, while Chen et al. (2019) only consider assets for which all firm characteristics are always fully observable (with no missing values), yielding an investment universe of 10,000 US-only assets. Other studies, such as Hou et al. (2020) explicitly account for the influence of micro-caps and also exclude all financial firms. This problem of *fuzziness* has recently gained more attention in the empirical asset pricing literature, in particular in the context of result replication (e.g. see Jensen et al. (2021)).

Our proposed methodology helps to structure some of the above-mentioned empirical difficulties. Since the estimated risk prices are aggregates of the partial derivatives, we could also estimate λ on subsamples, such as by industry, market capitalisation or any other grouping of interest. This estimation would yield an industry-specific (or market capitalisation-specific) price of risk, making our proposed methodology more flexible. Moreover, a subsample-specific estimation does not require a separate model re-fit, making the estimation computationally efficient, which is particularly appealing in the context of machine learning, where the estimation of g might be computationally very expensive. Furthermore, modern non-parametric models tend to be data-intensive, requiring as much training data as possible. Consequently, excluding entire industries or size classes from the training data set would reduce the training set dramatically, making complex model estimations more difficult. Thus, we may wish to re-write equation (15) as

$$\hat{\lambda}_{k,t}^{(s)}(\hat{\boldsymbol{\beta}}_t^{(k)}) = \bar{\boldsymbol{\lambda}}_{k,t}(\hat{\boldsymbol{\beta}}_t^{(k)})\big|_{i\in\mathbb{S}}, \tag{20}$$

where $\mathbb{S}$ denotes a subsample of choice, for example, industries or groups of market capitalisation, and the bar notation denotes an aggregate such as the expected value or the median. Appendix I provides empirical details regarding the concerns mentioned above. As an example, classic FM regressions using OLS yield an out-of-sample predictive R^2 of -0.35% when evaluated across all assets, with manufacturing assets – which make up the most considerable portion of assets – yielding an out-of-sample predictive R^2 of -0.22% while agricultural assets – which make up a minor portion of all assets – only yield an out-of-sample predictive R^2 of -2.15%. Without providing further empirical details at this point, this short excursion alone provides initial evidence that we may be interested in a more granular analysis than an overall aggregate across all assets. Even if we are only interested in an overall estimate of risk prices that apply to all assets at time t, our proposed methodology still offers valuable attributes due to the more granular model insights resulting from partial derivatives' distribution. For example, a separate risk price estimation by industry or market capitalisation may still help detect biases in the estimation, serve as a general sanity check, or even be helpful for code debugging.

While the proposed risk price estimation in equations (13), (15) and (20) fundamentally rely on aggregation, we may still be interested in the asset-specific partial derivatives as they offer valuable insights into the inner model mechanics and significantly improve model interpretability. Those asset-specific partial derivatives are summarised in the $N_t \times K$ matrix $\boldsymbol{\lambda}_t$ and, besides their interpretation of being marginal portfolio returns, capture an asset's return's sensitivity to changes in the k-th risk exposure. The possibilities of analyses using those derivatives are too extensive to be discussed in their entirety in this paper. However, possible avenues include, but are not limited to, pairwise locally-weighted regression analyses, unsupervised clustering to discover intrinsic sensitivity clusters, correlation analyses investigating cross-asset sensitivities, long-short portfolio construction based on sorted derivatives, or distributional analyses.

In this paper, we limit the interpretability discussion to pairwise locally-weighted regressions, distributional analyses, and investigate the k-th partial derivatives in relation to the k-th input variable, or in relation to the j-th input variable or partial derivative, where $j \neq k$. We are particularly interested in the general distributional properties such as skewness, asset-specific outliers, and the effect of the objective function which we discuss in section 4.12. Despite the fact that we do not pursue the idea any further, it is worth mentioning that the estimation and evaluation of risk prices on subsamples naturally leads to the notion of weak and strong risk factors, where we follow the definition of Lettau and Pelger (2020a) and define strong risk factors as factors that affect all underlying assets. In contrast, weak factors only affect a subset of the underlying assets. Moreover, time-varying risk compensation further closely relates to the topic of factor timing, which is particularly heavily researched by practitioners (e.g. see Dichtl et al. (2019)).

With this paper, we, therefore, intend to contribute to the on-going debate about the time-varying estimation of risk exposures and prices of risk and model interpretability. In particular, we do not report our findings as point estimates as it is typically done, but as time-varying estimates with confidence (respectively tolerance) intervals. Moreover, in section 3.3 we further introduce non-linear and non-parametric model selection which helps to navigate the *factor zoo* in current literature. Most importantly, the methodology above is intended to offer a wide variety of applications that is universal for a large class of estimator functions and that can be tailored to the specific means of the respective stakeholder. We discuss a class of estimator functions that fulfil these above-mentioned requirements in section 3. The overarching theme, however, is that we intend to move elements of model interpretability and communications of inner model mechanics to the foreground.

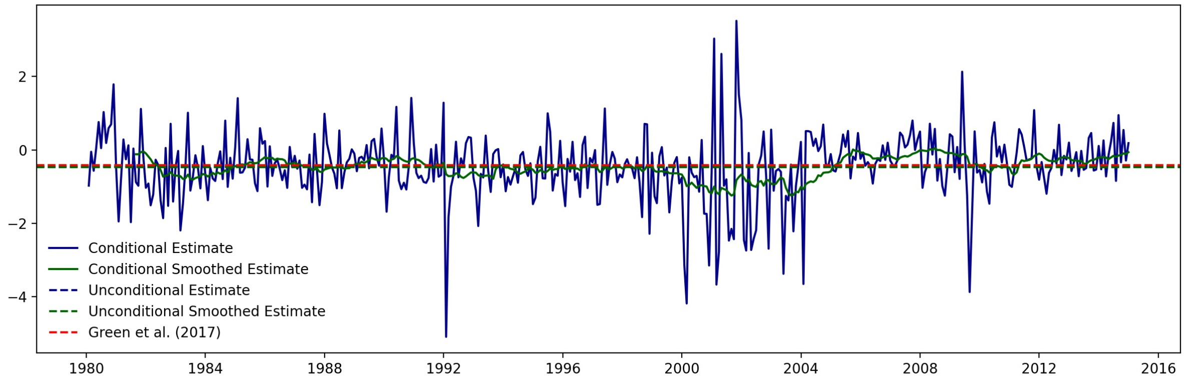


Figure 1:
Smoothing effect – example: asset-growth
The graph displays the smoothing effect exemplified by *asset growth*. The solid blue line corresponds to monthly Fama-Macbeth regressions, while the dotted blue line visualises the respective unconditional estimate. The solid green line displays the smoothed estimated through a rolling window average, with the green dotted line corresponding to the unconditional estimate from the smoothed estimate. For direct comparison, the dotted red line visualises the original estimate reported by Green et al. (2017).

2.4 Smoothing

There is considerable empirical evidence and rational economic intuition that the risk exposures and the prices of risk vary smoothly over time (e.g. see Adrian et al. (2015) or Ang and Kristensen (2012)). One of the returning arguments is that asset prices, and hence returns, are connected to economic cycles, as argued, for example, by Ferson et al. (1987), Fama and French (1989), Harvey (1989) or Campbell (1999), with economic cycles also evolving smoothly over time. As a consequence, we may wish to smooth out the risk price estimates to only allow for gradual changes over time. A complete discussion of various smoothing regimes is beyond the scope of this paper and shall not be the main focus. However, the previously introduced methodology for risk price (and exposure) estimation can also be smoothed using standard smoothing techniques, such as a rolling window estimations or backwards-looking rolling averages. Lewellen (2015), for example, reports a ten-year rolling average of Fama-MacBeth slopes, which are estimated using standard OLS regressions. As an example, consider the univariate OLS regression model, where we regress adjusted excess returns on the firm characteristic *asset growth* analogously to Green et al. (2017)[14]. The dotted lines in figure 1 display the unconditional risk price estimates, which are derived from the time-series average of the conditional estimates visualised by the solid lines. It can be seen that our unconditional estimates are almost identical to the original value reported by Green et al. (2017)[15]. However, it can also be seen that the conditional risk price estimate resulting from cross-sectional FM regressions for every $t = 1, ..., T$ are not smooth. In contrast, the solid green line visualises the smoothing effect of an exemplary 5-year rolling window average of the regression coefficients.

In this paper, we follow a similar approach. Section 2.3 discusses the fixed estimation of $\hat{\lambda}_{k,t}$, which

[14] See appendix A for a detailed variable definition

[15] See appendix E for further details on a replication study.

is the nonlinear equivalent to the conditional Fama-Macbeth estimates visualised by the solid blue line in figure 1. We smooth the risk price estimates, similarly to Lewellen (2015), by estimating $\hat{\lambda}_{k,t}$ through a five-year backwards looking rolling window approach. The rolling window approach is a balanced trade-off between achieving a desirable level of smoothness and taking potential economic cycles into account. We further investigate an expanding window approach as part of the robustness checks, with results reported in appendix I.

2.5 Empirical Inference

For empirical inference, we diverge from to the traditional approach of reporting unconditional risk prices as time-series averages following equation (11), where under the null risk prices are zero and the null hypothesis is traditionally tested using asymptotic test statistics (e.g. see Gibbons et al. (1989)). In contrast, we propose a data-driven approach to report the fixed response estimates along with time-varying confidence intervals. Thus, risk prices transition in and out of statistical significance over time, where statistical significance is defined as a confidence interval that does not include zero. This time-varying statistical significance offers the opportunity to analyse estimated prices of risk at different points in time, which may be of particular interest, for example, times of economic crisis.

In particular, we propose bootstrapping to obtain standard errors for the risk price estimates. However, we acknowledge that bootstrapping can be computationally expensive and, therefore, be an infeasible option in practice. Consequently, we also provide alternative estimation strategies, which are computationally inexpensive and receive more practical attention at the cost of being less rigorous. As pointed out by Cochrane (2009), modern asset pricing does not necessarily need to rely on asymptotic theory to calculate accurate standard errors. Monte Carlo simulations or bootstrapping offer competitive data-driven alternatives, especially in small sample, with Dixon and Polson (2019) proposing a similar approach.

Following Kapetanios (2008), we propose non-parametric bootstrapping and define the non-parametric bootstrap sample as $\{\boldsymbol{r}^*_{t+1}, \boldsymbol{c}_t^{*(1)}, \cdots, \boldsymbol{c}_t^{*(K)}\}$, where the star notation denotes some form of cross-sectional resampling from the original return and characteristics data[16]. Thus, for example, we denote $\boldsymbol{r}^*_{t+1} = (r_{j_1,t+1}, \cdots, r_{j_j,t+1}, \cdots, r_{j_{N_{t+1}},t+1})'$ a N_{t+1}-dimensional vector of re-sampled returns, where the vector of indices $(j_1, \cdots, j_{N_{t+1}})'$ is obtained by drawing with replacement from $(1, \cdots, N_{t+1})'$. The same vector of indices is used to draw from the betas as well. Kapetanios (2008) points out that the re-sampling can be adjusted for cross-sectional dependence. We further remark, that the re-sampling can also be adjusted to asset pricing specific data specifics: in section 4, we introduce the standard CRSP and Compustat US-only dataset, which is the gold standard in empirical asset pricing. It is well-known, however, that the dataset is strongly misbalanced, especially with regard to microcap stocks (e.g. see Hou et al.

[16] Note, that parametric bootstrapping would also be possible, but an in-depth discussion is beyond the scope of this paper.

(2020)). The above-mentioned re-sampling can, therefore, also follow a stratified approach, where the re-sampling is done by sub-groups, such as market capitalisation and (or) industry.

For each bootstrap iteration $b = 1, \cdots, B$, we re-fit g_t, such that $\hat{g}_{t,b}$ denotes the estimated function using the b-th re-sampled dataset and $\hat{\lambda}_{k,t}^{(b)}$ the b-th fixed response estimation. Subsequently, standard errors are $\mathrm{SE}(\hat{\lambda}_{k,t}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\hat{\lambda}_{k,t}^{(b)} - \bar{\lambda}_{k,t}}$, with $\bar{\lambda}_{k,t} = \frac{1}{B}\sum_{b=1}^{B}\hat{\lambda}_{k,t}^{(b)}$, see Efron and Tibshirani (1986). We propose confidence intervals to be $\mathrm{CI}_{k,t} = (-\rho \times \mathrm{SE}(\hat{\lambda}_{k,t}), \rho \times \mathrm{SE}(\hat{\lambda}_{k,t}))$, where $\rho \geq 3$ guided by Harvey et al. (2016) who argue that higher statistical hurdles are required to account for for multiple testing and data mining issues.

We acknowledge, that the above-introduced bootstrap approach may be infeasible in practice, if the estimation of g is computationally expensive and B is large, for example $B = 1,000$. With this paper, we intend to take real-life computational costs into account. In the following, we, therefore, introduce computationally less expensive alternatives at the cost of being less rigorous. A full description is beyond the scope of this paper and we refer readers to Kapetanios et al. (2019) for a more detailed discussion. The alternatives include:

1. Instead of re-fitting $\hat{g}$ for each bootstrap iteration, keep the original estimate and merely derive B fixed response estimates, using the re-sampled data only such that $\hat{\lambda}_{k,t}^{(b)} = \bar{\boldsymbol{\lambda}}_{k,t}(\boldsymbol{c}_{k,t}^{(b)})$ denotes the fixed response estimate derived from the b-th re-sampled characteristics data denoted by $\boldsymbol{c}_{k,t}^{(b)}$, with $\mathrm{SE}(\hat{\lambda}_{k,t}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\hat{\lambda}_{k,t}^{(b)} - \bar{\lambda}_{k,t}}$.

2. Instead of bootstrapping return and characteristics data to derive bootstrapped partial derivatives, we can also bootstrap directly from the distribution $\boldsymbol{\lambda}_{k,t}(\boldsymbol{c}_t^{(k)})$, such that $\lambda_{k,t}^{(b)}$ denotes the b-th cross-sectionally re-sampled N_{t+1}-vector of partial derivatives, with $\mathrm{SE}(\hat{\lambda}_{k,t}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\hat{\lambda}_{k,t}^{(b)} - \bar{\lambda}_{k,t}}$.

3. Instead of performing any type of bootstrap, estimate distributional percentiles directly from $\boldsymbol{\lambda}_{k,t}(\boldsymbol{c}_t^{(k)})$, if N_{t+1} is large.

4. Use the backwards-looking rolling standard deviation approach to construct *tolerance intervals* or *tolerance bands*, such that $\mathrm{TI}_{k,t} = (-\rho \times \sigma(\hat{\lambda}_{k,t}), \rho \times \sigma(\hat{\lambda}_{k,t}))$, where σ is estimated through a backwards-looking rolling window approach.

For all methods, except the fourth alternative, the re-sampling can either be done at time t, time s, where $s \neq t$, or over the entire sample. There are, however, certain real-life data issues which should be considered. If, for example, g is estimated using machine learning (which we discuss further in section 3 and is common in modern empirical asset pricing), the input data is typically pre-processed in a way that it is either cross-sectionally rank standardised or cross-sectionally rank-normalised. As a result, the data is equally distributed across all t. If in addition N_{t+1} is very large, it becomes irrelevant at what t the data was re-sampled. For the purpose of computational ease and simplicity, we follow the fourth alternative in our empirical application. Due to the real-life considerations, we utilise the term *tolerance interval* or *tolerance bands*, where risk premia are considered *empirically important* when the tolerance bands do not include zero. To impose a rigorous empirical hurdle, the tolerance bands are constructed with $\rho = 3$, in the style of Harvey et al. (2016). We appreciate that the derived empirical importance is an ad-hoc approach, but more robust estimates are infeasible in practice due to an exponentially increasing computational cost.

2.6 Comparison to Other Interpretability Measures

There exist a number of interpretability measures in current literature, in particular with regard to nonlinear estimator functions. We do not intend to provide a holistic overview here and concentrate on a brief comparison of our proposed methodology to the most commonly used measures and measures that most closely relate to our methodology, including partial dependence, individual conditional expectation, accumulated local effect, local interpretable model-agnostic explanations (LIME) and Shapley values. For a more detailed discussion, we refer readers to Doshi-Velez and Kim (2017), Molnar (2020) and Miller (2019). In general, there is a differentiation between model-agnostic and model specific interpretability, where model-agnostic interpretability is universally applicable and, thus, very flexible. We see our proposed methodology as a hybrid of model-agnostic and model specific interpretability tools, as it generalises to a large number of models under the conditions discussed in section 2.1 (e.g. see Ribeiro et al. (2016b)).

Partial dependence plots (PDP), introduced by Friedman (2001), are among the most popular graphical interpretability tools and show the functional relationship between model inputs and the model's predictions. In particular, they plot an average effect an input variable has on the prediction. Our methodology diverges from PDP in two key ways. First, our methodology is not concerned about an average effect, but instead offers asset-specific model insights, which is, therefore, much more granular. Second, partial dependence plots assume that the input variables are independent, an assumption that is not required for our methodology.

Individual conditional expectation (ICE) expand the idea of partial derivative plots to the individual item level, see Goldstein et al. (2015). In our setting, this would correspond to an individual line per asset, which can result in overcrowded plots if N is large. While this makes an analysis on asset-level possible, it does not allow for analyses in relation to other input variables. Moreover, our proposed methodology offers the opportunity to even investigate sensitivities between each other.

Neither partial dependence plots nor individual conditional expectations can deal with correlated input data. Accumulated local effect (ALE), introduced by Apley and Zhu (2020), offers a solution to this problem by visualising how model predictions change in a small window of the input data around a given point, by using differences in predictions rather than averages and taking changes in another variable into account. However, ALE does not allow for asset-level investigations.

Local interpretable model-agnostic explanations (LIME), proposed by Ribeiro et al. (2016a), intends to make any (machine learning) classifier or regressor interpretable through local interpretable approximations, which may not generalise globally, such as linear models or devision trees. While the local approximation offers an intuitive motivation for model interpretability is poses the problem of correctly defining the neighbourhood in which the approximation is taking place. With our methodology, we also offer great interpretability through the interpretability lens, but do not need to worry about local permutation and neighbourhood definitions.

Shapley values originates from game theory and coined by Shapley (1953), is a method for evaluation ”payouts” to ”players”, depending on their contribution to the ”total payout” and is defined as the average marginal contribution of a feature value across all possible coalitions. One of the key disadvantages is that the computational cost increases exponentially with the number of input variables. While the methodology has become increasingly popular, the model insights for asset pricing-specific interpretability is limited compared to the partial derivative approach, due to the computational cost (we typically have to deal with a large number of input variables) and the limited insight on asset-level.

Last but not least, Dimopoulos et al. (1995) propose a neural network specific interpretability tool that is fundamentally based on the sum of the squared norm of partial derivatives of a neural network with respect to its input. It is the tool that most closely relates to our methodology, but differs in two key ways: first, the measure is an aggregate, meaning that it does not allow for asset-specific model insights. Second, it does not allow for relative analyses in which we compare and contrast model sensitivities in relation to other sensitivities.

What separates our approach from all the above-mentioned alternatives is, that model interpretability is not a separate step and, therefore, does not involve a separate estimation. Instead, it is integral part of the estimation already as we use the partial derivatives to estimate risk prices. This makes our methodology readily available, computationally inexpensive and easily implementable. In addition, in section 3.3, we introduce an objective function that directly makes use of the partial derivatives as part of the optimisation. Furthermore, the possibility of further analyses does not stop with the partial derivatives themselves. As discussed in the previous section, we can use the partial derivatives and run, for example, unsupervised learning on top, which makes our methodology much more flexible to common alternatives.

Although we do not follow this approach further at this point, it is theoretically conceivable, that interpretability measures, such as Shapley values, are directly incorporated into an objective function, analogously to the Jacobian in equation (30). We appreciate that such objective function can be computationally expensive and, thus, infeasible in practice. However, our approach naturally eludes to the concept of penalising other forms of variable significance.

3 Machine Learning Estimation

In this section, we concentrate on the estimation of g_t and discuss a class of estimator functions $\mathcal{G}$ that satisfies the assumptions addressed in section 2 (e.g. see Farrell et al. (2021)). Although our proposed methodology generalises to a large number of possible estimator functions under the condition of differentiability, we particularly focus on deep neural networks in this section due to their recent popularity in empirical asset pricing (e.g. Messmer (2017a), Gu et al. (2020a) and Chen et al. (2019)). However, we are not necessarily plain supporters of neural networks, as their training can be computationally expensive and data-intensive, especially if their architecture is complex. Those aspects can be limiting in practice. In addition, practitioners frequently express their scepticism towards neural networks, partly because out-of-the-box architectures and training regimes do not necessarily provide significantly better (computational cost-adjusted) performances. In this context, De Prado (2018) points out what we can verify empirically: the architectural design and training of a competitively performing neural network requires in-depth domain knowledge in asset pricing, especially with regards to data-specific nuances, and machine learning theory, with plain vanilla out-of-the-box models performance being underwhelming, which we discuss in section 4.

We show that extensive hyperparameter tuning and asset pricing-specific training regimes are required to match standard and competitive benchmark models. Most importantly, however, our proposed methodology sheds light on the complex inner network mechanics that are often perceived as "black boxes". Our methodology, hence, significantly improves model interpretability and communication to various stakeholders. On the other hand, improved model interpretability can help eradicate some of the scepticism towards neural networks, making them a viable model choice.

The deep neural network literature is exceptionally vast and interdisciplinary, ranging from fields such as natural language processing over image recognition to asset pricing. Each domain requires specific tweaks to the data preprocessing, architectural design and training regimes. We find that asset pricing-specific solutions currently being underrepresented in literature, as most studies primarily apply out-of-the-box schemes. With this paper, we intend to transfer and combine interdisciplinary approaches, by, for example, incorporating estimation strategies stemming from the field of image recognition and giving them an interpretable meaning in asset pricing (see section 3.3). However, due to the sheer amount of literature, we cannot provide a holistic overview of all potential model architectures, training or regularisation schemes, so we primarily focus on fully connected feed-forward networks in this section. For a more detailed and general discussion of neural networks, we refer readers to Bishop (1995) and Goodfellow et al. (2016).

3.1 Deep Artificial Neural Networks

Fully connected feed-forward neural networks are comprised in the collective term *deep learning*, a subfield of machine learning[17]. They are among the most potent estimators due to their universal approximation capabilities, which means that they can approximate an arbitrary function with arbitrary accuracy (e.g. see Hornik et al. (1989) or Cybenko (1989)), with Farrell et al. (2021) showing that they are consistent estimators. In the empirical asset pricing literature and other disciplines, they are currently among the most popular model choices because of their competitive out-of-sample performances (e.g. see Messmer (2017a), Chen et al. (2019), and Gu et al. (2020a)). Section 1 introduced current challenges in modern asset pricing, including high-dimensional datasets and pronounced nonlinearities. Neural networks thrive in this environment through their extreme functional flexibility, which stems from the interconnection of (potentially many) so-called *layers*, enabling nonlinear estimations. Figure 2 visualises an arbitrary fully connected feed-forward neural network and displays its three core hierarchical elements: the input, hidden, and the output layer(s). While there does not exist an exact definition, the term *deep* typically refers to neural networks with more than one hidden layer. This biological imitation is the origin of the model's name. In simple terms, the input data flows from the left input layer through all hidden layers, where the nonlinear transformations take place in the neurons, to the output layer, which results in a final prediction – in our case, a return prediction analogously to equation (3).

More specifically, the model displayed consists of $L+2$ layers, where $l = 0$ denotes the K-dimensional input layer, with the dimension of the input layer corresponding to the input data dimension. Further, the displayed neural network consists of L hidden layers and an output layer, where the dimension of the output layer corresponds to the dimension of the target variable – in our case, an asset's return. Each hidden layer consists of H_l nodes that are all fully connected through weights with all H_{l-1} nodes from the preceding layer. In addition, the shaded nodes in figure 2 display the biases.

All nodes in the hidden layers perform a nonlinear transformation of their inputs through an *activation function*, where each node's input is a linear combination of the outputs from all preceding nodes plus a bias. There are many potential activation function candidates. Some of the most common activation functions are visualised in figure 3. Interestingly, and to the best of our knowledge, there does not yet exist an asset-pricing specific activation function. Thus, the activation functions displayed in figure 3 are universally applicable across domains, including asset pricing.

While there is no clear guideline for which activation function to use for asset pricing applications, rectified linear unit (ReLU) is the most commonly used activation function in recent empirical asset pricing literature. Furthermore, Farrell et al. (2021) show that neural networks designed with ReLU as activation functions in their hidden layers are consistent estimators. However, ReLU is vulnerable to the

[17] There exist other types of neural networks, such as recurrent or long-short term memory (LSTM) neural networks, which also find occasional applications in asset pricing, or at least finance or economics in general, but are beyond the scope of this paper.

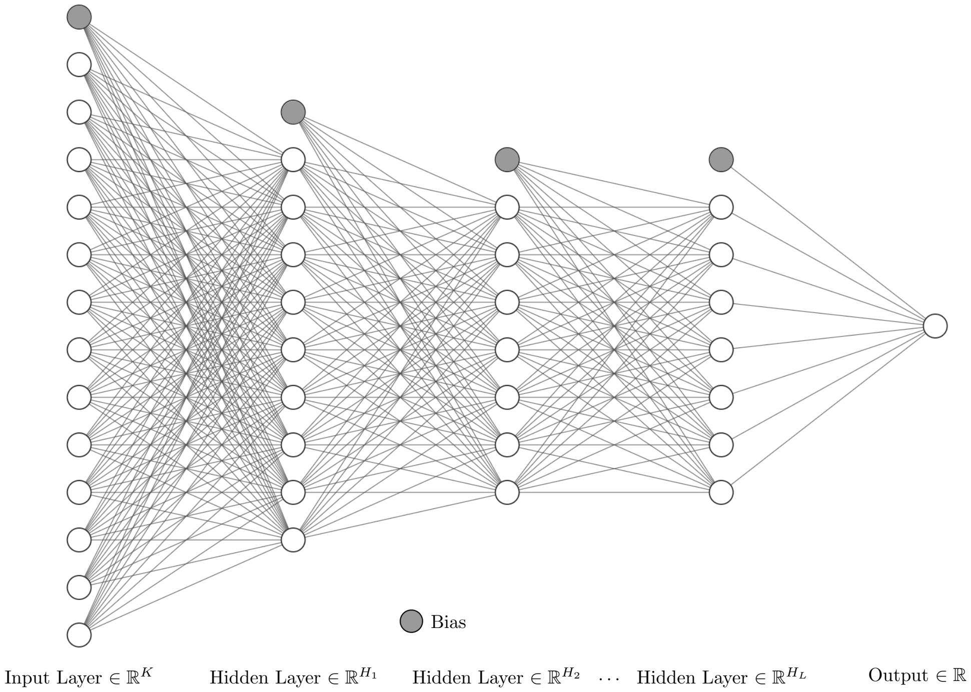
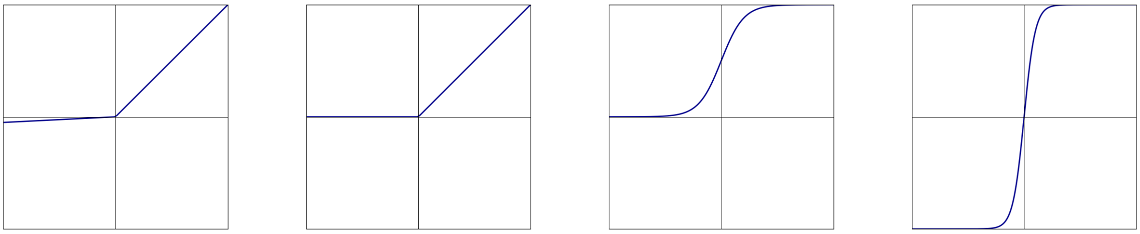

Bias
Input Layer $\in \mathbb{R}^K$
Hidden Layer $\in \mathbb{R}^{H_1}$
Hidden Layer $\in \mathbb{R}^{H_2}$
$\cdots$
Hidden Layer $\in \mathbb{R}^{H_L}$
Output $\in \mathbb{R}$

Figure 2:
Example of an arbitrary deep neural network architecture
The displayed fully-connected feedfworward neural network consists of an input layer with K input nodes (analogously to a K-dimensional input vector $\mathbf{c}_{i,t}$), $l = 1, ..., L$ hidden layers, where H_l denotes the number of nodes in the l-th layer, with $H_0 = K$ denoting the input nodes and a single output node. Furthermore, each hidden layer features a fully-connected bias, visualised by the shaded nodes.

vanishing gradient problem, which is sometimes referred to as *dying ReLU problem* (e.g. see Hu et al. (2021)). Section 3.2 introduces an exemplary optimisation algorithm, which can be used to train neural networks. What unites all training algorithms is that they are fundamentally based on gradient descent in combination with backpropagation. Thus, effective learning requires the existence of a gradient with respect to the model parameters, particularly with respect to the neural network's weight. However, due to the functional form of ReLU, the vanishing gradient problem describes a scenario in which the gradients of the network weights approach zero. Consequently, the optimisation is stuck as there is no clear direction for the next model parameter update. To the best of our knowledge, this well-known problem associated with ReLU has not yet been well-documented in the empirical asset pricing literature.

Leaky ReLU is an alternative activation function that is specifically designed to avoid the vanishing gradient problem. Instead of the output being zero (i.e. producing no model parameter gradient) when the input is negative, leaky ReLU produces a small positive slope (in PyTorch, for example, the default value is 0.01). That is, leaky ReLU computes $f(x) = \mathbb{1}(x < 0)(ax) + \mathbb{1}(x >= 0)(x)$, where $a = 0.01$. Leaky ReLU and other alternatives that are specifically designed to avoid the vanishing gradient problem have gained popularity in recent years in the general machine learning literature, with examples including, Maas et al. (2013) or Xu et al. (2015). Due to the popularity and easy implementation, the empirical application presented in section 4 also applies leaky ReLU. Besides ReLU and leaky ReLU, figure 3



(a) Leaky rectified linear unit (ReLU)

(b) Rectified linear unit (ReLU)

(c) Sigmoid

(d) Hyperbolic tangent (tanh)

Figure 3:
Common activation functions
From left to right, the figure displays the most commonly used activation functions in the general machine learning literature, and particularly in asset pricing. They range from leaky rectified linear unit (Leaky ReLU), rectified linear unit (ReLU), to sigmoid and hyperbolic tangent. Section 4 primarily focuses on an application using leaky ReLU.

summarises common activation functions in the literature, which includes the sigmoid and hyperbolic tangent. However, a full discussion of all possible activation function candidates is beyond the scope of this paper, and we refer readers to Nwankpa et al. (2018) for a more detailed discussion.

A natural critique for the choice of leaky ReLU in the context of this paper – and in particular with regard to sections 2.1 and 2.2 – is that leaky ReLU is not continuously differentiable. While mathematically correct, this issue is negligible in practice, as discussed by Goodfellow et al. (2016), since software implementations (for example, PyTorch) are prone to rounding errors. Those rounding errors make it very unlikely to land exactly on the singularity point. Additionally, even if the singularity point was reached, leaky ReLU is commonly implemented so that the right-hand side derivative is used, meaning that software implementations always produce a gradient.

The fact that there exists no clear guideline for the choice of the activation function introduces a fuzziness as part of the empirical application. The term fuzziness in this context describes that there is no right or wrong when it comes to the activation function choice in empirical asset pricing applications. In this context, section 2 introduces the term hyperparameter, which describes model parameters that cannot be optimised algorithmically as part of the optimisation procedure. Thus, hyperparameters must be set manually instead. Their (locally) optimal value can be found, for example, through cross-validation. The activation function choice is one of such hyperparameters. Theoretically, each node in each layer could apply a different activation function, which would require manual tuning to find a (local) optimum. However, to dramatically reduce the computational cost of the grid search finding (locally) optimal hyperparameter values, we assume that each node in each hidden layer applies the same activation function, namely leaky ReLU. Unfortunately, the list of hyperparameters does not only include the activation function and can be extensive. The list of potential hyperparameters includes but is not limited to: the learning rate, batch size, number of epochs, number of epochs before early stopping is activated, penalisation parameters or even architectural hyperparameters, such as the number of hidden layers or the number of nodes in each layer. Section 4.3 discusses the considered hyperparameters and their tuning regime in detail, with a particular focus on the computational cost, since the computational cost increases exponentially with the number of hyperparameters. While it is well-known that hyperparameter tuning is the computational bottleneck of neural networks, we find that current literature pays little attention to this crucial step. Thus, we intend to be as transparent as possible about our hyperparameter tuning regime to increase reproducibility.

Mathematically, figure 2 simplifies to

$$g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t) = (f_1 \circ \cdots \circ f_l \circ \cdots \circ f_L)(\mathbf{c}_{i,t};\ \mathbf{W}_t, \theta_t), \tag{21}$$

where $\mathbf{W}_t$ captures all learnable model parameters, such as weights and biases, while θ collects all remaining model (hyper-)parameters. For reasons of simplicity, we assume that the activation functions, denoted by σ, are the same in each hidden layer, such that $\sigma = \sigma_l, \forall l$. Zooming in on an arbitrary j-th node in the l-th hidden layer, the nonlinear transformation is summarised by

$$f_j^{(l)} = \sigma(\sum_{j=1}^{H_{l-1}} w_{ji}^{(l-1)} x_i^{(l-1)} + b_{ji}^{(l-1)}) = x_j^{(l)}, \tag{22}$$

where $w_{ji}^{(l-1)}$ denotes the weight connecting the i-th node from the preceding layer $l-1$ with the j-th node in the l-th hidden layer. The same notation applies for the biases. The output of the i-th node from the preceding $l-1$-th layer is denoted by $x_i^{(l-1)}$, where for the input layer ($l = 0$) $\mathbf{x}_0 = \mathbf{c}_{i,t}$. Consequently, we can rewrite equation (21) as

$$g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t) = \mathbf{W}^{(L)}\sigma(\cdots\sigma(\mathbf{W}^{(l)}(\cdots\sigma(\mathbf{W}^{(0)}\mathbf{c}_{i,t} + \mathbf{b}_0)) + \mathbf{b}_l)\cdots) + \mathbf{b}_L, \tag{23}$$

where $\mathbf{W}^{(l)}$ denotes a $H_{l-1} \times H_L$ matrix, capturing all weights connecting the $l-1$-th with the l-th layer. The same notational logic applies to the biases. Note that equation (23) means that we do not apply another nonlinear transformation through an activation function in the output node.

3.2 Standard Objective Function and Optimisation

This section reviews the standard objective function and stochastic optimisation algorithm that we use in this paper to train deep neural networks and are commonly used in the literature. In addition, we propose an alternative objective function that allows for non-linear model selection and sensitivity penalisation in section 3.3.

One of the main reasons neural networks have become popular in the empirical asset pricing literature is their data-driven estimation approach. This form of estimation does not impose any substantial direct restrictions on the functional form of the estimator function. During an iterative process, called *learning*, we minimise a stochastic objective function through first-order gradient-based optimisation to find an *optimal* set of (learnable[18]) model parameters[19].

Such a data-driven approach is particularly appealing in data-rich environments and when the underlying true data-generating process is unclear, as in return data. However, the main disadvantage is that fitting a neural network does not necessarily involve any economic or financial theory. Therefore, letting the data speak for itself is prone to data-snooping or allows for statistical estimators that have little economic intuition. We discuss solutions to these challenges in section 3.3.

We find that domain-specific knowledge is still crucial regarding the data pre-processing, the architectural design of the network or even the objective function. In this section, we review the standard procedure of finding an optimal set of learnable model parameters, where learnable parameters refer to model parameters that are not found manually but iteratively updated during the training procedure. At the heart of the training of neural networks and finding this set of model parameters is an objective function that is minimised iteratively through a stochastic optimisation algorithm, evaluated concerning the actual data generating process (DGP). Algebraically, this is summarised by

$$\mathbb{J}_t(\mathbf{W}_t, \theta_t) = \mathbb{E}_{\mathbf{c}_{i,t}, \mathbf{r}_{i,t+1} \sim p_{DGP}}[L(g(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t); \mathbf{r}_{i,t+1})], \tag{24}$$

where L denotes a loss function, and the time index indicates that the objective function corresponds to the T cross-sectional estimations (e.g. see Goodfellow et al. (2016)). However, the actual data generating process is unknown. Hence, we replace the underlying DGP with the empirical distribution (e.g. see Messmer (2017a)), such that equation (24) can be re-written as

$$\hat{\mathbb{J}}_t(\mathbf{W}_t, \theta_t) = \mathbb{E}_{\mathbf{c}_{i,t}, \mathbf{r}_{i,t+1} \sim p_{\mathbf{c}_{i,t}, \mathbf{r}_{i,t+1}}}[L(g(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t); \mathbf{r}_{i,t+1})], \tag{25}$$

with the objective function

$$\mathcal{L}(\mathbf{c}; \mathbf{W}_t, \theta_t) = \frac{1}{2N_t} \sum_{i=1}^{N_t} (r_{i,t+1} - g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t))^2 + \Omega(\mathbf{W}_t), \tag{26}$$

where we follow the standard assumption that the loss function L is the mean-squared-error. Equation (25) extended equation (24) with a penalty term, denoted by $\Omega(\mathbf{W}_t)$, which analogously to Friedman et al. (2010), is defined as

$$\Omega(\mathbf{W}_t) = \lambda \left((1-\alpha) \left\| \mathbf{W}_t \right\|_1^1 + \alpha \frac{1}{2} \left\| \mathbf{W}_t \right\|_2^2 \right). \tag{27}$$

[18] Not all model parameters are learnable and, therefore, found through training. The optimal values of hyperparameters, such as the learning rate or penalty terms, are found manually through cross-validation.

[19] The term *optimal* generally refers to model parameters that lead to a local minimum, as a global minimum is elusive.

The penalty term in equation (27) allows for L_1, L_2 or L_1 and L_2 norm weight penalisation (note that the penalty is only applied to the weights, not the biases). In general, any form of regularisation is aimed at preventing in-sample overfitting in order to achieve better out-of-sample generalisations. There exists a variety of ways to introduce regularisation, which we discuss further in section 4.3, and include early stopping or batch normalisation in addition to the weight penalties from equation (27).

The *optimal* set of weights is found first-order gradient-based optimisation in combination, along with backpropagation which is used to calculate the gradient of the loss function with respect to the learnable model parameters (e.g. see Rumelhart et al. (1986)), such that

$$\mathbf{W}_t^* = \min_{\mathbf{W}_t} \sum_{i=1}^{N_t} \big(r_{i,t+t} - g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta)\big)^2 + \Omega(\mathbf{W}_t) \tag{28}$$

denotes the (learnable) model parameters which minimise the objective function. Any additional model hyperparameters, such as the penalty parameter λ in equation (27), are tuned through, for example, cross-validation, which we discuss in section 4.3. It follows that the estimated function becomes $\hat{g} = g(\cdot \quad; \mathbf{W}_t^*)$.

The optimisation of the neural network's objective function is almost exclusively non-convex (e.g. see Messmer (2017a)). As a consequence, the objective function is optimised numerically, using gradient-based optimisers. One of the most common optimisers used in current literature is stochastic gradient descent (SGD). However, there are some challenges regarding the use of SGD, including the difficulty of finding an optimal learning rate (if the rate is too small, convergence is too slow, if the rate is too large, it can hinder convergence), and the fact that SGD applies the same learning rate to all learnable model parameters. To tackle those challenges, we focus on the Adaptive Moment Estimation or *Adam*, proposed by Kingma and Ba (2015)), as a gradient descent optimisation algorithm to solve equation (28) and find that Adam performs well empirically. The algorithm is summarised in algorithm 1.

Adam allows for adaptive learning rates for each learnable model parameter and learning rate decay, which is beneficial when a minimum is approached. The optimisation dynamics of Adam is often described with the analogy of a heavy ball with friction rolling down the loss function, see Heusel et al. (2017). Similarly to other optimisers such as Adadelta, Adam stores an exponentially decaying average of past partial derivatives of the loss function with respect to the model parameters in addition to an exponentially decaying average of past gradients, similarly to optimisers with momentum (e.g. Nesterov accelerated gradient descent). For a more detailed overview on gradient-based optimisation, see Ruder (2016).

Algorithm 1: Adam algorithm, analogously to Kingma and Ba (2015). g_i^2 indicates the element-wise square $g_i \odot g_i$. Empirically well-performing default values are $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. With β_1^i and β_2^i we denote β_1, respectively β_2 to the power i, where i is the iteration step.

```
Require: $\alpha$: stepwise
Require: $\beta_1, \beta_1 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(w)$: Stochastic objective function with learnable parameters $w$
Require: $w_0$: Initial parameter vector
$m_0 \leftarrow 0$ (Initialise 1st moment vector) $v_0 \leftarrow 0$ (Initialise 2nd moment vector)
$i \leftarrow 0$ (Initialise iteration step)
while $w_i$ not converged do
    $i \leftarrow i + 1$
    $g_i \leftarrow \nabla_w f_i(w_{i-1})$ (Get gradients w.r.t. stochastic objective at iteration $i$)
    $m_i \leftarrow \beta_1 \cdot m_{i-1} + (1 - \beta_1) \cdot g_i$ (Update biased first moment estimate)
    $v_i \leftarrow \beta_1 \cdot v_{i-1} + (1 - \beta_1) \cdot g_i^2$ (Update biased second raw moment estimate)
    $\hat{m}_i \leftarrow \frac{m_i}{1-\beta_1^i}$ (Compute bias-corrected first moment estimate)
    $\hat{v}_i \leftarrow \frac{v_i}{1-\beta_2^i}$ (Compute bias-corrected second raw moment estimate)
    $w_i \leftarrow w_{i-1} - \alpha \cdot \frac{\hat{m}_i}{\sqrt{\hat{v}_i}+\epsilon}$
end
return $w_i$ (Resulting parameters)
```

3.3 Jacobian Objective Function

Section 3.2 introduced a standard objective function and optimisation strategy. One of the key advantages of the objective function in equation (26) in combination with gradient descent-based optimisation is its purely data-driven approach: a local minimum is reached without imposing any economic or financial theory. The resulting statistical estimator is particularly advantageous in data-rich environments, where the true data-generating process is unknown, and we wish to let the data speak for itself. There are, however, asset-pricing specific conditions, which should be considered during training.

First, return data is characterised by a notoriously low signal-to-noise ratio. Therefore, a purely data-driven approach poses the danger of data-snooping and finding apparent statistical relationships even though they are not there. In combination with our previous discussion about model interpretability, this is particularly disadvantageous from a communication perspective. For example, a neural network might find a nonlinear statistical link between specific firm characteristics and returns for which it is difficult to find an economic justification. In turn, an investor might be sceptical about the model and decide to use a simpler model with more interpretable and economically meaningful firm characteristic interactions instead at the cost of inferior model performance.

Second, despite the abundance of potential regressors, there is considerable empirical evidence, such as presented by Kelly et al. (2018), Kozak et al. (2020) or Freyberger et al. (2020), in favour of an economically motivated sparse representation of the asset pricing model using a reduced number of regressors. Such a parsimonious representation can either be achieved through prior data pre-selection, where only a small number of regressors is hand-picked before training the neural network or through model selection analogously to, for example, Lasso.

Third, as part of the regularisation strategy, we wish to limit an individual asset's influence on the predictions to make the model performance more robust. As discussed previously, compared to other disciplines, the standard datasets used in empirical asset pricing comprise a comparatively small number of observations. This relative data sparsity is combined with the severe unbalancedness of financial datasets, where microcaps make up the vast majority of observations but are the least relevant stock group for investors due to their illiquidity and high transaction costs. We are, therefore, in a dilemma: do we keep all observations, including microcaps, in our training dataset, or do we throw out a significant portion of the dataset, making training complex models more difficult. Limiting an asset's influence on the model performance can help mitigate this dilemma, as it allows us to keep all observations while also ensuring that microcaps are not primarily driving model performance.

On this basis, we propose an objective function that helps mitigate those asset pricing specific challenges, particularly by allowing for model selection as part of the training algorithm. Our proposed objective function can be used for an arbitrary choice of g_t under the assumption that gradient descent optimisation is used in order to determine its optimal parameters and the standard assumptions discussed in section 2.

Let

$$\mathcal{L}_t(\mathbf{c}, \mathbf{W}_t, \theta_t; \mathbf{J}_t) = \frac{1}{2N_t}\sum_{i=1}^{N_t}(r_{i,t+1} - g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t))^2 + \Omega(\mathbf{J}_t) \tag{29}$$

be the objective function, where Ω denotes a "Jacobian"-based regulariser, defined as

$$\Omega(\mathbf{J}_t)) = \lambda\left((1-\alpha)\left\|\mathbf{J}_t\right\|_1^1 + \alpha\frac{1}{2}\left\|\mathbf{J}_t\right\|_2^2\right), \tag{30}$$

where $\|\cdot\|_p^p$ denotes the Frobenius norm of the "Jacobian" $\mathbf{J}_t \in \mathbb{R}^{N_t \times K}$. We use quotation marks to indicate that $\mathbf{J}_t$ is not strictly the Jacobian. Instead, each row in $\mathbf{J}_t$ corresponds to the partial derivatives of g_t with respect to the asset's K firm characteristics, such that the dimension of $\mathbf{J}_t$ correspond to the dimension of the input data. We acknowledge the nuances in the Jacobian definition, but for the purpose of simplicity we omit the quotation marks in the following.

Moreover, we differentiate between two type of Jacobian regularisation: element-wise and column-mean. In the case of element-wise Jacobian regularisation, each element of the Jacobian enters the penalty term, for example when $p = 2$, as $\|\mathbf{J}_t\|_2^2 = (\sum_{i=1}^{N_t}\sum_{k=1}^{K} d_{ik,t}^2)^{\frac{1}{2}}$, with $d_{ik,t} = \frac{\partial g_t(\cdot)}{\partial c_{ik,t}}$. This means that we penalise the sum of each element or partial derivative in $\mathbf{J}_t$. In contrast, columns-mean Jacobian regularisation, which we denote by $\bar{\mathbf{J}}_t$, penalises the sum of the column-mean for each firm characteristic, such that for example, when $p = 2$, such that $\left\|\bar{\mathbf{J}}_t\right\|_2^2 = (\sum_{k=1}^{K}(\frac{1}{N_t}\sum_{i=1}^{N_t} d_{ik,t})^2)^{\frac{1}{2}}$. Although we do pursue this path further in this paper, other variants of penalisations are imaginable, such as a value weighted average penalisation, where the weights correspond to the inverse of an asset's market capitalisation, meaning that microcaps are penalised more heavily. Although we do not pursue this idea further in the empirical application, it is conceivable that the objective function includes a penalty term for both – the Jacobian and the model parameters – such that

$$\mathcal{L}_t(\mathbf{c}, \mathbf{W}_t, \theta_t; \mathbf{J}_t) = \frac{1}{2N_t}\sum_{i=1}^{N_t}(r_{i,t+1} - g_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta_t))^2 + \Omega(\mathbf{J}_t, \mathbf{W}_t), \tag{31}$$

with

$$\Omega(\mathbf{J}_t, \mathbf{W}_t) = \lambda_W \left((1-\alpha_W)\left\|\mathbf{W}_t\right\|_1^1 + \alpha_W \frac{1}{2}\left\|\mathbf{W}_t\right\|_2^2\right) + \lambda_J \left((1-\alpha_J)\left\|\mathbf{J}_t\right\|_1^1 + \alpha_J \frac{1}{2}\left\|\mathbf{J}_t\right\|_2^2\right), \tag{32}$$

Gradient regularisation offers three key advantages. First, any form of regularisation is generally helpful for preventing the model from overfitting. As in equation (28), weight regularisation is intended to keep the outputs of each hidden layer away from the saturated regions of the activation function, as, for example, discussed by Goodfellow et al. (2016). On the other hand, gradient regularisation penalises large output changes due to small changes in input. Thus, Jacobian regularisation enforces a smoothness prior (e.g. see Drucker and Cun (1992) or Hoffman et al. (2019)) and thereby increases model robustness. In general, the idea of gradient regularisation as a method to increase model robustness is not new to machine learning literature. In the field of image recognition, in particular, it is a common strategy to deal with *adversarial examples*, where minimal changes in the input can lead to grossly wrong classifications (e.g. see, Goodfellow et al. (2015), Lyu et al. (2016), Varga et al. (2018) or Barrett and Dherin (2020)). With this paper, we propose Jacobian-based regularisation for asset pricing, where due to the low signal-to-noise ratio present in financial data, it is also desirable to be protected against large changes in output due to small changes in input.

Second, gradient regularisation is deeply rooted in economic theory and motivated by economic intuition. Unlike image recognition, where it is difficult to interpret the partial derivatives, they have a precise and well-defined economic meaning in our setup: risk prices. The penalty in equation (30) shrinks specific risk prices in equation (20) towards zero, or even precisely to zero depending on the norm and size of the penalty. There exists a large body of literature in asset pricing, such as Kelly et al. (2018), Kozak et al. (2020) or Freyberger et al. (2020), who argue for an economically motivated sparse representation of the asset pricing model using a reduced number of risk factors. Our proposed objective function makes precisely this possible and is, therefore, a generalisation of model selection, where model selection deals with the question of which firm characteristics have incremental predictive power, given all other characteristics. Moreover, due to the time-varying nature of our proposed methodology, we allow this variable selection to vary over time. Finally, our proposed objective function has a low computational cost and is directly implementable as part of the training algorithm.

3.4 Linear Benchmarks

To make the neural networks' empirical results more comparable, we benchmark them against a range of commonly used alternative models. In addition to the ordinary least-squares (OLS) and neural network estimations discussed in sections 2.1 and 3.1, we also investigate weighted least squares (WLS) and three penalised linear regressions, including Ridge, Lasso and Elastic Net. While the primary numeric evaluation metric constitutes the out-of-sample performance (see section 2.6 for performance definitions), we also consider a model's interpretability and general insights that we can gain from it to paint the complete model comparison picture. This total evaluation package includes, but is not limited to, the time-varying model performance or a model's performance consistency, the firm characteristics the model puts the most weight on, how those change over time, or the interpretability of a model's output.

3.4.1 Weighted Least Squares

In the empirical asset pricing literature, it is well-documented that market capitalisation significantly impacts model performance. For example, in the standard US-only dataset sourced from CRSP and Compustat, microcaps make up nearly 60% of all stock while their total market capitalisation merely represents around 3% of the total market capitalisation of all stocks on average. The two main problems arising from this are that OLS overstates the importance of microcaps, while at the same time, microcaps are much harder to trade on in practice due to their limited liquidity and higher trading costs. As a consequence, the least-squares loss function is replaced with

$$L(g(\cdot)) = \frac{1}{2N_t} \sum_{i=1}^{N_t} w_{i,t} (r_{i,t+1} - g_t(\mathbf{c}_{i,t}))^2, \tag{33}$$

where g_t is linear in firm characteristics, analogously to standard OLS, with $g_t = \boldsymbol{C}_t \boldsymbol{\lambda}_t$ and $\hat{\boldsymbol{\lambda}}_t = (\hat{\boldsymbol{C}}_t^{'} \boldsymbol{W} \hat{\boldsymbol{C}}_t)^{-1} \hat{\boldsymbol{C}}_t^{'} \boldsymbol{C} \boldsymbol{r}_{t+1}, \;\; \forall t = 1, ..., T$, and $\boldsymbol{W}$ is a diagonal matrix containing the weights.

We use a stock's market value of equity at time t as the weight $w_{i,t}$. The weighted least squares objective in equation (33) allows us to place more weight on statistically or economically informative observations. In particular, due to the limited liquidity and increased trading costs of microcaps, WLS tilts estimates towards higher liquidity and lower trading costs, as the objective function underweights stocks with a smaller market capitalisation in favour of large stocks. Therefore, it helps mitigate the challenges arising from the unbalancedness present in typical asset pricing datasets while being computationally very cheap at the cost of imposing a strong assumption about the linear functional form of g_t.

3.4.2 Penalised Linear Regressions

The main econometric disadvantage of the unregularised linear benchmarks (OLS and WLS) is that they fail in high-dimensional settings when the number of predictors K reaches the number of cross-sectional observations N_t, as the model becomes inefficient or even inconsistent. In a low signal-to-noise environment as present in return data, these concerns are particularly troublesome. In addition, and as pointed out by Han et al. (2019), ordinary or weighted least squares may suffer from the problem of overfitting in high-dimensional multivariate regressions. To avoid overfitting, parsimonious regression models with a reduced number of parameters are crucial. As briefly discussed in section 3.2, and more specifically in equation (26), the most common machine learning solution for imposing parameter parsimony is to append the objective function with a penalty term, such that the new objective function is

$$\mathcal{L}_t(\mathbf{C}_t, \boldsymbol{\lambda}_t) = \frac{1}{2N_t} \sum_{i=1}^{N_t} (r_{i,t+1} - g_t(\mathbf{c}_{i,t}))^2 + \Omega(\boldsymbol{\lambda}_t), \tag{34}$$

where $\Omega(\boldsymbol{\lambda}_t)$ denotes a penalty, defined as

$$\Omega(\boldsymbol{\lambda}_t) = \lambda \left((1-\alpha) \left\| \boldsymbol{\lambda}_t \right\|_1^1 + \alpha \frac{1}{2} \left\| \boldsymbol{\lambda}_t \right\|_2^2 \right) \tag{35}$$

and g_t is a linear estimator function. The hyperparameters, such as the penalty term λ and α, are found manually through cross-validation (e.g. see Friedman et al. (2010)). This form of regularisation mechanically diminishes a model's in-sample performance to boost its out-of-sample performance stability. As argued by Gu et al. (2020b), the model's performance stability is particularly improved if regularisation reduces the fit of noise while preserving the signal fit.

In equation (35), the case of $\alpha = 0$ corresponds to the lasso or L_1 penalisation, which is capable of model selection due to the geometry of the objective function (e.g. see Hastie et al. (2009)). Lasso, therefore, imposes sparsity. The case of $\alpha = 1$ corresponds to the ridge regression, or L_2 penalisation, which shrinks parameters towards (but not exactly) to zero. This form of shrinkage limits the magnitude of coefficients and, hence, prevents coefficients from becoming too large. All intermediate cases, where $0 < \alpha < 1$, correspond to the elastic net, a lasso and ridge regression mixture.

The penalised linear regression model and its objective function clarify the link between our proposed methodology and the linear regression case. The main difference between standard objective functions, such as equation (26), for neural networks and objective functions for the linear case, such as equation (34), is that in the linear case, the coefficients are penalised, while in the neural network case we penalise the network's weights, which are not equivalent to coefficients. On the other hand, our proposed objective function in section 3.3 uses the same analogy to the objective function in equation (34) as the penalisation of the partial derivatives is the non-linear generalisation of the penalisation of the coefficients.

Last but not least, we acknowledge that there exists a myriad of potential additional benchmarks such as principal components, partial least squares or instrumented principal components (e.g. see Kelly et al. (2018)). However, with this paper, we do not intend to present a comprehensive model horserace. Instead, we aim to shift the focus to our proposed methodology of increased interpretability using neural networks and keep the model benchmarks with OLS, WLS, Lasso, Ridge and Elastic Net to an extensive minimum.

4 An Empirical Study of U.S. Equities

This section provides a detailed overview of the investment universe we consider, consisting of monthly stock returns and 103 firm characteristics, which we construct analogously to Green et al. (2017). We show that deep neural networks are a reasonable model choice from a performance perspective and have substantial interpretability advantages. We explicitly do not wish to shift existing out-of-sample performance frontiers and are not plain supporters of deep neural networks. However, this section shows that conditional on the model choice, deep neural networks add significantly from an interpretability perspective leveraging our proposed methodology, compared to standard linear benchmarks. In particular, we show which risk factors help explain cross-sectional returns over time and how our results compare to existing research such as Green et al. (2017).

4.1 Data

We follow standard procedure, source monthly stock returns from CRSP for firms listed on NYSE, AMEX, and NASDAQ, and obtain firm-level data from Compustat and I/B/E/S. We use the Effective Federal Funds rate as a proxy for the risk-free rate of return, which we source from the Federal Reserve Economic Data repository (FRED) and obtain price level data directly from the US Bureau of Labor Statistics (BLS). Stock returns are adjusted for delistings and are matched in month t with firm characteristics that are most recently available to an investor at the beginning of month t. We assume that firm characteristics based on annual accounting information become available at least six months after the fiscal year ending and firm characteristics based on quarterly accounting data become available at least four months after the fiscal quarter ending to avoid information leakage.

The sample period spans over 492 months from January 1980 to December 2020, yielding over four decades of data. The sample period choice in empirical asset pricing is somewhat arbitrary. This paper primarily focuses on fitting non-linear machine learning models whose training is data-intensive and, therefore, requires a minimum amount of data. However, data availability varies considerably over time. For example, the NASDAQ only began its operations in the early 1970s, adding a significant number of stocks to the sample at that time. Moreover, I/B/E/S data only became available in the late 1980s and early 1990s. Thus, our sample choice yields a balanced trade-off between data availability, comparability to other existing research and timespan over economically relevant and heterogeneous periods. Figure 4 visualises the time-series properties of an unrestricted US-only investment universe.

We consider stocks with share code[20] 10 or 11 only and impose no restrictions on the company size, industry or time since a company has been listed on an exchange to minimise data-snooping biases (e.g. see Lo and MacKinlay (1990)). We adopt standard company size definitions based on market capitalisation analogously to Fama and French (1993) and define microcaps as stocks whose market capitalisation is smaller than the 20th percentile of the market equity of NYSE stocks. In contrast, the market capitalisation of small stocks is greater than the 20th percentile but smaller than the median (50th percentile) market capitalisation for NYSE stocks. Subsequently, large stocks are all remaining stocks whose market capitalisation is larger than the median market equity of NYSE listed stocks. Figure 4 and table 1 reveal that microcaps merely represent around 3% of aggregated market capitalisation while making up around 60% of the total number of stocks. This well-known misbalance poses specific difficulties. Conversely, the general machine learning mantra that more data is better than less data also holds in asset pricing. From this point of view, microcaps provide the vast majority of training data. On the other hand, empirical asset pricing frequently faces criticism from practitioners as many findings heavily rely on microcaps, but microcaps are generally less liquid and more expensive to trade. We acknowledge the existence of such problems and explicitly discuss those in our empirical analysis.

We build a large collection of stock-level firm characteristics based on the cross-section of stock returns literature. Similarly to Gu et al. (2020b), and for comparability, we adopt the firm characteristic definitions of Green et al. (2017) and compute 103 predictive firm characteristics, with 63 of those being based on annual data. In contrast, 15 are based on quarterly data, and 25 are of monthly frequency[21]. Appendix A provides a detailed description of the variable definitions and assumptions that go into the firm characteristics calculations. Moreover, we acknowledge that Fama and French (1993) or Hou et al. (2020) propose slightly divergent definitions for certain characteristics such as book-to-market. However, we intend to primarily focus on our proposed methodology and leave diverging characteristic definitions open to the reader.

4.2 Data Cleaning and Preprocessing

We follow Green et al. (2017) and impose minor restrictions on our investment universe. More specifically, we only consider stocks with observable month-end market capitalisation and non-missing common equity and one-month momentum. While those restrictions downsize the unrestricted universe, our investment

[20] US-based common stocks are identified with share codes 10 and 11, where the second digit refers to securities that have not been further defined (0), and securities that need not be further defined (1), see https://wrds-www.wharton.upenn.edu/data-dictionary/form_metadata/crsp_a_stock_msf_identifyinginformation/shrcd/.

[21] We are particularly grateful that Jeremiah Green publishes his SAS code on his website, https://drive.google.com/file/d/0BwwEXkCgXEdRQWZreUpKOHBXOUU/view. We *translated* his SAS into Python. In particular, appendix A shows that our data has a median correlation of 98.8% with Green's dataset and explains the origin of minor differences.

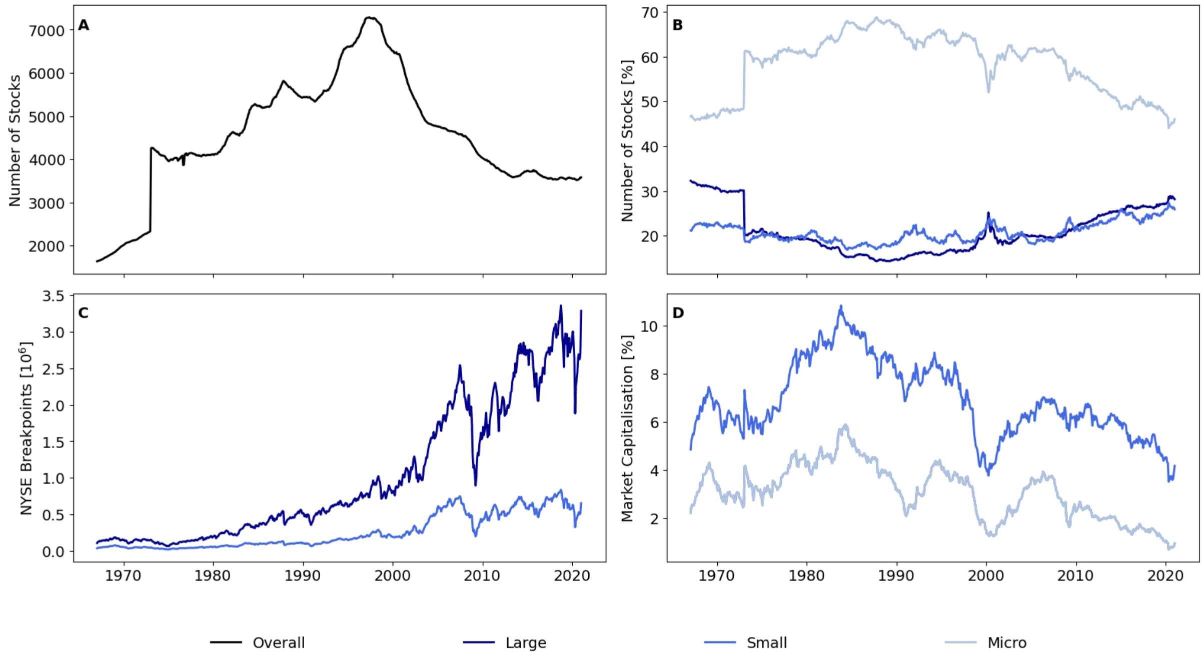


Figure 4:
Summary of unrestricted investment universe, January 1967 - December 2020
Microcaps are smaller than the 20th percentile of market equity of NYSE stocks, while small stocks have a market cap of larger than the 20th percentile of NYSE stocks but smaller than the median (50th percentile), large stocks are larger than the median of NYSE market equity. Analogously to Hou et al. (2020), we report the time-series properties of the size clusters from January 1967 to December 2020. Panel A shows the total number of investable stocks per month. Panel B shows the the time-series of the total number of microcap, small and big stocks over time as a fraction of the total number of stocks, expressed in percentage. Panel C displays the NYSE breakpoints for microcaps and small stocks. Panel D plots the total market capitalisation of microcaps and small stocks as a fraction of total market capitalisation, expressed in percentage. Panel B and D show that on average, microcaps make up nearly 60% of the total number of stocks while only representing circa 3% of aggregated market capitalisation.

universe still includes over 19,600 individual stocks, averaging 4,473 per month. Each month, to control outliers, we winsorise continuous characteristics at the 1st and 99th percentile (and positively bounded characteristics at the 99the percentile only). Moreover, we consider two different data preprocessing regimes. First, we cross-sectionally standardise all characteristics and replace missing values with the post-standardisation mean of zero. The advantage of this approach is the convenient interpretation of coefficients as they correspond to the expected change in returns, given a unit change in standard deviation in the characteristic of interest, keeping all other characteristics constant. Additionally, it makes our results directly comparable to those of Green et al. (2017).

While this standardisation imposes a common mean of zero and a standard deviation of one for each characteristic, it does not affect the vastly different magnitudes of each characteristic. However, machine learning models are known to perform best when their input features (in this case, firm characteristics) are identically distributed and of the same magnitudes. Therefore, firm characteristics are typically

	Number of firms	% of total market cap.	Value-weighted adj. excess returns		Equal-weighted adj. excess returns		Cross-sectional std of adj. excess returns
			Mean	Std	Mean	Std	
Market	4473	100.00	0.69	4.49	0.87	5.83	17.05
Large	916	90.78	0.69	4.42	0.76	4.93	8.81
Small	932	6.60	0.82	5.96	0.82	6.04	11.95
Micro	2625	2.62	0.68	6.34	0.95	6.51	20.40
All-but-micro	1848	97.38	0.69	4.47	0.79	5.41	10.57

Table 1:
Data summary – average monthly values:
The table shows averages of monthly value- and equal-weighted average returns, monthly cross-sectional standard deviations (Std) of returns for all stocks (Market) and microcaps (Micro), small, big, and all-but-micro stocks. The table also shows the average number of stocks and the average percentage of the aggregate market capitalisation in each size group each month.

cross-sectionally rank normalised such that

$$\tilde{c}_{ik,t-1} = \frac{\text{rank}(c_{ik,t-1})}{N_t + 1} \tag{36}$$

is the rank-normalised characteristic and lies in $[0, 1]$, examples include Gu et al. (2020b) or Freyberger et al. (2020). The advantage of this approach is an increased model performance, at the cost of a more difficult interpretation of the coefficients as the firm characteristics in equation (36) are uniformly distributed. In order to benefit from the advantages of both worlds, we cross-sectionally rank-normalise the data following

$$\tilde{c}_{ik,t-1} = \Phi^{-1}\left(\frac{\text{rank}(c_{ik,t-1} - \bar{c}_{k,t-1})}{N_t + 1}\right), \tag{37}$$

where $\bar{c}_{k,t-1}$ denotes the cross-sectional mean of the k-th characteristic at time $t-1$, and Φ^{-1} denotes the truncated normal quantile function. Similarly to equation (36), the cross-sectionally normalised characteristics in equation (37) are also bounded as we force them to lie in $[-3, 3]$. However, because they are (nearly) normally distributed, the interpretation of the coefficients is easier as they correspond to a unit change in standard deviation in rank of the characteristic of interest, keeping all else constant.

4.3 Estimation strategy

The key advantage of using machine learning compared to classical linear models, such as OLS, in FM-like regressions is the possibility to actively mitigate in-sample overfitting through regularisation, which yields much more robust results. We apply a variety of regularisation techniques, with gradient regularisation (see section 3.3) being one of them. In addition, we also apply early stopping (e.g. see Finnoff et al. (1993)), learning rate shrinkage as part of the Adam optimiser (Kingma and Ba (2015)), dropout (e.g. see Srivastava et al. (2014)) and batch normalisation (e.g. see Ioffe and Szegedy (2015)). Regularisation strategies vary across disciplines and depend on the use case and the network type. In general, there does not seem to be a standardised consensus for which regularisation strategy generally works best.

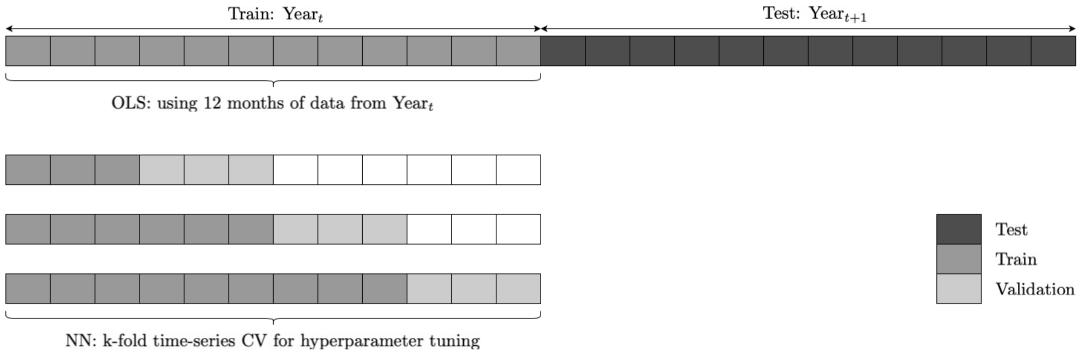

Train: Year_t
Test: Year_{t+1}
OLS: using 12 months of data from Year_t
Test
Train
Validation
NN: k-fold time-series CV for hyperparameter tuning

Figure 5:
Sample splitting strategy
This figure displays the sample splitting strategy used for cross-validation as part of the hyperparameter tuning. We follow a time-series approach in which the timely order of the data is preserved in order to prevent information leakage. The displayed splitting strategy is utilised for all neural networks and penalised benchmark regressions, including Ridge, LASSO and Elastic Net, which are introduced in section 3.4. In section 4.4, we discuss in greater detail that we differentiate between splitting the data by calendar versus financial year.

For example, Srivastava et al. (2014) argue to insert batch normalisation before the activation function, whereas Bianchi et al. (2021) apply batch normalisation after the activation but before dropout. Li et al. (2019) argue that conflicting problems between batch normalisation and dropout are best mitigated by applying dropout after batch normalisation, conditional on a low dropout probability. van Laarhoven (2017) shows that weight decay combined with batch normalisation does not have a regularisation effect but influences the effective learning rate instead. We apply batch normalisation after the activation function but before dropout[22].

We follow standard machine learning practice and split our data into three subsamples: a training set which is used for training the model, a validation set which we use for evaluating the model (particularly as part of the hyperparameter tuning), and a test set which we use for out-of-sample assessment. There exists a variety of different schemes for splitting the data. Gu et al. (2020b) or Bianchi et al. (2021), for example, use an expanding window approach. We intend to resemble FM-like regressions as closely as possible. For that reason, we do not use an expanding window approach but refit our model periodically using a fixed window. However, we diverge from the original FM approach[23] and refit our model annually in order to reduce computational cost.

Concerning the data splitting, the size of the training and validation split is ultimately an empirical question and largely depends on data availability (e.g. see Arlot and Celisse (2010)). The type of data splitting (for example, k-fold, stratified k-fold, or shuffle split) depends on the empirical application. With financial data, we must avoid information leakage from the future when splitting the data. Therefore, we apply k-fold time-series splitting, which is visualised in figure 5.

[22] Note that backtesting a large number of different strategies is itself a form of overfitting (e.g. see De Prado (2018)), in combination with low dropout probabilities. In order to minimise problematic issues arising from this backtest overfitting, we only used a small subsample of randomly chosen 24 consecutive months to evaluate several strategies before picking the one that worked best empirically.

[23] The standard FM approach is to refit the model monthly.

We periodically refit our model on data from year t, using the described k-fold time-series cross-validation, and forecast excess returns for all months in year $t+1$. In particular, we use $k=3$ to find a balanced trade-off between data availability and computational cost, with the smallest observed training split containing circa 13,330 observations. This scheme allows us to make results comparable to existing research that uses traditional FM regressions. To further increase comparability, we replicate the results from Green et al. (2017) by refitting the model monthly and annually (as done in their paper). The results can be found in appendix E.

Moreover, we are particularly interested in the time-series aspect of risk-premia estimations, feature selections and performance. Therefore, we allow for different network specifications at each time as we periodically tune all hyperparameters applying random grid search. By doing so, we follow Bergstra and Bengio (2012) and draw the respective hyperparameters independently from the distributions summarised in table 2. Random grid search has advantages over brute-force grid search, as it dramatically reduces the computational cost. Bergstra and Bengio (2012) show that a small subset of all hyperparameter combinations is sufficient to minimise the validation error crucially.

In addition to hyperparameters such as learning rate or parameter penalties, we are also interested in tuning what we call *architectural* hyperparameters. These include the number of layers, number of nodes in each hidden layer and the layer structure. There exists no deterministic rule for deriving an optimal neural network design (e.g. see Heaton (2008)). Therefore, a kind of arbitrariness exists regarding a network's layer structure and design, which is frequently criticised. We generally differentiate between a *tapered* and *constant* architecture type, which are described in detail in appendix D). For the constant architecture type, the number of nodes remains the same in each hidden layer and is drawn from the distribution shown in table 2. For the tapered-type architecture, a starting number of nodes for the first hidden layer is drawn from the distribution shown in table 2, denoted N_1, where all subsequent layers follow the rule $N_l = \left\lceil \frac{N_{l-1}}{2^l} \right\rceil$, with the boundary condition $N_1/2^{\mathrm{HL}} \geq 1$. Most importantly, however, the networks' architectural structure directly correlates with the model complexity, which we can capture numerically, for example, through the total number of model parameters.

Regarding the tuning of architectural hyperparameters, our paper takes a different approach than other studies, such as Gu et al. (2020b) or Bianchi et al. (2021). While different architectures are commonly benchmarked in a horserace, they typically remain constant over time. In the end, an overall (or on average) architectural structure is picked as the best. In this paper, on the other hand, we are fundamentally interested in the time-varying model complexity, particularly during times of crisis. Therefore, we allow the neural network to flexibly change its architecture every time we refit as part of the hyperparameter tuning.

Hyperparameters – general					
weight penalty $\lambda^{(w)} = 10^a$	weight penalty ratio $\alpha^{(w)}$	gradient penalty $\lambda^{(d)} = 10^a$	gradient penalty ratio $\alpha^{(d)}$	dropout probability D	learning rate $\text{LR} = 10^a$
$a \sim U(-8, -4)$	$\in B^*$	$a \sim U(-8, -4)$	$\in B^*$	$\sim U(0.05, 0.25)$	$a \sim U(-5, -3)$

Hyperparameters – architecture		
layer structure LS	# of nodes $N = \lfloor e^a \rfloor$	# of hidden layers $\text{HL} \in [1, a], \text{HL} \in \mathbb{Z}_+$
$\in$ {tapered, constant}	$a \sim U\left(\log(\frac{K}{2}), (1.1 \times K)\right)^{**}$	$a = \left\lfloor \frac{\log(K/2)}{\log(2)} \right\rfloor^{**}$

* where $B = \{0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$.
** where K is the number of factors / characteristics.

Table 2:
Hyperparameter tuning
Each time we refit a model that requires hyperparameter tuning, we perform a random grid search, analogously to Bergstra and Bengio (2012). In particular, we independently draw 80 different combinations from the distributions shown above. We find that this number is a good trade-off between validation error minimisation and computational cost.

4.4 Calendar Year vs. Fiscal Year

Section 4.3 introduces the estimation strategy, including the proposed data splitting protocol. In order to reduce computational cost while simultaneously retaining the typical Fama-Macbeth approach, we re-fit the model under consideration annually rather than monthly. This approach is supported from a computation cost point of view and a data dynamics perspective. In section 4.1 and in appendix A, we discuss that 63 of the considered characteristics are based on annual accounting information. Therefore, despite the cross-sectional rank normalisation, which can result in small monthly changes even in annual firm characteristics due to a change in rank, a significant number of the firm characteristics under consideration only changes considerably once a year – namely when a new annual report with novel fundamental accounting data is released. The fact that most firm characteristics are constructed based on information that merely becomes available once a year also supports the idea of re-fitting the model annually. Subsequently, the key question is what *year* the re-fitting is referring to. The calendar year or the fiscal year? The fiscal year appears to be the natural choice due to the aforementioned connection of annual firm characteristics to annual reports. However, not every company's fiscal year ends in the same month.

Figure 6 shows that the average change in book-to-market (which serves as a placeholder representative for annual characteristics) is not evenly distributed over all months but peaks in July. Most companies' fiscal year in our investment universe ends in December, causing this uneven distribution. The time lag between December and July is explained by the publication lag as discussed in section 4.1 and the fact that we match monthly returns with lagged firm characteristics as discussed in section 2. Subsequently, we define the *fiscal year* (from an information availability perspective) to span from July

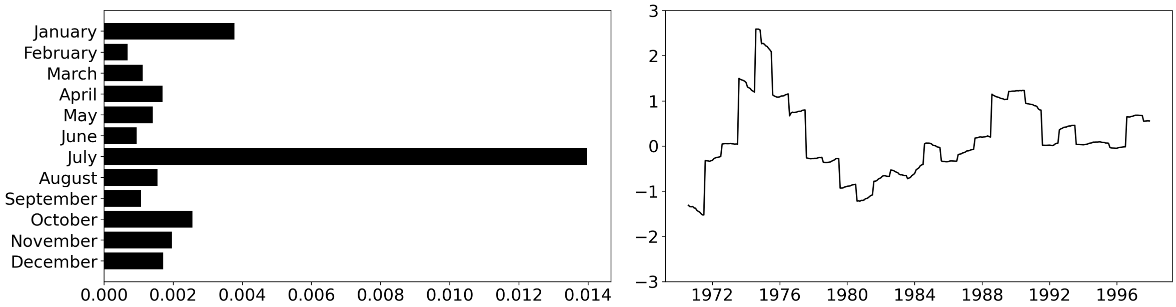


Figure 6:
Average annual changes, book-to-market
On the left hand side, the figure shows the average change in book-to-market (as an arbitrarily chosen annual firm characteristic) per month over the entire sample. It can be seen that, on average, the most significant changes happen in July, meaning that after aligning returns with lagged firm characteristics and including a publication lag of six months, that for most companies included in the investment universe, the fiscal year seems to end in December. On the right hand side, the rank-normalised book-to-market of an arbitrarily chosen asset (Beverly Enterprises, permno = 47992) confirms that the observations tend to change only once every 12 months, emphasised by the step-like pattern.

to June, as this is true for the average company in our investment universe. As an example, figure 6 also shows the time-series of rank-normalised book-to-market values of an arbitrarily chosen asset – Beverly Enterprises (permno=47992). The step-like pattern confirms that significant changes merely occur once a year. Therefore, we consider both cases in our analysis: re-fitting by calendar year spanning from January to December and by fiscal year, which spans from July to June.

4.5 Model Comparison

We evaluate model performances using two key measures: the out-of-sample cross-sectional mean R^2 (XS-R^2) and the out-of-sample predictive R^2. In general, assessing an asset pricing model should include evaluating how well the model describes systematic risk and how well the model explains risk compensation. We follow Chen et al. (2019) and define[24] the cross-sectional mean R^2 – which indicates the model's ability to describe common variation in realised returns across stocks – as

$$\text{XS-}R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\left(\frac{1}{T_i}\sum_{t\in\mathcal{T}_{\text{oos}}}\hat{\epsilon}_{i,t+1}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\left(\frac{1}{T_i}\sum_{t\in\mathcal{T}_{\text{oos}}}r_{i,t+1}\right)^2} \tag{38}$$

where $\mathcal{T}_{\text{oos}}$ denotes the out-of-sample test split, with $\hat{\epsilon}_{i,t+1} = r_{i,t+1} - \hat{r}_{i,t+1}$ and $\hat{r}_{i,t+1} = \hat{g}_t(\mathbf{c}_{i,t}; \mathbf{W}_t, \theta)$. Analogously to Gu et al. (2020b), we do not de-mean the denominator in equation (38) due to the non-stationarity and noise in the mean estimation. Note that the cross-sectional mean R^2 differs slightly from a conventional total R^2 measure, which is also commonly used in empirical asset pricing. Since we

[24] On page 19 in the original paper by Chen et al. (2019), equation (38) appears to be using the estimated return in the denominator. However, analogously to the common R-squared definition, and the *Github* repository `https://github.com/jasonzy121/Deep_Learning_Asset_Pricing/blob/6c26b9dad01e76b214ab8f5566c42a29e99677c9/src/model/model_utils.py#L57`, we follow (38) and use the actual excess return instead of an estimated return.

only impose very mild restrictions on our investment universe, including no restrictions on the time a stock must have been listed on an exchange, our empirical dataset is unbalanced[25]. This unbalancedness means that while observations range over decades for some stocks, they can be extremely limited for others. Taking the time-series average of the residuals first and weighting the estimated means by their convergence rate accounts for differences in precision and, therefore, makes it less prone to outliers.

Next, we report the predictive R^2, which indicates the model's ability to explain cross-sectional differences in expected returns, defined as

$$\text{Predictive } R^2 = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_{\text{oos}}}\left(r_{i,t+1} - \hat{\beta}'_{i,t}\hat{\lambda}_t\right)^2}{\sum_{(i,t)\in\mathcal{T}_{\text{oos}}}\left(r_{i,t+1}\right)^2}. \tag{39}$$

The predictive R^2 measures the explained variation in $r_{i,t+1}$ due to $\hat{\beta}'_{i,t}\hat{\lambda}_t$, where $\hat{\lambda}_t$ denotes a vector of conditional risk price estimates, using only information that is available up until $t-1$. Equation (39) follows a restricted approach, in which we assume that the model's ability to describe risk compensation is solely based on the exposures to systematic risk (for a more detailed discussion on the topic, see for example Kelly et al. (2019)).

Note that compared to Kelly et al. (2019) we do not use the unconditional but the conditional risk price estimate. The conditional risk price estimate in combination with annual re-fitting yields conservative out-of-sample performances compared to existing literature. However, we are particularly concerned about invalid future information leakages as part of the empirical backtesting exercise. Thus, we intend to follow the mantra of the general machine learning literature and do not allow any information spillovers from the test dataset into the training or validation set but in an asset pricing setting. Section 4.4 discusses that we re-fit our models annually, rather than monthly (as is typically done with linear models), to maintain the general spirit of the original Fama-Macbeth regressions while dramatically reducing the computational cost. Thus, in order to estimate a neural network on data from year t, *all* data from year t must have been observable for estimation at the time. For example, if λ_t refers to the risk prices estimate for September in year t, and to avoid invalid information leakage from the future, all data from year t must have been observable before the risk price for September could have been estimated. In the example of re-fitting the model by calendar year, the corresponding risk price in September could have only been estimated at the end of December – the point in time when all of the data was observable. In contrast, Kelly et al. (2019), for example, apply the unconditional risk price estimate to evaluate the predictive R^2. While predictions are performed at each point in time, the unconditional risk price estimate induces an unclean information leakage, as the unconditional risk price estimate is based on data from the entire sample, not just on data observable at time t.

[25] Our approach in this regard differs from other studies, such as Chen et al. (2019), who only consider stocks for which all characteristics are fully observable.

Furthermore, in section 2.4 we discuss the smoothening of the risk price estimates by using a five-year backwards-looking rolling window approach. Thus, we evaluate the predictive R^2 based on the year-end risk price estimate and return predictions for the entire year $t+1$. This procedure is clean from an information leakage perspective, as it only uses information that would have been observable at the time. In the example of re-fitting our models by calendar year, this equates to a risk price estimate based on available data until December of year t. Without the backwards-looking smoothening of the risk price estimates, and in a scenario where the models would be re-fitted monthly, the described procedure would induce a heavy *December bias*. However, since we smooth the risk price estimate using all information from the previous five years, there is no unwanted December bias while being clean from a general machine learning perspective. However, since the risk price estimates are only updated once a year, the out-of-sample prediction task is much more difficult than a monthly update, as new information is incorporated at a much lower frequency. The increased difficulty of the prediction task resulting from this methodology, is one of the reasons why our empirical performances are conservative compared to existing literature yet competitive.

4.6 Robustness

The previous sections 4.2 and 4.4 discuss the different data pre-processing and splitting regimes under consideration. In addition, we also examine a more restricted universe in which we exclude all microcap stocks from the sample. In a different approach, only a smaller subset of 49 characteristics is considered, which we call *core* characteristics and which comprise the most commonly used characteristics in the empirical literature (see appendix B for a description). This manual pre-selection reduces the dimensionality of the investment universe by nearly 50%. However, reporting all results for all scenario combinations would be too expansive and confusing. We, therefore, primarily focus on results for the baseline case: re-fitting by calendar year, using all and the core rank-normalised characteristics, and estimating risk prices on a rolling window basis. Additionally, appendix G summarises all scenarios' results as a robustness analysis. We show that our results are qualitatively stable across all scenarios.

There are, however, minor quantitative differences that we discuss in appendix C. For example, splitting by fiscal year is a more "difficult" out-of-sample task for any model as the test dataset predominantly includes new accounting data from newly released annual reports, making predictions harder. We also show that one metric alone (either XS-R^2 or predictive R^2) is frequently insufficient to evaluate a model's overall performance. There are also minor differences in performance regarding the conditional risk price estimates, which depend on an expanding or rolling window estimation. While appendix I presents a more detailed discussion, these examples highlight that many assumptions directly impact model performances, posing specific difficulties. In particular, out-of-sample model performances are the first sanity check for any empirical analysis and precede an economic interpretation of the results (if a model does not perform well out-of-sample, do we trust its outcomes?). Consequently, the assumptions that go into the model fitting process also directly impact inference. Controversially, and in most cases, there is no clear right or wrong regarding these assumptions, such as splitting the data by calendar or fiscal year.

4.7 The Cross-section of Stock Returns

We compare fifteen different models and report their out-of-sample cross-sectional mean R^2 and predictive R^2 in table 3. In particular, we investigate ten different neural network types that we benchmark against five commonly used linear benchmarks, including ordinary least squares, weighted least squares, Elastic Net, Ridge and Lasso. While all neural networks can take any of the flexible architectural forms discussed in section 4.3, they differ in the type of regularisation that is used in their respective objective function. We consider neural networks with no regularisation in their objective function (NN), neural networks with weight constraints, including L_1 (NN-W1), L_2 (NN-W2), and both L_1 and L_2 (NN-W1W2) regularisation, element-wise Jacobian constraints, including L_1 (NN-J1), L_2 (NN-J2), and both L_1 and L_2 (NN-J1J2) regularisation, as well as column-mean Jacobian constraints, including L_1 (NN-J1-m), L_2 (NN-J2-m), and both L_1 and L_2 (NN-J1J2-m) regularisation.

The left panel in table 3 reports R^2's at the individual stock level for the case of using all 103 firm characteristics. The table shows that a model's ability to describe the total variations in returns by the common variation among returns (i.e. describing risk) does not necessarily correlate with a model's ability to explain cross-sectional differences in expected returns (i.e. describing risk compensation). This difference in performance measure is evidenced, for example, by a positive cross-sectional mean R^2's for OLS or Ridge, but a corresponding negative predictive R^2. Consequently, we evaluate model performances hereafter following a combined approach, taking both performance measures into account.

Table 3 shows that OLS, Lasso, Ridge and Elastic Net yield positive cross-sectional mean R^2's of 5.57%, 13.79%, 5.58%, and 13.81% (with WLS being the only model producing a negative XS-R^2 of -1.96%), but they fail to produce positive and non-zero predictive R^2's, indicating that they are less effective in estimating risk prices – at least in the FM-like setup of this paper. This finding may not be surprising for unregularised models, such as OLS or WLS, as they are prone to in-sample overfitting in a high-dimensional environment. However, even highly regularised linear models such as Lasso, Ridge, or Elastic Net do not seem to capture risk prices satisfactorily, as they produce out-of-sample predictive R^2's of 0.00%, -0.35%, and 0.00% respectively and are, therefore, dominated by (or on par with) a naive return forecast of zero to all stocks. Overall, our results question the usage of typical linear Fama-Macbeth regressions when they are analysed from an out-of-sample perspective.

In contrast, neural networks produce positive R^2's across all performance measures, with the XS-R^2's ranging from 0.92% (NN) to 16.12% (NN-J1-m), and with the predictive R^2's ranging from 0.04% (NN-J1J2) to 0.12% (NN-W2). Despite the apparent outperformance of neural networks, valuable and

	All Characteristics		Core Characteristics	
	XS-R² [%]	**Pred. R²** [%]	**XS-R²** [%]	**Pred. R²** [%]
OLS	5.57	−0.35	6.91	−0.07
WLS	−1.96	−1.81	7.20	−0.37
Lasso	13.79	0.00	13.81	0.00
Ridge	5.58	−0.35	6.91	−0.07
Elastic Net	13.81	0.00	13.67	0.00
NN	0.92	0.11	14.75	0.10
NN-W1	9.17	0.11	13.42	0.10
NN-W2	9.74	0.12	17.20	0.12
NN-W1W2	11.16	0.07	13.96	0.12
NN-J1	14.88	0.05	16.15	0.04
NN-J2	14.32	0.09	15.72	0.08
NN-J1J2	8.15	0.04	13.45	0.03
NN-J1-m	16.12	0.10	15.83	0.09
NN-J2-m	14.99	0.07	14.03	0.05
NN-J1J2-m	14.78	0.08	15.77	0.06

Table 3:
Out-of-sample performance summary:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by calendar year, firm characteristics are cross-sectionally rank-normalised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

transferable insights can be gained from the linear models' results. In particular, model selection (or even regularisation in general), as in the case of Lasso, significantly improves the estimation of systematic risk and the prices of risk. This improvement becomes evident when comparing the XS-R^2's of OLS with Lasso (5.57% – 13.79%) and the XS-R^2's of NN with NN-J1 (0.92% – 14.88%). The results show that applying an element-wise L_1 penalisation to the Jacobian during training yields a cross-sectional mean R^2 that is 16 times higher than using an unregularised objective function.

The general insight that sparse model representations, for example, through the implementation of L_1 norm regularisation, are favourable – especially in a data-rich environment, as is the case with using all 103 firm characteristics – is not new to the empirical asset pricing literature (e.g. see Kozak et al. (2020), or Freyberger et al. (2020)). However, to the best of our knowledge, we provide the first evidence that this insight can be confirmed using Jacobian regularisation as part of a non-linear model's objective function. In particular, with 16.12%, NN-J1-m produces the highest cross-sectional mean R^2 and the third-highest predictive R^2 with 0.10%, providing strong evidence in favour of Jacobian regularisation.

The right panel in table 3 reports R^2's at the individual stock level for the case of using a subset of the 49 core characteristics only. Like the all-characteristics case, neural networks dominate all linear benchmarks and consistently yield positive R^2's on both performance metrics. As expected, due to the reduced dimensionality, the performances of the unregularised linear models have improved relative to the higher dimensional setting. For example, while WLS only produces a negative XS-R^2 of -1.96% in the all-characteristics case, it yields a positive XS-R^2 of 7.20% using the core characteristics only. A similar pattern emerges for neural networks, where the XS-R^2 of NN improved from 0.92% to 14.75%. However, even in an environment with reduced dimensionality, no linear model produces positive and non-zero predictive R^2's.

Consequently, as expected, the relative importance of L_1 penalisation is reduced for all models in the core characteristics case. For example, while all Jacobian-regularised neural networks still produce competitive out-of-sample R^2's, penalising the network's weights[26] only appears to yield the most robust results for the core-characteristics case, with NN-W2 producing the highest XS-R^2 (17.20%) and the highest predictive R^2 (0.12%). Overall, table 3 shows that neural networks consistently produce positive out-of-sample R^2's that appear to be stable across the all-characteristics and core-characteristics case. Moreover, neural networks trained with objective functions applying Jacobian regularisation are among the most robust models, with neural networks, in general, being the best-performing models overall.

In addition to the overall model performances in table 3, figure 7 summarises model performances by company size for the all-characteristics case only, with an analogous analysis for the core characteristics case in appendix G. A differentiation by company size, measured by market capitalisation (see section 4), is essential since there is empirical evidence that risk factor performance is dependent on the investment universe definition and the inclusion of micro-caps. Bartram et al. (2021) argue that micro-caps are characterised by lower liquidity, higher idiosyncratic volatility, and pose more significant short-selling frictions. Systematic risk factors that empirically depend on micro-caps are frequently deemed irrelevant by institutional investors, as, for example, trading on their signal is capital-intensive. At the same time, and as discussed before, micro-caps provide the vast majority of data in asset pricing. The commonly used investment universe of US stocks only consists of 60% micro-caps, meaning they provide the largest share of training data. From this point of view, the evolution of modern machine learning in asset pricing towards more complex and computationally more intensive models relies heavily on micro-caps simply from a data providing perspective.

This paper does not intend to replicate an institutional investor's mandate and does not fully account for real-world frictions since our primary focus lies on the universally applicable methodology. However, we take the criticism seriously. In particular, we are cautious about identifying small scale inefficiencies that are merely driven by illiquidity.

The top panel in figure 7 shows the cross-sectional mean R^2 for all models in percentage. The performance is based on estimated models using all stocks but focuses on fits among the size class subsamples. A similar pattern to the overall performances from table 3 emerges as linear models fare

[26] Note, that penalising the weights does not yield in model selection as it does for Jacobian regularisation.

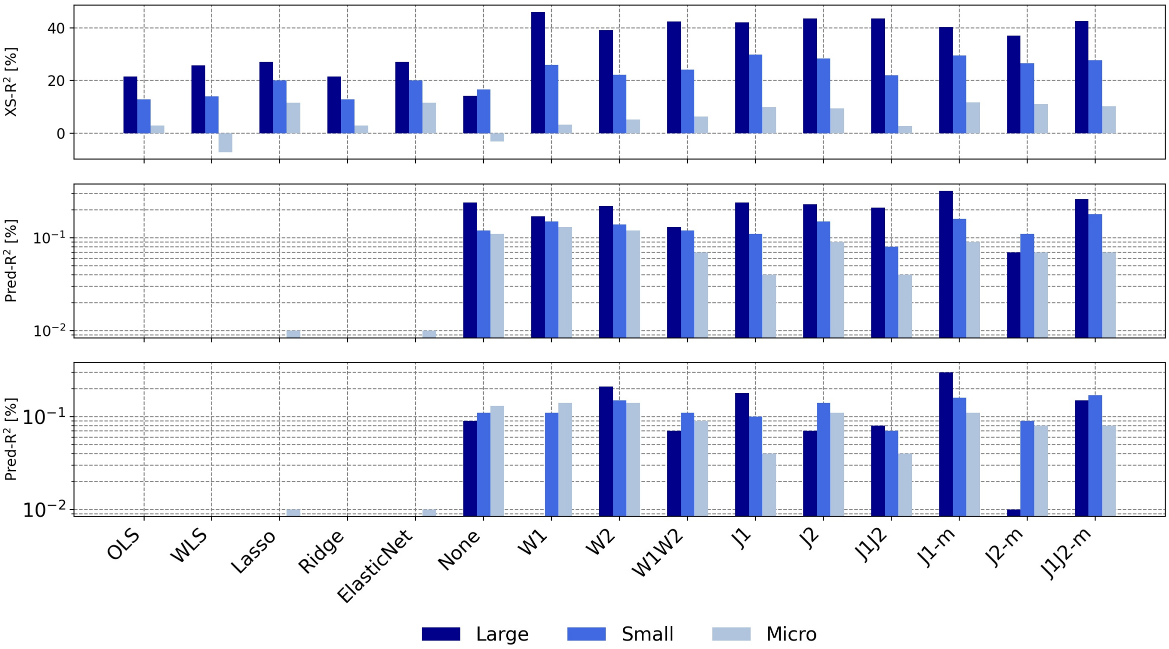


Figure 7:
Performances summary by size
The top panel shows the out-of-sample XS-R^2 in percentage for all models grouped by size (i.e. Large, Small, and Micro). The middle panel displays the predictive R^2 on a logarithmic scale, where the predictive R^2 is estimated using risk prices that are estimated over all assets. The bottom panel shows the predictive R^2 on a logarithmic scale, where a separate risk price estimate is used for each size class.

poorly, while neural networks are the best performing models, especially among large stocks, with XS-R^2s ranging from 14.27% (NN) to 45.93% (NN-W1). The panel in the middle reports the predictive R^2, where the performance is again based on a risk price estimate using all stocks. The results are reported on a logarithmic scale to make even small differences visible. Neural networks are the only models producing positive R^2's for all size classes. While the performance of large stocks is still particularly successful, with predictive R^2's ranging from 0.07% to 0.32%, the dichotomy between model performances of large and non-large stocks is less distinct compared to the performance differences measured by the cross-sectional mean R^2, as micro-caps produce predictive R^2's ranging from 0.04% to 0.13%.

With the criticism expressed by Bartram et al. (2021), the question becomes: how reliable are these estimates? Are we potentially overstating the performances of large stocks? As a form of robustness check, the bottom panel in figure 7, reports predictive R^2's, where lambda is estimated separately for each size class. This approach does not require any model re-fitting and is, therefore, computationally inexpensive. Due to the linear nature of OLS, WLS, Lasso, Ridge and Elastic Net, the risk price estimate does not change, and their performances continue to fare poorly. For the nonlinear neural networks, however, the size class-specific lambda estimate differs from the overall estimate. Ideally, the performances between the panel in the middle and the bottom do not change significantly, indicating robust estimates.

Interestingly, in the high-dimensional all-characteristics case, neural networks with weight penalisation show less robust results, as the large stock performance of NN-W1 drops from 0.17% to −0.02%. L_2 weight penalisation seems to offer more robust results as the performance of large stocks remains constant (0.22% to 0.21%). The most robust model is NN-J1-m, which produces the highest predictive R^2 for the case of using an overall lambda estimate (0.32%) and a size class-specific estimate (0.30%). The bottom panel supports the argument of including L_1 norm Jacobian regularisation in the objective function in a high-dimensional environment, as NN-J1 and NN-J1J2-m also produce competitive and stable predictive R^2's (together with NN-W2).

The superior out-of-sample performances shown by neural networks in this section suggest that nonlinear models estimate systematic risk and risk compensation more effectively. In addition, and as expected, we find that the importance of including forms of regularisation that enable model selection increases with the dimensionality of the input data. More specifically, this paper provides further evidence for the ongoing debate in empirical asset pricing that systematic risk and risk compensation are intrinsically nonlinear. The neural networks under consideration show favourable inner model mechanics that produce consistent and robust estimates — consequently, understanding *why* a model produces a particular outcome is of utmost importance.

4.8 The Impact of the Objective Function

The primary focus of this paper is the estimation of time-varying risk premia and the focus on model interpretability while simultaneously offering competitive and stable model performances. As a consequence, we are particularly interested in the partial derivatives' distributions as they directly relate to the topic variable importance (see section 4.9), offer valuable insights about the degree of certainty regarding the risk premia estimation (see section 4.11) and the inner model mechanics (see section 4.12). Ideally, we observe partial derivatives that are clear of extreme outliers and offer meaningful tolerance bands for the risk premia estimation. In this section, we show that the choice of the objective function directly impacts the objectives mentioned above and must become an integral part of the training regime design of neural networks.

Sections 3.2 and 3.3 mathematically introduce the two different types of objective functions that include either the penalisation of a neural network's weights or the penalisation of the input-output Jacobian. In general, any form of regularisation – including weight or Jacobian penalisation – primarily serves the purpose of reducing overfitting, e.g. see Goodfellow et al. (2016). In this section, we further show that the objective function choice directly impacts economic interpretability. Therefore, the objective function choice is relevant from a performance point of view and crucial in the domain-specific application in empirical asset pricing. Section 3.3 discusses how the idea of Jacobian regularisation is generally not new to the broader machine learning literature and has so far been predominantly used in the field of image recognition, for example, in the context of adversarial examples. However, we argue that Jacobian regularisation becomes economically interpretable in asset pricing, besides boosting performance, as it enables time-varying variable selection by setting the influence (or in an economically interpretable term: risk premium) of certain input variables close to or precisely to zero. This feature is not straightforward in the case of weight regularisation. Further, in this section, we specifically only focus on the regularisation that is directly implemented through the objective function and do not discuss other forms of regularisation, such as dropout or early stopping and refer to section 4.3.

Starting with the regularisation of the model weights, the objective function in equation (24) forces the weights closer to the origin by adding the regularisation term to the objective function motivated by the intention to improve the model's ability to generalise better out-of-sample or, equally, to reduce overfitting, e.g. see Goodfellow et al. (2016). Empirically, we confirm the theoretical properties of the weight penalisation strategy. Both L_1 and L_2 norm weight regularisation can have a positive impact on the out-of-sample performance. As shown in table 3, the cross-sectional mean R^2 for neural networks using either L_1 or L_2 weight penalisation is nearly ten times higher in the case of using all 103 characteristics. Compared to the high-dimensional case of using all 103 characteristics, this performance boost is less pronounced in the case of using the core characteristics only, where L_2 weight regularisation improves the XS-R^2 and the predictive R^2 from 14.75% to 17.20%, respectively from 0.10% to 0.12%, compared to a neural network without any weight penalisation in the objective function. We conclude that weight penalisation has its most considerable influence in high dimensional settings, as expected.

While weight penalisation improves out-of-sample performance, it is more difficult to interpret its effect economically. The economic interpretability is particularly difficult when weight penalisation is combined with other forms of regularisation, such as batch normalisation. For example, combining weight penalisation with batch normalisation merely affects the effective learning rate, as discussed by van Laarhoven (2017). Unlike ridge or lasso regressions, where the penalisation directly impacts the regression coefficients (economically interpretable), weight penalisation does not intentionally and directly penalise the coefficients (in the form of the derivatives) or even variable selection. More importantly, due to the intrinsic nonlinearity of g, penalising the weights does not necessarily impose any boundaries on the distribution of the partial derivatives, meaning that they are not clear of (potentially extreme) outliers[27]. To see why this is, consider the single-hidden layer neural network $g(x, w) = \sigma(w^T x + b)$, such that $\frac{\partial g}{\partial x} = \sigma' w$, where σ' denotes the derivative of the activation function, σ, with respect to its input.

Jacobian regularisation, on the other hand, and as discussed in section 3.3, does precisely this: due to the imposed penalty in the form of the Frobenius norm on the Jacobian, the magnitude of the derivatives is bounded and dependent on the penalty forced towards zero, allowing for risk premia selection. Since the partial derivatives are used to estimate time-varying risk premia, the Jacobian regularisation is

[27] We discuss the importance of outliers and how they are related to the notion of *stable* predictions in section 4.12

economically interpretable and particularly useful in asset pricing. Moreover, in section 4.12, we further discuss the benefits of the Jacobian objective function in the context of model stability.

To illustrate the practical effects of the various objective functions, we present the empirical distributions of the partial derivatives, measured as the overall, cross-characteristic time-series interquartile ranges average, the cross-sectional time-series average of the maximum number of outliers[28] and the maximum cross-sectional spread between the maximum and minimum value of derivatives, which are summarised in figure 8. The maximum max-min spread and all time-series averages are both based on monthly values. The interquartile range serves as a first impression of the spread of partial derivatives, while the outliers indicate the frequency of extreme values. Lastly, the max-min spread serves as an indication for the spread of the most extreme values.

Due to the cross-sectionality and the summarising nature of the aggregated interquartile range, granularity is lost. Nonetheless, figure 8 shows that the partial derivatives' distribution depends on the objective function. In particular, those neural networks trained with an objective function using element-wise L_1 norm Jacobian regularisation (i.e. J1, J1J2, J1-m, J1J2-m) appear to be characterised by the most narrow distribution. Moreover, when compared to the objective function without any additional penalties (NN, neither weight nor Jacobian penalty), the impact of the element-wise Jacobian penalty is most significant for the all-characteristics case. This empirical finding is valuable insight, as we ideally wish to reduce the dimensionality of the input data. The effect is smaller for the already dimensionality reduced core characteristics case. In addition, figure 8 shows that all neural networks that used Jacobian regularisation as part of their objective function produce less extreme values, as their maximum max-min spread is smaller compared to the spread of any neural network that was trained with weight regularisation.

However, figure 8 does not account for the time-variation in the distributions of the partial derivatives. This time-varying distribution is not only different for each neural network. It also differs across all characteristics. A holistic analysis of the time-varying distributional properties of each characteristic and model is beyond the scope of this paper. Instead, figure 9 exemplary shows the interquartile ranges as an indication for the distribution of the partial derivatives for the characteristic *return-on-assets*, as it is one of the most relevant variables, which we discuss in section 4.9. The left column of figure 9 refers to the year 2008, and the right column to the year 2019. We intentionally picked a crisis and non-crisis year, as defined by the NBER recession indicator. The bottom panel displays the same box plot as the top panel but includes a visualisation of outliers.

Figure 8 shows that the spread of the partial derivatives on characteristic level is indeed time-dependent. Moreover, the magnitude of the partial derivatives can be vastly different across model and time. Without further discussing the distributions concerning this particular characteristic, our

[28] Where we define an *outlier* as a partial derivative that lies outside $[\text{Q1} - 1.5 \times \text{IQR}, \text{Q3} + 1.5 \times \text{IQR}]$, with Q1 and Q3 being the $25th$, respectively $75th$ percentile and IQR the interquartile range, defined as $\text{IQR} = Q1 - Q3$.

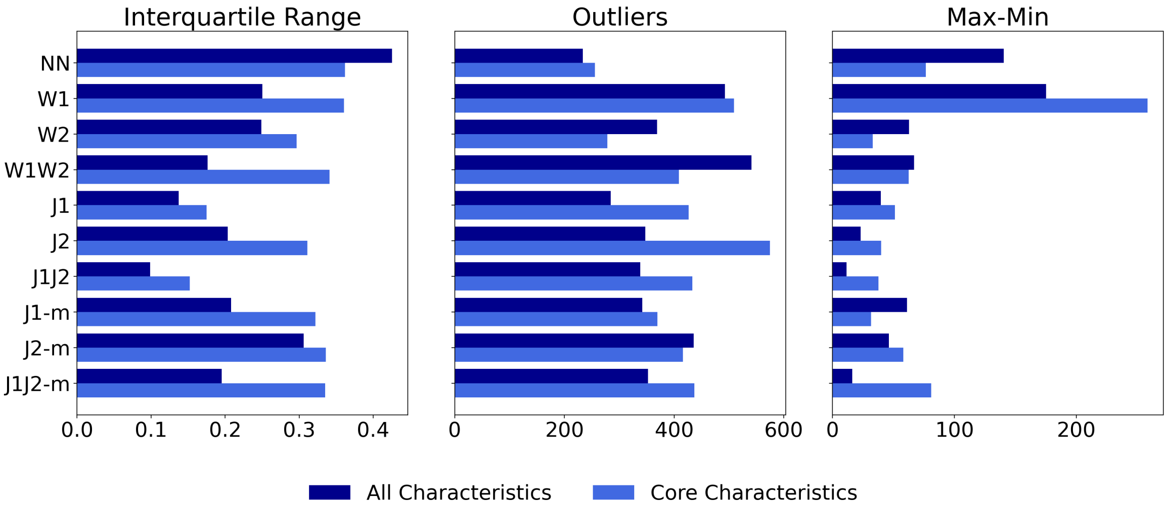


Figure 8:
Overall distributional summary of partial derivatives
The left panel shows the overall, cross-sectional time-series average of the interquartile range for. The middle panel displays the time-series average of the maximum number of outliers across characteristics. The right panel displays the maximum max-min spread across characteristics. The distributional properties are shown for each model and the all and core-characteristics case.

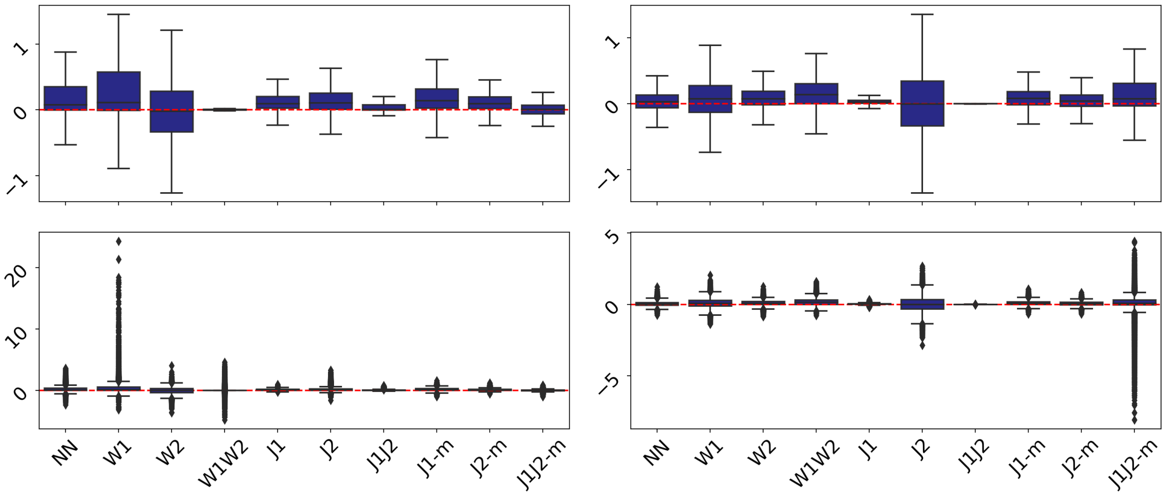


Figure 9:
Exemplary distributional properties: *return-on-assets* (2008, 2016)
The left column shows the distributional properties, summarised by the interquartile range, of the partial derivatives of the characteristic return-on-assets for the crisis year 2008, and the right column refers to the year 2019. In addition, the visualisation of in the bottom row is identical to the top row, but it additionally visualises outliers, which are not included in the top row.

critical finding is that out-of-sample performance alone constitutes an incomplete metric on which model selection should be based. Due to the economic interpretability of the partial derivatives, the distribution, including the spread, outliers and consistency, should also be considered when picking a model for further analyses. Therefore, we propose considering secondary and interpretable model qualities such as the distributional properties of the partial derivatives as part of the model selection exercise. To the best of our knowledge, this paper is the first to propose considering such a secondary and indirect measure for the process of model selection in asset pricing.

One of the critical problems is that there is no unambiguous way to value the various derivative distributions. A too sizeable interquartile range across all characteristics is undesirable as it does not allow for clear variable selection and induces more significant uncertainty. Equally, an interquartile range that is too small may pose limitations due to a variable selection that is too strong. We find that NN-W2 and NN-J1-m portray a reasonable middle ground as these models offer strong out-of-sample performances and desirable partial derivative distributions, especially in the core-characteristics case. Moreover, they represent a mix of models that use model weight and Jacobian regularisation as part of their objective function. Therefore, in section 4.11, we primarily focus on those two models to make our analysis more straightforward and clear, with more analytical output in appendix I.

4.9 Which Firm Characteristics Matter?

The competitive model performances of neural networks presented in section 4.7 provide empirical evidence in favour of a nonlinear model choice. In section 4.8 we show that the objective function not only has a profound impact on model performance but also on the interpretability of the partial derivatives. In particular, we conclude that NN-W2 and NN-J1-m provide competitive model performances in combination with desirable partial derivative distributions. They are the primary models we focus on in this section, with more empirical results presented in appendix H.

In this section, we investigate the relative importance of individual firm characteristics (variable importance) for the performance of each model. We build on existing research such as Ruck et al. (1990), or Leray and Gallinari (1999) and measure variable importance as the absolute median of the partial derivatives at time t, such that variable importance is defined as

$$\mathrm{VI}_{t,k} = \left|\bar{\boldsymbol{d}}_{t,k}\right| , \tag{40}$$

where $\boldsymbol{d}_{t,k}$ denotes an N_t-dimensional vector of partial derivatives with respect to the k-th firm characteristics at t, and the bar notation in equation (40) indicates the median. Thus, the variable importance definition in equation (40) is closely related to the estimation of risk premia, which is conceptually meaningful: For a given risk exposure, a higher risk compensation is more important. However, in contrast to the risk price estimation, the estimated variable importance is not smoothed. Equation (40) yields a time-series of variable importances measured in absolute median values, which we rank at each t. Thus, we can further analyse variable importance, for example, for specific periods or the entire sample. For example, table 4 reports the overall most important input variables and the most important input variables over the most recent five years, where the aggregated variable importances are the median of the time-series variable importances over the respective period. Using the median rather than the mean is justified because the time-series of median partial derivatives can vary significantly over time. These significant changes in absolute value result from the periodic re-fitting, the architectural freedom of the neural networks, and their nonlinearity. Therefore, to reduce the effect of extreme outliers in variable importance, which may be caused by a relatively short period, where an input variable becomes empirically very important compared to the remaining periods, we use the median to aggregate the time-series of variable importances into a fixed estimate. In addition to the examples discussed in this section, appendix H provides further empirical details on the time-varying variable importances.

Empirically, we confirm the well-documented fact that variable importance varies over time and depends on the time horizon over which it is estimated (e.g. see Gu et al. (2020b) who also report time-varying variable importance). Our main contribution is two-fold. First, the objective function profoundly influences which input variables turn out to be the most important. To the best of our knowledge, we are the first to document the importance of the objective function in the context of variable importance in empirical asset pricing estimated by deep neural networks. Secondly, we empirically confirm the theoretical variable selection ability of Jacobian regularisation and show that, for example, in an overall estimation, the total number of firm characteristics considered influential for NN-J1-m (28) is reduced by nearly 25%, compared to NN-W2 (38).

Variable or feature importance and model interpretability is a fundamental part of modern machine learning literature in general, not just in asset pricing (e.g. see Molnar et al. (2020)). The most commonly used methodologies to assess variable importance include the R^2 reduction from setting all values of a given predictor to zero (e.g. see Gu et al. (2020b)), the sum of squared sensitivities (e.g. see Dimopoulos et al. (1995)), partial dependence plots (e.g. see Friedman (2001)), individual conditional expectations (e.g. Goldstein et al. (2015), accumulated local effects (e.g. see Apley and Zhu (2020)), Shapley values (e.g. see Shapley (1953)) and absolute derivatives (e.g. see Ruck et al. (1990) or Leray and Gallinari (1999)). The literature in this field is vast, so we do not claim to provide a holistic overview. However, one shortcoming of all these measures is that they are typically reported as an aggregated summary over the entire sample. However, aggregated point estimates make time-dependent analyses or economic model output interpretations difficult. In asset pricing, particularly, stakeholders may be interested in asset-level model insights that are also time-varying. One of the key reasons for this is that financial markets are dynamic and, thus, constantly evolving. Under the assumption of efficient markets, tradable risk factors that may earn an investor abnormal returns at a certain point in time should be arbitraged away. Moreover, regulatory changes, geopolitical events or even technological advancements, such as high-frequency (or at least electronic) trading, support the idea that variable importance should vary with time because the underlying risks are changing themselves. For this reason, we estimate variable influence estimated over the entire sample, as well as time-varying. In particular, we separately report an overall and most recent variable importance, which we estimate over the most recent five years in our data sample. Further analysis is presented in appendix H, which serves as a robustness check and confirms our general findings.

Variable importance is crucial from an interpretability point of view and an asset pricing and domain-specific perspective. Variable importance in the context of model interpretability helps to understand which input features are primarily driving predictions. Improved model interpretability, as a consequence, is beneficial for any form of communication about the model and can help detect biases or even debug software. In asset pricing, in particular, variable importance may directly translate into trading strategies. While there is a myriad of different trading strategies – too large to be discussed in their entirety in this section – a common strategy involves the construction of long-short portfolios based on the most influential characteristics. In another exemplary approach, portfolios may be constructed based on return prediction, where an investor buys stocks with high return predictions and shorts stocks with low predictions[29].

In this domain-specific context, the link between partial derivatives, variable importance and Jacobian regularisation is intriguing. Although we do not pursue the issue further at this point, it is conceivable that an objective function applies semi-automated Jacobian regularisation. Semi-automated regularisation could mean that the penalty is not necessarily applied freely or entirely data-driven but is manually imposed instead. This imposition suggests that an investor may wish to manually penalise the influence of specific characteristics or even individual assets more than others. Two possible practical applications may be economic, social and governance (ESG) investing, where an investor may wish to limit the influence of non-ESG assets.

Another example may be 1-month momentum, which can be an expensive signal to trade on due to high turnovers and potentially high transaction costs, as it is mainly driven by illiquid and expensive to trade microcaps. An intuitive counter-argument would be to exclude particular assets or firm characteristics from the training sample. However, as we discuss in section 1, at a monthly frequency, financial data is relatively scarce compared to other domains in which machine learning models thrive, making training ultra-complex neural networks or other machine learning models difficult. In this setting, making use of all available data is crucial. Therefore, semi-automated Jacobian regularisation may be a way for institutional investors, for example, to use all available data but to manually limit the signal extracted from certain assets, industries, or even entire firm characteristics.

Table 4 summarises the most important characteristics by model, measured as the absolute median partial derivative and estimated over the entire sample (left panel) and the most recent five years only (right panel). Our definition for variable importance, summarised in equation (40) is not new to the general machine learning literature. Similar approaches are proposed, for example, by Ruck et al. (1990) or Leray and Gallinari (1999). As a consequence, we do not claim to propose a generally new concept

[29] In this context, the model must not necessarily be good in predicting returns but predict the rank of future returns. Nonetheless, model performance is reliant on the most influential firm characteristics.

	Most important characteristics: entire sample					Most important characteristics: 2016-2020				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
NN	beta	mom1m	mve	std_turn	mom12m	ep	mom12m	invest	sp	pchsale_pchinvt
NN-W1	mom1m	std_turn	mom12m	turn	retvol	mom36m	beta	roavol	retvol	mom1m
NN-W2	mom12m	mom1m	beta	ill	retvol	roavol	turn	pchsale_pchrect	indmom	ear
NN-W1W2	mom12m	mom1m	std_turn	retvol	mve	mom12m	ep	sgr	gma	currat
NN-J1	std_turn	mom1m	maxret	retvol	mve	lev	roavol	currat	cfp	turn
NN-J2	mom12m	roavol	mom1m	roaq	mve	mom12m	salecash	ep	indmom	roavol
NN-J1J2	mom12m	beta	roavol	mom1m	mve	mve	roavol	agr	maxret	sgr
NN-J1-m	mom12m	mom1m	mom36m	roaq	ill	indmom	mve	dy	beta	pchsale_pchinvt
NN-J2-m	mom12m	std_turn	beta	mom1m	mve	chpmia	beta	mom36m	ill	chinv
NN-J1J2-m	mom1m	beta	mom12m	std_turn	mve	cashpr	indmom	salecash	lgr	ill

Legend:

momentum	profitability	trading frictions	intangibles	value vs growth	investments

Table 4:
Most important characteristics:
The table summarises the most important characteristics measured in absolute median partial derivatives. The left panel reports the most important characteristics estimated over the entire sample, while the right panel focuses on the most recent five years (2016-2020). The categories by which the characteristics are grouped by follow the definitions of Hou et al. (2020).

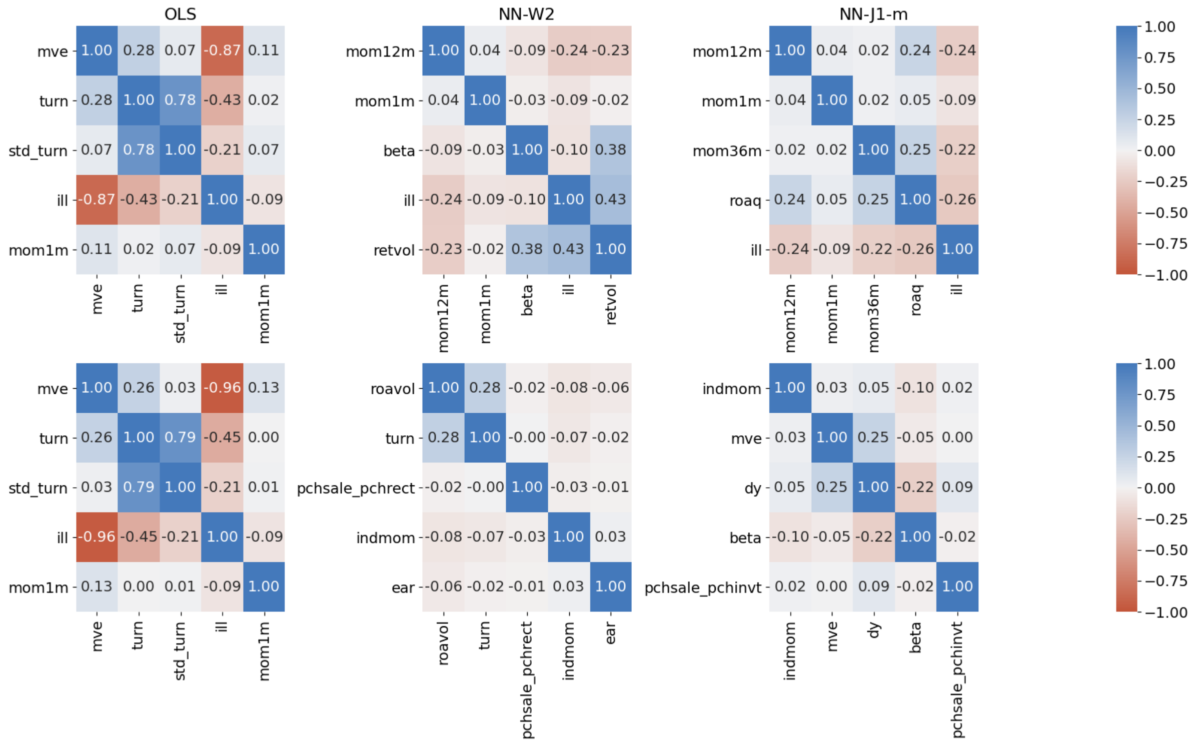
in this section. Instead, we point out that due to the economic interpretation of the partial derivatives in the domain-specific setting of empirical asset pricing, variable importance is closely related to the estimation of risk premia (see section 2.2 and 2.4). Moreover, while table 4 reports aggregated variable importances (for the overall sample, or the most recent five years), in section 4.12 we further discuss that the partial derivatives on asset-level provide even more granular insight into the inner model mechanics.

The table serves two purposes. First, it offers a guide for which characteristics we investigate further in subsequent sections. Second, it provides initial evidence for the time-varying nature of variable importance, despite being reported as an aggregate. For example, while momentum and trading friction characteristics dominate variable importances that have been estimated over the entire sample, characteristics that fall into the value vs growth, intangibles or investments category seem to have moved to the foreground in more recent years indicating that the estimated variable importance is time-dependent. At the same time, table 4 serves as a sanity check, as the firm characteristics that can be found in the table are also frequently discovered by other literature. For example, table 4 includes nine of the ten overall most important firm characteristics that are discovered by the neural network designed by Gu et al. (2020b). Differences in the order of characteristics or the absence of specific characteristics in table 4 in comparison to existing literature may be caused by diverging data preprocessing, network architectures, objective function designs, or even training regimes (e.g. including/excluding batch-normalisation, dropout or learning rate shrinkage). Nonetheless, the fact that table 4 includes commonly referenced firm characteristics boosts the confidence in our findings.

The findings summarised in table 4 are particularly interesting with the dataset-specific dynamics discussed in section 4.4 and allow for a critical evaluation. The fact that firm characteristics that fall into the momentum and trading frictions category dominate in an overall estimation may not be surprising. The reason for this is that momentum and trading friction characteristics tend to be more volatile than the majority of other firm characteristics, as they are updated more frequently than firm characteristics that are predominantly based on annual report information. The visualisation of the time-series of the rank-normalised book-to-market values shown in figure 6, for example, illustrates the limited volatility of firm characteristics that are predominantly based on annual report information. Thus, in order to increase model performance, it may be desirable to increase the volatility of the input variables, for example, through firm characteristic interactions with other economic or return time-series (e.g. see Gu et al. (2020b)). Since we do not claim to provide a new out-of-sample performance benchmark but primarily focus on the proposed methodology, we do not include such an empirical study in this paper. However, the observation that *fast-moving* firm characteristics tend to come out on top in the empirical asset pricing literature compared to relatively *slow-moving* firm characteristics is currently not discussed enough. This paper provides further empirical evidence in this direction.

A common and reoccurring criticism in the context of variable importance is the correlation structure of the most influential input variables. If a model's most influential input variables are highly correlated with each other, it is unclear how much independent *signal* each is adding. Equally, if two different models pick different firm characteristics, but they are themselves highly correlated, it is again unclear if it was simply by chance that one model picked one variable over the other. In high-dimensional settings, such as in contemporary empirical asset pricing, (minor) multicollinearity is typically present in the input data. Therefore, the model's top selections should ideally be uncorrelated and uncorrelated to other models' selections, where the other models provide similar performance benchmarks.

Figure 10 visualises the spearman correlations of the top five most influential firm characteristics for NN-W2 and NN-J1-m. The left column includes OLS as a linear benchmark. The top panel refers to the overall estimate, while the bottom panel refers to the top picks of the most recent five years. The figure clearly shows that the firm characteristics picked by OLS are much stronger correlated than those picked by the two neural networks. The top five most influential input firm characteristics for NN-W2 and NN-J1-m in the over and most recent five years estimation are only very weakly correlated. An exception are illiquidity and return volatility, which are moderately correlated in the overall estimation for NN-W2 with a Spearman correlation of 0.43. We conclude that NN-W2 and NN-J1-m are better capable of extracting independent signals from different firm characteristics when compared to linear regressions. In addition, figure 11 investigates the correlations of the overall most influential firm characteristics across models, including OLS, NN-W2 and NN-J1-m. The left panel shows the correlation matrix estimated over the entire sample, and the right panel shows the correlation matrix estimated using only data from the most recent five years. Moreover, the most influential characteristics for OLS are boxed in the top left corner. It can be seen that the firm characteristics that are considered most influential for NN-W2 and NN-J1-m are only weakly to moderately correlated when analysed over the entire sample and much less correlated in recent years. The fact that the most influential firm characteristics are largely weakly correlated (or at least only moderately) strengthens the confidence in the signal extracted from each of the characteristics due to the low multicollinearity. In combination with previous sections, we conclude



OLS

	mve	turn	std_turn	ill	mom1m
mve	1.00	0.28	0.07	-0.87	0.11
turn	0.28	1.00	0.78	-0.43	0.02
std_turn	0.07	0.78	1.00	-0.21	0.07
ill	-0.87	-0.43	-0.21	1.00	-0.09
mom1m	0.11	0.02	0.07	-0.09	1.00

NN-W2

	mom12m	mom1m	beta	ill	retvol
mom12m	1.00	0.04	-0.09	-0.24	-0.23
mom1m	0.04	1.00	-0.03	-0.09	-0.02
beta	-0.09	-0.03	1.00	-0.10	0.38
ill	-0.24	-0.09	-0.10	1.00	0.43
retvol	-0.23	-0.02	0.38	0.43	1.00

NN-J1-m

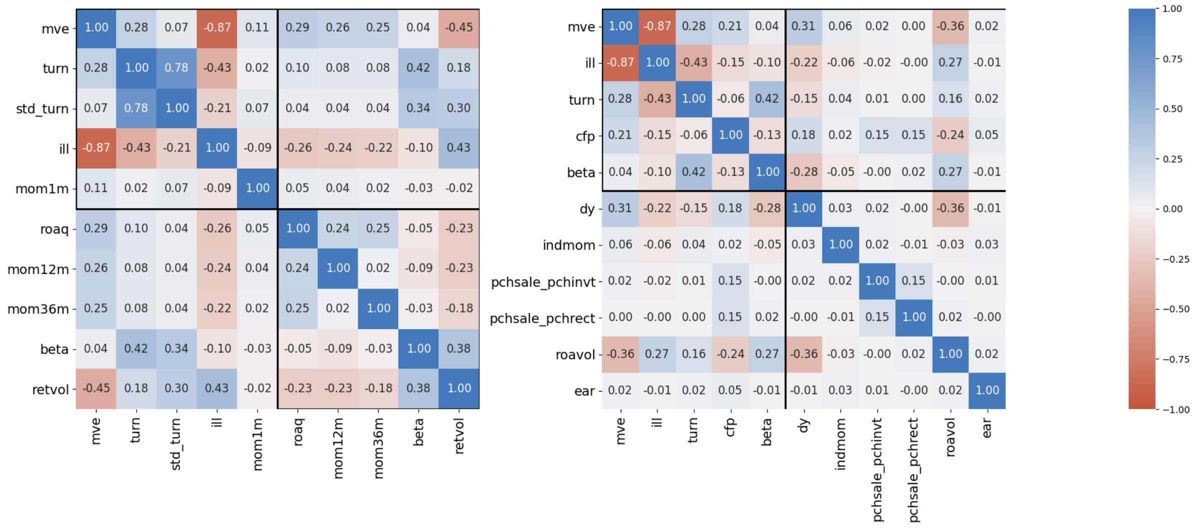
	mom12m	mom1m	mom36m	roaq	ill
mom12m	1.00	0.04	0.02	0.24	-0.24
mom1m	0.04	1.00	0.02	0.05	-0.09
mom36m	0.02	0.02	1.00	0.25	-0.22
roaq	0.24	0.05	0.25	1.00	-0.26
ill	-0.24	-0.09	-0.22	-0.26	1.00

	mve	turn	std_turn	ill	mom1m
mve	1.00	0.26	0.03	-0.96	0.13
turn	0.26	1.00	0.79	-0.45	0.00
std_turn	0.03	0.79	1.00	-0.21	0.01
ill	-0.96	-0.45	-0.21	1.00	-0.09
mom1m	0.13	0.00	0.01	-0.09	1.00

	roavol	turn	pchsale_pchrect	indmom	ear
roavol	1.00	0.28	-0.02	-0.08	-0.06
turn	0.28	1.00	-0.00	-0.07	-0.02
pchsale_pchrect	-0.02	-0.00	1.00	-0.03	-0.01
indmom	-0.08	-0.07	-0.03	1.00	0.03
ear	-0.06	-0.02	-0.01	0.03	1.00

	indmom	mve	dy	beta	pchsale_pchinvt
indmom	1.00	0.03	0.05	-0.10	0.02
mve	0.03	1.00	0.25	-0.05	0.00
dy	0.05	0.25	1.00	-0.22	0.09
beta	-0.10	-0.05	-0.22	1.00	-0.02
pchsale_pchinvt	0.02	0.00	0.09	-0.02	1.00

Figure 10:
Spearman correlations of most significant firm characteristics within models
The graph displays the spearman correlation matrices of the top five most influential firm characteristics by model, including OLS, which serves as a linear benchmark, NN-W2 and NN-J1-m. The top row displays the correlation matrices estimated over the entire sample, and the bottom row is estimated over the most recent five years only.



	mve	turn	std_turn	ill	mom1m	roaq	mom12m	mom36m	beta	retvol
mve	1.00	0.28	0.07	-0.87	0.11	0.29	0.26	0.25	0.04	-0.45
turn	0.28	1.00	0.78	-0.43	0.02	0.10	0.08	0.08	0.42	0.18
std_turn	0.07	0.78	1.00	-0.21	0.07	0.04	0.04	0.04	0.34	0.30
ill	-0.87	-0.43	-0.21	1.00	-0.09	-0.26	-0.24	-0.22	-0.10	0.43
mom1m	0.11	0.02	0.07	-0.09	1.00	0.05	0.04	0.02	-0.03	-0.02
roaq	0.29	0.10	0.04	-0.26	0.05	1.00	0.24	0.25	-0.05	-0.23
mom12m	0.26	0.08	0.04	-0.24	0.04	0.24	1.00	0.02	-0.09	-0.23
mom36m	0.25	0.08	0.04	-0.22	0.02	0.25	0.02	1.00	-0.03	-0.18
beta	0.04	0.42	0.34	-0.10	-0.03	-0.05	-0.09	-0.03	1.00	0.38
retvol	-0.45	0.18	0.30	0.43	-0.02	-0.23	-0.23	-0.18	0.38	1.00

	mve	ill	turn	cfp	beta	dy	indmom	pchsale_pchinvt	pchsale_pchrect	roavol	ear
mve	1.00	-0.87	0.28	0.21	0.04	0.31	0.06	0.02	0.00	-0.36	0.02
ill	-0.87	1.00	-0.43	-0.15	-0.10	-0.22	-0.06	-0.02	-0.00	0.27	-0.01
turn	0.28	-0.43	1.00	-0.06	0.42	-0.15	0.04	0.01	0.00	0.16	0.02
cfp	0.21	-0.15	-0.06	1.00	-0.13	0.18	0.02	0.15	0.15	-0.24	0.05
beta	0.04	-0.10	0.42	-0.13	1.00	-0.28	-0.05	-0.00	0.02	0.27	-0.01
dy	0.31	-0.22	-0.15	0.18	-0.28	1.00	0.03	0.02	-0.00	-0.36	-0.01
indmom	0.06	-0.06	0.04	0.02	-0.05	0.03	1.00	0.02	-0.01	-0.03	0.03
pchsale_pchinvt	0.02	-0.02	0.01	0.15	-0.00	0.02	0.02	1.00	0.15	-0.00	0.01
pchsale_pchrect	0.00	-0.00	0.00	0.15	0.02	-0.00	-0.01	0.15	1.00	0.02	-0.00
roavol	-0.36	0.27	0.16	-0.24	0.27	-0.36	-0.03	-0.00	0.02	1.00	0.02
ear	0.02	-0.01	0.02	0.05	-0.01	-0.01	0.03	0.01	-0.00	0.02	1.00

Figure 11:
Spearman correlations of most significant firm characteristics across models
The graph displays the spearman correlation matrices of the top five most influential firm characteristics across models, including OLS, which serves as a linear benchmark, NN-W2 and NN-J1-m. The left panel refers to the overall estimated correlation matrix, while the right panel refers to the most recent five years only.

that the nonlinear estimation through neural networks offers competitive out-of-sample performances and desirable properties regarding the distribution of the partial derivatives.

4.10 Asset Selection

Table 5 summarises the effect of the objective function on the dimensionality reduction, where partial derivatives within the empirical range of $[-5e^{-4}, 5e^{-4}]$ are considered empirically less critical or unimportant. The table 5 reports the cross-characteristic and time-series average percentage[30] of partial derivatives that are empirically unimportant. Due to the aggregated nature of table 5, granularity is lost. Therefore, appendix H provides further detail on the time-varying nature of the empirical variable selection. Moreover, due to the evaluation on the asset-level, we use the term *asset selection* and the more general machine learning term *dimensionality reduction* interchangeably in this section. From this perspective, table 5 reports the percentage of de-selected assets, that empirically do not drive model predictions.

Our critical empirical findings are four-fold. First, the asset selection or dimensionality reduction effect is most pronounced in the case of Jacobian regularisation, compared to penalising the weights in the objective function. While this theoretical property of the Jacobian regularisation is evident, due to the nature of the objective function, table 5 confirms it empirically. For example, in the case of training neural networks on the sub-selection of the 49 core characteristics only, L_1 norm Jacobian regularisation reduces the dimensionality by 49.6%. In contrast, in the case of no regularisation in the objective function, merely 26.4% of the partial derivatives are considered empirically unimportant.

Second, the variable selection strength is most substantial in the case of element-wise Jacobian regularisation. In the example of using the core characteristics only, element-wise L_1 norm Jacobian regularisation empirically deselects 49.6% of the assets on average, compared to 29.6% in the case of column-wise L_1 norm Jacobian regularisation. Due to the nature of the different Jacobian penalisations, the difference in the strength of asset selection is expected. For example, penalising the column means towards zero does not necessarily mean that every asset's partial derivative must be close to zero, as in the case of element-wise penalisation. Third, we empirically find that the variable selection effect of Jacobian regularisation is more pronounced in higher than lower dimensions. For example, in the case of training the neural networks on all 103 firm characteristics, with column-wise L_1 norm Jacobian regularisation, 40.6% of the partial derivatives are considered empirically less critical, compared to only 29.7% in the case of the 49 core characteristics only. This relative difference in the strength of dimensionality reduction, dependent on the dimensionality of the input, is expected, as we expect the actual but unknown data-generating process to be sparse (e.g. Kozak et al. (2020)). Thus, we find that Jacobian regularisation yields empirical results that are in line with prior expectations.

Fourth, however, we find no significant difference in average asset selection across size classes—. For example, for column-wise L_1 and L_2 Jacobian regularisation, and in the core characteristics case,

[30] At each month, we calculate the percentage of the asset-level partial derivatives that lie within the empirical threshold of $[-5e^{-4}, 5e^{-4}]$, average the percentage across firm characteristics. In the last step, we report the time-series average of the time-varying cross-characteristic average in table 5.

	Core Characteristics				All Characteristics			
	Overall	Large	Small	Micro	Overall	Large	Small	Micro
NN	26.4	28.4	27.2	25.2	31.6	35.1	32.5	29.8
NN-W1	26.6	29.8	27.9	25.1	31.0	33.8	31.1	29.8
NN-W2	27.9	30.0	28.4	26.7	32.5	35.5	33.8	30.8
NN-W1W2	30.0	32.4	30.8	28.9	44.8	49.5	45.7	42.8
NN-J1	49.6	52.9	51.0	48.1	60.0	62.3	60.7	58.8
NN-J2	35.6	37.4	36.0	34.8	36.3	41.2	38.0	34.0
NN-J1J2	56.1	59.5	57.3	54.3	58.4	63.7	59.8	56.0
NN-J1-m	29.7	31.8	30.3	28.8	37.2	40.6	37.8	35.7
NN-J2-m	30.9	33.0	31.5	29.8	34.5	36.1	34.8	33.9
NN-J1J2-m	34.0	35.4	34.0	33.6	39.7	43.0	40.3	38.2

Table 5:
Average dimensionality reduction [%]:
The table reports the cross-characteristic and time-series percentage average of partial derivatives that are considered less empirically important, where partial derivatives within $[-5e^{-4}, 5e^{-4}]$ are considered less important.

the overall dimensionality reduction is 34.0%, 35.4%, 34.0% and 33.6% for large, small, respectively microcaps. This finding further generalises across all types of objective functions considered in this paper. We have previously discussed the unbalancedness of financial datasets and the practical problems that are imposed by microcaps. Thus, the marginal influence of information extracted from microcaps would ideally be less than for small or large stocks. It is conceivable to impose stricter penalties on microcaps manually to counter this empirical behaviour of the Jacobian regularisation. However, an empirical evaluation of this alternative fitting strategy is beyond the scope of this paper.

This brief excursion in the effects of the objective function on the influence on the asset level demonstrates the strength of our proposed methodology. To the best of our knowledge, there exist limited empirical insights in current literature into which particular assets drive asset pricing models in an (almost) unrestricted investment universe. A further model investigation on the asset level is beyond the scope of this paper. However, it is conceivable to further evaluate which particular assets or industries are the most dominant drivers in the models under consideration. We return to an asset-level model insight in section 4.12.

4.11 Risk Compensation

In this section, we focus on the risk price estimation, with the methodology introduced in section 2. In particular, this section emphasises the advantages of our proposed methodology, as risk prices can be estimated by economically meaningful subgroups such as industries or size classes. In addition, the derived tolerance bands allow particularly useful insights into the estimated time-variability of risk prices, which is far superior to aggregated point estimates reported in tables.

Similarly to Lewellen (2015), figure 12 summarises the time-varying risk premia, estimated by NN-W2 and NN-J1-m. The risk premia are estimated following the methodology presented in section 2.2

and 2.4. For clarity, we again focus on results from models trained on the core characteristics only. Additional risk premia estimates, such as for models trained on all firm characteristics, different forms of data-preprocessing and risk premia estimations when microcaps are excluded entirely, can be found in appendix I.

The figure shows that our proposed methodology yields a time-varying risk premium estimate with tolerance bands. We consider risk premia to be empirically important if the tolerance bands do not include zero. To make our results directly comparable to existing literature, we also report risk price estimates following standard Fama-Macbeth regression (e.g. see section 2.1). In particular, the linear risk price estimates are reported as a constant time-series average, which is common in literature (e.g. see Green et al. (2017)). To increase the comparability of results, we diverge slightly from the methodology presented in Green et al. (2017) and apply the same cross-sectional rank-normalisation as data preprocessing for the linear estimates. Consequently, the input data for all neural network estimates is the same as in the case of linear regressions. Secondly, while Green et al. (2017) also consider an estimation using all firm characteristics, we are primarily focussing on the core characteristics case in this section. However, this core selection diverges slightly from the methodology presented in Green et al. (2017). Third, analogously to the training of the neural networks (which is done annually to reduce computational cost), the linear regressions are re-fitted annually to make the results from linear regressions directly comparable to the neural networks' output. Due to these slight differences compared to the original paper presented by Green et al. (2017), there are minor differences in the risk price estimates.

The signs of the linearly replicated risk price estimates are broadly identical to those reported in Green et al. (2017), boosting the confidence in our empirical results. There are only four firm characteristics, including change in 6-month momentum (chmom), current ratio (currat), illiquidity (ill) and 12-month momentum (mom12m), for which our linear estimates seem to be of the opposite sign, compared to the signs reported in Green et al. (2017). These minor differences may be due to the diverging data preprocessing, annual re-fitting and slightly diverging firm characteristic subset selection. Interestingly, the sign of the risk premium for 12-month momentum estimated linearly is contrary to the economic intuition. The risk premium estimated by the neural networks, on the other hand, is economically plausible. Most importantly, however, the signs of the risk premia earned by each firm characteristic, estimated by neural networks, are also broadly in line with existing research, such as Lewellen (2015) or Green et al. (2017). This alignment with existing research is one of the most important findings: the risk premia estimated through our proposed methodology are economically meaningful and in line with existing research. This critical finding fills our previous argument about model interpretability with life, as economically meaningful estimates are essential when communicating model output.

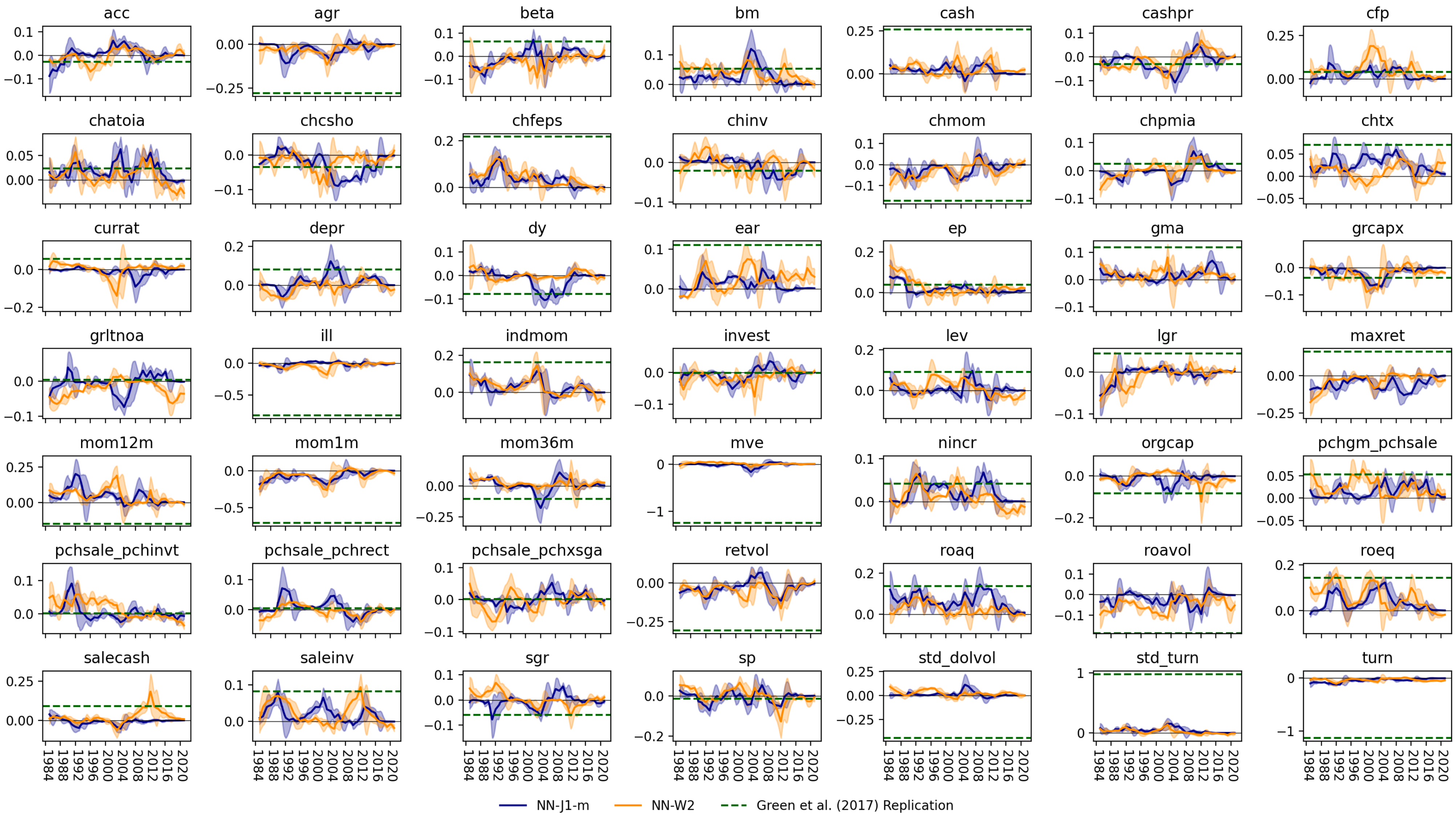


Figure 12:
Time-varying risk premia, by size class – NN-J1-m, NN-W2, core characteristics
The graph plots the time-varying risk premia estimations for NN-J1-m (blue) and NN-W2 (orange), where the dotted green line represents the analogous linear estimation as a benchmark or sanity check.

Further, figure 12 shows that for a small number of risk premia, such as illiquidity (ill), size (mve), volatility of liquidity (std_turn), and turnover (turn), the estimates by the linear benchmark are extreme outliers. On the other hand, none of the risk premia estimated by the neural networks can be classified as extreme relative to each other. This *stability* is a strong advantage of the neural network estimation compared to the linear benchmark. In particular, the opposite signs of estimated risk premia for the volatility of liquidity (std_turn) and turnover (turn) in the case of linear estimation is particularly worrying as both firm characteristics are highly positively correlated (see appendix F). Therefore, we confirm the well-documented fact that linear regressions can be unstable in the presence of multicollinearity. In contrast, both neural networks (NN-J1-m and NN-W2) provide much more stable estimates.

Overall, figure 12 emphasises the strength of our proposed methodology as it can be seen that risk premia are time-varying and transitioning in and out of empirical importance. This time-variation suggests that a time-dependent risk premia consideration offers richer insights than an aggregated perspective, for example, if risk premia are estimated as time-series averages. Further, figure 12 demonstrates that the choice of the objective function directly impacts the risk price estimation. To exemplify the importance of the objective function, figure I.1 in appendix I plots the risk premia estimates for neural networks with no weight or Jacobian regularisation (NN) and the element-wise L_1 norm Jacobian regularisation. As discussed in section 4.10, element-wise L_1 norm Jacobian regularisation is arguably the strongest and, therefore, not necessarily the ideal regularisation due to an overly strong asset selection. The other extreme – no regularisation – is not beneficial either. As a result, the figure shows that most risk premia estimated by NN-J1 are empirically insignificant, with only a few exceptions. On the other hand, no regularisation hardly yields any variable selection, with most risk premia remaining empirically significant over time. To the best of our knowledge, the importance of the objective function in the context of risk premia estimation is not yet well-documented in current literature. This section intends to provide first empirical evidence supporting an ongoing debate about an economically meaningful objective function choice.

The risk premia displayed in figure 12 are estimated across all size classes. However, in section 4.1, we discuss how the investment universe under consideration consists to 60% of microcaps. Further, microcaps are expensive to trade due to their illiquidity and increased transaction costs. To check the robustness of our estimates across firm size classes (large, small and microcaps), figures 13 and 14 plot the risk premia estimated separately by size class. As discussed in section 2.3, our proposed methodology allows for size class-specific estimations without the need to re-fit the model. Figures 13 and 14 show that the risk premia estimates provided by neural networks are robust across size classes, indicated by only minor differences in each separately estimated risk premium.

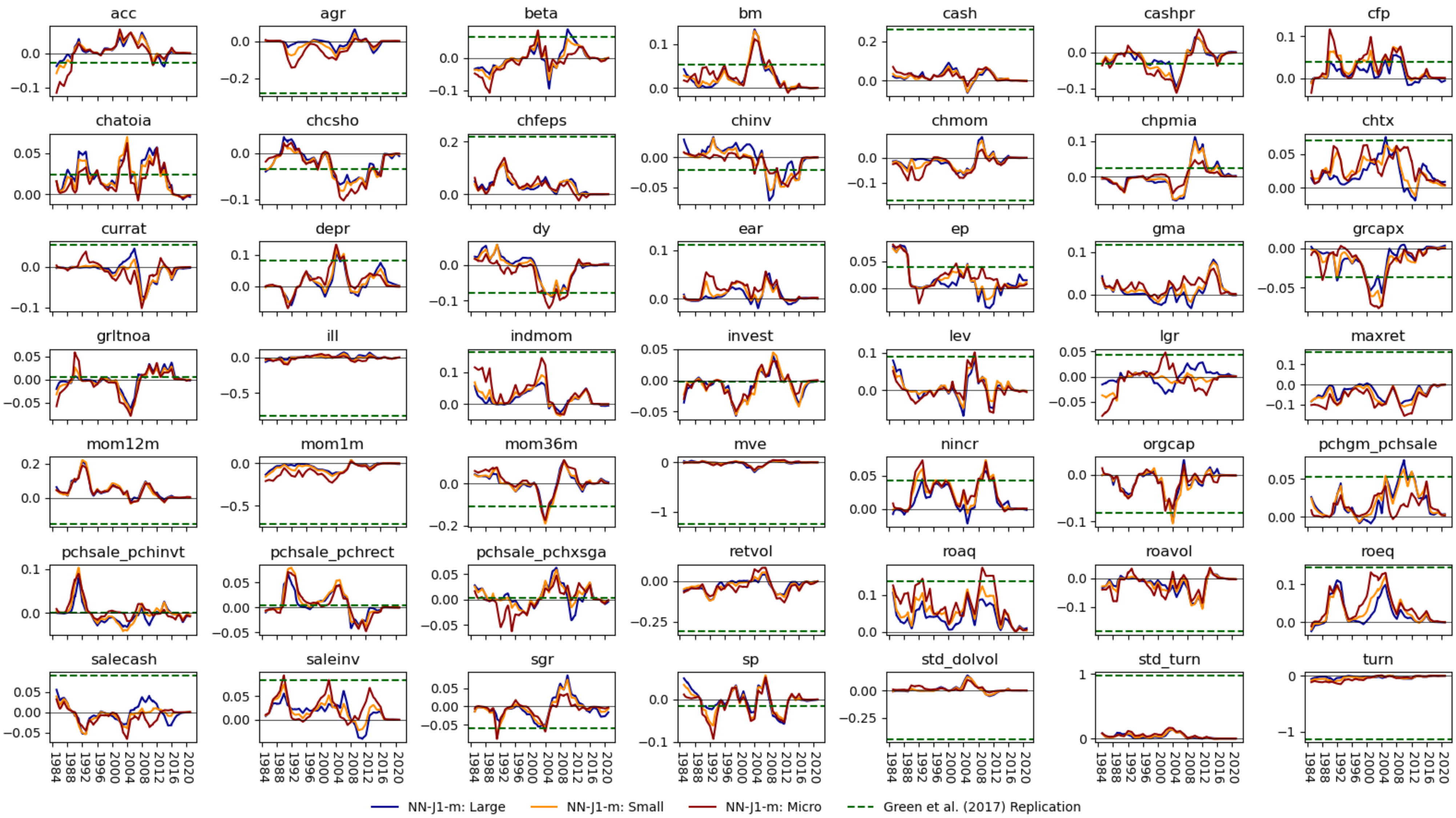


Figure 13:
Time-varying risk premia, by size class – NN-J1-m, core characteristics
The graph plots the time-varying risk premia estimations by size class for NN-J1-m (blue = large, orange = small, red = microcaps), where the dotted green line represents the analogous linear estimation as a benchmark or sanity check.

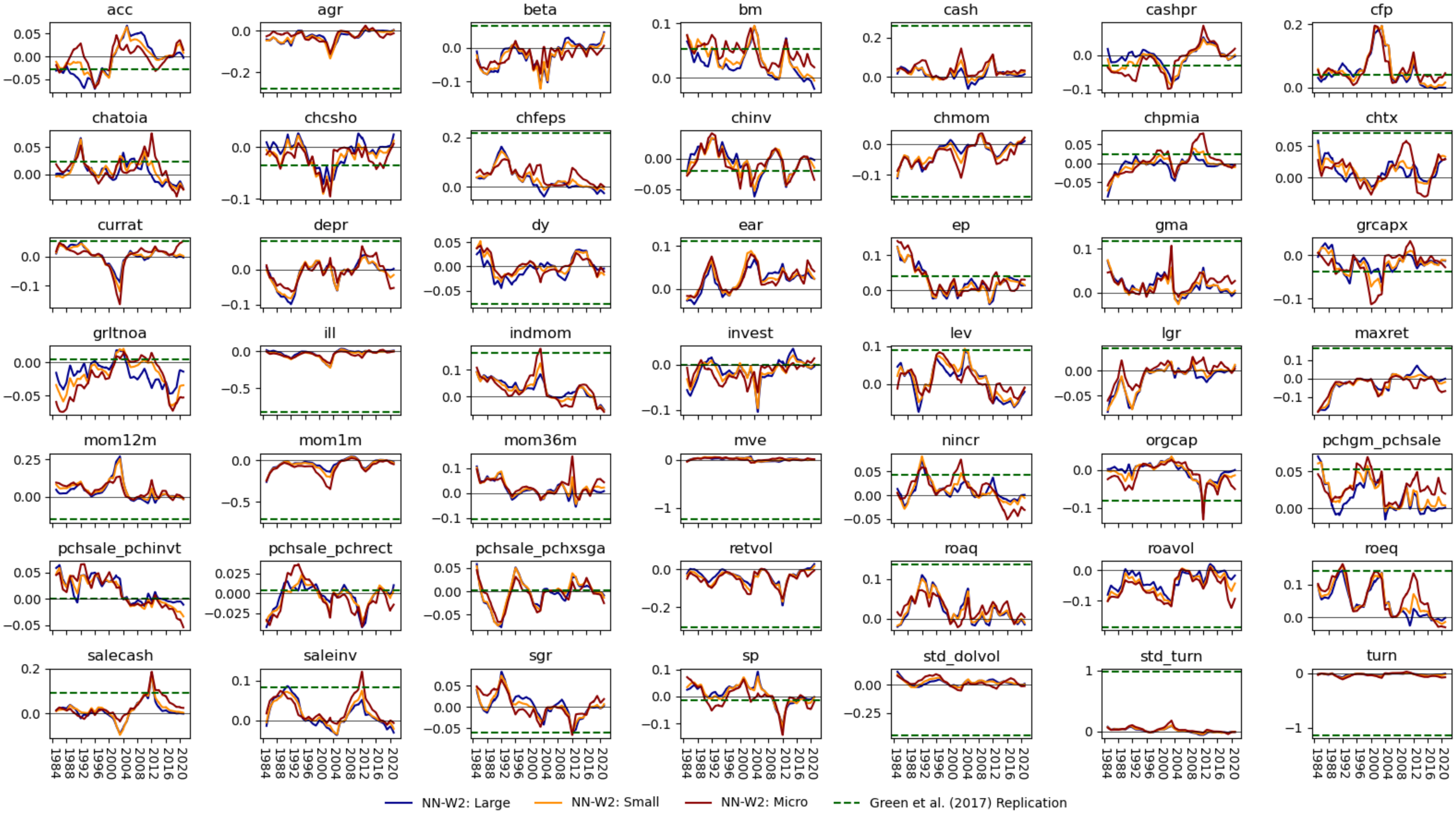


Figure 14:
Time-varying risk premia, by size class – NN-W2, core characteristics
The graph plots the time-varying risk premia estimations by size class for NN-W2 (blue = large, orange = small, red = microcaps), where the dotted green line represents the analogous linear estimation as a benchmark or sanity check.

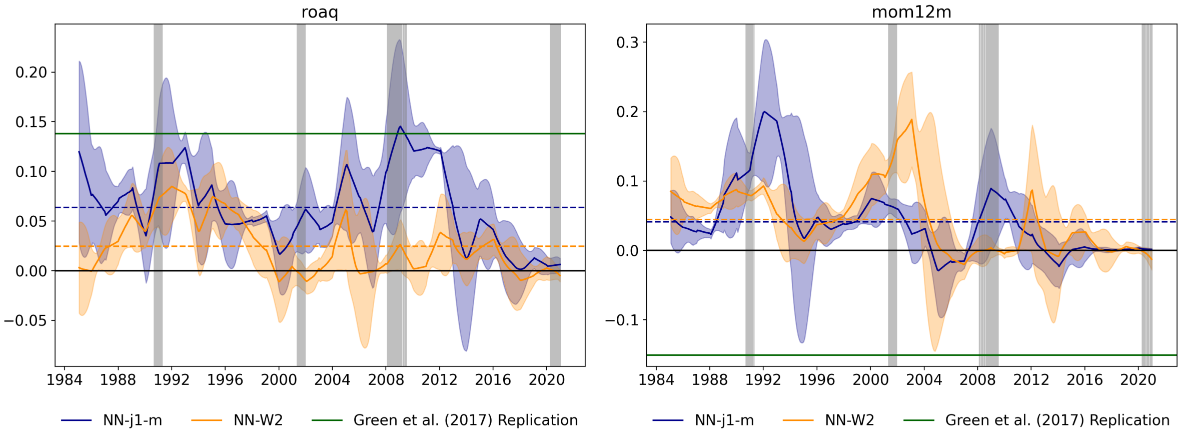


Figure 15:
Time-varying risk premia: NN-J1-m and NN-W2 (roaq, mom12m)
The left panel shows the estimated time-varying risk premium for being exposed to the firm characteristic return-on-assets. The right panel shows the estimated time-varying risk premium for being exposed to the firm characteristic 12-month momentum. the blue line with tolerance bands refer to the estimation by NN-J1-m, and the orange line with tolerance bands refer to the estimation by NN-W2. The green line indicates the linear replication, analogously to Green et al. (2017), but translated into our setting to make the results directly comparable.

Figures 12 to 14 respectively plot all 49 firm characteristics, which are included in the core characteristics sub-selection. An individual and separate discussion of each of the 49 estimated risk premia is beyond the scope of this paper. Instead, we handpick two firm characteristics as examples and discuss the risk premium exposure to those firm characteristics earns over time. Those two include return-on-assets, which is among the overall most influential characteristics for NN-J1-m (see section 4.9), and 12-month momentum, which is the overall most influential characteristics for NN-W2 and NN-J1-m. Figure 15 plots the time-varying risk premia estimates for being exposed to each of the two firm characteristics return-on-assets and 12-month momentum. Similar to the previous plots, the solid green line refers to the linear replication of risk price estimates, analogously to Green et al. (2017), but translated into the same data preprocessing and annual re-fitting as the neural networks to make the results directly comparable. The figure shows that the unconditional risk price estimate, analogously to equation (11), are far inferior to the time-varying risk price estimates, especially in combination with the empirical tolerance bands.

We do not claim that all estimated risk premia by neural networks are always consistent with economic theory. The examples discussed above, however, show that neural networks can indeed provide economically meaningful estimates. Due to their time-varying nature, an investor may use this additional information for further research, such as market timing. While a complete discussion is beyond the scope of this paper, and there is compelling evidence showing severe difficulties investors are facing when timing the market (e.g. Dichtl et al. (2019)), the discussion in this section shows that approaching the risk price estimation from a time-varying perspective, an investor may have more meaningful information at hand, compared to an aggregated time-series average.

First, we focus on the left panel in figure 15 which refers to the risk premium estimated by NN-J1-m and NN-W2 an investor is estimated to receive for being exposed to the return-on-assets characteristic. The graph shows that the risk premium over time slowly shrinks towards zero but lies almost entirely on the positive side of the x-axis, suggesting a positive relationship between the risk exposure and expected returns. This positive relation is also well-documented in the literature, for example, by Balakrishnan et al. (2010). Similar to Lewellen (2015) who identifies a similar pattern, the slight decrease in magnitude in compensation suggests that past risk premia estimates are likely to overstate the cross-sectional dispersion in actual expected returns going forward. This diminishing effect seems to be more pronounced in the case of J1-m, compared to NN-W2, for which the tolerance bands more often include zero and are, therefore, more frequently considered empirically less important.

An interesting pattern in the case of the risk price estimation based on NN-J1-m emerges during the great financial crises. As argued by Balakrishnan et al. (2010), return-on-assets is particularly indicative regarding losses. In particular, they argue that if past earnings announcements report a loss, there is an increased probability of loss reports going forward. Since there is a positive relationship between earnings (losses) and returns, it is interesting to see that this relationship seems to be strengthening during the financial crisis in the case of NN-J1-m – a period characterised by significant abnormal losses. We do not claim that all risk premia estimated by the neural networks follow stringent economic reasoning. Moreover, a complete discussion of all risk premia is beyond the scope of this paper. However, the example of return-on-assets shows that it is possible to find economic meaning in the neural network's estimations. It is important to note that in the example above, the period of the great financial crisis was particularly turbulent, and return-on-assets is a quarterly variable, meaning that new information becomes available more frequently compared to characteristics of annual frequency. It is not necessarily clear if the frequency also influences the increased magnitude of risk prices.

Second, the right panel in figure 15 refers to the risk premium estimated by NN-J1-m and NN-W2 an investor is estimated to receive for being exposed to the 12-month momentum characteristic. Surprisingly, the linear estimation is on the *wrong* side of the x-axis, as economic intuition suggests a positive relationship between risk premia and expected returns (e.g. see Jegadeesh (1990)). Similarly to the previous discussion about return-on-assets, the risk premia estimated by the two neural networks are on the *right* side of the x-axis. However, the magnitude seems to be diminishing over time. It can also be seen that the two different objective functions yield different time-dependent risk premia.

A potential criticism of the analysis presented above is that the time-varying nature of the estimated risk prices is not exclusive to nonlinear models such as neural networks. In fact, instead of estimating the risk prices unconditionally – as is typically done in classical, linear Fama-Macbeth regressions – one could report the conditional risk premia estimates of OLS. A complete comparison of all conditional risk premia across all models and firm characteristics is beyond the scope of this section. Instead, figure 16 exemplarily focuses on the time-varying risk premia estimates by NN-J1-m, NN-J1, and NN-W2 for an

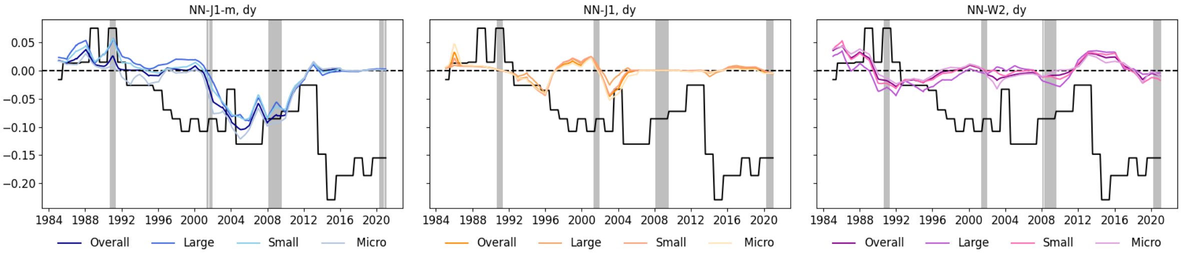


Figure 16:
Conditional risk premia comparison – dividend-yield
The graph compares the conditional risk premia estimates of the linear benchmark OLS with three different neural networks, including NN-J1-m (left), NN-J1 (middle), and NN-W2 (right).

exposure to dividend-yield, where, for clarity, no tolerance bands are reported. The figure provides solid empirical evidence favouring our nonlinear methodology: Due to the annual re-fitting and the intrinsic linearity of OLS, the conditional risk prices merely change once a year (when a new model is re-fitted). In contrast, due to the nonlinearity of neural networks, despite the annual re-fitting strategy, the estimated risk prices are still varying. In addition, figure 16 shows that with nonlinear neural networks, a risk price estimation by market capitalisation is possible without the need to re-fit the model. Such insight cannot be gained from linear regressions unless the model is re-fitted on a subgroup only.

In the example of dividend yield, the risk prices estimated by NN-J1-m are broadly in line with the linear estimation until the early 2010s (except for the late 1990s and early 2000s, where the risk premium, estimated by NN-J1-m, is near zero). However, in more recent years, the two estimates diverge starkly, as the risk premia estimated by NN-J1-m (and NN-J1 and NN-W2) are zero, while the linear model estimates negative premia. The middle panel of figure 16 shows the risk premia estimated by NN-J1. It can be seen that compared to the other two neural networks, NN-J1 much more strictly estimates no risk premia resulting from exposure to the dividend yield. The risk premia estimated by NN-W2, displayed in the right panel, are less rigorously set to zero than NN-J1 but diverge slightly from the risk premia estimated by NN-J1-m. For example, OLS and NN-J1-m estimate negative risk premia associated with exposures to dividend yield in the mid-2000s, while NN-W2 does not seem to estimate such a negative relationship. However, what unites all three neural networks is the possibility of a separate risk premia estimation by market capitalisation, which can be very insightful. For example, it can be seen that in the early 1990s, NN-J1-m estimates opposite signs for the risk premia expected to be earned by large stocks and microcaps. A total economic discussion is beyond the scope of this paper. However, figure 16 provides an insightful example for the usefulness of our proposed methodology as it allows for much more granular analyses compared to regular linear modelling.

We conclude that the risk premia estimated by the two neural networks presented in this section are robust across size classes, are economically meaningful, and broadly in line with existing research. Moreover, the time-varying estimation in combination with tolerance bands offers helpful insights that are superior to constant time-series averages, resulting from traditional Fama-Macbeth regressions. The choice of the objective function should become an integral part of model design. Further robustness checks and results are presented in appendix I.

4.12 Model Interpretability and Inner Model Mechanics

This section explores the inner model mechanics of neural networks further and shows that leveraging the partial derivatives offers valuable model insights that help improve model interpretability and explainability. In particular, we present empirical evidence suggesting strong nonlinear input variable interactions. We exemplify how the neural networks' return sensitivities for changes in firm characteristics vary nonlinearly across assets, given return sensitivities for changes in other firm characteristics. The crucial finding is that these nonlinear sensitivity interactions are time-dependent and depend on the assets' market capitalisation. An analogous analysis by industry or market capitalisation and industry is also conceivable. However, a complete discussion about inner model mechanics by industry is beyond the scope of this section. Instead, we provide an exemplary analysis in appendix J.

Further, we discuss how analysing the partial derivatives of neural networks with respect to firm characteristics on asset level relative to the firm characteristic inputs can offer relevant insights into the functioning of the objective function and can help with software debugging. Moreover, the detection of extreme partial derivative outliers can also be worthy. While individual assets usually do not significantly influence the overall model performance, mainly when the model performance is evaluated based on thousands or tens of thousands of assets, an investor may still be interested in how the model handles individual assets. Extreme partial derivatives help identify individual assets that the model does not seem to be handling well. In particular, extreme partial derivatives can indicate unstable return predictions, where the term unstable in this context describes the circumstance that even minor changes in the input variables can lead to substantially varying return predictions. This form of model instability may be undesirable for an investor.

Similar to previous sections, reporting all empirical results across all models, different data preprocessing regimes, or points in time is beyond the scope of this section. Thus, we primarily focus on the neural networks NN-J1-m and NN-W2 and draw our analysis to a small selection of different time points to exemplify our general findings and refer readers to appendix J for further empirical analyses. Figures 17 and 18 visualise how the neural networks' return sensitivities for changes in firm characteristics vary nonlinearly across assets, given return sensitivities for changes in other firm characteristics. The nonlinear sensitivity interactions are estimated by market capitalisation (i.e. large, small and micro stocks) following the locally weighted regression method proposed by Cleveland (1979). Most importantly, the figures only refer to a single point in time. Since the nonlinear sensitivity interactions are time-varying, the findings in figures 17 and 18 cannot necessarily be generalised to other points in time. Further,

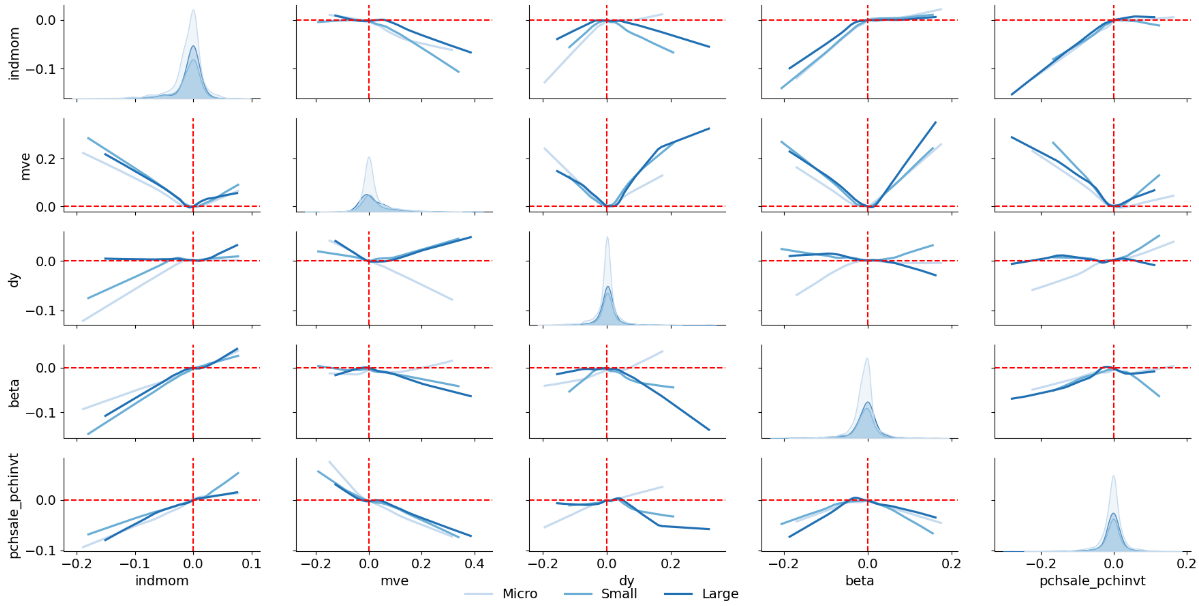


Figure 17:
Nonlinear sensitivity interactions NN-J1-m – exemplary date 2017-06-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are estimated by market capitalisation. The diagonal displays the distribution of sensitivities by market capitalisation. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J1-m, analogously to table 4.

NN-J1-m and NN-W2 were both estimated using 49 firm characteristics. However, figures 17 and ?? merely visualise the nonlinear sensitivity interactions of the most influential firm characteristics over the most recent five years for each model, analogously to table 4.

In simple terms, the figures show that for a given level of return prediction sensitivity to the firm characteristics on the x-axis, the lines represent the expected return prediction sensitivity to changes in the firm characteristics listed on the y-axis. As a specific example, consider the top panel in the center column of figure 17. The panel shows that for the neural network NN-J1-m and date 2017-06-30, assets that show a positive return prediction sensitivity to changes in dividend yield (dy) are expected to be negatively sensitive to changes in industry-momentum (indmom). Moreover, the greater this return prediction sensitivity is to changes in dividend yield, the more negative we expect the sensitivity to be to changes in industry momentum, with this relationship being stronger for small and large stocks. However, for microcaps, this relationship is relatively weakened or even of the opposite sign, where the expected sensitivity to industry momentum is positive, given a positive sensitivity to changes in dividend yield. For stocks for which the model estimates a negative return prediction sensitivity to changes in dividend yield, the effect on the return prediction sensitivity to changes in industry momentum is expected to remain negative, with this relationship being more pronounced for microcaps.

A complete discussion of all interactions presented in figures 17 and 18 is beyond the scope of this section, which is why we merely concentrate on the methodology to estimate interaction in the following.

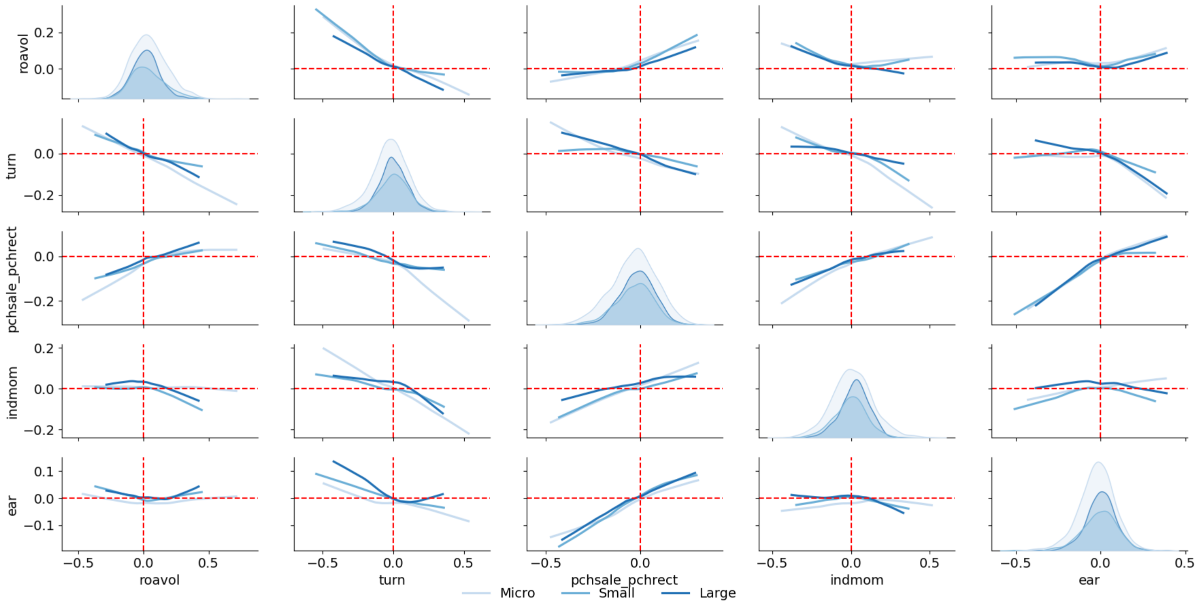


Figure 18:
Nonlinear sensitivity interactions NN-W2 – exemplary date 2017-06-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-W2, analogously to table 4.

To the best of our knowledge, we are the first to propose the analysis of nonlinear derivative interactions in empirical asset pricing to improve model interpretability and explainability. Input feature interaction and the detection of nonlinearities in the model estimation mechanism are widely studied fields in the general machine learning literature. In particular, in fields such as asset pricing, where machine learning models are designed for critical decision making, an understanding of the inner model mechanics is paramount (e.g. see Goodman and Flaxman (2017) for an interdisciplinary approach to the issue of model interpretability). The idea of leveraging the input gradients to increase model interpretability is generally not new to the general machine learning literature and in the field of image recognition in particular. Exemplary approaches leveraging the input gradients include Simonyan et al. (2013), Ross et al. (2017) or Sundararajan et al. (2017). The methodology shown in figures 17 and 18 differs from the common approaches used in image recognition in that they are concerned with the question of how the sensitivity to changes in a characteristic is expected to be for a given asset and given the sensitivity to a change in another characteristic.

Besides the nonlinear sensitivity interactions, a neural network's interpretability and explainability can further be improved by investigating the relationship between the partial derivatives of a neural network with respect to an input characteristic and the input characteristic itself. Since the partial derivatives are calculated on asset level, this approach helps to confirm the implications of an objective function and visually discovers outliers. For example, we examine the firm characteristic 1-month mo-

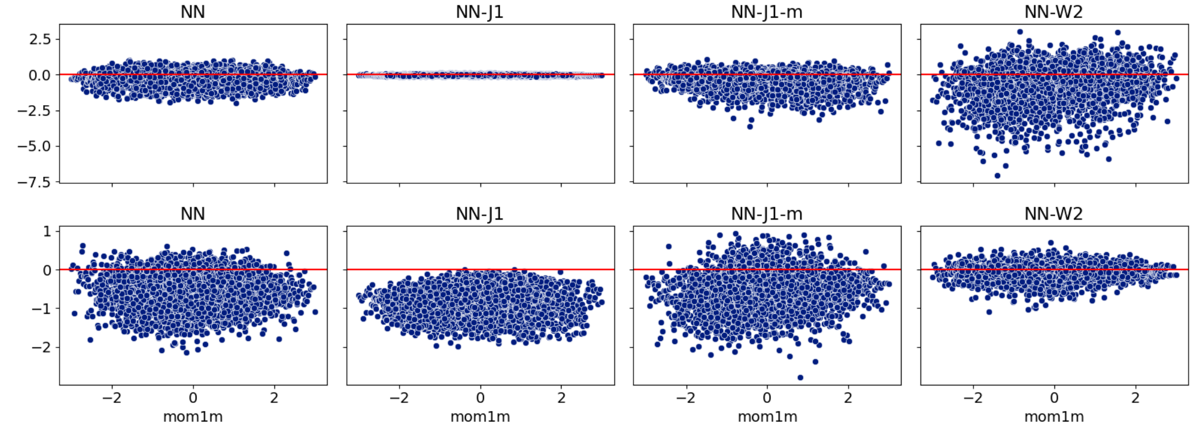


Figure 19:
Input gradients – 1-month momentum:
The top panel shows the input gradients with respect to 1-month momentum plotted against the input values themselves. The top row refers to 1999-08-31 and the bottom row to 2002-02-28. The plot shows the distributional properties of the input gradients of neural networks with no regularisation (NN), element-wise, respectively column-wise, L_1 norm Jacobian regularisation (NN-J1, NN-J1-m) and L_2 norm weight regularisation (NN-W1).

mentum further – a characteristic frequently reported as one of the overall most essential input features. Table 4 shows that 1-month momentum is also estimated to be among the top five most influential firm characteristics when the variable importance is estimated over the entire sample for NN, NN-W1, NN-W2, NN-W1W2, NN-J1-m, NN-J2-m and NN-J1J2-m, making it a relevant characteristic in our empirical analysis. Figure 19 visualises the input gradient of four different neural networks for 1-month momentum on the y-axis (the input gradients are multiplied by 100), relative to the rank-normalised 1-month momentum input on the x-axis. It is important to note that the distributional patterns shown in figure 19 do not necessarily generalise to other points in time. To emphasise the time-varying nature of the input gradient, the top row in figure 19 refers to the date 1999-08-31, and the bottom row to the date 2002-02-28. Each dot in the figure represents a single asset (identified by a unique stock ID).

The first and most noticeable takeaway is that the distribution of partial derivatives varies significantly depending on time and objective function. For example, figure 19 shows that for NN-W2, the input gradients are much more widespread in 1999-08-31, with the partial derivatives ranging from -7.08 to 3.03, compared to an input gradient spread ranging from -1.08 to 0.70 in 2002-02-28. Furthermore, in 1999-08-31, the element-wise L_1 norm Jacobian regularisation yields an extremely narrow distributional band around zero, meaning that 1-month momentum was not selected as an essential feature for that particular neural network then. Even though this feature selection does not generalise to all points in time, it is an empirical confirmation of the functioning of the Jacobian penalty as part of the objective function. Interestingly, in 2002-02-28, NN-J1 yields almost exclusively negative partial derivatives, which is in line with economic theory. A similar but less pronounced pattern becomes apparent for NN-J1-m in 2002-02-28.

The previously discussed but rather naive approach of analysing the distributional properties in general (see also section 4.8), figure 19 naturally alludes to the notion of a section-wise analysis, where, for example, the distributional properties in the extreme or equally in common values are of interest. For example, in figure 19, in 2002-02-28, for the neural network with no regularisation, the input gradients are relatively evenly distributed across the input values. In comparison, for NN-J1-m, the distribution is the widest for typical input values of 1-month momentum and much more narrowly distributed for extreme input values.

After highlighting the general finding that the partial derivatives are time-varying and dependent on the objective function, we focus on the possibility and usefulness of investigating marginal model sensitivity on asset-level in the following. As an example, figure 20 visualises the input gradients with respect to the firm characteristic *size* in 1992-06-30 plotted against the rank-normalised firm characteristic input (the input gradients are multiplied by 100). The graph summarises the distributional shapes of the four exemplary neural networks NN, NN-W1, NN-J1-m and NN-J2. It can be seen that in the case of NN-W1, some of the input gradients are extreme outliers. In particular, the graph highlights a single asset, marked in red. The partial derivative corresponds to the stock *Frontier Adjusters America Inc.* (permno = 10628, in the CRSP universe). The size of the stock is in the bottom 25th percentile, and the company classifies as a microcap. Although the market capitalisation of *Frontier Adjusters America Inc.* is not an extreme value, the question becomes why the return prediction sensitivity to changes in size for NN-W1 is so extreme or, more importantly, why would a stakeholder care about it. In addition, it can be seen from figure 20 that the prediction sensitivity to changes in size is much smaller, for example, for NN-J1-m and NN-J1.

There may be different reasons for why one would care about a stock-level analysis. For example, if an investor was to construct long-short portfolios based on sorted partial derivatives (see section 4.13 for further analyses on the topic), the investor may want to know what stocks end up in the portfolio. Alternatively, one may also be interested in limiting the disproportionate marginal influence of a single asset on the prediction. An analysis on asset level as in figure 20, therefore, offers the opportunity to validate the design of the objective function and serves as a sanity check if the objective function achieves its purpose.

To exemplarily investigate the case of *Frontier Adjusters America Inc.* further, figure 21 displays the partial derivatives of NN-W2 with respect to all 49 firm characteristics at 1992-06-30 (the input gradients are multiplied by 100). It can be seen that the model does not seem to be handling this particular asset well, as a large proportion of the partial derivatives are of a large magnitude. For the example of the partial derivative with respect to size, and to illustrate why it is essential to be aware of such model outliers if all other input values are kept constant for *Frontier Adjusters America Inc.*, and only the input values for the firm characteristic size are changed minimally, the return predictions become unstable, as indicated by the large input gradient. More specifically, when the firm characteristics size is only varied

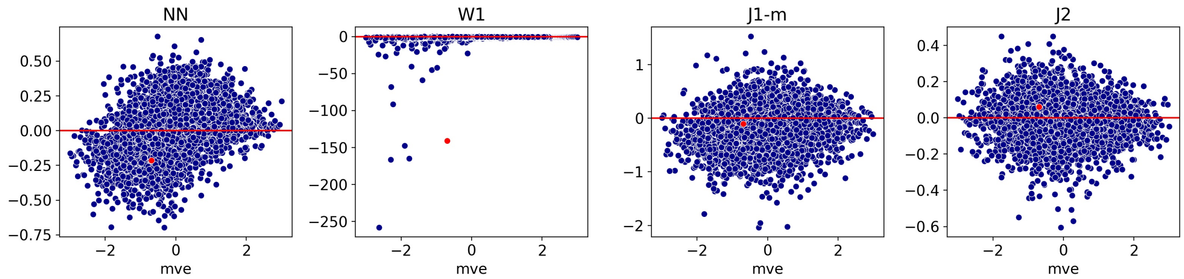


Figure 20:
Partial derivatives − Size:
The plot displays the partial derivatives of three different neural networks with respect to the firm characteristic *size* in 1994. The left panel refers to a neural network with L_1 weight, the panel in the middle to a neural network with L_1 Jacobian penalisation and the right panel to a neural network with L_2 Jacobian penalisation in its objective function. Each dot represents a single asset, with the red dot exemplarily highlighting the company *Illinois Central Corp.*

by ±1% in rank, which equates to a range of 23.17th - 25.38th percentile and is a tiny change in the magnitude of the input variable, the return predictions vary by up to almost 10%. For comparisons, the same variation in size merely yields a variation in return prediction of 0.007% for NN-J1-m. This example shows that extreme outliers in the partial derivatives can be an indicator for unstable predictions. In practice, such small changes in the input variables can quickly occur due to rounding errors or erroneous data pre-processing. If a stakeholder is cautious about prediction stability, the analysis mentioned above can be a tool to evaluate the effectiveness of the desired functioning of the objective function of choice.

We do not claim that neural networks trained with Jacobian regularisation as part of their objective function are entirely exempt from the issues related to extreme model sensitivities. Instead, we intend to bring to the attention that simple tools such as the analysis of the partial derivatives, which offers model insights on the asset level, may help detect such issues.

So far, we have primarily focused the distributional analysis on the asset level of the input gradients relative to the respective firm characteristic inputs, such as plotting the partial derivatives of a neural network with respect to 1-month momentum against the very 1-month momentum input values. There is, however, no intuitive justification for limiting this type of analysis to this case. Instead, an investor or other stakeholder may also be interested in the distributional properties of the input gradients relative to another firm characteristic input, or even in a multi-dimensional setting, relative to multiple other firm characteristic inputs. An in-depth discussion of all possible angles to this approach is beyond the scope of this section. However, to illustrate the possibility of this further analysis, figure 22 exemplarily visualises the distributional properties of the partial derivatives of NN-J1-m with respect to book-to-market (bm) plotted against the overall top five most influential firm characteristics for NN-J1-m. The figure shows that distributions may differ depending on the input characteristic.

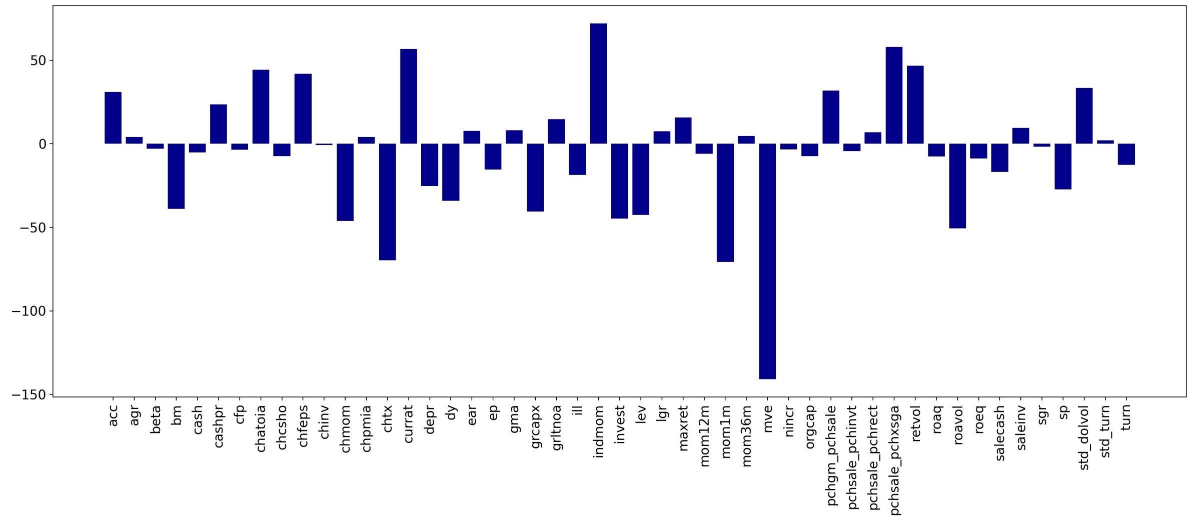


Figure 21:
Partial derivatives – Size:
The plot displays the partial derivatives of three different neural networks with respect to the firm characteristic *size* in 1994. The left panel refers to a neural network with L_1 weight, the panel in the middle to a neural network with L_1 Jacobian penalisation and the right panel to a neural network with L_2 Jacobian penalisation in its objective function. Each dot represents a single asset, with the red dot exemplarily highlighting the company *Illinois Central Corp.*

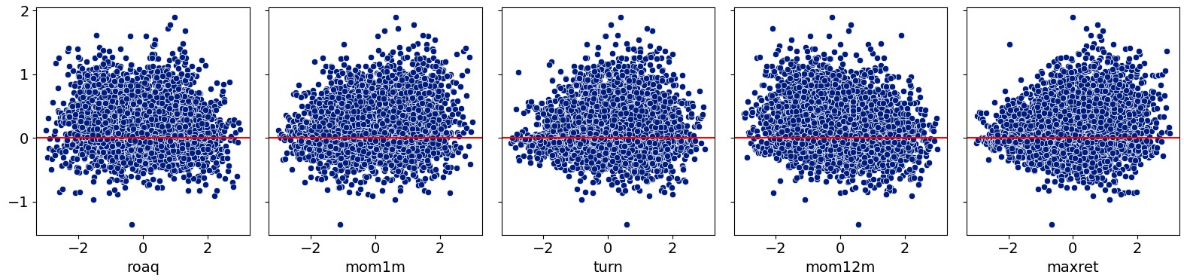


Figure 22:
Input gradient (book-to-market) against most influential firm characteristics (NN-J1-m):
The graphs shows the input gradients with respect to book-to-market plotted against the top five most influential input characteristics in 1999-08-31 for NN-J1-m.

4.13 Double-Sorted Portfolios

Last but not least, we discuss the implications of model sensitivities in the context of (double-) sorted portfolios. Our main contribution in this section is the introduction of the concept of sensitivity-sorted portfolios. Characteristic sorted portfolios, in general, are widely used in the empirical asset pricing literature. Specifically, they belong to the overarching umbrella of literature exploring *anomalies*, where portfolios are formed based on sorted characteristics. The most common procedure in the empirical literature consists of sorting stocks by a given characteristic, such as size or book-to-market, followed by grouping the stocks into N portfolios, where N is typically 3, 5 or 10. This paper constructs quintile portfolios (i.e. $N = 5$) as quintile portfolios offer a balanced trade-off between diversification and spread. Subsequently, equal-weighted or value-weighted portfolio returns are computed for each portfolio and at each point in time. An investor may decide to go long the top quintile and short the bottom quintile.

The fundamental economic theory, or economic conjecture, behind characteristic-sorted portfolios, is the hypothesis that expected returns should be increasing (or decreasing) in some characteristic, e.g. see Patton and Timmermann (2008). The literature in this field is vast, and we do not claim to provide a holistic overview in this section. However, the most relevant literature in the field of single-sorted portfolios includes Basu (1977, 1983), Fama and French (1992, 1993, 2006), Banz (1981), Reinganum (1981), Ang et al. (2006a) or Jegadeesh and Titman (1993). Moreover, in recent years, a growing body of literature has evolved, discussing the statistical properties of portfolio sorts. The informal recognition that portfolio sorts can be a nonparametric alternative to imposing linearity on the relationship between expected returns and firm characteristics has fuelled this recent development, as discussed by Fama and French (2008) or Cochrane (2011). Examples of the discussion of the statistical properties include, but are not limited to, Patton and Timmermann (2008) or Cattaneo et al. (2020).

This paper does not claim to discover previously undiscovered relationships between firm characteristics and expected returns based on portfolio sorts. Instead, we introduce the conceptual framework that an investor may consider model sensitivities as part of the portfolio construction exercise, especially risk management. We propose the following portfolio sorting procedure. First, we follow a standard methodology of sorting assets into five portfolios based on firm characteristics, such as return-on-assets. In a second step, we further sort the assets in each portfolio into five more portfolios based on the out-of-sample estimated model sensitivity to changes in the firm characteristic, yielding a total of 25 double-sorted portfolios. The choice of constructing quintile portfolios offers a balanced trade-off between diversification (with an average of 180 stocks in each double-sorted portfolio) and spread of the firm characteristics. The nuance on the out-of-sample sensitivity estimation is subtle but essential: the portfolios are constructed only on information available to an investor at time t. Thus, using out-of-sample estimated return prediction sensitivities rather than in-sample estimates avoids information leakage. In a final step, we construct equal-weighted or value-weighted portfolio returns, where the value corresponds to the market value of equity at time $t-1$.

The reasoning behind the proposed methodology is that with single-sorted portfolios, an investor is *betting* on the continuation of a back-tested empirical relationship between firm characteristics and expected returns. Under the assumption that an estimation function g generalises well out-of-sample (where we intentionally use the somewhat loosely defined term *well* to appreciate that in real-life, even state-of-the-art machine learning models can only explain a small fraction of future returns), sorting on model sensitivity further introduces a bet on *volatility*. For example, if a model's return prediction for a particular asset is highly sensitive to changes in a given firm characteristic, even a slight divergence in the expected value of the future value of that firm characteristic can lead to significant volatilities in return prediction (see section 4.12).

As a consequence, the second sort can be interpreted as a *bet* on volatility, where a low return prediction sensitivity – given a model g that is generalising well out-of-sample – corresponds to less volatile return predictions and a higher sensitivity to a higher return prediction volatility. In other words, return predictions for assets for which the model prediction sensitivities to changes in specific firm characteristics is high are less stable if there are unexpected changes in the firm characteristic. This circumstance can be understood in that the model is more particular about some assets than others. Consequently, we would expect portfolio volatilities to be lower for portfolios sorted on low sensitivities compared to portfolios sorted on high sensitivities.

In addition to the above-mentioned portfolio-sorting procedure, we report the single sorted portfolio returns as a simple benchmark. However, a direct comparison between the two portfolio construction, especially regarding portfolio volatility, is limited since the single sorted portfolios are much more diversified and consist of an average of 900 assets per basket, compared to 180 for the double-sorted portfolios. Therefore, the single-sorted portfolios are by definition alone much more diversified. This effect is amplified in an investment universe that includes microcaps, which are much more volatile in general (see table A.1). To increase comparability, we also report portfolio performances for an investment universe excluding microcaps entirely.

Similarly to previous sessions, a holistic overview of all possible portfolio sorts, including estimated out-of-sample sensitivities from all models, data preprocessing regimes (such as data normalisation or the inclusion or exclusion of microcaps), is beyond the scope of this section. To make this section consistent with the previous sections, we merely focus on the two neural networks NN-W2 and NN-J1-m and the two characteristics, 12-month momentum and return-on-assets. 12-month momentum appears in the top five overall most influential characteristics for both models and is the overall most influential characteristic for NN-W2, see section 4.9). Return-on-assets is among the overall most influential characteristic for NN-J1-m, see section 4.9). In particular, table 6 summarises the annualised monthly average return, volatilities and Sharpe ratios of the single-sorted and double-sorted quintile portfolios, where the single-sorted portfolios serve as a benchmark. The table further differentiates between equal and value-weighted portfolio returns. Additional portfolios summaries, including portfolios that are constructed on an investment universe that excludes microcaps can be found in appendix K along with the corresponding tables regarding the portfolios constructed on the 12-month momentum signal.

	Annualised Returns [%]						Annualised Volatility [%]						Annualised Sharpe Ratio					
	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High
Panel A: Value-Weighted Portfolios – NN-W2																		
Low	6.72	10.36	6.79	4.28	8.81	6.75	27.53	31.79	28.48	28.18	29.15	30.01	0.24	0.31	0.23	0.15	0.29	0.22
Q2	6.60	7.50	6.65	7.53	6.89	6.12	21.24	23.77	22.61	23.36	23.20	23.87	0.30	0.31	0.29	0.31	0.29	0.25
Q3	8.59	9.73	9.05	8.73	7.01	7.58	17.19	18.38	17.77	19.23	18.35	18.85	0.48	0.51	0.49	0.44	0.37	0.39
Q4	8.95	11.16	9.19	10.63	8.63	6.33	14.47	16.55	15.54	15.86	15.62	18.13	0.59	0.64	0.57	0.64	0.53	0.34
High	11.25	12.11	14.52	10.46	9.22	12.03	15.83	19.56	16.72	16.47	17.69	20.02	0.68	0.59	0.82	0.61	0.50	0.57
Panel B: Equal-Weighted Portfolios – NN-W2																		
Low	9.32	9.33	11.49	8.06	8.26	7.42	32.27	32.40	31.18	31.23	32.26	33.40	0.28	0.28	0.35	0.25	0.25	0.21
Q2	9.83	11.73	9.74	9.95	10.40	7.04	21.24	22.61	20.95	21.13	22.18	23.14	0.44	0.49	0.45	0.45	0.45	0.29
Q3	11.75	14.18	12.08	11.16	11.14	10.86	16.75	18.25	17.05	16.87	17.50	18.98	0.67	0.73	0.67	0.63	0.61	0.55
Q4	12.92	14.91	14.27	12.84	12.43	10.25	17.88	18.54	17.63	17.94	18.31	20.25	0.68	0.75	0.76	0.68	0.64	0.48
High	13.97	15.47	13.61	14.51	13.20	13.64	18.95	19.85	18.67	19.06	19.71	21.06	0.69	0.73	0.69	0.71	0.63	0.61
Panel C: Value-Weighted Portfolios – NN-J1-m																		
Low	6.72	5.07	9.04	6.24	6.50	3.58	27.53	27.49	29.47	31.06	30.74	30.91	0.24	0.18	0.29	0.20	0.21	0.11
Q2	6.60	7.48	9.40	6.33	3.56	5.09	21.24	22.04	23.12	22.47	22.82	23.38	0.30	0.33	0.39	0.27	0.15	0.21
Q3	8.59	8.76	9.90	9.23	7.59	5.92	17.19	17.64	18.16	18.56	18.54	18.46	0.48	0.48	0.52	0.48	0.40	0.31
Q4	8.95	8.79	9.31	9.41	10.44	9.41	14.47	15.24	15.97	15.74	15.27	18.66	0.59	0.55	0.56	0.57	0.65	0.48
High	11.25	11.14	11.69	9.41	12.99	9.16	15.83	17.04	16.48	16.80	17.92	19.59	0.68	0.62	0.67	0.54	0.68	0.45
Panel D: Equal-Weighted Portfolios – NN-J1-m																		
Low	9.32	11.22	12.51	10.43	7.21	6.05	32.27	31.67	33.50	33.31	34.04	32.39	0.28	0.34	0.35	0.30	0.21	0.18
Q2	9.83	12.88	11.34	9.89	8.00	7.33	21.24	21.33	21.37	21.96	21.76	23.20	0.44	0.57	0.51	0.43	0.35	0.31
Q3	11.75	12.15	12.32	12.61	11.38	9.69	16.75	17.31	16.78	17.33	17.24	18.07	0.67	0.67	0.70	0.69	0.63	0.51
Q4	12.92	13.21	12.27	13.17	13.02	12.59	17.88	17.51	18.18	18.11	18.36	19.84	0.68	0.71	0.64	0.69	0.67	0.60
High	13.97	15.27	14.23	13.33	14.63	13.27	18.95	19.17	19.15	19.32	18.98	21.41	0.69	0.75	0.70	0.65	0.72	0.59

Table 6:
Double-sorted portfolios – return-on-assets:
The table summarises the annualised average monthly returns, volatilities and Sharpe ratios of double-sorted quintile portfolios. The portfolios are benchmarked against single-sorted quintile portfolios. Double-sorted portfolios are sorted on the characteristic first and by the out-of-sample sensitivity with respect to that sensitivity second.

Table 6 presents an empirical analysis of the portfolio sorting strategy. Portfolio performances, thus, heavily rely on the out-of-sample accuracy of the model sensitivity estimations provided by NN-W2 and NN-J1-m. Moreover, certain firm characteristics may be more important to one model than another, meaning they are intrinsically handled differently across models. For those reasons, we do not claim that the sorting strategy necessarily yields consistent results across all models and firm characteristics. However, exemplary excerpts from table 6 (and the additional tables in appendix K) paint at least an interesting picture. For example, single-sorted portfolios on return-on-assets table 6 confirm the economic intuition that the top quintile portfolio should earn a higher return than the bottom quintile, evidenced by an annualised Sharpe ratio of 0.69 for the top quintile and 0.28 for the bottom quintile portfolio. In addition, the table shows that if a second sorting is introduced, portfolios which are sub-sorts of the top quintile portfolio that are sorted on low sensitivities in addition can improve the Sharpe ratio even further. For example, the annualised Sharpe ratio of the double-sorted quintile portfolio (high characteristic, low sensitivity) is 0.73 compared to 0.69 in the case of the single sort. A similar pattern also emerges in the case of portfolios sorted on the sensitivities estimated by NN-J1-m. In particular, the middle panel of table 6 shows that the volatility tends to increase with high sensitivities.

While we do not claim that these patterns necessarily generalise across all characteristics and neural networks, they are at least noticeable. To the best of our knowledge, we are the first to propose such an out-of-sample portfolios construction approach, where the second sort is based on estimated out-of-sample sensitivities. The recognition that higher sensitivities may lead to higher volatilities and lower returns naturally leads to the concept of constructing long-short portfolios. In the case of single-sorts, the procedure is straight forward: an investor would go long the top quintile portfolio while shorting the bottom quintile. In the case of the double-sorted portfolios, we propose to go long the portfolio sorted on high firm characteristics but low sensitivities, and go short the opposing portfolio sorted on low characteristic values but high sensitivities.

Figure 23 plots the cumulative returns of such long-short portfolio returns, comparing the portfolio returns of those portfolios constructed on the sensitivities of NN-J1-m (top panel) and NN-W2 (bottom panel). In addition, the cumulative returns are benchmarked against the cumulative returns of a broad market portfolio, which we source from Kenneth French's website[31]. The graph displays the different returns of equal-weighted (EW) and value-weighted (VW, by market capitalisation) returns, further including portfolios that are constructed using an investment universe excluding microcaps. The graph underlines that we do not necessarily have discovered previously unknown anomalies, since the portfolio returns heavily rely on microcaps. Nonetheless, it can be seen that, for example, in the case of NN-J1-m, an equally-weighted portfolio without microcaps performed much better during the great financial crises in the 2000s. A full discussion of all portfolio returns is beyond the scope of this section, and we refer

[31] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, last accessed on December 1, 2021.

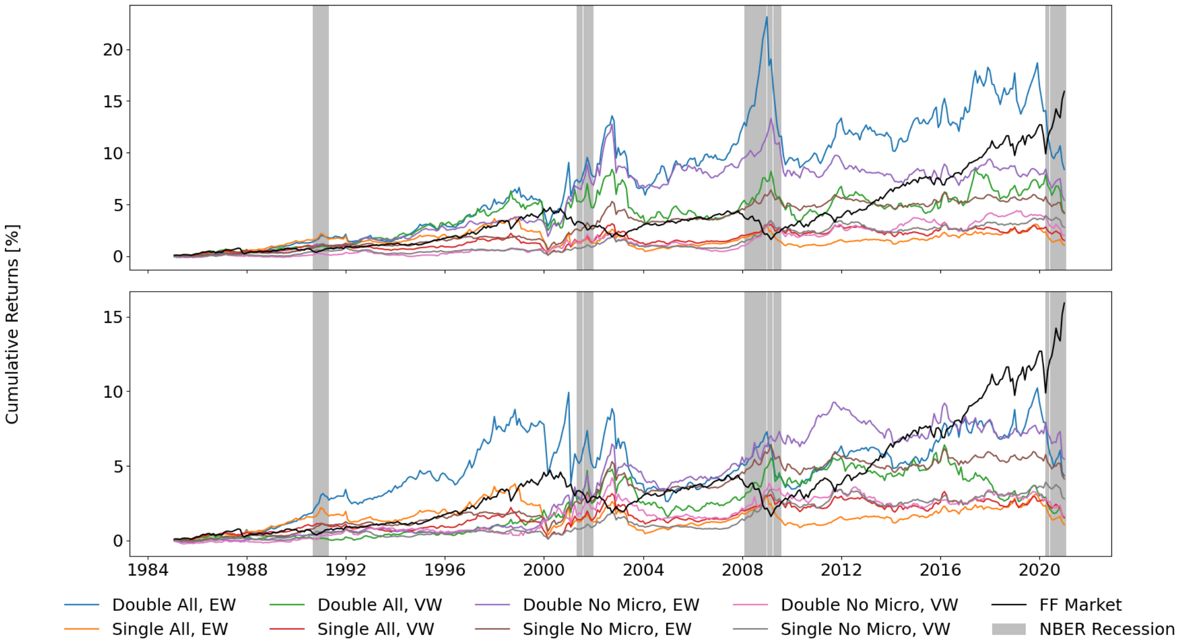


Figure 23:
Cumulative portfolio returns – return-on-assets
The graphs shows the cumulative portfolio returns of single and double-sorted portfolio returns, where the out-of-sample sensitivities of the double-sorted portfolios are estimated by NN-J1-m (top) and NN-W2 (bottom). The cumulative portfolio returns are benchmarked against the cumulative market return, where the market portfolio is sourced from Kenneth French's website.

readers to appendix K for additional information. However, in addition to figure 23, table 7 summarises the results of the classic Fama-French regressions, where the portfolio returns are regressed on a market portfolio and the three commonly used Fama-French benchmark portfolios high-minus-low (HML), small-minus-big (SMB) and momentum (UMD). Moreover, the table merely refers to portfolios constructed on the signal from return-on-assets and we refer readers to appendix K for further results. The table differentiates between single and double-sorted, and equal and value-weighted portfolios. Furthermore, two different investment universes are considered, where one excludes all microcaps. It can be seen that the double-sorted portfolios tend to outperform the single-sorted measured by annualised Sharpe ratio. In addition, the adjusted R^2 tend to be lower for double-sorted portfolios, indicating that the test assets considered in this empirical application are less suitable to explain the returns of double-sorted portfolios compared to single-sorted. The portfolios presented in this section primarily focused on portfolio sorts concerning a single firm characteristic and the corresponding return sensitivity to changes in the same characteristic. However, there is no intuitive justification for limiting the procedure to this simplistic approach. Instead, it is further conceivable to design the sorting procedure of the second sort with sensitivities to changes in other characteristics. In the interest of clarity, we do not pursue this path further, but an alternative method such as sorting on sensitivity to changes in another firm characteristic is conceivable. Moreover, we do not claim to have discovered any new anomalies. Instead, this section intends to introduce the notion of sensitivity-sorted portfolios.

	All stocks				No microcaps			
	NN-J1-m equal-weighed		NN-J1-m value-weighed		NN-J1-m equal-weighed		NN-J1-m value-weighed	
	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort
intercept (α)	0.006**	0.002	0.008***	0.005**	0.006***	0.006***	0.005*	0.006***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Mktrf (β_1)	−0.126**	−0.070	−0.396***	−0.297***	0.163	−0.020	−0.098	−0.290**
	(0.060)	(0.061)	(0.062)	(0.045)	(0.162)	(0.150)	(0.156)	(0.134)
HML (β_2)	0.430***	0.405***	0.166	0.196*	−0.254***	−0.199***	−0.300***	−0.282***
	(0.136)	(0.141)	(0.109)	(0.103)	(0.070)	(0.053)	(0.082)	(0.059)
SMB (β_3)	−0.648***	−0.788***	−0.881***	−0.889***	−0.533***	−0.533***	−0.507***	−0.532***
	(0.096)	(0.135)	(0.085)	(0.078)	(0.142)	(0.179)	(0.112)	(0.134)
UMD (β_4)	0.437***	0.411***	0.258***	0.229***	0.257**	0.152	0.260***	0.182**
	(0.088)	(0.119)	(0.080)	(0.061)	(0.101)	(0.095)	(0.089)	(0.082)
Observations	432		432		432		432	
R^2	0.324	0.357	0.387	0.512	0.332	0.331	0.273	0.415
Adjusted R^2	0.317	0.351	0.382	0.508	0.326	0.325	0.266	0.410
Annualised Return [%]	8.72	4.28	7.35	4.26	6.70	5.57	4.73	4.73
Annualised Volatility [%]	20.38	20.10	22.23	17.91	16.30	13.30	17.86	13.68
Annualised Sharpe Ratio	0.41	0.21	0.32	0.23	0.40	0.41	0.26	0.34
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 7:
Portfolio summary – J1-m, return-on-assets:
The table summarises standard Fama-French regressions, following

$$R_{p,t} = \alpha + \beta_1 \mathrm{Mktrf}_t + \beta_2 \mathrm{HML}_t + \beta_3 \mathrm{SMB}_t + \beta_4 \mathrm{UMD}_t + \epsilon_t,$$

where we regress the respective portfolio returns on the Fama-French 3 Factor model plus a market factor. The market and Fama-French portfolios are sourced directly from Fama's website through WRDS. The constructed portfolio returns are in excess of the risk-free rate.

5 Conclusion

We propose a new way to estimate time-varying risk premia for nonlinear and non-parametric estimator functions. The crucial innovation of our proposed methodology is the nonlinear generalisation of the linear Fama-Macbeth regressions. For this purpose, we estimate risk premia through partial derivatives of a nonlinear estimator function with respect to its input. Our methodology is universally applicable to a large number of estimator functions under the condition of differentiability. We show how the proposed methodology nests the linear Fama-Macbeth regressions as a special case and can be used to estimate risk factor exposures and risk premia.

The second crucial innovation is the introduction of Jacobian regularisation into empirical asset pricing. The newly introduced objective function allows for nonlinear and non-parametric model selection and can be understood as a generalisation of the linear equivalents LASSO, Ridge or Elastic Net. Most importantly, however, due to the economic interpretation of the partial derivatives, Jacobian regularisation is particularly appealing in asset pricing. Our estimation allows us to understand key firm characteristics that drive cross-sectional returns and provides empirical evidence that the inner model mechanics are strongly nonlinear.

Our primary conclusions are four-fold. First, we emphasise the potential of deep neural networks in empirical asset pricing and can show that all neural networks under consideration outperform the linear benchmarks. Second, we show and quantify the importance of regularising the input gradients as part of the objective function. Third, time-varying risk premia estimates provide a much richer insight into cross-sectional returns compared to single-point estimates. Fourth, partial derivatives offer valuable model insights on the asset level and help understand complex model mechanisms better and enable software debugging or the detection of unwanted model biases.

Our empirical findings have direct practical benefits for asset pricing practitioners that go beyond our empirical analysis. First, the conceptual introduction of Jacobian regularisation in asset pricing opens future possibilities for semi-automated penalisation, where practitioners may impose a manual prior. Second, we introduce the concept of sensitivity-sorted portfolios, which can improve portfolio construction from a risk management perspective.

Last but not least, this paper is intended to stimulate a general discussion about model interpretability and explainability of machine learning empirical asset pricing models. The fundamentally different signal-to-noise ratio and the intrinsically ever-changing data dynamics, combined with relative data scarcity, make a direct comparison of model interpretability and explainability to other disciplines such as image recognition difficult. Given our results, model insights on the individual asset level can be crucially important to various stakeholders.

References

Abarbanell, J.S., Bushee, B.J., 1998. Abnormal returns to a fundamental analysis strategy. Accounting Review , 19–45.

Adrian, T., Crump, R.K., Moench, E., 2015. Regression-based estimation of dynamic asset pricing models. Journal of Financial Economics doi:10.1016/j.jfineco.2015.07.004.

Agrawal, P., 2020. The Recent Decade of Drawdown in Value: A Diagnosis and an Enhancement .

Ali, A., Hwang, L.S., Trombley, M.A., 2003. Arbitrage risk and the book-to-market anomaly. Journal of Financial Economics 69, 355–373.

Almeida, H., Campello, M., 2007. Financial constraints, asset tangibility, and corporate investment. The Review of Financial Studies 20, 1429–1460.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets 5, 31–56.

Amihud, Y., Mendelson, H., 1989. The effects of beta, bid-ask spread, residual risk, and size on stock returns. The Journal of Finance 44, 479–486.

Anderson, C.W., Garcia-Feijoo, L., 2006. Empirical evidence on capital investment, growth options, and security returns. The Journal of Finance 61, 171–194.

Ang, A., Chen, J., Xing, Y., 2006a. Downside risk. doi:10.1093/rfs/hhj035.

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006b. The cross-section of volatility and expected returns. Journal of Finance doi:10.1111/j.1540-6261.2006.00836.x.

Ang, A., Kristensen, D., 2012. Testing conditional factor models. Journal of Financial Economics 106, 132–156.

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82, 1059–1086.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys doi:10.1214/09-SS054, arXiv:0907.4728.

Asness, C.S., Porter, R.B., Stevens, R.L., 2000. Predicting stock returns using industry-relative firm characteristics. Available at SSRN 213872 .

Avramov, D., Chordia, T., 2006. Asset pricing models and financial market anomalies. Review of Financial Studies doi:10.1093/rfs/hhj025.

Bai, J., Zhou, G., 2015. Fama-MacBeth two-pass regressions: Improving risk premia estimates. Finance Research Letters doi:10.1016/j.frl.2015.08.001.

Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. Journal of Accounting and Economics 50, 20–41.

Bali, T.G., Cakici, N., Whitelaw, R.F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of financial economics 99, 427–446.

Bandyopadhyay, S.P., Huang, A.G., Wirjanto, T.S., 2010. The Accrual Volatility Anomaly. SSRN Electronic Journal , 1–36URL: http://papers.ssrn.com/abstract=1793364.

Bansal, R., Yaron, A., 2004. Risks for the long run: A potential resolution of asset pricing puzzles. doi:10.1111/j.1540-6261.2004.00670.x.

Banz, R.W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics doi:10.1016/0304-405X(81)90018-0.

Barbee Jr, W.C., Mukherji, S., Raines, G.A., 1996. Do sales–price and debt–equity explain stock returns better than book–market and firm size? Financial Analysts Journal 52, 56–60.

Barillas, F., Shanken, J., 2018. Comparing Asset Pricing Models. Journal of Finance doi:10.1111/jofi.12607.

Barrett, D.G., Dherin, B., 2020. Implicit gradient regularization. arXiv:2009.11162.

Barth, M.E., Elliott, J.A., Finn, M.W., 1999. Market rewards associated with patterns of increasing earnings. Journal of Accounting Research 37, 387–413.

Bartram, S.M., Lohre, H., Pope, P.F., Ranganathan, A., 2021. Navigating the factor zoo around the world: an institutional investor perspective. Journal of Business Economics , 1–49.

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. The journal of Finance 32, 663–682.

Basu, S., 1983. The relationship between earnings' yield, market value and return for NYSE common stocks. Further evidence. Journal of Financial Economics doi:10.1016/0304-405X(83)90031-4.

Bauman, W.S., Dowen, R., 1988. Growth projections and commin stock returns. Financial Analysts Journal 44, 79.

Belo, F., Lin, X., Bazdresch, S., 2014. Labor hiring, investment, and stock return predictability in the cross section. Journal of Political Economy 122, 129–177.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics doi:10.1214/aos/1013699998.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research .

Bhandari, L.C., 1988. Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence. The Journal of Finance doi:10.1111/j.1540-6261.1988.tb03952.x.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond Risk Premiums with Machine Learning. The Review of Financial Studies doi:10.1093/rfs/hhaa062.

Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford university press.

Black, F., 1972. Capital Market Equilibrium with Restricted Borrowing. The Journal of Business doi:10.1086/295472.

Black, F., Jensen, M.C., Scholes, M.S., 1972. The Capital Asset Pricing Model: Some Empirical Tests. Journal of Business .

Brandt, M.W., Kishore, R., Santa-Clara, P., Venkatachalam, M., 2008. Earnings announcements are full of surprises. SSRN eLibrary .

Brown, D.P., Rowe, B., 2007. The productivity premium in equity returns. Available at SSRN 993467 .

Bryzgalova, S., Pelger, M., Zhu, J., 2019. Forest Through the Trees: Building Cross-Sections of Stock Returns. SSRN Electronic Journal doi:10.2139/ssrn.3493458.

Campbell, J.Y., 1999. Asset prices, consumption, and the business cycle. Handbook of macroeconomics 1, 1231–1303.

Campbell, J.Y., Cochrane, J.H., 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. Journal of Political Economy doi:10.1086/250059.

Cattaneo, M.D., Crump, R.K., Farrell, M.H., Schaumburg, E., 2020. Characteristic-sorted portfolios: Estimation and inference. Review of Economics and Statistics 102, 531–551.

Chaieb, I., Langlois, H., Scaillet, O., 2018. Time-Varying Risk Premia in Large International Equity Markets. SSRN Electronic Journal doi:10.2139/ssrn.3103752.

Chan, K.C., fu Chen, N., Hsieh, D.A., 1985. An exploratory investigation of the firm size effect. Journal of Financial Economics doi:10.1016/0304-405X(85)90008-X.

Chandrashekar, S., Rao, R., 2009. The Productivity of Corporate Cash Holdings and the Cross-Section of Expected Stock Returns. SSRN Electronic Journal doi:10.2139/ssrn.1334162.

Chen, L., Pelger, M., Zhu, J., 2019. Deep Learning in Asset Pricing. SSRN Electronic Journal doi:10.2139/ssrn.3350138, arXiv:1904.00745.

Chen, L., Zhang, L., 2010. A better three-factor model that explains more anomalies. Journal of Finance 65, 563–595.

Chen, N.F., 1991. Financial Investment Opportunities and the Macroeconomy. The Journal of Finance doi:10.2307/2328835.

Chen, N.F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. Journal of business , 383–403.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econometrics Journal doi:10.1111/ectj.12097.

Chordia, T., Goyal, A., Shanken, J.A., 2015. Cross-Sectional Asset Pricing with Individual Stocks: Betas versus Characteristics. SSRN Electronic Journal doi:10.2139/ssrn.2549578.

Chordia, T., Subrahmanyam, A., Anshuman, V.R., 2001. Trading activity and expected stock returns. Journal of financial Economics 59, 3–32.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association 74, 829–836.

Cochrane, J.H., 2009. Asset pricing: (Revised Edition).

Cochrane, J.H., 2011. Presidential Address: Discount Rates. Journal of Finance doi:10.1111/j.1540-6261.2011.01671.x.

Cooper, M.J., Gulen, H., Schill, M.J., 2008. Asset growth and the cross-section of stock returns. the Journal of Finance 63, 1609–1651.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems doi:10.1007/BF02551274.

Daniel, K., Titman, S., 1997. Evidence on the characteristics of cross sectional variation in stock returns. Journal of Finance doi:10.1111/j.1540-6261.1997.tb03806.x.

Datar, V.T., Naik, N.Y., Radcliffe, R., 1998. Liquidity and stock returns: An alternative test. Journal of financial markets 1, 203–219.

De Prado, M.L., 2018. Advances in financial machine learning. John Wiley & Sons.

Desai, H., Rajgopal, S., Venkatachalam, M., 2004. Value-glamour and accruals mispricing: One anomaly or two? The Accounting Review 79, 355–385.

Dichtl, H., Drobetz, W., Lohre, H., Rother, C., Vosskamp, P., 2019. Optimal timing and tilting of equity factors. Financial Analysts Journal 75, 84–102.

Diether, K.B., Malloy, C.J., Scherbina, A., 2002. Differences of opinion and the cross section of stock returns. The Journal of Finance 57, 2113–2141.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters 2, 1–4.

Dixon, M.F., Polson, N.G., 2019. Deep fundamental factor models. arXiv preprint arXiv:1903.07677 .

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Drucker, H., Cun, Y.L., 1992. Improving Generalization Performance Using Double Backpropagation. IEEE Transactions on Neural Networks doi:10.1109/72.165600.

Eberhart, A.C., Maxwell, W.F., Siddique, A.R., 2004. An examination of long-term abnormal stock returns and operating performance following R and D increases. The Journal of Finance 59, 623–650.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science , 54–75.

Eisfeldt, A.L., Papanikolaou, D., 2013. Organization capital and the cross-section of expected returns. The Journal of Finance 68, 1365–1406.

Elgers, P.T., Lo, M.H., Pfeiffer Jr, R.J., 2001. Delayed security price adjustments to financial analysts' forecasts of annual earnings. The Accounting Review 76, 613–632.

Fairfield, P.M., Whisenant, S., Yohn, T.L., 2003. The differential persistence of accruals and cash flows for future operating income versus future profitability. Review of Accounting Studies 8, 221–243.

Fama, E.F., 1976. Foundations of Finance : Portfolio Decisions and Securities Prices. Blackwell.

Fama, E.F., French, K.R., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics doi:10.1016/0304-405X(89)90095-0.

Fama, E.F., French, K.R., 1992. The Cross-Section of Expected Stock Returns. The Journal of Finance doi:10.2307/2329112.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics doi:10.1016/0304-405X(93)90023-5.

Fama, E.F., French, K.R., 1996. Multifactor explanations of asset pricing anomalies. Journal of Finance doi:10.1111/j.1540-6261.1996.tb05202.x.

Fama, E.F., French, K.R., 2006. Profitability, investment and average returns. Journal of Financial Economics doi:10.1016/j.jfineco.2005.09.009.

Fama, E.F., French, K.R., 2008. Dissecting anomalies. Journal of Finance doi:10.1111/j.1540-6261.2008.01371.x.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics doi:10.1016/j.jfineco.2014.10.010.

Fama, E.F., French, K.R., 2020. Comparing Cross-Section and Time-Series Factor Models. Review of Financial Studies doi:10.1093/rfs/hhz089.

Fama, E.F., MacBeth, J.D., 1973. Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy doi:10.1086/260061.

Farrell, M.H., Liang, T., Misra, S., 2021. Deep neural networks for estimation and inference. Econometrica 89, 181–213.

Feng, G., Giglio, S., 2017. Taming the Factor Zoo. SSRN Electronic Journal doi:10.2139/ssrn.2934020.

Feng, G., Polson, N., Xu, J., 2020. Deep Learning in Characteristics-Sorted Factor Models. SSRN Electronic Journal doi:10.2139/ssrn.3243683.

Feng, G., Polson, N.G., Xu, J., 2018. Deep Learning in Characteristics-Sorted Factor Models. arXiv preprint arXiv:1805.01104 .

Ferson, W.E., Harvey, C.R., 1991. The Variation of Economic Risk Premiums. Journal of Political Economy doi:10.1086/261755.

Ferson, W.E., Harvey, C.R., 1993. The Risk and Predictability of International Equity Returns. The Review of Financial Studies doi:10.1093/rfs/6.3.527.

Ferson, W.E., Kandel, S., Stambaugh, R.F., 1987. Tests of asset pricing with time-varying expected risk premiums and market betas. The Journal of Finance 42, 201–220.

Ferson, W.E., Korajczyk, R.A., 1995. Do Arbitrage Pricing Models Explain the Predictability of Stock Returns? The Journal of Business doi:10.1086/296667.

Finnoff, W., Hergert, F., Zimmermann, H.G., 1993. Improving model selection by nonconvergent methods. Neural Networks doi:10.1016/S0893-6080(05)80122-4.

Francis, J., LaFond, R., Olsson, P.M., Schipper, K., 2004. Costs of equity and earnings attributes. The accounting review 79, 967–1010.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting Characteristics Nonparametrically. Review of Financial Studies doi:10.1093/rfs/hhz123.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software doi:10.18637/jss.v033.i01.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

Gagliardini, P., Ossola, E., Scaillet, O., 2016. Time-Varying Risk Premium in Large Cross-Sectional Equity Data Sets. Econometrica doi:10.3982/ecta11069.

Gettleman, E., Marks, J.M., 2006. Acceleration strategies. SSRN Electronic Journal .

Ghysels, E., 1998. On stable factor structures in the pricing of risk: Do time-varying betas help or hurt? Journal of Finance doi:10.1111/0022-1082.224803.

Gibbons, M.R., 1982. Multivariate tests of financial models. A new approach. Journal of Financial Economics doi:10.1016/0304-405X(82)90028-9.

Gibbons, M.R., Ross, S.A., Shanken, J., 1989. A Test of the Efficiency of a Given Portfolio. Econometrica doi:10.2307/1913625.

Giglio, S., Xiu, D., 2016. Asset Pricing with Omitted Factors. doi:10.2139/ssrn.2865922.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. journal of Computational and Graphical Statistics 24, 44–65.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. volume 1. MIT press Cambridge.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks, in: Advances in Neural Information Processing Systems.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. arXiv:1412.6572.

Goodman, B., Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a “right to explanation”. AI magazine 38, 50–57.

Goyal, A., 2012. Empirical cross-sectional asset pricing: A survey. doi:10.1007/s11408-011-0177-7.

Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average u.s. monthly stock returns. Review of Financial Studies doi:10.1093/rfs/hhx019.

Gu, S., Kelly, B., Xiu, D., 2020a. Autoencoder asset pricing models. Journal of Econometrics doi:10.1016/j.jeconom.2020.07.009.

Gu, S., Kelly, B., Xiu, D., 2020b. Empirical Asset Pricing via Machine Learning. Review of Financial Studies doi:10.1093/rfs/hhaa009.

Guo, R., Lev, B., Shi, C., 2006. Explaining the Short-and Long-Term IPO Anomalies in the US by R and D. Journal of Business Finance and Accounting 33, 550–579.

Hafzalla, N., Lundholm, R., Matthew Van Winkle, E., 2011. Percent accruals. The Accounting Review 86, 209–236.

Han, Y., He, A., Rapach, D., Zhou, G., 2019. Firm Characteristics and Expected Stock Returns. SSRN Electronic Journal .

Hansen, L.P., Richard, S.F., 1987. The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. Econometrica doi:10.2307/1913601.

Harvey, C.R., 1989. Time-varying conditional covariances in tests of asset pricing models. Journal of Financial Economics doi:10.1016/0304-405X(89)90049-4.

Harvey, C.R., Liu, Y., 2014. Lucky Factors. doi:10.2139/ssrn.2528780.

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the Cross-Section of Expected Returns. Review of Financial Studies doi:10.1093/rfs/hhv059.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction. doi:10.1007/b94608.

Hawkins, E.H., Chamberlin, S.C., Daniel, W.E., 1984. Earnings expectations and security prices. Financial Analysts Journal 40, 24–38.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. American Economic Review doi:10.1257/aer.103.2.732.

Heaton, J., 2008. Introduction to neural networks with Java. Heaton Research, Inc.

Heaton, J.B., Polson, N.G., Witte, J.H., 2017. Deep learning for finance: deep portfolios. doi:10.1002/asmb.2209.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30.

Hoffman, J., Roberts, D.A., Yaida, S., 2019. Robust learning with Jacobian regularization. arXiv:1908.02729.

Holthausen, R.W., Larcker, D.F., 1992. The prediction of stock returns using financial statement information. Journal of accounting and economics 15, 373–411.

Hong, H., Kacperczyk, M., 2009. The price of sin: The effects of social norms on markets. Journal of financial economics 93, 15–36.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks doi:10.1016/0893-6080(89)90020-8.

Hou, K., Moskowitz, T.J., 2005. Market frictions, price delay, and the cross-section of expected returns. The Review of Financial Studies 18, 981–1020.

Hou, K., Robinson, D.T., 2006. Industry concentration and average stock returns. The Journal of Finance 61, 1927–1956.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies doi:10.1093/rfs/hhu068.

Hou, K., Xue, C., Zhang, L., 2020. Replicating Anomalies. Review of Financial Studies doi:10.1093/rfs/hhy131.

Hu, Z., Zhang, J., Ge, Y., 2021. Handling vanishing gradient problem using artificial derivative. IEEE Access 9, 22371–22377.

Huang, A.G., 2009. The cross section of cashflow volatility and expected stock returns. Journal of Empirical Finance 16, 409–429.

Ilmanen, A., Israel, R., Moskowitz, T.J., Thapar, A., Wang, F., 2019. How Do Factor Premia Vary Over Time? A Century of Evidence. SSRN Electronic Journal .

Imajo, K., Minami, K., Ito, K., Nakagawa, K., 2020. Deep portfolio optimization via distributional prediction of residual factors. arXiv:2012.07245.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: 32nd International Conference on Machine Learning, ICML 2015. arXiv:1502.03167.

Jagannathan, R., Wang, Z., 1996a. The conditional CAPM and the cross-section of expected returns. Journal of Finance doi:10.1111/j.1540-6261.1996.tb05201.x.

Jagannathan, R., Wang, Z., 1996b. The Conditional CAPM and the Cross-Section of Expected Returns. The Journal of Finance 51, 3–53. URL: http://www.jstor.org/stable/2329301, doi:10.2307/2329301.

Jagannathan, R., Wang, Z., 1998. An asymptotic theory for estimating beta-pricing models using cross-sectional regression. Journal of Finance doi:10.1111/0022-1082.00053.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. The Journal of finance 45, 881–898.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance 48, 65–91.

Jensen, M.C., 1968. The Performance of Mutual Funds in the Period 1945-1964. The Journal of Finance doi:10.2307/2325404.

Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2021. Is There A Replication Crisis In Finance? Technical Report.

Jiang, G., Lee, C.M.C., Zhang, Y., 2005. Information uncertainty and expected returns. Review of Accounting Studies 10, 185–221.

Kama, I., 2009. On the market reaction to revenue and earnings surprises. Journal of Business Finance and Accounting 36, 31–50. doi:10.1111/j.1468-5957.2008.02121.x.

Kapetanios, G., 2007. Measuring conditional persistence in nonlinear time series. Oxford Bulletin of Economics and Statistics 69, 363–386.

Kapetanios, G., 2008. A bootstrap procedure for panel data sets with many cross-sectional units. The Econometrics Journal 11, 377–395.

Kapetanios, G., Papailias, F., Taylor, A.M.R., 2019. A generalised fractional differencing bootstrap for long memory processes. Journal of Time Series Analysis 40, 467–492.

Kelly, B., Pruitt, S., 2015. The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics 186, 294–316.

Kelly, B.T., Pruitt, S., Su, Y., 2018. Instrumented Principal Component Analysis. SSRN Electronic Journal doi:10.2139/ssrn.2983919.

Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics doi:10.1016/j.jfineco.2019.05.001.

Kim, B., Khanna, R., Koyejo, O.O., 2016. Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems 29.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. arXiv:1412.6980.

Kozak, S., Nagel, S., Santosh, S., 2018. Interpreting Factor Models. Journal of Finance doi:10.1111/jofi.12612.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics doi:10.1016/j.jfineco.2019.06.008.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R., 2016. Ask me anything: Dynamic memory networks for natural language processing, in: 33rd International Conference on Machine Learning, ICML 2016. arXiv:1506.07285.

van Laarhoven, T., 2017. L2 Regularization versus Batch and Weight Normalization. arXiv:1706.05350.

Lakonishok, J., Shleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation, and risk. The journal of finance 49, 1541–1578.

Leray, P., Gallinari, P., 1999. Feature selection with neural networks. Behaviormetrika 26, 145–166.

Lerman, A., Livnat, J., Mendenhall, R.R., 2007. The high-volume return premium and post-earnings announcement drift .

Lettau, M., Pelger, M., 2020a. Estimating latent asset-pricing factors. Journal of Econometrics doi:10.1016/j.jeconom.2019.08.012.

Lettau, M., Pelger, M., 2020b. Factors That Fit the Time Series and Cross-Section of Stock Returns. Review of Financial Studies doi:10.1093/rfs/hhaa020.

Leung, E., Lohre, H., Mischlich, D., Shea, Y., Stroh, M., 2021. The promises and pitfalls of machine learning for predicting stock returns. The Journal of Financial Data Science 3, 21–50.

Lev, B., Nissim, D., 2004. Taxable income, future earnings, and equity values. The accounting review 79, 1039–1074.

Lewellen, J., 2015. The Cross-section of Expected Stock Returns. Critical Finance Review doi:10.1561/104.00000024.

Li, X., Chen, S., Hu, X., Yang, J., 2019. Understanding the disharmony between dropout and batch normalization by variance shift, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi:10.1109/CVPR.2019.00279, arXiv:1801.05134.

Lintner, J., 1965. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. The Review of Economics and Statistics doi:10.2307/1924119.

Litzenberger, R.H., Ramaswamy, K., 1982. The effects of dividends on common stock prices tax effects or information effects? The Journal of Finance 37, 429–443.

Liu, W., 2006. A liquidity-augmented capital asset pricing model. Journal of financial Economics 82, 631–671.

Lo, A.W., MacKinlay, A.C., 1990. Data-Snooping Biases in Tests of Financial Asset Pricing Models. Review of Financial Studies doi:10.1093/rfs/3.3.431.

Loughran, T., Ritter, J.R., 1995. The new issues puzzle. The Journal of finance 50, 23–51.

Lyu, C., Huang, K., Liang, H.N., 2016. A unified gradient regularization family for adversarial examples, in: Proceedings - IEEE International Conference on Data Mining, ICDM. doi:10.1109/ICDM.2015.84.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, Citeseer. p. 3.

Markowitz, H., 1952. Portfolio Selection. The Journal of Finance doi:10.2307/2975974.

Mclean, R.D., Pontiff, J., 2016. Does Academic Research Destroy Stock Return Predictability? Journal of Finance doi:10.1111/jofi.12365.

Merton, R.C., 1973. An Intertemporal Capital Asset Pricing Model. Econometrica doi:10.2307/1913811.

Messmer, M., 2017a. Deep Learning and the Cross-Section of Expected Returns. SSRN Electronic Journal doi:10.2139/ssrn.3081555.

Messmer, M., 2017b. The (Adaptive) Lasso in the Zoo - Firm Characteristic Selection in the Cross-Section of Expected Returns. SSRN Electronic Journal doi:10.2139/ssrn.2930436.

Michaely, R., Thaler, R.H., Womack, K.L., 1995. Price reactions to dividend initiations and omissions: Overreaction or drift? the Journal of Finance 50, 573–608.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267, 1–38.

Mohanram, P.S., 2005. Separating winners from losers among lowbook-to-market stocks using financial statement analysis. Review of accounting studies 10, 133–170.

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 417–431.

Moskowitz, T.J., Grinblatt, M., 1999. Do industries explain momentum? The Journal of finance 54, 1249–1290.

Mossin, J., 1966. Equilibrium in a Capital Asset Market. Econometrica doi:10.2307/1910098.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116, 22071–22080.

Newey, W.K., West, K.D., 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent. Technical Report.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics doi:10.1016/j.jfineco.2013.01.003.

Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S., 2018. Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378 .

Oh, D.H., Patton, A.J., 2018. Time-Varying Systemic Risk: Evidence From a Dynamic Copula Model of CDS Spreads. Journal of Business and Economic Statistics doi:10.1080/07350015.2016.1177535.

Ou, J.A., Penman, S.H., 1989. Financial statement analysis and the prediction of stock returns. Journal of accounting and economics 11, 295–329.

Palazzo, B., 2012. Cash holdings, risk, and expected returns. Journal of Financial Economics 104, 162–185.

Patton, A.J., Timmermann, A., 2008. Portfolio Sorts and Tests of Cross-Sectional Patterns in Expected Returns. .

Petkova, R., Zhang, L., 2005. Is value riskier than growth? Journal of Financial Economics 78, 187–202.

Piotroski, J.D., 2000. Value investing: The use of historical financial statement information to separate winners from losers. Journal of Accounting Research , 1–41.

Pohl, W., Schmedders, K., Wilms, O., 2018. Higher order effects in asset pricing models with long-run risks. The Journal of Finance 73, 1061–1111.

Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. The Journal of Finance 63, 921–945.

Rapach, D.E., Strauss, J.K., Zhou, G., 2013. International stock return predictability: What is the role of the United States? Journal of Finance doi:10.1111/jofi.12041.

Reinganum, M.R., 1981. Misspecification of capital asset pricing. Journal of Financial Economics doi:10.1016/0304-405x(81)90019-2.

Rendleman Jr, R.J., Jones, C.P., Latane, H.A., 1982. Empirical anomalies based on unexpected earnings and the importance of risk adjustments. Journal of Financial economics 10, 269–287.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016a. " Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016b. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 .

Richardson, S.A., Sloan, R.G., Soliman, M.T., Tuna, I., 2005. Accrual reliability, earnings persistence and stock prices. Journal of Accounting and Economics 39, 437–485. doi:10.1016/j.jacceco.2005.04.005.

Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. The Journal of Portfolio Management doi:10.3905/jpm.1985.409007.

Ross, A.S., Hughes, M.C., Doshi-Velez, F., 2017. Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 .

Ross, S.A., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory doi:10.1016/0022-0531(76)90046-6.

Ruck, D.W., Rogers, S.K., Kabrisky, M., 1990. Feature selection using a multilayer perceptron. Journal of Neural Network Computing 2, 40–48.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 .

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature doi:10.1038/323533a0.

Scherbina, A., 2008. Suppressed negative information and future underperformance. Review of Finance 12, 533–565.

Shanken, J., 1985. Multivariate tests of the zero-beta CAPM. Journal of Financial Economics doi:10.1016/0304-405X(85)90002-9.

Shanken, J., 1992. On the Estimation of Beta-Pricing Models. Review of Financial Studies doi:10.1093/rfs/5.1.1.

Shapley, L.S., 1953. A value for n-person games. Princeton University Press.

Sharpe, W.F., 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. The Journal of Finance doi:10.1111/j.1540-6261.1964.tb02865.x.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. Nature doi:10.1038/nature16961.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 .

Sloan, R.G., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Accounting Review .

Sokolić, J., Giryes, R., Sapiro, G., Rodrigues, M.R.D., 2017. Robust large margin deep neural networks. IEEE Transactions on Signal Processing 65, 4265–4280.

Soliman, M.T., 2008. The use of DuPont analysis by market participants. The accounting review 83, 823–853.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research .

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR. pp. 3319–3328.

Thomas, J., Zhang, F.X., 2011. Tax expense momentum. Journal of Accounting Research 49, 791–821.

Thomas, J.K., Zhang, H., 2002. Inventory changes and future returns. Review of Accounting Studies 7, 163–187.

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) doi:10.1111/j.2517-6161.1996.tb02080.x.

Titman, S., Wei, K.C., Xie, F., 2004. Capital investments and stock returns. Journal of Financial and Quantitative Analysis doi:10.1017/s0022109000003173.

Tuzel, S., 2010. Corporate real estate holdings and the cross-section of stock returns. The Review of Financial Studies 23, 2268–2302.

Ullah, A., 1988. Non-parametric estimation of econometric functionals. Canadian Journal of Economics , 625–658.

Umlandt, D., 2020. Likelihood-Based Dynamic Asset Pricing: Learning Time-Varying Risk Premia from Cross-Sectional Models. SSRN Electronic Journal doi:10.2139/ssrn.3666324.

Valta, P., 2016. Strategic default, debt structure, and stock returns. Journal of Financial and Quantitative Analysis 51, 197–229.

Varga, D., Csiszárik, A., Zombori, Z., 2018. Gradient Regularization Improves Accuracy of Discriminative Models. Schedae Informaticae doi:10.4467/20838476SI.18.003.10408, arXiv:1712.09936.

Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 .

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology doi:10.1111/j.1467-9868.2005.00503.x.

[!h]

A List of Firm Characteristics

This appendix provides further details about the list of firm characteristics considered in this paper, their source and frequency. Table A.1 provides a summary.

Table A.1:
Description, sources and frequency of all 95 firm characteristics
We construct and report a universe of firm characteristics analogously to Green et al. (2017) or Gu et al. (2020b). For the purpose of direct comparison, we adopt the firm characteristic definitions of Green et al. (2017) while simultaneously acknowledging that certain characteristics could be defined differently as documented by Hou et al. (2020). In comparison to Green et al. (2017), our data is characterised by some subtle differences. First, we convert the SAS code published by Green et al. (2017) into Python to improve code integration and version control. Second, we calculate industry-adjusted variables only after the Compustat and CRSP datasets have been merged (instead of calculating the industry mean before merging both datasets), leading to slightly different results. The reasoning behind our methodology is that we want the adjustment to be only influenced by stocks that are included in the investment universe and not by all Compustat stocks.

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
1	absacc	Absolute accruals	Bandyopadhyay et al. (2010)	Compustat	Annual
2	acc	Accruals	Sloan (1996)	Compustat	Annual
3	aeavol	Abnormal earnings announcement volume	Lerman et al. (2007)	Compustat+CRSP	Quarterly
4	age	Age since first Compustat coverage	Jiang et al. (2005)	Compustat	Annual
5	agr	Asset growth	Cooper et al. (2008)	Compustat	Annual
6	baspread	Bid-ask spread	Amihud and Mendelson (1989)	CRSP	Monthly
7	beta	Market beta	Fama and MacBeth (1973)	CRSP	Monthly
8	betasq	Market beta squared	Fama and MacBeth (1973)	CRSP	Monthly
9	bm	Book-to-market	Rosenberg et al. (1985)	Compustat	Annual
10	bm_ia	Industry-adjusted book-to-market	Asness et al. (2000)	Compustat	Annual
11	cash	Cash holdings	Palazzo (2012)	Compustat	Quarterly
12	cashdebt	Cash-flow to debt	Ou and Penman (1989)	Compustat	Annual
13	cashpr	Cash productivity	Chandrashekar and Rao (2009)	Compustat	Annual
14	cfp	Cash-flow to price	Desai et al. (2004)	Compustat	Annual
15	cfp_ia	Industry-adjusted cash-flow to price	Asness et al. (2000)	Compustat	Annual
16	chatoia	Industry-adjusted change in asset turnover	Soliman (2008)	Compustat	Annual
17	chcsho	Change in common shares outstanding	Pontiff and Woodgate (2008)	Compustat	Annual
18	chempia	Industry-adjusted change in employees	Asness et al. (2000)	Compustat	Annual
19	chfeps	Change in analysts' mean earnings forecast	Hawkins et al. (1984)	I/B/E/S	Monthly
20	chinv	Change in inventory	Thomas and Zhang (2002)	Compustat	Annual
21	chmom	Change in 6-month momentum	Gettleman and Marks (2006)	CRSP	Monthly
22	chnanalyst	Change in analyst coverage	Scherbina (2008)	I/B/E/S	Monthly
23	chpmia	Industry-adjusted change in profit margin	Soliman (2008)	Compustat	Annual
24	chtx	Change in tax expenses	Thomas and Zhang (2011)	Compustat	Quarterly

(continued)

Table A.1:
Description, sources and frequency of all 95 firm characteristics
We construct and report a universe of firm characteristics analogously to Green et al. (2017) or Gu et al. (2020b). For the purpose of direct comparison, we adopt the firm characteristic definitions of Green et al. (2017) while simultaneously acknowledging that certain characteristics could be defined differently as documented by Hou et al. (2020). In comparison to Green et al. (2017), our data is characterised by some subtle differences. First, we convert the SAS code published by Green et al. (2017) into Python to improve code integration and version control. Second, we calculate industry-adjusted variables only after the Compustat and CRSP datasets have been merged (instead of calculating the industry mean before merging both datasets), leading to slightly different results. The reasoning behind our methodology is that we want the adjustment to be only influenced by stocks that are included in the investment universe and not by all Compustat stocks.

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
25	cinvest	Corporate investment	Titman et al. (2004)	Compustat	Quarterly
26	convind	Convertible debt indicator	Valta (2016)	Compustat	Annual
27	currat	Current ratio	Ou and Penman (1989)	Compustat	Annual
28	depr	Depreciation	Holthausen and Larcker (1992)	Compustat	Annual
29	disp	Volatility in analysts' forecasts	Diether et al. (2002)	I/B/E/S	Monthly
30	divi	Dividend initiation	Michaely et al. (1995)	Compustat	Annual
31	divo	Dividend omission	Michaely et al. (1995)	Compustat	Annual
32	dolvol	Dollar trading volume	Chordia et al. (2001)	CRSP	Monthly
33	dy	Dividend yield	Litzenberger and Ramaswamy (1982)	Compustat	Annual
34	ear	Earnings announcement return	Brandt et al. (2008)	Compustat+CRSP	Quarterly
35	egr	Growth in common shareholder equity	Richardson et al. (2005)	Compustat	Annual
36	ep	Earnings-to-price	Basu (1977)	Compustat	Annual
37	fgr5yr	Most recently available analyst forecasted 5-year growth	Bauman and Dowen (1988)	I/B/E/S	Quarterly
38	gma	Gross profitability	Novy-Marx (2013)	Compustat	Annual
39	grcapx	Growth in capital expenditures	Anderson and Garcia-Feijoo (2006)	Compustat	Annual
40	grltnoa	Growth in long-term net operating assets	Fairfield et al. (2003)	Compustat	Annual
41	herf	Industry sales concentration	Hou and Robinson (2006)	Compustat	Annual
42	hire	Employee growth rate	Belo et al. (2014)	Compustat	Annual
43	idiovol	Idiosyncratic return volatility	Ali et al. (2003)	CRSP	Monthly
44	ill	Illiquidity	Amihud (2002)	CRSP	Monthly
45	indmom	Industry momentum	Moskowitz and Grinblatt (1999)	CRSP	Monthly
46	invest	Capital expenditures and inventory	Chen and Zhang (2010)	Compustat	Annual
47	ipo	IPO year indicator	Loughran and Ritter (1995)	CRSP	Monthly
48	lev	Leverage	Bhandari (1988)	Compustat	Annual
49	lgr	Growth in long-term debt	Richardson et al. (2005)	Compustat	Annual

(continued)

Table A.1:
Description, sources and frequency of all 95 firm characteristics
We construct and report a universe of firm characteristics analogously to Green et al. (2017) or Gu et al. (2020b). For the purpose of direct comparison, we adopt the firm characteristic definitions of Green et al. (2017) while simultaneously acknowledging that certain characteristics could be defined differently as documented by Hou et al. (2020). In comparison to Green et al. (2017), our data is characterised by some subtle differences. First, we convert the SAS code published by Green et al. (2017) into Python to improve code integration and version control. Second, we calculate industry-adjusted variables only after the Compustat and CRSP datasets have been merged (instead of calculating the industry mean before merging both datasets), leading to slightly different results. The reasoning behind our methodology is that we want the adjustment to be only influenced by stocks that are included in the investment universe and not by all Compustat stocks.

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
50	maxret	Maximum daily return	Bali et al. (2011)	CRSP	Monthly
51	mom12m	12-month momentum	Jegadeesh (1990)	CRSP	Monthly
52	mom1m	1-month momentum	Jegadeesh and Titman (1993)	CRSP	Monthly
53	mom36m	36-month momentum	Jegadeesh and Titman (1993)	CRSP	Monthly
54	mom6m	6-month momentum	Jegadeesh and Titman (1993)	CRSP	Monthly
55	ms	Financial statement score	Mohanram (2005)	Compustat	Quarterly
56	mve	Log-size	Banz (1981)	CRSP	Monthly
57	mve_ia	Industry adjusted log-size	Asness et al. (2000)	Compustat	Annual
58	nanalyst	Number of analyst forecasts	Elgers et al. (2001)	I/B/E/S	Monthly
59	nincr	Number of earnings increases	Barth et al. (1999)	Compustat	Quarterly
60	operprof	Operating profitability	Fama and French (2015)	Compustat	Annual
61	orgcap	Organisational capital	Eisfeldt and Papanikolaou (2013)	Compustat	Annual
62	pchcapx_ia	Industry-adjusted %-change in current ratio	Abarbanell and Bushee (1998)	Compustat	Annual
63	pchcurrat	%-change in current ratio	Ou and Penman (1989)	Compustat	Annual
64	pchdepr	%-change in depreciation	Holthausen and Larcker (1992)	Compustat	Annual
65	pchgm_pchsale	%-change in gross-margin - %-change in sales	Abarbanell and Bushee (1998)	Compustat	Annual
66	pchquick	%-change in quick ratio	Ou and Penman (1989)	Compustat	Annual
67	pchsale_pchinvt	%-change in sales - %-change in inventory	Abarbanell and Bushee (1998)	Compustat	Annual
68	pchsale_pchrect	%-change in sales - %-change in A/R	Abarbanell and Bushee (1998)	Compustat	Annual
69	pchsale_pchxsga	%-change in sales - %-change in SG&A	Abarbanell and Bushee (1998)	Compustat	Annual
70	pchsaleinv	%-change in sales-to-inventory	Ou and Penman (1989)	Compustat	Annual
71	pctacc	Percent accruals	Hafzalla et al. (2011)	Compustat	Annual
72	pricedelay	Price delay	Hou and Moskowitz (2005)	CRSP	Monthly
73	ps	Financial statement score	Piotroski (2000)	Compustat	Annual
74	quick	Quick ratio	Ou and Penman (1989)	Compustat	Annual

(continued)

Table A.1:
Description, sources and frequency of all 95 firm characteristics
We construct and report a universe of firm characteristics analogously to Green et al. (2017) or Gu et al. (2020b). For the purpose of direct comparison, we adopt the firm characteristic definitions of Green et al. (2017) while simultaneously acknowledging that certain characteristics could be defined differently as documented by Hou et al. (2020). In comparison to Green et al. (2017), our data is characterised by some subtle differences. First, we convert the SAS code published by Green et al. (2017) into Python to improve code integration and version control. Second, we calculate industry-adjusted variables only after the Compustat and CRSP datasets have been merged (instead of calculating the industry mean before merging both datasets), leading to slightly different results. The reasoning behind our methodology is that we want the adjustment to be only influenced by stocks that are included in the investment universe and not by all Compustat stocks.

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
75	rd	R&D increase	Eberhart et al. (2004)	Compustat	Annual
76	rd_mve	R&D to market capitalisation	Guo et al. (2006)	Compustat	Annual
77	rd_sale	R&D to sales	Guo et al. (2006)	Compustat	Annual
78	realestate	Real estate holdings	Tuzel (2010)	Compustat	Annual
79	retvol	Return volatility	Ang et al. (2006b)	CRSP	Monthly
80	roaq	Return on assets	Balakrishnan et al. (2010)	Compustat	Quarterly
81	roavol	Earnings volatility	Francis et al. (2004)	Compustat	Quarterly
82	roeq	Return on equity	Hou et al. (2015)	Compustat	Quarterly
83	roic	Return on invested capital	Brown and Rowe (2007)	Compustat	Annual
84	rsup	Revenue surprise	Kama (2009)	Compustat	Quarterly
85	salecash	Sales-to-cash	Ou and Penman (1989)	Compustat	Annual
86	saleinv	Sales-to-inventory	Ou and Penman (1989)	Compustat	Annual
87	salerec	Sales-to-receivables	Ou and Penman (1989)	Compustat	Annual
88	secured	Secured debt	Valta (2016)	Compustat	Annual
89	securedind	Secured debt indicator	Valta (2016)	Compustat	Annual
90	sfe	Analysts mean annual earnings forecast	Elgers et al. (2001)	I/B/E/S	Quarterly
91	sgr	Sales growth	Lakonishok et al. (1994)	Compustat	Annual
92	sgrvol	Revenue surprise volatility	Green et al. (2017)	Compustat	Quarterly
93	sin	Sin stocks	Hong and Kacperczyk (2009)	Compustat	Annual
94	sp	Sales-to-price	Barbee Jr et al. (1996)	Compustat	Annual
95	std_dolvol	Volatility of liquidity (dollar trading volume)	Chordia et al. (2001)	CRSP	Monthly
96	std_turn	Volatility of liquidity (share turnover)	Chordia et al. (2001)	CRSP	Monthly
97	stdacc	Accrual volatility	Bandyopadhyay et al. (2010)	Compustat	Quarterly
98	stdcf	Cash-flow volatility	Huang (2009)	Compustat	Quarterly
99	sue	Unexpected quarterly earnings	Rendleman Jr et al. (1982)	Compustat	Quarterly

(continued)

Table A.1:

Description, sources and frequency of all 95 firm characteristics

We construct and report a universe of firm characteristics analogously to Green et al. (2017) or Gu et al. (2020b). For the purpose of direct comparison, we adopt the firm characteristic definitions of Green et al. (2017) while simultaneously acknowledging that certain characteristics could be defined differently as documented by Hou et al. (2020). In comparison to Green et al. (2017), our data is characterised by some subtle differences. First, we convert the SAS code published by Green et al. (2017) into Python to improve code integration and version control. Second, we calculate industry-adjusted variables only after the Compustat and CRSP datasets have been merged (instead of calculating the industry mean before merging both datasets), leading to slightly different results. The reasoning behind our methodology is that we want the adjustment to be only influenced by stocks that are included in the investment universe and not by all Compustat stocks.

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
100	tang	Debt capacity / firm tangibility	Almeida and Campello (2007)	Compustat	Annual
101	tb	Tax income to book income	Lev and Nissim (2004)	Compustat	Annual
102	turn	Share turnover	Datar et al. (1998)	CRSP	Monthly
103	zerotrade	Zero trading days	Liu (2006)	CRSP	Monthly

(continued)

B List of Core Characteristics

In appendix A, we list all 103 firm characteristics. However, as commonly done in the empirical asset pricing literature, we also focus on a core characteristics case, which reduces the dimensionality of the all characteristics case from 103 to 49, where we only consider 49 core characteristics that are frequently identifies as most relevant in the literature (e.g. see, for example, Lewellen (2015), Green et al. (2017) or Gu et al. (2020b)).

No.	Acronym	Firm Characteristic	Characteristic Source	Data Source	Frequency
1	acc	Accruals	Sloan (1996)	Compustat	Annual
2	agr	Asset growth	Cooper et al. (2008)	Compustat	Annual
3	beta	Market beta	Fama and MacBeth (1973)	CRSP	Monthly
4	bm	Book-to-market	Rosenberg et al. (1985)	Compustat	Annual
5	cash	Cash holdings	Palazzo (2012)	Compustat	Quarterly
6	cashpr	Cash productivity	Chandrashekar and Rao (2009)	Compustat	Annual
7	cfp	Cash-flow to price	Desai et al. (2004)	Compustat	Annual
8	chatoia	Industry-adjusted change in asset turnover	Soliman (2008)	Compustat	Annual
9	chcsho	Change in common shares outstanding	Pontiff and Woodgate (2008)	Compustat	Annual
10	chfeps	Change in analysts' mean earnings forecast	Hawkins et al. (1984)	I/B/E/S	Monthly
11	chinv	Change in inventory	Thomas and Zhang (2002)	Compustat	Annual
12	chmom	Change in 6-month momentum	Gettleman and Marks (2006)	CRSP	Monthly
13	chpmia	Industry-adjusted change in profit margin	Soliman (2008)	Compustat	Annual
14	chtx	Change in tax expenses	Thomas and Zhang (2011)	Compustat	Quarterly
15	currat	Current ratio	Ou and Penman (1989)	Compustat	Annual
16	depr	Depreciation	Holthausen and Larcker (1992)	Compustat	Annual
17	dy	Dividend yield	Litzenberger and Ramaswamy (1982)	Compustat	Annual
18	ear	Earnings announcement return	Brandt et al. (2008)	Compustat+CRSP	Quarterly
19	ep	Earnings-to-price	Basu (1977)	Compustat	Annual
20	gma	Gross profitability	Novy-Marx (2013)	Compustat	Annual
21	grcapx	Growth in capital expenditures	Anderson and Garcia-Feijoo (2006)	Compustat	Annual
22	grltnoa	Growth in long-term net operating assets	Fairfield et al. (2003)	Compustat	Annual
23	ill	Illiquidity	Amihud (2002)	CRSP	Monthly
24	indmom	Industry momentum	Moskowitz and Grinblatt (1999)	CRSP	Monthly
25	invest	Capital expenditures and inventory	Chen and Zhang (2010)	Compustat	Annual
26	lev	Leverage	Bhandari (1988)	Compustat	Annual
27	lgr	Growth in long-term debt	Richardson et al. (2005)	Compustat	Annual
28	maxret	Maximum daily return	Bali et al. (2011)	CRSP	Monthly
29	mom12m	12-month momentum	Jegadeesh (1990)	CRSP	Monthly
30	mom1m	1-month momentum	Jegadeesh and Titman (1993)	CRSP	Monthly
31	mom36m	36-month momentum	Jegadeesh and Titman (1993)	CRSP	Monthly
32	mve	Log-size	Banz (1981)	CRSP	Monthly
33	nincr	Number of earnings increases	Barth et al. (1999)	Compustat	Quarterly
34	orgcap	Organisational capital	Eisfeldt and Papanikolaou (2013)	Compustat	Annual
35	pchgm_pchsale	%-change in gross-margin - %-change in sales	Abarbanell and Bushee (1998)	Compustat	Annual
36	pchsale_pchinvt	%-change in sales - %-change in inventory	Abarbanell and Bushee (1998)	Compustat	Annual
37	pchsale_pchrect	%-change in sales - %-change in A/R	Abarbanell and Bushee (1998)	Compustat	Annual
38	pchsale_pchxsga	%-change in sales - %-change in SG&A	Abarbanell and Bushee (1998)	Compustat	Annual
39	retvol	Return volatility	Ang et al. (2006b)	CRSP	Monthly
40	roaq	Return on assets	Balakrishnan et al. (2010)	Compustat	Quarterly
41	roavol	Earnings volatility	Francis et al. (2004)	Compustat	Quarterly
42	roeq	Return on equity	Hou et al. (2015)	Compustat	Quarterly
43	salecash	Sales-to-cash	Ou and Penman (1989)	Compustat	Annual
44	saleinv	Sales-to-inventory	Ou and Penman (1989)	Compustat	Annual
45	sgr	Sales growth	Lakonishok et al. (1994)	Compustat	Annual
46	sp	Sales-to-price	Barbee Jr et al. (1996)	Compustat	Annual
47	std_dolvol	Volatility of liquidity (dollar trading volume)	Chordia et al. (2001)	CRSP	Monthly
48	std_turn	Volatility of liquidity (share turnover)	Chordia et al. (2001)	CRSP	Monthly
49	turn	Share turnover	Datar et al. (1998)	CRSP	Monthly

Table B.1:
Description, sources and frequency of the 49 core characteristics
The table summarises the selection of the 49 core characteristics.

C Data Validation

Appendix A provides an overview over the 103 firm characteristics we use in our empirical analysis. We base our calculations on Green et al. (2017) and translate his SAS code into Python. Our dataset achieves an overall mean and median correlation with Green's data of 89%, respectively 99%, which makes our results directly comparable to existing research – see table C.2 for further details. Minor differences in the data, which lessen the mean and median correlation, are primarily the result of slightly different variable definitions. Out of 103 characteristics, 13 show an absolute correlation of less than 0.70. These characteristics and the reason for why they differ from Green's data are summarised in table C.1. Note, however, that we do not claim to push existing forecasting frontiers with this paper and firm characteristic definitions are, therefore, less important to us. The code documenting the data download and cleaning can be found on `https://github.com/fkempf92/FactorData`.

Table C.2 summarises the raw data of each characteristic in the dataset used by Green et al. (2017), where we source the data from Jeremiah Green's website[32]. In particular, we only consider stocks that appear in both datasets at the respective point in time. Note that for the purpose of a clear presentation, we round the summary statistics to one decimal place, such that some detail is lost. For example, the raw data for the characteristic illiquidity is very small in magnitude, which is why they only appear as zero in the rounded summary table C.2. For direct comparability, the sample ranges from January 1980 to December 2014.

Abbreviation	$\lvert\rho\rvert$	**Explanation for divergence**
cfp_ia	0.00	Industry adjustment after investment universe is formed.
sgrvol	0.01	Benchmark is not winsorised.
bm_ia	0.02	Industry adjustment after investment universe is formed.
pchcapx_ia	0.08	Industry adjustment after investment universe is formed.
tb	0.21	Industry adjustment after investment universe is formed.
IPO	0.24	First 12 months since listed, not the beginning of the sample.
pricedelay	0.30	Expanding window vs. fixed window.
chpmia	0.38	Industry adjustment after investment universe is formed.
salerec	0.39	Restriction to be strictly positive.
operprof	0.48	Different definition of *xsga0*.
herf	0.61	Industry adjustment after investment universe is formed.
cashdebt	0.61	Restriction to be strictly positive.
ms	0.63	Medians calculated after investment universe is formed.

Table C.1:
Diverging firm characteristics
There are 13 firm characteristics with an absolute correlation of smaller than 0.70. The main reason for why those characteristics differ from the benchmark is that we perform all industry adjustments after the investment universe if formed to reduce arbitrariness.

[32] `https://drive.google.com/file/d/0BwwEXkCgXEdRQWZreUpKOHBXOUU/view?resourcekey=0-1xjZ8fAc0sTybVC6RADDCA`, last accessed in November 2021.

Table C.2:
Data comparison
In order to make our dataset as comparable as possible, we benchmark it against the original dataset used by Green et al. (2017). For each firm characteristics, the table shows the minimum, maximum, mean, standard deviation and number of non-missing observations per characteristic, where the first row refers to the dataset used by Green et al. (2017) and the second row refers to the dataset we use. In addition, we report the correlation of each factor over the entire sample for observations that are available in both datasets.

	Green et al. (2017)					**Kapetanios and Kempf (2021)**					**Correlation**
	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Corr
absacc	0.0	1.1	0.1	0.1	1662807	0.0	1.1	0.1	0.1	1629133	1.0
acc	−1.1	0.5	0.0	0.1	1662807	−1.1	0.5	0.0	0.1	1629133	1.0
aeavol	−1.0	22.5	0.9	2.1	1709241	−1.0	22.2	0.9	2.1	1656590	1.0
age	1.0	40.0	10.7	8.5	1933709	1.0	53.0	12.3	10.4	1905887	0.9
agr	−0.7	6.1	0.2	0.5	1802012	−0.7	6.6	0.2	0.5	1763922	1.0
baspread	0.0	0.9	0.1	0.1	1933663	0.0	0.9	0.1	0.1	1914238	1.0
beta	−0.8	4.0	1.1	0.7	1911760	−0.8	3.6	1.0	0.6	1902405	1.0
betasq	0.0	15.8	1.6	1.9	1911760	0.0	13.3	1.4	1.7	1902405	1.0
bm	−2.6	7.6	0.7	0.7	1933709	−2.4	7.5	0.7	0.7	1905879	1.0
bm_ia	−1307.0	17188.4	29.4	806.7	1933709	−169.2	11.1	−0.9	11.7	1905879	0.0
cash	−0.1	1.0	0.2	0.2	1711537	0.0	1.0	0.2	0.2	1838052	1.0
cashdebt	−99.7	2.2	0.0	1.3	1864545	0.0	7.5	0.4	0.8	1701272	−0.6
cashpr	−537.3	600.3	−1.3	58.1	1913623	−519.5	610.8	−1.3	58.3	1886229	1.0
cfp	−4.7	37.5	0.1	0.8	1775262	−3.6	2.6	0.0	0.3	1714419	1.0
cfp_ia	−292.6	7043.8	15.8	348.2	1775262	−5.8	24.8	0.1	1.0	1714419	0.0
chatoia	−1.4	1.2	0.0	0.2	1657372	−1.0	1.2	0.0	0.2	1613689	1.0
chcsho	−0.8	2.7	0.1	0.3	1801298	−0.8	2.6	0.1	0.3	1763484	1.0
chempia	−11.1	3.8	−0.1	0.5	1797829	−15.2	3.2	−0.1	0.5	1902056	0.9
chfeps	−16.5	12.4	0.0	0.4	1000891	−1259.5	2898.6	0.0	3.2	1212688	1.0
chinv	−0.3	0.4	0.0	0.1	1753974	−0.3	0.4	0.0	0.1	1717278	1.0
chmom	−9.1	8.7	0.0	0.6	1791987	−9.1	8.4	0.0	0.6	1889792	1.0
chnanalyst	−42.0	39.0	0.0	1.5	1454778	−41.0	28.0	0.0	1.4	1203045	0.8
chpmia	−548.2	116.4	0.2	12.2	1774684	−100.7	64.5	0.1	5.8	1738220	0.4
chtx	−0.1	0.1	0.0	0.0	1687139	−0.1	0.1	0.0	0.0	1743920	1.0

(continued)

Table C.2:
Data comparison

In order to make our dataset as comparable as possible, we benchmark it against the original dataset used by Green et al. (2017). For each firm characteristics, the table shows the minimum, maximum, mean, standard deviation and number of non-missing observations per characteristic, where the first row refers to the dataset used by Green et al. (2017) and the second row refers to the dataset we use. In addition, we report the correlation of each factor over the entire sample for observations that are available in both datasets.

	Green et al. (2017)					Kapetanios and Kempf (2021)					Correlation
	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Corr
cinvest	−16.4	13.3	0.0	0.7	1683575	−17.7	14.3	0.0	0.9	1702671	0.8
convind	0.0	1.0	0.1	0.3	1933709	0.0	1.0	0.1	0.3	1905887	0.7
currat	0.1	55.8	3.3	4.8	1866688	0.0	56.4	3.2	4.7	1840598	1.0
depr	−1.0	5.8	0.3	0.4	1849772	0.0	5.8	0.3	0.4	1825876	1.0
disp	0.0	9.5	0.2	0.4	825466	0.0	48.2	0.2	0.4	925720	1.0
divi	0.0	1.0	0.0	0.2	1802057	0.0	1.0	0.1	0.2	1905887	1.0
divo	0.0	1.0	0.0	0.2	1802057	0.0	1.0	0.0	0.2	1905887	1.0
dolvol	−3.1	19.0	11.2	3.0	1859863	0.0	19.0	11.2	3.0	1849444	1.0
dy	−3.3	0.3	0.0	0.0	1928699	0.0	0.4	0.0	0.0	1900890	1.0
ear	−0.5	0.5	0.0	0.1	1720839	−0.5	0.5	0.0	0.1	1667564	1.0
egr	−3.8	9.0	0.1	0.7	1801856	−3.3	9.4	0.1	0.7	1763813	1.0
ep	−8.2	0.4	0.0	0.4	1933709	−8.0	0.5	0.0	0.4	1905879	1.0
fgr5yr	−16.8	85.0	16.7	9.7	760603	−29.1	100.0	16.7	10.2	1101004	1.0
gma	−1.0	1.8	0.4	0.4	1797413	−0.8	1.8	0.4	0.4	1759498	1.0
grcapx	−12.9	62.0	1.0	3.7	1614468	−12.9	51.3	0.9	3.5	1579720	1.0
grltnoa	−0.6	1.1	0.1	0.2	1348101	−0.6	1.1	0.1	0.2	1320055	1.0
herf	0.0	1.0	0.1	0.1	1933699	0.0	6.7	0.7	0.8	1905877	0.6
hire	−0.8	4.2	0.1	0.4	1797829	−0.8	3.6	0.1	0.3	1902056	1.0
idiovol	0.0	0.3	0.1	0.0	1911760	0.0	0.3	0.1	0.0	1902405	1.0
ill	0.0	0.0	0.0	0.0	1871109	0.0	0.0	0.0	0.0	1852025	1.0
indmom	−0.7	3.5	0.2	0.3	1933552	−0.6	2.8	0.2	0.3	1914251	0.9
invest	−0.6	1.5	0.1	0.2	1740666	−0.6	1.5	0.1	0.2	1706022	1.0
ipo	0.0	1.0	0.1	0.3	1933709	0.0	1.0	0.0	0.1	1914283	0.2
lev	0.0	75.6	2.3	4.9	1928155	0.0	72.9	2.3	4.9	1900545	1.0
lgr	−0.8	10.6	0.3	0.8	1795769	−0.8	10.5	0.3	0.8	1757961	1.0

(continued)

Table C.2:
Data comparison

In order to make our dataset as comparable as possible, we benchmark it against the original dataset used by Green et al. (2017). For each firm characteristics, the table shows the minimum, maximum, mean, standard deviation and number of non-missing observations per characteristic, where the first row refers to the dataset used by Green et al. (2017) and the second row refers to the dataset we use. In addition, we report the correlation of each factor over the entire sample for observations that are available in both datasets.

	Green et al. (2017)					**Kapetanios and Kempf (2021)**					**Correlation**
	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Corr
maxret	0.0	0.9	0.1	0.1	1933708	0.0	0.9	0.1	0.1	1914247	1.0
mom12m	−1.0	12.2	0.1	0.6	1791987	−1.0	11.6	0.1	0.6	1890346	1.0
mom1m	−0.7	2.2	0.0	0.2	1933709	−0.7	2.2	0.0	0.2	1913671	1.0
mom36m	−1.0	16.1	0.3	1.0	1503482	−1.0	16.5	0.3	1.0	1615526	1.0
mom6m	−0.9	8.1	0.1	0.4	1875841	−0.9	7.9	0.1	0.4	1909176	1.0
ms	0.0	8.0	3.8	1.7	1723021	0.0	8.0	3.0	1.6	1870869	0.6
mve	2.4	18.5	11.8	2.2	1933709	3.7	18.5	11.8	2.2	1914283	1.0
mve_ia	−13464.0	93388.7	−225.3	5010.7	1933709	−21176.3	90695.7	−535.5	5005.4	1905879	1.0
nanalyst	0.0	56.0	4.9	6.7	1479704	1.0	56.0	6.8	7.1	1223989	1.0
nincr	0.0	8.0	1.0	1.4	1723021	0.0	8.0	1.0	1.3	1871676	1.0
operprof	−6.3	9.2	0.8	1.1	1797257	−4.6	4.8	0.2	0.6	1759389	0.5
orgcap	0.0	0.1	0.0	0.0	1421950	0.0	0.1	0.0	0.0	106344	1.0
pchcapx_ia	−237.4	1673.3	7.9	81.9	1751019	−102.8	27.1	−0.4	4.5	1721446	0.1
pchcurrat	−0.9	6.4	0.1	0.6	1732910	−0.9	6.7	0.1	0.6	1697069	1.0
pchdepr	−0.9	7.8	0.1	0.6	1714968	−0.9	7.8	0.1	0.6	1681431	1.0
pchgm_pchsale	−12.3	4.8	−0.1	1.0	1778494	−12.2	4.8	−0.1	1.0	1741837	1.0
pchquick	−0.9	8.9	0.1	0.7	1722088	−0.9	8.2	0.1	0.7	1686671	1.0
pchsale_pchinvt	−11.6	3.8	−0.1	0.9	1427967	−11.5	3.5	−0.1	0.9	1402916	1.0
pchsale_pchrect	−7.9	3.4	−0.1	0.6	1727526	−7.0	3.3	−0.1	0.6	1692626	1.0
pchsale_pchxsga	−1.5	5.1	0.0	0.4	1498925	−1.5	4.2	0.0	0.4	1469847	1.0
pchsaleinv	−121.0	33.2	0.2	1.2	1409292	0.0	32.7	0.4	1.0	1384444	0.9
pctacc	−64.8	68.9	−0.9	6.0	1662795	−65.2	69.1	−0.8	5.9	1714379	1.0
pricedelay	−15.6	13.4	0.2	1.1	1911731	−10.4	13.1	0.2	1.0	1897374	0.3
ps	0.0	9.0	4.2	1.7	1802057	0.0	9.0	4.0	1.7	1905887	0.8
quick	0.1	49.3	2.6	4.3	1856675	0.0	49.2	2.6	4.2	1830916	1.0

(continued)

Table C.2:
Data comparison

In order to make our dataset as comparable as possible, we benchmark it against the original dataset used by Green et al. (2017). For each firm characteristics, the table shows the minimum, maximum, mean, standard deviation and number of non-missing observations per characteristic, where the first row refers to the dataset used by Green et al. (2017) and the second row refers to the dataset we use. In addition, we report the correlation of each factor over the entire sample for observations that are available in both datasets.

	Green et al. (2017)					**Kapetanios and Kempf (2021)**					**Correlation**
	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Corr
rd	0.0	1.0	0.1	0.4	1802057	0.0	1.0	0.1	0.3	1905887	0.9
rd_mve	0.0	2.4	0.1	0.1	933103	0.0	2.2	0.1	0.1	920259	1.0
rd_sale	−218.7	117.3	0.6	4.0	919541	0.0	114.7	0.6	3.8	906909	1.0
realestate	0.0	0.9	0.3	0.2	800471	0.0	0.9	0.3	0.2	788910	1.0
retvol	0.0	0.3	0.0	0.0	1933638	0.0	0.3	0.0	0.0	1914214	1.0
roaq	−0.5	0.2	0.0	0.1	1719749	−0.5	0.2	0.0	0.1	1837073	1.0
roavol	0.0	0.7	0.0	0.1	1452795	0.0	0.5	0.0	0.0	1826689	0.9
roeq	−1.4	1.2	0.0	0.1	1719494	−1.4	1.2	0.0	0.2	1836659	1.0
roic	−13.4	1.0	−0.1	0.9	1848261	−14.1	1.0	−0.1	0.9	1821692	1.0
rsup	−4.1	1.5	0.0	0.2	1708311	−3.8	1.5	0.0	0.2	1752159	1.0
salecash	−1230.9	2320.6	57.5	175.6	1917448	0.0	2234.4	57.3	173.2	1889982	1.0
saleinv	−35.4	979.0	28.0	69.7	1526327	0.0	947.1	27.9	69.0	1507878	1.0
salerec	−21796.0	240.9	11.5	60.7	1865361	0.0	239.1	11.6	24.9	1839024	0.4
secured	0.0	4.9	0.6	0.5	1132250	0.0	4.8	0.6	0.5	1113070	1.0
securedind	0.0	1.0	0.5	0.5	1933709	0.0	1.0	0.5	0.5	1905887	1.0
sfe	−105.0	3.0	−0.2	3.0	988947	−213.9	443.2	−0.4	5.0	1220454	0.9
sgr	−0.9	8.7	0.2	0.6	1778687	−0.9	8.1	0.2	0.6	1742029	1.0
sgrvol	0.0	52348.9	1.3	209.2	1451046	0.0	2.4	0.1	0.2	1730368	0.0
sin	0.0	1.0	0.0	0.1	1933709	0.0	1.0	0.0	0.1	1905887	1.0
sp	−4.1	40.5	2.1	3.4	1928114	0.0	37.8	2.1	3.3	1900450	1.0
std_dolvol	0.0	2.9	0.9	0.4	1867918	0.0	2.9	0.9	0.4	1848913	1.0
std_turn	0.0	104.3	4.1	6.4	1872873	0.0	126.9	4.1	6.6	1853743	1.0
stdacc	0.0	554.8	3.8	27.2	1213961	0.0	573.4	3.4	22.4	1511684	0.9
stdcf	0.0	1144.0	8.5	60.6	1213961	0.0	991.4	7.7	50.0	1511670	0.9
sue	−8.8	2.0	0.0	0.1	1710703	−10.3	2.0	0.0	0.2	1790611	0.9

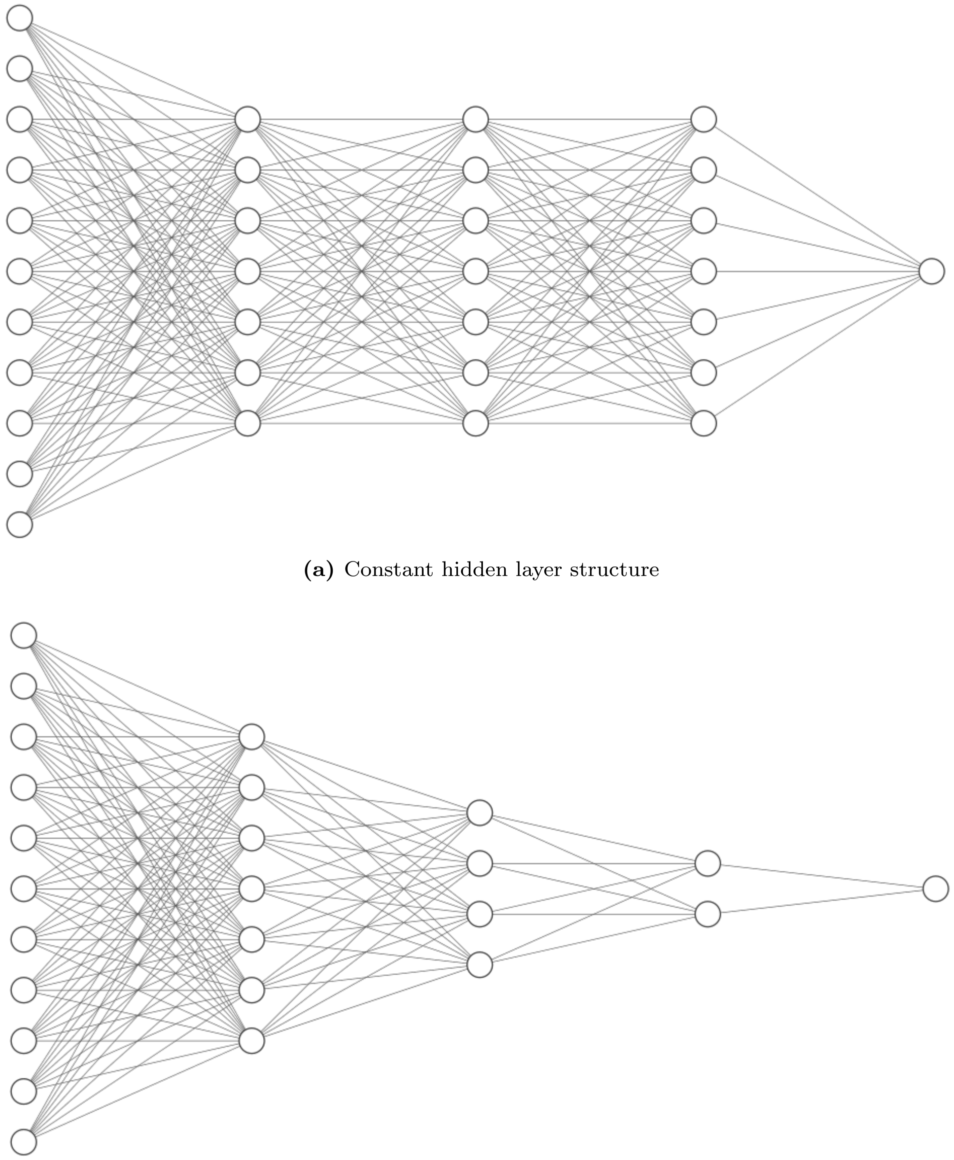
(continued)

Table C.2:
Data comparison
In order to make our dataset as comparable as possible, we benchmark it against the original dataset used by Green et al. (2017). For each firm characteristics, the table shows the minimum, maximum, mean, standard deviation and number of non-missing observations per characteristic, where the first row refers to the dataset used by Green et al. (2017) and the second row refers to the dataset we use. In addition, we report the correlation of each factor over the entire sample for observations that are available in both datasets.

	Green et al. (2017)					**Kapetanios and Kempf (2021)**					**Correlation**
	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Corr
tang	0.0	1.0	0.5	0.2	1853217	0.0	1.0	0.5	0.2	1828573	1.0
tb	−27.5	11.3	−0.1	1.7	1703181	−40.2	20.4	−0.1	1.4	1635468	0.2
turn	0.0	42.7	1.1	1.4	1861007	0.0	14.3	1.0	1.3	1848800	1.0
zerotrade	0.0	20.0	1.5	3.5	1871141	0.0	20.0	1.5	3.5	1852057	1.0

(continued)

D Neural Network Architectures



(a) Constant hidden layer structure

(b) Tapered hidden layer structure

Figure D.1:
Architectural structures
The graph displays a schematic representation of the two different neural network architectures under consideration. The top panel shows the constant hidden layer structure, while the bottom panel displays the tapered hidden layer structure.

We differentiate between two types of neural network architectures, which differ in the architectural form of the hidden layer structure. We refer to the two different architectural designs as *constant* and *tapered*. For the constant architectural structure, the number of nodes remains constant in each hidden layer, where the number of nodes is drawn from the distribution shown in table 2. The two different architectural structures are exemplarily displayed in figure D.1. There exists no deterministic way in the domain-specific context of empirical asset pricing through which an ideal network structure can be derived. To the best of our knowledge, the architectural design of the neural networks applied in empirical asset pricing is currently underrepresented. With this paper, we intend to bring the discussion about architectural decisions to the foreground.

We find, that the number of learnable model parameters change considerably over time, as visualised in figures D.2 and D.3. However, we also find that the average number of learnable model parameters is similar across models. Further, we find that most neural networks contain 4 to 5 hidden layers over time, with exceptions. Moreover, for the the majority of neural networks, the tapered architectural form is found to work best empirically, with the exact reason for this being unclear. The figures merely show an extract of all models for the purpose of clarity. Further results can be requested from the authors. Table D.1 provides further details.

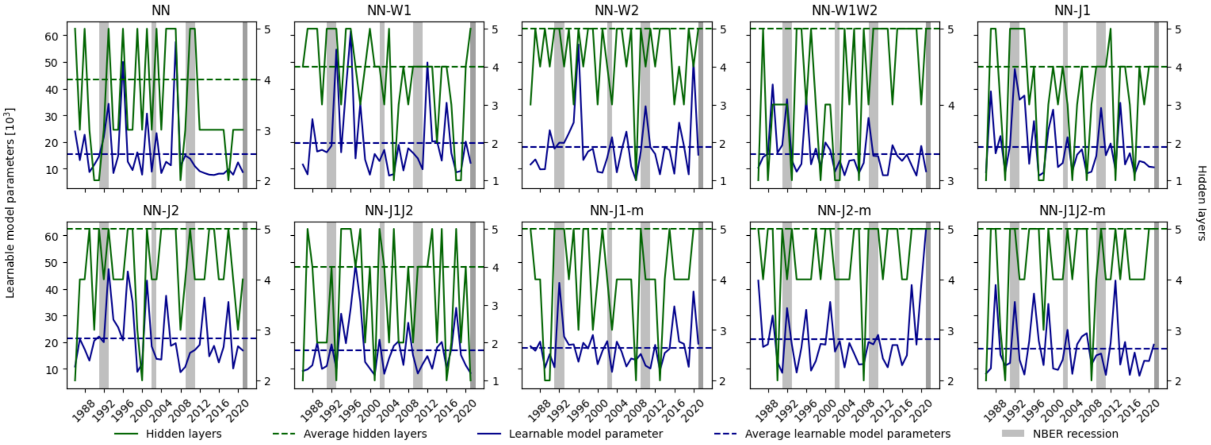


Figure D.2:
Time-varying neural network complexity – rank-normalised, calendar year, all characteristics
The blue line displays the number of learnable parameters over time in thousands, while the green line refers to the number of hidden layers in each neural network. The dotted lines visualise the repspective average.

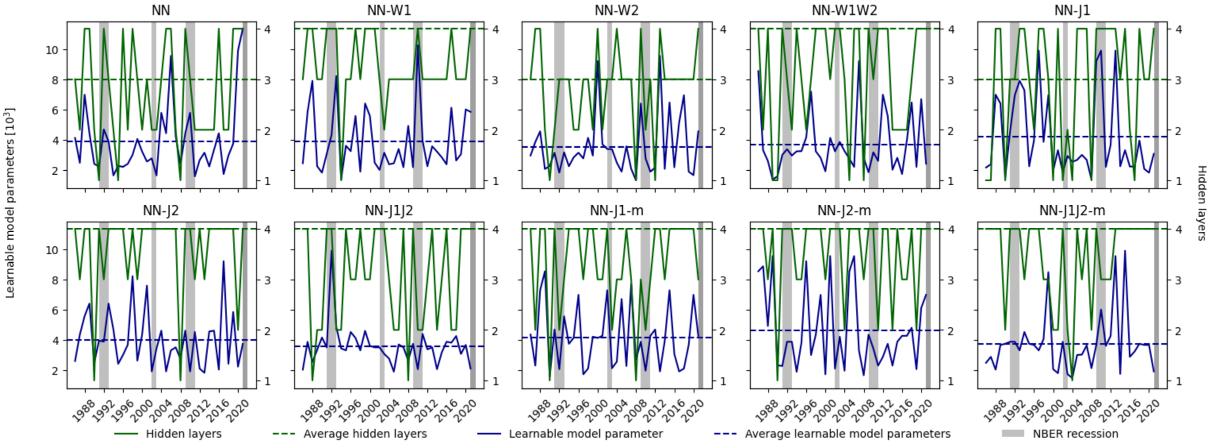


Figure D.3:
Time-varying neural network complexity – rank-normalised, calendar year, core characteristics
The blue line displays the number of learnable parameters over time in thousands, while the green line refers to the number of hidden layers in each neural network. The dotted lines visualise the repspective average.

Table D.1:
Architectural Summary
The table summarises the average number of learnable model parameters, percentage of times of constant hidden-layer architectural form, and the mode of the number of hidden layers over time. The table further differentiates between all data preprocessing strategies considered in this paper, annual re-fitting regimes and firm characteristics considered.

Network	Data pre-processing	Calendar year all characteristics			Calendar year core characteristics			Fiscal year all characteristics			Fiscal year core characteristics		
		Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL
NN	rank norm	15608.5	22.2	3	3910.8	25.0	2	15034.2	30.6	3	3701.1	27.8	4
	rank norm (nm)	10824.7	19.4	1	3068.6	69.4	1	12597.9	22.2	1	2652.3	75.0	1
	std	16973.4	38.9	5	4456.2	36.1	4	16400.7	25.0	3	4064.0	16.7	4
	std (nm)	12630.0	25.0	1	3547.6	77.8	1	10280.4	8.3	3	3072.7	77.8	1
W1	rank norm	19707.6	38.9	4	3913.9	50.0	3	16635.4	38.9	5	3266.0	30.6	3
	rank norm (nm)	21820.8	66.7	5	3383.9	33.3	1	14497.2	52.8	4	2665.9	36.1	1
	std	20024.4	50.0	4	3153.0	25.0	3	14074.8	41.7	5	3457.7	38.9	3
	std (nm)	19296.4	66.7	1	3326.2	41.7	1	22409.1	72.2	5	3199.2	47.2	1
W2	rank norm	18189.5	58.3	5	3522.4	38.9	3	16478.3	55.6	5	3780.2	38.9	3
	rank norm (nm)	20100.6	61.1	5	3407.5	47.2	1	23744.9	52.8	4	3554.4	38.9	1
	std	24882.4	58.3	5	3508.9	27.8	3	16111.3	38.9	5	3862.2	33.3	4
	std (nm)	20602.7	52.8	5	3486.6	44.4	1	19353.7	63.9	1	3294.4	55.6	1
W1W2	rank norm	15531.2	11.1	5	3684.3	36.1	4	16415.9	22.2	5	2990.4	19.4	4
	rank norm (nm)	12693.5	41.7	1	3100.7	66.7	1	15079.0	50.0	1	2751.2	75.0	1
	std	19224.1	30.6	5	3186.0	30.6	4	17461.5	22.2	5	3572.8	27.8	4
	std (nm)	11311.1	36.1	1	3447.6	61.1	1	12333.7	41.7	1	2725.4	75.0	1
J1	rank norm	18290.6	44.4	4	4199.6	47.2	4	18526.2	50.0	5	4080.4	58.3	4
	rank norm (nm)	14055.5	52.8	1	2919.5	11.1	1	12539.5	77.8	1	2936.6	13.9	1
	std	18264.6	47.2	4	4157.6	44.4	4	17196.1	50.0	4	4387.2	50.0	4
	std (nm)	12883.4	50.0	1	3000.4	27.8	1	12634.2	61.1	1	3089.2	25.0	1

(continued)

Table D.1:
Architectural Summary
The table summarises the average number of learnable model parameters, percentage of times of constant hidden-layer architectural form, and the mode of the number of hidden layers over time. The table further differentiates between all data preprocessing strategies considered in this paper, annual re-fitting regimes and firm characteristics considered.

Network	Data pre-processing	Calendar year all characteristics			Calendar year core characteristics			Fiscal year all characteristics			Fiscal year core characteristics		
		Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL	Avg model params.	% of const arch	Mode HL
J2	rank norm	21325.9	44.4	4	3984.3	22.2	4	20192.5	44.4	5	3966.0	27.8	4
	rank norm (nm)	14387.1	13.9	4	3437.5	27.8	4	15647.8	33.3	4	3278.3	30.6	1
	std	20499.5	47.2	5	4610.4	38.9	4	19870.8	47.2	4	3480.4	16.7	4
	std (nm)	15114.3	36.1	2	3153.9	22.2	1	13150.1	30.6	4	3558.9	30.6	4
J1J2	rank norm	17029.9	38.9	2	3583.1	22.2	4	16975.9	41.7	4	3652.5	41.7	4
	rank norm (nm)	11431.9	16.7	1	4830.3	61.1	4	12827.6	19.4	1	3880.7	77.8	1
	std	18601.4	41.7	5	4103.2	38.9	4	19998.7	41.7	4	3766.5	36.1	4
	std (nm)	11260.1	19.4	1	3778.8	52.8	1	12808.1	27.8	1	4602.0	72.2	4
J1-m	rank norm	17819.0	33.3	4	4141.8	30.6	4	20101.3	36.1	4	4645.4	36.1	4
	rank norm (nm)	15084.9	25.0	4	3375.7	36.1	1	14124.7	22.2	4	3748.6	33.3	4
	std	18206.3	25.0	5	4086.4	27.8	4	18606.6	36.1	4	4396.5	30.6	4
	std (nm)	13016.3	36.1	3	3501.9	13.9	1	15626.4	22.2	4	3321.2	22.2	1
J2-m	rank norm	21116.2	36.1	5	4655.4	33.3	4	18786.0	36.1	5	4079.3	25.0	4
	rank norm (nm)	14609.2	16.7	4	3191.4	27.8	1	16301.5	33.3	4	3648.3	27.8	4
	std	17642.4	30.6	4	4139.5	47.2	4	19910.8	33.3	4	4125.8	33.3	4
	std (nm)	12456.0	19.4	4	3919.8	30.6	1	16230.1	38.9	3	3764.8	30.6	1
J1J2-m	rank norm	17551.6	33.3	5	3733.5	16.7	4	17316.2	36.1	5	3623.2	13.9	4
	rank norm (nm)	14860.5	27.8	4	4092.5	44.4	4	16004.2	30.6	4	4696.6	58.3	4
	std	18589.8	27.8	5	3686.4	19.4	4	19786.5	41.7	5	3741.1	16.7	4
	std (nm)	15475.9	30.6	4	5027.3	66.7	3	16929.6	36.1	4	4385.9	61.1	2

(continued)

E Regression Replication

To further the validity of the data collection and preprocessing, we replicate table 4 in Green et al. (2017), with the results reported in table E.1. The purpose of the replication is two-fold. First, we find that our replication is largely in line with the results reported by Green et al. (2017), boosting the confidence in our investment universe. Secondly, even though the results summarised in table E.1 are the product of a slightly diverging methodology compared to the methodology presented in this paper (i.e. univariate regressions and monthly-refitting), they serve as a sanity check for the economic interpretability of our empirical results presented in section 4.11. Note, however, that table E.1 closely resembles the investment universe used in Green et al. (2017), meaning that the sample ends in December 2014. In our empirical application, however, we use all available data and our sample ends in December 2020. In appendix I, we report the linear risk price estimation analogously to those reported in Green et al. (2017), but with the data-preprocessing and investment universe selection of this paper.

Table E.1:
Regression Replication
All stocks Green replication.

	(A) Single characteristics, no benchmark model						(B) Single characteristic, Carhart benchmark						(C) Single characteristic, 5-factor benchmark						(D) Single characteristic, q-factor benchmark					
	All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
agr	−0.22	−3.17	−0.26	−4.52	−0.46	−8.26	−0.20	−3.48	−0.23	−5.01	−0.37	−8.52												
bm	0.18	1.13	0.15	1.88	0.34	5.02													0.13	0.93	0.11	1.51	0.17	2.40
mom12m	0.33	2.54	0.26	2.32	0.19	1.70							0.30	2.43	0.26	2.45	0.22	2.24	0.31	2.45	0.25	2.31	0.21	2.13
mve	−0.09	−0.88	−0.05	−0.95	−0.33	−2.99																		
operprof	0.07	1.21	0.08	1.58	0.05	0.62	0.09	2.17	0.11	2.65	0.13	2.18							0.06	1.44	0.07	1.96	0.12	2.49
roeq	0.23	2.79	0.16	2.70	0.15	1.71	0.26	3.62	0.15	3.11	0.20	2.99	0.26	4.14	0.15	3.36	0.22	3.59						
absacc	−0.04	−0.40	−0.09	−1.40	−0.01	−0.10	−0.08	−0.98	−0.09	−1.97	−0.04	−0.62	−0.01	−0.08	−0.04	−0.95	0.00	−0.08	0.01	0.15	−0.04	−0.77	−0.01	−0.24
acc	−0.22	−2.48	−0.12	−2.50	−0.21	−3.13	−0.13	−1.97	−0.08	−2.21	−0.16	−3.00	−0.19	−2.49	−0.09	−2.46	−0.13	−2.79	−0.20	−2.49	−0.10	−2.38	−0.13	−2.70
aeavol	0.00	0.04	0.00	−0.15	0.10	5.24	−0.03	−0.56	−0.01	−0.66	0.05	2.65	0.00	0.07	0.00	0.15	0.07	3.89	0.00	0.05	0.00	−0.02	0.06	3.57
age	0.00	0.00	0.08	1.13	0.10	1.33	0.05	0.86	0.11	2.14	0.27	5.99	−0.04	−0.61	0.02	0.38	0.19	4.08	−0.02	−0.31	0.04	0.64	0.22	3.80
baspread	−0.43	−1.28	−0.22	−1.56	0.23	1.53	−0.64	−2.29	−0.29	−2.33	0.13	0.97	−0.40	−1.33	−0.18	−1.37	0.14	1.07	−0.42	−1.34	−0.19	−1.35	0.16	1.17
beta	−0.08	−0.49	−0.11	−0.72	−0.08	−0.58	−0.17	−1.11	−0.16	−1.17	−0.06	−0.43	−0.03	−0.21	−0.06	−0.41	0.00	0.01	−0.05	−0.28	−0.07	−0.46	−0.01	−0.08
betasq	−0.10	−0.53	−0.12	−0.83	−0.09	−0.65	−0.19	−1.15	−0.17	−1.29	−0.08	−0.69	−0.05	−0.28	−0.07	−0.53	−0.03	−0.21	−0.06	−0.32	−0.08	−0.57	−0.04	−0.29
bm_ia	−0.04	−0.65	0.00	−0.10	0.18	3.11	0.00	0.01	0.02	0.26	0.07	0.72	0.01	0.12	0.04	0.65	0.10	0.92	−0.03	−0.49	0.00	−0.13	0.08	1.86
cash	0.11	1.09	0.01	0.13	0.09	0.80	0.11	1.39	0.02	0.20	0.11	1.19	0.19	2.02	0.08	0.83	0.14	1.57	0.16	1.56	0.06	0.53	0.11	1.13
cashdebt	0.00	0.01	−0.07	−1.09	−0.03	−0.31	0.03	0.29	−0.04	−0.75	0.00	−0.02	0.08	0.65	−0.02	−0.38	0.02	0.26	0.05	0.33	−0.04	−0.60	0.00	−0.05
cashpr	−0.08	−1.77	−0.10	−2.40	−0.16	−3.55	−0.04	−1.09	−0.04	−1.69	−0.05	−1.56	−0.06	−1.66	−0.05	−2.04	−0.04	−1.40	−0.07	−1.86	−0.08	−2.26	−0.08	−1.90
cfp	0.14	1.42	0.15	2.24	0.11	1.40	0.10	1.56	0.13	2.76	0.12	2.01	0.07	1.06	0.09	2.11	0.11	2.03	0.06	0.73	0.09	1.60	0.10	1.65
cfp_ia	0.08	1.71	0.12	4.23	0.15	2.95	0.05	1.32	0.10	5.03	0.13	3.81	0.03	0.82	0.08	4.25	0.13	4.29	0.05	1.10	0.09	3.84	0.12	3.51
chatoia	0.11	2.51	0.09	4.23	0.09	5.42	0.08	2.13	0.07	3.76	0.08	4.89	0.04	1.07	0.04	2.05	0.05	3.27	0.04	1.26	0.04	2.07	0.05	2.98
chcsho	−0.13	−3.03	−0.18	−3.57	−0.26	−6.35	−0.10	−3.06	−0.13	−3.53	−0.17	−5.80	−0.08	−2.66	−0.06	−2.42	−0.07	−2.66	−0.09	−2.86	−0.08	−2.84	−0.08	−3.01
chempia	−0.05	−1.16	−0.08	−2.09	−0.18	−4.17	−0.04	−1.18	−0.06	−1.81	−0.12	−3.10	0.05	1.18	0.03	1.03	0.01	0.36	0.04	1.01	0.04	1.12	0.01	0.34
chfeps	0.25	2.88	0.11	2.91	0.18	5.60	0.18	2.47	0.09	2.51	0.17	5.69	0.24	2.90	0.11	2.97	0.19	5.64	0.24	2.76	0.11	2.89	0.18	5.79
chinv	−0.16	−3.18	−0.14	−4.26	−0.26	−6.60	−0.11	−2.83	−0.11	−4.24	−0.20	−5.79	−0.08	−1.81	−0.07	−2.50	−0.11	−3.75	−0.09	−2.09	−0.07	−2.59	−0.12	−3.84

(continued)

Table E.1:
Regression Replication
All stocks Green replication.

	(A) Single characteristics, no benchmark model						(B) Single characteristic, Carhart benchmark						(C) Single characteristic, 5-factor benchmark						(D) Single characteristic, q-factor benchmark					
	All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
chmom	−0.31	−2.86	−0.11	−1.68	−0.21	−3.42	−0.28	−3.12	−0.09	−1.52	−0.18	−3.06	−0.31	−3.52	−0.12	−2.17	−0.21	−4.08	−0.31	−3.30	−0.10	−1.80	−0.19	−3.55
chnanalyst	0.02	0.74	0.00	−0.06	0.00	0.01	0.02	0.45	−0.01	−0.26	0.02	0.41	0.02	0.51	0.02	0.53	0.05	1.04	0.02	0.61	0.01	0.27	0.03	0.79
chpmia	0.01	0.27	−0.02	−0.48	−0.02	−0.46	0.01	0.14	−0.02	−0.68	−0.02	−0.63	0.01	0.22	−0.01	−0.22	0.00	−0.11	0.01	0.16	−0.01	−0.31	0.00	0.07
chtx	0.09	1.81	0.06	1.68	0.14	5.23	0.05	1.29	0.01	0.42	0.12	4.78	0.13	3.71	0.09	3.21	0.18	6.95	0.09	2.01	0.05	1.46	0.15	6.11
cinvest	0.02	0.60	−0.01	−0.44	0.00	−0.07	0.01	0.31	−0.01	−0.61	0.00	−0.12	0.01	0.15	−0.02	−0.64	−0.01	−0.47	−0.01	−0.16	−0.02	−0.92	−0.02	−1.07
convind	−0.04	−1.72	−0.05	−1.93	−0.06	−2.96	−0.04	−1.69	−0.04	−1.65	−0.03	−1.50	−0.03	−1.32	−0.03	−1.36	−0.03	−1.18	−0.03	−1.29	−0.02	−1.04	−0.02	−1.01
currat	−0.10	−1.51	−0.11	−2.19	−0.04	−1.14	−0.09	−1.74	−0.09	−2.19	−0.06	−1.61	−0.04	−0.89	−0.05	−1.38	−0.04	−1.18	−0.08	−1.49	−0.08	−1.89	−0.05	−1.58
depr	0.01	0.12	0.00	0.07	0.12	1.68	0.00	0.04	0.01	0.26	0.10	1.79	0.04	0.44	0.03	0.76	0.09	1.76	0.04	0.52	0.03	0.59	0.08	1.49
disp	−0.14	−0.94	−0.13	−2.36	−0.22	−3.43	−0.16	−1.31	−0.13	−2.63	−0.25	−4.68	−0.13	−0.91	−0.12	−2.37	−0.26	−4.68	−0.11	−0.76	−0.10	−2.09	−0.24	−4.47
divi	−0.01	−0.16	−0.07	−1.94	−0.08	−3.79	−0.04	−0.94	−0.07	−2.26	−0.10	−4.76	−0.01	−0.32	−0.06	−1.78	−0.08	−4.32	−0.01	−0.34	−0.06	−1.88	−0.08	−4.74
divo	0.03	0.77	−0.01	−0.63	0.02	0.95	0.00	−0.11	−0.02	−0.97	0.00	−0.28	0.04	1.45	0.01	0.41	0.01	0.86	0.04	1.46	0.01	0.31	0.01	1.01
dolvol	−0.14	−1.14	−0.08	−1.31	−0.29	−3.36	−0.21	−0.76	−0.14	−1.04	0.17	0.73	0.04	0.14	0.03	0.16	0.37	1.42	0.02	0.05	0.01	0.06	0.33	1.25
dy	0.03	0.25	0.06	0.76	0.04	0.66	0.01	0.17	0.01	0.22	0.05	1.14	−0.05	−0.60	−0.04	−0.66	0.00	0.05	−0.01	−0.07	0.01	0.16	0.04	0.62
ear	0.14	2.35	0.12	5.58	0.18	6.69	0.08	1.78	0.08	5.43	0.14	7.43	0.14	2.58	0.12	5.72	0.17	7.14	0.14	2.36	0.11	5.17	0.15	6.43
egr	−0.19	−2.91	−0.19	−3.96	−0.26	−5.40	−0.18	−3.29	−0.17	−4.24	−0.20	−5.42	−0.12	−2.27	−0.09	−2.91	−0.06	−2.33	−0.07	−1.86	−0.05	−2.18	−0.04	−1.53
ep	0.21	0.93	0.06	0.80	−0.12	−1.06	0.24	1.46	0.07	1.25	−0.04	−0.48	0.14	0.85	0.04	0.64	−0.02	−0.24	0.13	0.67	0.03	0.48	−0.04	−0.47
fgr5yr	−0.05	−0.32	−0.11	−0.87	−0.05	−0.42	−0.10	−0.70	−0.12	−1.11	−0.06	−0.64	0.05	0.34	0.00	0.01	0.04	0.41	0.01	0.07	−0.04	−0.30	−0.01	−0.07
gma	0.09	1.15	0.05	1.14	0.06	1.37	0.15	2.49	0.10	2.64	0.09	2.17	0.20	2.90	0.13	3.21	0.17	4.07	0.13	1.64	0.08	1.83	0.14	3.13
grcapx	−0.17	−2.75	−0.15	−4.31	−0.20	−7.34	−0.17	−3.16	−0.14	−4.59	−0.18	−7.48	−0.08	−2.21	−0.08	−3.37	−0.10	−5.13	−0.08	−2.07	−0.08	−3.15	−0.10	−5.15
grltnoa	−0.17	−3.86	−0.19	−4.31	−0.34	−6.74	−0.15	−3.92	−0.16	−4.28	−0.27	−6.28	−0.06	−1.25	−0.06	−1.90	−0.09	−2.53	−0.04	−0.71	−0.05	−1.44	−0.08	−2.23
herf	0.03	1.07	0.03	0.62	−0.02	−0.38	0.03	1.15	0.03	0.77	−0.03	−0.74	0.03	1.00	0.03	0.68	−0.03	−0.64	0.04	1.18	0.03	0.62	−0.02	−0.47
hire	−0.12	−2.00	−0.15	−3.28	−0.28	−7.35	−0.11	−2.40	−0.13	−3.44	−0.22	−6.95	0.02	0.38	0.01	0.26	−0.05	−2.50	0.02	0.40	0.01	0.42	−0.06	−2.69
idiovol	−0.15	−0.64	−0.19	−1.35	0.02	0.10	−0.40	−1.74	−0.30	−2.29	−0.20	−1.30	−0.13	−0.52	−0.15	−1.11	−0.11	−0.72	−0.14	−0.51	−0.16	−1.08	−0.12	−0.75
ill	−0.08	−0.57	−0.07	−2.98	0.41	5.38	−0.17	−1.73	−0.07	−3.21	0.33	5.41	−0.26	−2.60	−0.09	−3.94	0.29	4.87	−0.26	−2.78	−0.09	−3.69	0.32	5.32

(continued)

Table E.1:
Regression Replication
All stocks Green replication.

	(A) Single characteristics, no benchmark model						(B) Single characteristic, Carhart benchmark						(C) Single characteristic, 5-factor benchmark						(D) Single characteristic, q-factor benchmark					
	All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
indmom	0.09	1.24	0.15	1.69	0.32	3.40	0.03	0.53	0.09	1.42	0.28	3.84	0.08	1.26	0.15	1.91	0.31	3.61	0.07	1.03	0.14	1.67	0.31	3.41
invest	−0.14	−2.67	−0.20	−4.19	−0.36	−7.41	−0.13	−3.24	−0.18	−4.63	−0.30	−7.46	−0.06	−1.04	−0.09	−2.25	−0.15	−3.50	−0.05	−0.77	−0.09	−2.07	−0.15	−3.34
ipo	−0.10	−2.96	−0.08	−2.80	−0.12	−4.62	−0.08	−2.67	−0.08	−2.86	−0.11	−4.82	−0.09	−2.97	−0.07	−2.81	−0.11	−4.77	−0.10	−3.14	−0.08	−2.90	−0.12	−4.98
lev	0.06	0.51	0.09	1.09	0.11	1.26	0.01	0.06	0.02	0.34	−0.01	−0.08	0.01	0.16	0.02	0.35	−0.01	−0.11	0.04	0.37	0.07	0.86	0.05	0.56
lgr	−0.18	−2.94	−0.17	−4.08	−0.30	−9.04	−0.16	−3.28	−0.15	−4.78	−0.26	−9.88	0.07	1.37	0.05	1.67	−0.01	−0.23	0.05	1.04	0.05	1.30	−0.02	−0.59
maxret	−0.23	−1.13	−0.24	−2.20	−0.18	−1.46	−0.40	−2.28	−0.29	−3.14	−0.35	−3.47	−0.26	−1.43	−0.22	−2.27	−0.33	−3.39	−0.27	−1.41	−0.22	−2.17	−0.33	−3.29
mom1m	−0.11	−1.07	−0.13	−1.89	−0.63	−6.76	−0.26	−2.91	−0.21	−3.20	−0.69	−7.87	−0.22	−2.27	−0.19	−2.66	−0.68	−7.91	−0.20	−1.96	−0.17	−2.34	−0.66	−7.64
mom36m	−0.07	−0.74	−0.17	−2.40	−0.28	−2.98	−0.09	−1.19	−0.13	−2.21	−0.16	−2.14	−0.05	−0.62	−0.09	−1.65	−0.10	−1.55	−0.06	−0.82	−0.11	−1.98	−0.13	−2.01
mom6m	0.09	0.77	0.18	1.90	0.09	0.96	−0.30	−2.52	−0.03	−0.38	−0.08	−1.13	0.08	0.70	0.18	1.95	0.10	1.20	0.08	0.68	0.18	1.86	0.10	1.21
ms	0.06	1.10	0.07	1.36	0.09	1.73	0.10	2.73	0.12	2.63	0.19	3.97	0.11	2.50	0.11	2.37	0.20	3.93	0.07	1.50	0.09	1.75	0.17	3.22
mve_ia	−0.02	−1.04	−0.04	−1.01	−0.04	−0.89	0.01	0.52	0.05	1.54	0.20	5.49	−0.02	−1.09	−0.01	−0.30	0.14	3.78	−0.03	−1.39	−0.02	−0.43	0.16	4.23
nanalyst	0.00	0.02	0.05	1.05	−0.05	−0.76	0.03	0.60	0.14	2.50	0.23	3.63	0.02	0.46	0.11	1.99	0.19	2.94	0.02	0.37	0.13	2.24	0.22	3.52
nincr	0.09	2.87	0.13	4.16	0.18	7.14	0.08	3.20	0.10	3.91	0.20	8.30	0.10	3.82	0.15	5.21	0.23	9.32	0.08	2.98	0.12	4.21	0.19	8.49
orgcap	0.08	0.91	0.08	1.45	0.13	1.91	0.09	0.95	0.09	1.62	0.08	1.03	0.08	0.80	0.07	1.28	0.07	0.91	0.06	0.58	0.05	0.89	0.04	0.53
pchcapx_ia	−0.01	−0.14	0.00	−0.07	−0.05	−0.95	−0.02	−0.39	0.00	−0.04	−0.04	−0.87	0.01	0.35	0.02	0.54	0.01	0.20	0.04	0.82	0.04	0.75	0.01	0.16
pchcurrat	−0.09	−2.55	−0.07	−3.29	−0.09	−3.11	−0.09	−2.80	−0.07	−3.70	−0.09	−3.11	−0.07	−2.09	−0.02	−1.29	−0.03	−1.11	−0.07	−2.14	−0.02	−1.55	−0.03	−1.18
pchdepr	0.03	0.45	0.02	0.58	0.02	0.51	−0.01	−0.10	0.01	0.40	0.00	−0.12	0.05	0.98	0.02	0.85	−0.03	−0.89	0.05	0.99	0.02	0.76	−0.03	−1.03
pchgm_pchsale	0.08	2.45	0.08	3.05	0.09	3.07	0.08	2.94	0.08	3.59	0.10	4.16	0.05	1.66	0.05	2.49	0.06	2.92	0.05	1.51	0.05	2.34	0.07	3.06
pchquick	−0.05	−1.36	−0.05	−2.13	−0.07	−2.35	−0.06	−1.68	−0.05	−2.64	−0.07	−2.47	−0.02	−0.67	0.00	0.27	−0.01	−0.23	−0.03	−0.90	0.00	−0.11	−0.01	−0.39
pchsale_pchinvt	0.06	1.43	0.09	3.73	0.14	5.34	0.05	1.31	0.08	4.25	0.13	5.17	0.03	0.91	0.06	2.89	0.10	4.04	0.02	0.64	0.05	2.11	0.09	3.45
pchsale_pchrect	0.01	0.36	0.04	1.95	0.08	4.28	0.00	0.14	0.03	1.76	0.08	4.68	−0.04	−0.98	−0.02	−1.10	0.01	0.43	−0.04	−0.86	−0.02	−0.88	0.01	0.32
pchsale_pchxsga	−0.09	−1.65	−0.08	−2.70	−0.08	−3.06	−0.10	−1.80	−0.07	−2.91	−0.06	−2.41	−0.03	−0.77	−0.03	−1.53	0.00	−0.06	−0.04	−0.94	−0.04	−1.53	−0.01	−0.33
pchsaleinv	−0.02	−0.32	−0.03	−1.01	−0.04	−0.99	−0.05	−1.31	−0.04	−1.85	−0.07	−1.84	0.01	0.16	−0.01	−0.52	−0.04	−1.07	0.01	0.28	−0.01	−0.40	−0.05	−1.36
pctacc	−0.02	−0.60	−0.06	−2.24	−0.11	−3.90	−0.01	−0.24	−0.04	−1.92	−0.09	−4.25	−0.01	−0.14	−0.03	−1.38	−0.08	−3.67	−0.01	−0.36	−0.03	−1.41	−0.08	−3.41

(continued)

Table E.1:
Regression Replication
All stocks Green replication.

	(A) Single characteristics, no benchmark model						(B) Single characteristic, Carhart benchmark						(C) Single characteristic, 5-factor benchmark						(D) Single characteristic, q-factor benchmark					
	All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS		All WLS		No micro OLS		All OLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
pricedelay	0.00	−0.05	0.01	0.35	0.05	2.46	0.02	0.27	0.02	1.06	0.01	0.59	0.00	0.07	0.02	0.77	0.01	0.55	0.00	−0.06	0.01	0.52	0.01	0.40
ps	0.05	1.23	0.11	2.55	0.09	1.69	0.06	1.51	0.12	3.25	0.16	4.66	0.03	0.75	0.08	2.49	0.14	4.99	0.01	0.35	0.07	2.17	0.14	4.52
quick	−0.11	−1.78	−0.10	−1.89	−0.04	−0.96	−0.10	−2.24	−0.09	−2.01	−0.04	−1.10	−0.05	−1.16	−0.04	−1.02	−0.02	−0.55	−0.09	−1.70	−0.06	−1.42	−0.03	−0.89
rd	0.05	1.74	0.07	1.97	0.18	2.64	0.04	1.45	0.06	2.24	0.16	2.82	0.01	0.55	0.05	1.59	0.12	2.31	0.01	0.41	0.04	1.28	0.11	1.97
rd_mve	0.10	0.73	0.21	2.14	0.47	3.83	0.02	0.26	0.17	1.87	0.38	3.52	−0.01	−0.07	0.17	1.75	0.34	3.40	0.03	0.31	0.20	2.11	0.37	3.80
rd_sale	−0.23	−1.80	−0.08	−1.21	−0.03	−0.43	−0.19	−1.98	−0.06	−1.13	0.00	−0.07	−0.14	−1.36	−0.02	−0.32	0.02	0.33	−0.15	−1.39	−0.02	−0.29	0.01	0.23
realestate	0.10	1.62	0.09	1.33	0.02	0.38	0.12	2.05	0.09	1.64	0.04	0.78	0.10	1.73	0.07	1.27	0.02	0.48	0.10	1.71	0.07	1.23	0.03	0.51
retvol	−0.27	−1.12	−0.27	−2.10	−0.05	−0.33	−0.48	−2.33	−0.34	−3.12	−0.21	−1.82	−0.28	−1.29	−0.24	−2.07	−0.18	−1.59	−0.29	−1.30	−0.25	−2.02	−0.18	−1.55
roaq	0.34	2.88	0.17	2.32	0.15	1.28	0.40	4.46	0.18	3.10	0.21	2.32	0.40	4.78	0.18	3.21	0.24	2.77	0.28	2.44	0.11	1.48	0.17	1.77
roavol	−0.17	−1.35	−0.12	−1.36	−0.09	−0.74	−0.21	−2.21	−0.12	−1.92	−0.14	−1.52	−0.10	−0.94	−0.05	−0.75	−0.09	−0.98	−0.10	−0.85	−0.06	−0.76	−0.09	−1.01
roic	0.32	2.77	0.14	2.20	0.02	0.15	0.39	4.43	0.16	3.02	0.04	0.49	0.37	4.24	0.12	2.21	0.03	0.40	0.31	3.38	0.10	1.75	0.03	0.43
rsup	0.01	0.16	0.02	0.63	−0.06	−1.00	−0.01	−0.13	0.01	0.20	−0.03	−0.66	0.05	0.65	0.06	1.65	0.05	1.05	0.05	0.65	0.05	1.40	0.02	0.48
salecash	0.05	0.86	0.03	0.87	0.03	0.82	0.02	0.52	0.03	0.98	0.00	−0.07	0.00	0.10	0.01	0.26	−0.02	−0.62	0.01	0.22	0.01	0.35	−0.01	−0.32
saleinv	−0.01	−0.35	0.02	0.64	0.03	1.05	−0.02	−0.67	0.01	0.50	0.03	1.61	−0.02	−0.85	0.01	0.22	0.03	1.30	−0.01	−0.30	0.02	0.54	0.03	1.44
salerec	0.08	1.66	0.06	1.45	0.04	1.07	0.07	1.84	0.06	1.59	0.04	0.97	0.07	1.62	0.05	1.27	0.02	0.62	0.07	1.58	0.05	1.17	0.03	0.72
secured	−0.16	−1.05	−0.07	−1.14	0.02	0.28	−0.21	−1.32	−0.12	−2.10	−0.11	−2.43	−0.18	−1.02	−0.05	−0.92	−0.09	−1.70	−0.19	−1.05	−0.08	−1.16	−0.10	−1.92
securedind	0.01	0.23	−0.02	−0.58	−0.01	−0.21	0.01	0.19	−0.02	−0.61	−0.06	−1.46	0.01	0.40	0.00	0.17	−0.03	−0.75	0.01	0.52	0.00	−0.11	−0.03	−0.79
sfe	−0.13	−0.97	0.00	−0.03	−0.01	−0.17	−0.12	−1.09	−0.01	−0.17	−0.01	−0.09	−0.13	−1.11	−0.03	−0.60	0.00	0.06	−0.14	−1.10	−0.04	−0.69	−0.01	−0.15
sgr	−0.26	−2.75	−0.16	−2.89	−0.25	−6.67	−0.23	−3.24	−0.14	−3.42	−0.20	−6.54	−0.08	−1.30	−0.02	−0.50	−0.05	−1.78	−0.10	−1.50	−0.02	−0.46	−0.06	−2.02
sgrvol	0.05	0.42	0.04	1.03	0.18	2.72	−0.11	−1.46	−0.02	−0.54	0.00	0.03	−0.09	−1.11	0.01	0.15	0.02	0.28	−0.03	−0.29	0.03	0.77	0.05	0.96
sin	0.04	1.69	0.04	1.83	0.03	1.62	0.04	1.95	0.03	1.81	0.04	2.17	0.03	1.67	0.04	1.90	0.04	2.11	0.03	1.69	0.03	1.72	0.04	1.93
sp	0.28	1.82	0.17	2.48	0.29	4.25	0.11	1.24	0.10	1.91	0.11	1.82	0.08	0.84	0.08	1.57	0.09	1.47	0.17	1.33	0.13	2.01	0.15	2.29
std_dolvol	0.05	0.44	0.06	1.55	0.16	2.81	−0.09	−1.02	0.05	1.27	−0.11	−1.41	−0.09	−0.96	0.03	0.73	−0.14	−1.74	−0.08	−0.83	0.03	0.79	−0.14	−1.62
std_turn	−0.01	−0.08	−0.05	−0.62	−0.02	−0.31	−0.12	−1.45	−0.09	−1.51	−0.05	−0.93	0.01	0.15	−0.01	−0.12	0.02	0.33	0.02	0.14	−0.01	−0.14	0.01	0.23

(continued)

Table E.1:
Regression Replication
All stocks Green replication.

	(A) Single characteristics, no benchmark model						(B) Single characteristic, Carhart benchmark						(C) Single characteristic, 5-factor benchmark						(D) Single characteristic, q-factor benchmark					
	All		No micro		All		All		No micro		All		All		No micro		All		All		No micro		All	
	WLS		OLS		OLS		WLS		OLS		OLS		WLS		OLS		OLS		WLS		OLS		OLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
stdacc	−0.15	−1.95	−0.08	−2.02	−0.08	−1.42	−0.18	−3.29	−0.09	−2.98	−0.09	−2.17	−0.13	−2.30	−0.05	−1.58	−0.06	−1.58	−0.12	−2.00	−0.05	−1.41	−0.06	−1.58
stdcf	−0.16	−2.10	−0.09	−2.12	−0.08	−1.44	−0.18	−3.38	−0.10	−3.10	−0.08	−1.96	−0.13	−2.40	−0.05	−1.65	−0.05	−1.37	−0.12	−2.13	−0.05	−1.44	−0.05	−1.33
sue	0.21	2.41	0.10	3.35	0.20	5.82	0.14	1.70	0.07	2.44	0.19	6.68	0.20	2.53	0.09	3.32	0.21	7.19	0.16	1.97	0.07	2.47	0.14	4.96
tang	0.02	0.35	−0.01	−0.08	0.07	0.95	0.01	0.34	0.00	−0.05	0.05	0.80	0.03	0.54	0.02	0.25	0.05	0.85	0.02	0.28	0.00	0.03	0.04	0.53
tb	0.05	1.32	0.06	1.68	0.04	0.77	0.09	2.57	0.07	2.72	0.11	3.25	0.07	1.87	0.05	1.73	0.10	3.29	0.06	1.46	0.04	1.37	0.09	2.93
turn	−0.03	−0.29	−0.16	−1.39	−0.28	−3.30	−0.13	−1.35	−0.21	−2.32	−0.25	−2.95	0.00	0.02	−0.09	−0.88	−0.13	−1.40	0.00	−0.02	−0.10	−0.89	−0.14	−1.48
zerotrade	−0.01	−0.15	−0.03	−1.36	0.09	1.74	−0.12	−1.57	−0.03	−1.52	−0.07	−0.97	−0.17	−2.17	−0.04	−1.90	−0.10	−1.39	−0.17	−2.10	−0.04	−1.83	−0.07	−0.95

(continued)

F Correlation Analysis

Figures F.1 to F.4 exhibit the correlation of all stocks and firm characteristics included in our analysis from January 1980 to December 2020. We use the rank-normalised, winsorised and pooled dataset to calculate the correlation coefficients. We further differentiate between the core characteristics only, all characteristics, and empirical correlation matrices for an investment universe, in which microcaps are excluded. We find that the empirical correlation structure is fairly heterogenous. The average firm characteristic correlation for the core characteristics across all stocks is 2.2%, 2.4% for core characteristics without microcaps, 2.6% for all characteristics across all stocks, and 3.1% for all characteristics without microcaps.

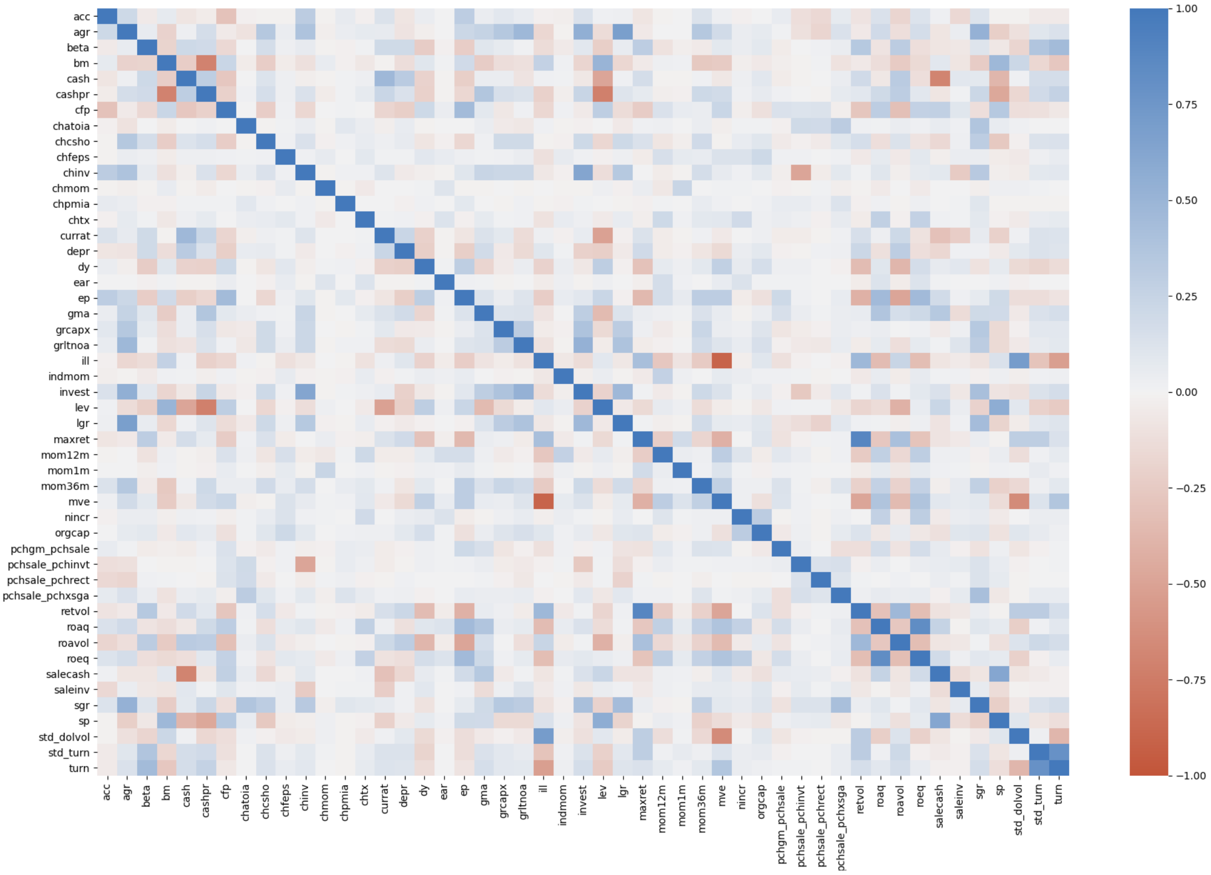


Figure F.1:
Empirical correlation matrix – rank-normalised data, core characteristics
The graph shows the empirical correlation matrix of the core characteristics, from January 1980 to December 2020.

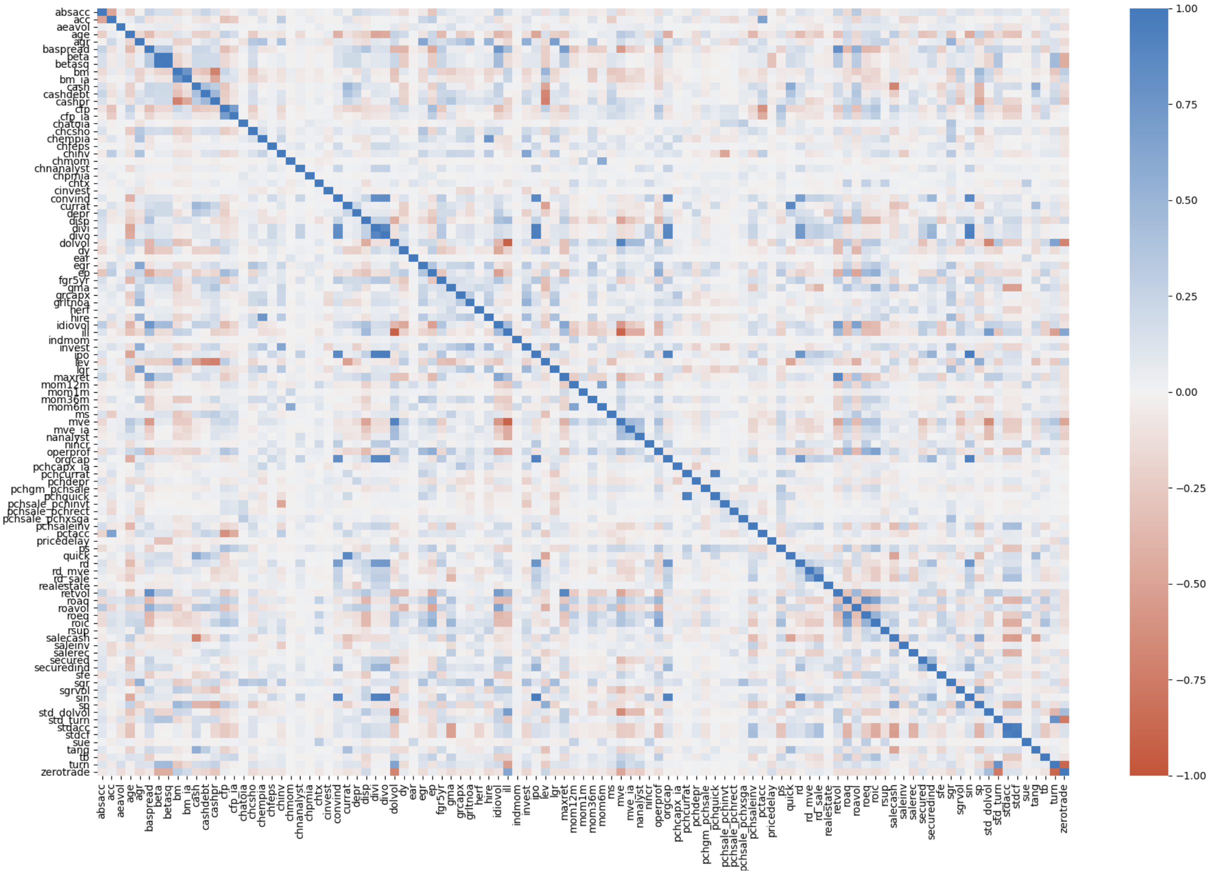


Figure F.2:
Empirical correlation matrix– rank-normalised data, all characteristics
The graph shows the empirical correlation matrix of all characteristics, from January 1980 to December 2020.

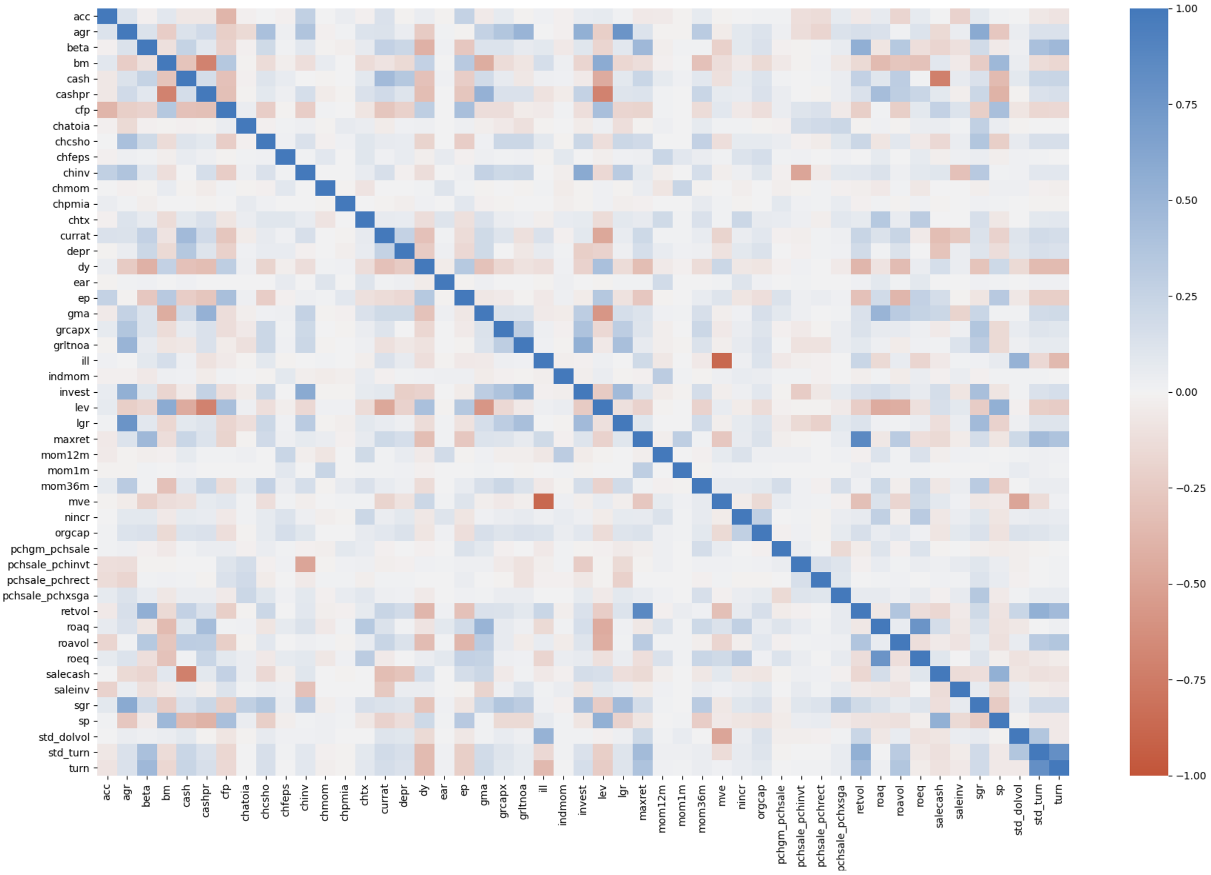


Figure F.3:
Empirical correlation matrix – rank-normalised data, core characteristics (no microcaps)
The graph shows the empirical correlation matrix of the core characteristics, excluding microcaps, from January 1980 to December 2020.

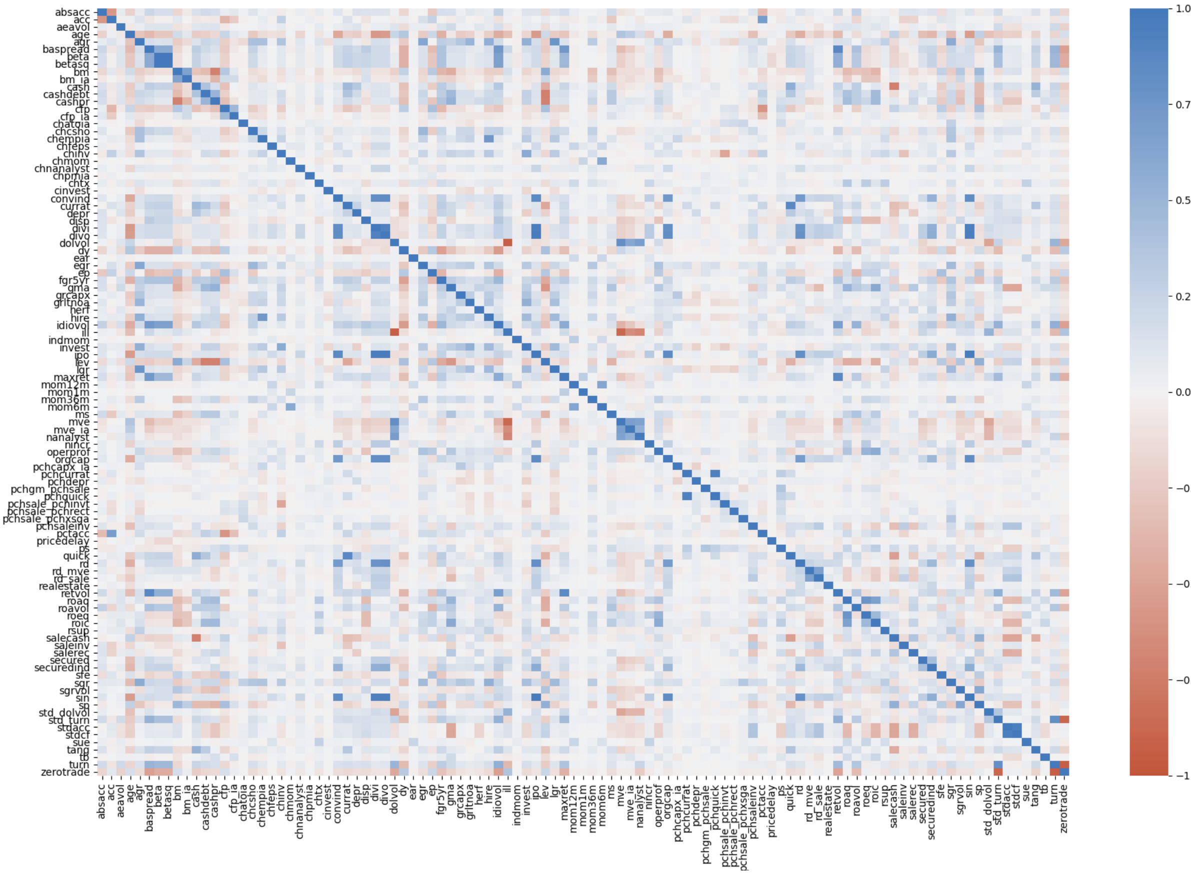


Figure F.4:
Empirical correlation matrix – rank-normalised data, all characteristics (no microcaps)
The graph shows the empirical correlation matrix of all characteristics, excluding microcaps, from January 1980 to December 2020.

G Performance Summary

This section in the appendix provides further details about the empirical out-of-sample performances measured in cross-sectional mean R^2 and predictive R^2 (see equations (38) and (39)). In particular, we report a more detailed summary and include results from all different types of data pre-processing (rank-normalisation vs. standardisation), re-fit frequencies (calendar vs. fiscal year), and investment universe restrictions (all stocks vs. excluding microcaps). We further report results by market capitalisation, yielding a large number of results. A full discussion is beyond the scope of this appendix. However, we find that the neural network performances are stable across data-preprocessing regimes, with neural networks tending to offer robust performances for large stocks, which are the most liquid and relevant for institutional investors.

Model	Subgroup	Calendar Year – Rank Normalisation			
		All Characteristics		Core Characteristics	
		XS-R^2 [%]	Pred.-R^2 [%]	XS-R^2 [%]	Pred.-R^2 [%]
OLS	Overall	−0.35	5.57	−0.07	6.91
	Large	−3.75	21.50	−2.96	22.76
	Small	−1.29	12.96	−0.90	14.61
	Micro	−0.05	2.97	0.21	4.25
WLS	Overall	−1.81	−1.96	−0.37	7.20
	Large	−3.40	25.84	−1.00	31.57
	Small	−2.30	13.96	−0.55	19.68
	Micro	−1.62	−7.09	−0.28	2.99
Lasso	Overall	0.00	13.79	0.00	13.81
	Large	0.00	27.03	0.00	26.77
	Small	−0.01	20.01	0.00	20.30
	Micro	0.00	11.61	0.00	11.60
Ridge	Overall	−0.35	5.58	−0.07	6.91
	Large	−3.74	21.49	−2.96	22.76
	Small	−1.28	12.96	−0.90	14.62
	Micro	−0.05	2.98	0.21	4.25
Elastic Net	Overall	0.00	13.81	0.00	13.67
	Large	0.00	27.09	0.00	26.58
	Small	−0.01	20.06	0.00	20.11
	Micro	0.00	11.62	0.00	11.47
NN	Overall	0.11	0.92	0.10	14.75
	Large	0.20	14.27	0.02	38.61
	Small	0.11	16.59	0.09	26.48
	Micro	0.11	−3.11	0.11	10.72
NN-W1	Overall	0.11	9.17	0.10	13.42
	Large	0.10	45.93	0.11	40.72
	Small	0.12	25.90	0.12	25.79
	Micro	0.12	3.22	0.11	9.02
NN-W2	Overall	0.12	9.74	0.12	17.20
	Large	0.18	39.13	0.33	45.80
	Small	0.13	22.14	0.18	30.81
	Micro	0.12	5.18	0.11	12.46
NN-W1W2	Overall	0.07	11.16	0.12	13.96
	Large	0.08	42.32	0.18	40.01
	Small	0.10	24.15	0.14	26.58
	Micro	0.08	6.35	0.13	9.60
NN-J1	Overall	0.05	14.88	0.04	16.15
	Large	0.21	41.99	0.10	39.51
	Small	0.10	29.89	0.05	26.90
	Micro	0.04	9.97	0.04	12.35
NN-J2	Overall	0.09	14.32	0.08	15.72
	Large	0.17	43.50	0.24	41.34
	Small	0.12	28.42	0.12	27.47
	Micro	0.09	9.44	0.07	11.56
NN-J1J2	Overall	0.04	8.15	0.03	13.45
	Large	0.17	43.53	0.15	39.54
	Small	0.07	22.10	0.07	26.41
	Micro	0.04	2.85	0.02	9.02
NN-J1-m	Overall	0.10	16.12	0.09	15.83
	Large	0.28	40.32	0.16	45.01
	Small	0.15	29.44	0.10	29.43
	Micro	0.10	11.76	0.09	11.05
NN-J2-m	Overall	0.07	14.99	0.05	14.03
	Large	0.00	37.06	0.08	35.21
	Small	0.09	26.56	0.07	24.35
	Micro	0.08	11.12	0.06	10.47
NN-J1J2-m	Overall	0.08	14.78	0.06	15.77
	Large	0.21	42.55	0.02	34.19
	Small	0.16	27.79	0.06	27.26
	Micro	0.07	10.22	0.06	12.19

Table G.1:
Out-of-sample performance summary – calendar year, rank-normalisation:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by calendar year, firm characteristics are cross-sectionally rank-normalised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Calendar Year – Rank Normalisation (No Microcaps)			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2	Pred.-R^2	XS-R^2	Pred.-R^2
OLS	Overall	−0.93	8.58	−0.54	8.93
	Large	−1.41	29.15	−0.80	30.81
	Small	−0.69	3.45	−0.40	3.48
WLS	Overall	−2.33	8.08	−0.88	10.38
	Large	−2.79	22.00	−1.31	25.25
	Small	−2.13	4.61	−0.64	6.68
Lasso	Overall	−0.01	8.25	0.00	8.22
	Large	−0.01	18.54	0.00	18.84
	Small	−0.01	5.69	0.00	5.57
Ridge	Overall	−0.96	8.59	−0.54	8.94
	Large	−1.45	29.17	−0.80	30.81
	Small	−0.71	3.47	−0.40	3.49
Elastic Net	Overall	−0.01	8.13	0.00	8.22
	Large	−0.01	18.23	0.00	18.70
	Small	−0.01	5.61	0.00	5.60
NN	Overall	−0.03	8.35	0.00	15.03
	Large	−0.05	30.90	−0.08	35.25
	Small	−0.02	2.73	0.06	9.98
NN-W1	Overall	−0.02	5.86	−0.06	9.74
	Large	−0.05	30.77	−0.14	25.85
	Small	0.01	−0.35	0.00	5.73
NN-W2	Overall	0.03	12.77	0.01	9.10
	Large	0.01	34.71	−0.06	31.59
	Small	0.05	7.31	0.05	3.49
NN-W1W2	Overall	−0.17	−3.86	0.02	12.00
	Large	−0.40	10.77	−0.03	27.86
	Small	−0.04	−7.51	0.05	8.04
NN-J1	Overall	−0.03	14.43	−0.03	13.13
	Large	0.03	31.14	−0.01	30.78
	Small	−0.05	10.26	−0.03	8.73
NN-J2	Overall	0.00	11.44	0.01	12.47
	Large	−0.07	31.89	−0.02	27.51
	Small	0.05	6.34	0.03	8.73
NN-J1J2	Overall	−0.01	14.09	−0.01	9.18
	Large	0.03	35.69	−0.02	23.91
	Small	−0.03	8.71	−0.01	5.51
NN-J1-m	Overall	−0.01	11.94	0.02	10.06
	Large	−0.06	31.57	0.02	24.43
	Small	0.03	7.05	0.03	6.48
NN-J2-m	Overall	0.01	12.61	0.02	13.47
	Large	0.02	30.87	0.00	32.44
	Small	0.02	8.06	0.04	8.75
NN-J1J2-m	Overall	0.01	13.33	0.04	16.44
	Large	−0.01	29.05	0.05	31.58
	Small	0.02	9.41	0.04	12.67

Table G.2:
Out-of-sample performance summary – calendar year, rank-normalisation, no microcaps:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by calendar year, firm characteristics are cross-sectionally rank-normalised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Fiscal Year – Rank Normalisation			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2	Pred.-R^2	XS-R^2	Pred.-R^2
OLS	Overall	−0.27	2.20	0.02	3.81
	Large	−3.15	12.90	−2.35	12.81
	Small	−1.09	7.10	−0.69	8.40
	Micro	−0.03	0.46	0.24	2.26
WLS	Overall	−1.76	−8.02	−0.53	1.88
	Large	−3.34	15.21	−1.78	18.26
	Small	−2.02	6.87	−0.70	12.20
	Micro	−1.51	−12.62	−0.43	−1.33
Lasso	Overall	0.03	12.49	0.03	12.43
	Large	0.01	24.60	0.00	23.94
	Small	−0.01	19.06	−0.01	18.89
	Micro	0.03	10.32	0.04	10.33
Ridge	Overall	−0.27	2.21	0.02	3.82
	Large	−3.14	12.90	−2.35	12.81
	Small	−1.09	7.11	−0.69	8.41
	Micro	−0.03	0.47	0.24	2.26
Elastic Net	Overall	0.03	12.50	0.03	12.44
	Large	0.01	24.71	0.00	24.00
	Small	0.00	19.07	−0.01	18.91
	Micro	0.03	10.33	0.04	10.34
NN	Overall	0.13	2.39	0.08	7.70
	Large	−0.12	21.57	−0.24	24.63
	Small	0.05	11.91	0.02	14.77
	Micro	0.15	−0.87	0.10	5.08
NN-W1	Overall	0.08	3.84	0.09	−1.18
	Large	−0.18	29.78	−0.29	24.52
	Small	0.01	15.34	0.00	9.50
	Micro	0.10	−0.29	0.12	−5.14
NN-W2	Overall	0.04	6.93	0.08	−10.86
	Large	−0.12	30.08	−0.17	−33.96
	Small	0.01	15.80	−0.01	−17.61
	Micro	0.05	3.52	0.11	−7.86
NN-W1W2	Overall	0.07	5.47	0.04	6.15
	Large	−0.30	29.52	−0.14	28.46
	Small	−0.04	17.58	0.00	16.28
	Micro	0.10	1.35	0.06	2.54
NN-J1	Overall	0.04	6.08	0.05	6.84
	Large	−0.37	11.71	−0.05	18.85
	Small	−0.03	11.18	0.02	14.62
	Micro	0.08	4.67	0.06	4.44
NN-J2	Overall	0.07	7.02	0.07	5.04
	Large	−0.16	21.73	−0.19	25.61
	Small	0.01	15.15	0.01	14.23
	Micro	0.09	4.37	0.09	1.75
NN-J1J2	Overall	0.04	8.88	0.05	7.56
	Large	−0.04	30.38	0.09	22.86
	Small	0.00	17.82	0.04	17.99
	Micro	0.06	5.57	0.04	4.40
NN-J1-m	Overall	0.08	4.53	0.06	7.43
	Large	−0.01	22.40	−0.10	24.72
	Small	0.04	12.19	0.01	15.42
	Micro	0.09	1.73	0.07	4.62
NN-J2-m	Overall	0.07	4.39	0.06	4.80
	Large	−0.13	27.30	−0.08	24.14
	Small	0.03	13.90	0.03	11.87
	Micro	0.09	0.87	0.07	2.01
NN-J1J2-m	Overall	0.09	8.31	0.07	7.30
	Large	−0.14	30.31	−0.05	29.20
	Small	0.02	17.62	0.02	17.23
	Micro	0.11	4.89	0.08	3.77

Table G.3:
Out-of-sample performance summary – fiscal year, rank-normalisation:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by fiscal year, firm characteristics are cross-sectionally rank-normalised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Fiscal Year – Rank Normalisation (No Microcaps)			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2	Pred.-R^2	XS-R^2	Pred.-R^2
OLS	Overall	−0.85	9.03	−0.36	7.37
	Large	−1.32	22.56	−0.70	19.39
	Small	−0.60	5.65	−0.16	4.38
WLS	Overall	−2.06	0.98	−0.90	3.67
	Large	−2.72	8.57	−1.15	10.24
	Small	−1.73	−0.91	−0.75	2.03
Lasso	Overall	0.00	9.77	0.00	10.24
	Large	0.01	15.65	0.01	16.64
	Small	0.00	8.30	0.00	8.65
Ridge	Overall	−0.85	9.03	−0.36	7.38
	Large	−1.35	22.57	−0.70	19.39
	Small	−0.60	5.65	−0.16	4.38
Elastic Net	Overall	0.00	9.78	0.00	10.39
	Large	0.01	15.66	0.01	16.88
	Small	0.00	8.31	0.00	8.77
NN	Overall	−0.06	10.62	0.02	6.30
	Large	−0.23	25.81	−0.07	18.85
	Small	0.03	6.84	0.08	3.17
NN-W1	Overall	0.03	3.75	0.01	8.51
	Large	−0.10	20.12	−0.13	26.06
	Small	0.10	−0.34	0.09	4.13
NN-W2	Overall	−0.01	8.06	0.02	5.07
	Large	−0.08	29.32	−0.05	16.00
	Small	0.04	2.76	0.07	2.34
NN-W1W2	Overall	−0.02	3.74	0.03	7.93
	Large	−0.19	22.67	−0.04	19.18
	Small	0.08	−0.98	0.07	5.12
NN-J1	Overall	0.01	9.18	0.01	9.58
	Large	−0.02	21.05	0.01	22.83
	Small	0.03	6.23	0.02	6.28
NN-J2	Overall	0.00	9.54	0.01	5.46
	Large	−0.10	28.63	−0.06	17.15
	Small	0.06	4.79	0.05	2.54
NN-J1J2	Overall	−0.06	10.17	0.04	9.41
	Large	−0.08	29.03	0.03	22.15
	Small	−0.05	5.46	0.04	6.24
NN-J1-m	Overall	0.03	5.98	0.05	7.50
	Large	−0.04	17.34	0.02	22.71
	Small	0.08	3.14	0.07	3.71
NN-J2-m	Overall	0.05	7.91	0.00	9.83
	Large	0.02	23.17	−0.02	23.40
	Small	0.07	4.10	0.01	6.45
NN-J1J2-m	Overall	0.06	9.62	0.03	11.82
	Large	−0.01	26.70	−0.01	22.04
	Small	0.10	5.36	0.05	9.28

Table G.4:
Out-of-sample performance summary – fiscal year, rank-normalisation, no microcaps:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by fiscal year, firm characteristics are cross-sectionally rank-normalised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Calendar Year – Standardisation			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2 [%]	Pred.-R^2 [%]	XS-R^2 [%]	Pred.-R^2 [%]
OLS	Overall	−0.27	5.60	0.05	8.53
	Large	−3.70	26.35	−1.70	26.17
	Small	−1.12	14.03	−0.53	17.82
	Micro	0.00	2.45	0.23	5.43
WLS	Overall	−2.05	−21.71	−0.92	−3.06
	Large	−2.95	33.15	−1.27	35.85
	Small	−1.84	15.74	−0.58	22.22
	Micro	−2.02	−33.03	−0.93	−10.83
Lasso	Overall	0.00	13.81	0.00	14.14
	Large	−0.02	28.12	0.00	31.03
	Small	−0.01	20.53	0.00	22.03
	Micro	0.01	11.46	0.00	11.36
Ridge	Overall	−0.27	5.61	0.06	8.53
	Large	−3.70	26.35	−1.70	26.18
	Small	−1.11	14.03	−0.53	17.82
	Micro	0.00	2.45	0.23	5.43
Elastic Net	Overall	0.00	13.83	0.00	14.19
	Large	−0.02	28.07	0.00	30.97
	Small	−0.01	20.60	0.00	22.10
	Micro	0.01	11.48	0.00	11.43
NN	Overall	0.11	14.79	0.08	13.62
	Large	0.08	41.15	0.16	42.91
	Small	0.10	27.77	0.09	26.78
	Micro	0.12	10.34	0.09	8.92
NN-W1	Overall	0.04	8.29	0.13	10.53
	Large	−0.02	29.63	0.15	37.01
	Small	0.02	19.74	0.12	23.93
	Micro	0.05	4.50	0.14	5.98
NN-W2	Overall	0.06	10.02	0.11	12.44
	Large	−0.05	26.73	0.03	35.92
	Small	0.05	19.72	0.14	23.48
	Micro	0.09	6.90	0.13	8.57
NN-W1W2	Overall	0.09	12.36	0.02	15.61
	Large	0.16	41.00	0.10	43.73
	Small	0.11	27.41	0.03	27.88
	Micro	0.09	7.34	0.02	11.17
NN-J1	Overall	0.00	15.55	0.02	12.31
	Large	0.15	38.21	0.02	38.02
	Small	0.05	27.60	0.02	24.43
	Micro	−0.01	11.54	0.02	8.08
NN-J2	Overall	0.03	15.86	0.08	13.68
	Large	0.04	35.97	0.14	34.31
	Small	0.05	26.23	0.12	23.57
	Micro	0.04	12.36	0.08	10.24
NN-J1J2	Overall	−0.01	13.93	0.04	15.04
	Large	−0.04	36.44	0.01	35.57
	Small	0.02	25.01	0.03	24.94
	Micro	−0.01	10.13	0.04	11.61
NN-J1-m	Overall	0.10	15.61	0.05	14.38
	Large	0.26	43.79	0.18	40.83
	Small	0.14	28.55	0.09	26.50
	Micro	0.10	11.03	0.05	10.09
NN-J2-m	Overall	0.08	12.96	0.11	13.08
	Large	0.15	38.62	0.16	35.44
	Small	0.08	23.23	0.15	24.35
	Micro	0.09	9.08	0.11	9.25
NN-J1J2-m	Overall	0.04	15.23	0.07	15.00
	Large	−0.01	36.73	0.14	41.26
	Small	0.03	26.82	0.11	27.62
	Micro	0.05	11.40	0.07	10.62

Table G.5:
Out-of-sample performance summary – calendar year, standardisation:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by calendar year, firm characteristics are cross-sectionally standardised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Calendar Year – Standardisation (No Microcaps)			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2 [%]	Pred.-R^2 [%]	XS-R^2 [%]	Pred.-R^2 [%]
OLS	Overall	−0.75	9.16	−0.38	10.32
	Large	−1.16	31.86	−0.53	29.95
	Small	−0.52	3.51	−0.31	5.43
WLS	Overall	−2.79	4.26	−0.99	8.53
	Large	−2.91	27.50	−1.17	25.98
	Small	−2.74	−1.54	−0.88	4.18
Lasso	Overall	0.00	9.12	0.00	9.18
	Large	0.00	19.96	0.00	19.77
	Small	0.00	6.42	0.00	6.54
Ridge	Overall	−0.75	9.17	−0.38	10.32
	Large	−1.16	31.86	−0.53	29.95
	Small	−0.52	3.51	−0.31	5.43
Elastic Net	Overall	0.00	9.11	0.00	9.04
	Large	0.00	19.80	0.00	19.50
	Small	0.00	6.45	0.00	6.43
NN	Overall	0.00	11.07	−0.03	1.09
	Large	−0.04	33.71	−0.02	23.23
	Small	0.03	5.42	−0.03	−4.43
NN-W1	Overall	−0.18	1.57	0.02	13.82
	Large	−0.23	19.91	−0.01	30.28
	Small	−0.15	−3.01	0.04	9.71
NN-W2	Overall	−0.10	−11.73	0.03	6.77
	Large	−0.13	3.84	−0.02	24.56
	Small	−0.07	−15.61	0.05	2.33
NN-W1W2	Overall	−0.08	7.31	0.00	−12.18
	Large	−0.12	30.90	−0.04	15.33
	Small	−0.06	1.42	0.02	−19.04
NN-J1	Overall	0.03	9.92	0.00	11.28
	Large	0.02	31.47	0.01	24.51
	Small	0.05	4.55	0.00	7.98
NN-J2	Overall	−0.03	13.32	0.04	7.78
	Large	−0.05	28.11	0.09	26.02
	Small	−0.01	9.63	0.02	3.23
NN-J1J2	Overall	−0.09	9.21	0.01	13.37
	Large	−0.08	28.24	0.03	29.82
	Small	−0.09	4.47	0.01	9.27
NN-J1-m	Overall	−0.01	11.51	0.02	10.42
	Large	−0.05	26.79	−0.01	30.41
	Small	0.01	7.70	0.05	5.44
NN-J2-m	Overall	−0.02	8.99	−0.04	12.00
	Large	0.04	30.16	−0.09	23.81
	Small	−0.04	3.71	0.00	9.06
NN-J1J2-m	Overall	−0.06	7.81	0.02	12.69
	Large	−0.11	26.63	0.01	30.01
	Small	−0.03	3.11	0.03	8.38

Table G.6:
Out-of-sample performance summary – calendar year, standardisation, no microcaps:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by calendar year, firm characteristics are cross-sectionally standardised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

Model	Subgroup	Fiscal Year − Standardisation			
		All Characteristics		Core Characteristics	
		XS-R^2 [%]	Pred.-R^2 [%]	XS-R^2 [%]	Pred.-R^2 [%]
OLS	Overall	−0.12	2.71	0.02	5.25
	Large	−2.92	19.19	−2.10	16.15
	Small	−0.90	9.56	−0.55	11.56
	Micro	0.10	0.18	0.21	3.22
WLS	Overall	−2.49	−30.72	−1.19	−8.84
	Large	−5.09	24.62	−1.01	23.31
	Small	−2.11	8.88	−0.76	15.63
	Micro	−2.32	−42.49	−1.22	−15.97
Lasso	Overall	0.04	11.91	0.04	12.15
	Large	−0.02	24.89	−0.02	23.93
	Small	−0.02	19.71	−0.02	19.67
	Micro	0.05	9.45	0.05	9.82
Ridge	Overall	−0.12	2.72	0.02	5.25
	Large	−2.92	19.19	−2.10	16.16
	Small	−0.89	9.57	−0.55	11.56
	Micro	0.10	0.18	0.21	3.22
Elastic Net	Overall	0.04	11.80	0.04	12.21
	Large	−0.02	24.70	−0.02	24.04
	Small	−0.02	19.64	−0.02	19.71
	Micro	0.05	9.33	0.05	9.89
NN	Overall	0.11	6.95	0.09	5.79
	Large	−0.01	23.27	−0.10	24.27
	Small	0.06	16.78	0.03	16.29
	Micro	0.13	3.84	0.10	2.40
NN-W1	Overall	0.12	4.10	0.15	2.05
	Large	−0.18	24.33	−0.22	26.34
	Small	0.04	14.13	−0.01	15.20
	Micro	0.15	0.67	0.19	−2.29
NN-W2	Overall	0.04	3.40	0.11	6.57
	Large	−0.17	21.54	−0.13	22.50
	Small	0.00	13.80	0.02	15.45
	Micro	0.06	0.04	0.14	3.67
NN-W1W2	Overall	0.07	1.26	0.10	6.45
	Large	−0.14	16.15	−0.12	22.59
	Small	0.03	8.21	0.03	14.51
	Micro	0.08	−1.19	0.12	3.70
NN-J1	Overall	0.03	7.35	0.06	5.43
	Large	−0.09	25.22	0.03	14.64
	Small	−0.01	16.31	0.04	12.40
	Micro	0.04	4.30	0.06	3.40
NN-J2	Overall	0.13	1.31	0.05	6.01
	Large	−0.17	16.54	−0.01	15.08
	Small	0.05	7.24	0.03	12.79
	Micro	0.15	−0.95	0.06	4.03
NN-J1J2	Overall	0.03	3.92	0.03	6.62
	Large	−0.14	17.06	−0.01	21.55
	Small	−0.01	10.23	0.00	13.41
	Micro	0.05	1.73	0.03	4.21
NN-J1-m	Overall	0.06	7.45	0.12	4.74
	Large	−0.15	27.69	−0.19	18.23
	Small	0.00	15.13	−0.01	13.69
	Micro	0.08	4.48	0.15	2.02
NN-J2-m	Overall	0.08	2.12	0.10	8.84
	Large	−0.08	18.70	−0.06	27.85
	Small	0.02	10.50	0.05	17.83
	Micro	0.10	−0.72	0.12	5.69
NN-J1J2-m	Overall	0.08	2.44	0.07	6.55
	Large	−0.28	15.38	−0.13	22.01
	Small	0.01	8.55	0.01	15.08
	Micro	0.11	0.31	0.09	3.76

Table G.7:
Out-of-sample performance summary − fiscal year, standardisation:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by fiscal year, firm characteristics are cross-sectionally standardised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

		Fiscal Year – Standardisation (No Microcaps)			
Model	Subgroup	All Characteristics		Core Characteristics	
		XS-R^2 [%]	Pred.-R^2 [%]	XS-R^2 [%]	Pred.-R^2 [%]
OLS	Overall	−0.72	9.11	−0.31	8.41
	Large	−1.08	28.51	−0.57	19.85
	Small	−0.54	4.27	−0.16	5.56
WLS	Overall	−1.96	−3.31	−1.06	−0.81
	Large	−2.26	21.58	−1.16	9.46
	Small	−1.80	−9.52	−0.99	−3.37
Lasso	Overall	0.00	10.92	0.00	8.96
	Large	0.00	20.32	0.01	14.90
	Small	0.00	8.57	0.00	7.48
Ridge	Overall	−0.72	9.11	−0.31	8.41
	Large	−1.08	28.51	−0.57	19.85
	Small	−0.53	4.28	−0.16	5.56
Elastic Net	Overall	0.00	10.66	0.00	9.07
	Large	0.00	19.11	0.01	15.08
	Small	0.00	8.55	0.00	7.57
NN	Overall	0.01	5.94	0.01	6.41
	Large	−0.12	22.22	−0.12	19.99
	Small	0.09	1.89	0.08	3.03
NN-W1	Overall	−0.12	−0.38	−0.03	3.62
	Large	−0.25	16.41	−0.12	15.44
	Small	−0.04	−4.56	0.02	0.68
NN-W2	Overall	−0.04	−8.56	−0.02	7.55
	Large	−0.14	9.56	−0.06	21.06
	Small	0.02	−13.08	0.01	4.18
NN-W1W2	Overall	−0.12	2.37	0.00	2.04
	Large	−0.27	21.76	−0.10	13.84
	Small	−0.03	−2.46	0.06	−0.90
NN-J1	Overall	0.02	7.53	0.01	6.17
	Large	0.01	21.16	0.01	18.12
	Small	0.02	4.14	0.01	3.19
NN-J2	Overall	−0.01	6.62	0.03	2.82
	Large	−0.10	16.95	−0.03	12.44
	Small	0.05	4.04	0.06	0.42
NN-J1J2	Overall	−0.01	5.53	0.01	7.61
	Large	−0.03	13.97	0.02	17.23
	Small	−0.01	3.43	0.01	5.21
NN-J1-m	Overall	0.01	6.29	0.05	2.25
	Large	−0.05	21.32	0.00	10.60
	Small	0.05	2.54	0.08	0.17
NN-J2-m	Overall	0.04	1.86	0.04	8.87
	Large	0.02	17.03	−0.02	18.11
	Small	0.05	−1.92	0.08	6.57
NN-J1J2-m	Overall	−0.08	7.87	0.03	8.18
	Large	−0.18	22.83	−0.01	22.64
	Small	−0.01	4.14	0.05	4.58

Table G.8:
Out-of-sample performance summary – fiscal year, standardisation, no microcaps:
The table summarises the cross-sectional mean R^2 and predictive R^2 in percentage for all fifteen models under consideration. The left panel displays model performances using all 103 characteristics. The right panel displays model performances using a subset of 49 core characteristics. All models are periodically re-fitted by fiscal year, firm characteristics are cross-sectionally standardised, and $\hat{\lambda}_t$ is estimated on a 5-year backward-looking rolling window basis.

H Variable Importance

Tables H.1-H.4 summarise the most influential firm characteristics analogously to section 4.9. It can be seen that the variable importance for all neural networks remains largely the same, regardless of the data pre-processing (rank-normalisation vs. standardisation). Moreover, the estimated variable importance appears to be robust in the face of the two different reoccurring re-fitting strategies, namely by calendar year and fiscal year. We observe some differences in estimated variable importance when microcaps are excluded. This is not surprising since microcaps make up nearly 60% of all data. While there are minor differences, we still observe the dominance of trading frictions such as momentum and the ability of Jacobian-regularised neural networks to find independent signal in different firm characteristics. Overall, we conclude that our findings are robust.

	Most important characteristics: entire sample					**Most important characteristics: 2016-2020**				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Rank-normalised, by calendar year, all characteristics										
NN	mom1m	mve	mom12m	chfeps	retvol	mve	lev	bm	currat	mom36m
NN-W1	mom1m	std_turn	turn	mve	mom12m	beta	mom12m	turn	roavol	chpmia
NN-W2	mom1m	std_turn	mom12m	maxret	turn	beta	ill	turn	grcapx	retvol
NN-W1W2	mom1m	mom12m	beta	std_turn	retvol	beta	turn	ep	mve	roavol
NN-J1	mom12m	mom1m	cfp	mom36m	sp	mve	cfp	ill	std_dolvol	mom36m
NN-J2	mom1m	mom12m	maxret	chmom	indmom	cfp	chfeps	agr	chpmia	indmom
NN-J1J2	mom12m	maxret	roaq	beta	retvol	mve	ill	retvol	ep	maxret
NN-J1-m	mom1m	mom12m	std_turn	chmom	cfp	cashpr	acc	mom1m	chpmia	lgr
NN-J2-m	mom1m	roavol	mom12m	roaq	maxret	salecash	ep	mom12m	mve	cfp
NN-J1J2-m	mom12m	mom1m	maxret	retvol	beta	cashpr	lev	pchsale_pchinvt	maxret	beta
Rank-normalised, by calendar year, core characteristics										
NN	beta	mom1m	mve	std_turn	mom12m	ep	mom12m	invest	sp	pchsale_pchinvt
NN-W1	mom1m	std_turn	mom12m	turn	retvol	mom36m	beta	roavol	retvol	mom1m
NN-W2	mom12m	mom1m	beta	ill	retvol	roavol	turn	pchsale_pchrect	indmom	ear
NN-W1W2	mom12m	mom1m	std_turn	retvol	mve	mom12m	ep	sgr	gma	currat
NN-J1	std_turn	mom1m	maxret	retvol	mve	lev	roavol	currat	cfp	turn
NN-J2	mom12m	roavol	mom1m	roaq	mve	mom12m	salecash	ep	indmom	roavol
NN-J1J2	mom12m	beta	roavol	mom1m	mve	mve	roavol	agr	maxret	sgr
NN-J1-m	mom12m	mom1m	mom36m	roaq	ill	indmom	mve	dy	beta	pchsale_pchinvt
NN-J2-m	mom12m	std_turn	beta	mom1m	mve	chpmia	beta	mom36m	ill	chinv
NN-J1J2-m	mom1m	beta	mom12m	std_turn	mve	cashpr	indmom	salecash	lgr	ill
Rank-normalised, by calendar year, no microcaps, all characteristics										
NN	mom1m	mom12m	lev	chmom	bm	mom1m	cfp	currat	retvol	mom12m
NN-W1	mom1m	mom12m	mve	bm	turn	mom1m	cashpr	mve	ep	indmom
NN-W2	mom1m	beta	turn	chmom	bm	beta	sgr	lev	std_dolvol	invest
NN-W1W2	mom1m	chmom	turn	mom12m	mve	grltnoa	std_dolvol	chinv	sp	salecash
NN-J1	mom12m	bm	beta	mom1m	lev	pchsale_pchrect	turn	gma	sgr	currat
NN-J2	mom1m	mom12m	beta	turn	bm	beta	cfp	mom1m	saleinv	chatoia
NN-J1J2	turn	maxret	mom1m	mom12m	lev	beta	retvol	turn	roavol	currat
NN-J1-m	mom1m	mom12m	chmom	chpmia	indmom	mom36m	salecash	lev	sgr	std_turn
NN-J2-m	beta	mom12m	mom1m	turn	chfeps	cfp	pchgm_pchsale	saleinv	roeq	cashpr
NN-J1J2-m	mom1m	chmom	beta	turn	lev	gma	cash	mve	mom1m	lgr
Rank-normalised, by calendar year, no microcaps, core characteristics										
NN	mom12m	beta	mom1m	chmom	mve	mom1m	pchgm_pchsale	maxret	ep	mom12m
NN-W1	mom12m	beta	mom1m	turn	chmom	mom1m	orgcap	beta	depr	cfp
NN-W2	mom12m	beta	mom1m	chmom	maxret	mom1m	lev	beta	cashpr	maxret
NN-W1W2	beta	mom12m	mom1m	chmom	retvol	beta	indmom	lev	mom12m	chcsho
NN-J1	maxret	bm	retvol	turn	beta	mve	maxret	cash	sp	roavol
NN-J2	beta	mom1m	chmom	mom36m	orgcap	cfp	mve	beta	orgcap	maxret
NN-J1J2	beta	agr	ep	mom12m	retvol	lev	cashpr	ill	nincr	bm
NN-J1-m	mom12m	beta	chmom	std_turn	mom1m	mom1m	beta	mve	chmom	gma
NN-J2-m	mom12m	beta	mom1m	retvol	chmom	mom1m	retvol	currat	ear	pchgm_pchsale
NN-J1J2-m	mom12m	mom1m	beta	chmom	turn	ep	acc	mom1m	roaq	sp

Table H.1:
Most important characteristics – rank-normalised input, calendar year:
The table summarises the most important characteristics measured in absolute median partial derivatives. The left panel reports the most important characteristics over the entire sample, while the right panel only focuses on the most recent five years. The table refers to all neural networks that use rank-normalised input data and that are trained by calendar year. We consider the case of including and excluding microcaps.

	Most important characteristics: entire sample					**Most important characteristics: 2016-2020**				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Rank-normalised, by fiscal year, all characteristics										
NN	mom1m	mve	std_turn	mom12m	indmom	mom1m	chcsho	mve	retvol	ill
NN-W1	mom1m	mve	std_turn	mom12m	indmom	std_turn	mve	mom12m	mom1m	turn
NN-W2	mom1m	mve	mom12m	std_turn	indmom	acc	indmom	invest	chtx	chcsho
NN-W1W2	mom1m	mve	mom12m	std_turn	indmom	pchsale_pchxsga	mve	indmom	grltnoa	turn
NN-J1	mve	mom1m	std_turn	mom12m	mom36m	maxret	roeq	beta	retvol	turn
NN-J2	mom1m	mve	std_turn	indmom	retvol	roavol	cashpr	saleinv	roaq	turn
NN-J1J2	mve	mom1m	std_turn	bm	roeq	mve	invest	retvol	mom36m	roavol
NN-J1-m	mom1m	mve	chmom	std_turn	beta	beta	chfeps	gma	orgcap	mom1m
NN-J2-m	mom1m	mve	std_turn	indmom	mom12m	mom1m	mve	mom36m	maxret	grltnoa
NN-J1J2-m	mom1m	chmom	cash	mve	mom12m	chpmia	bm	sgr	chtx	orgcap
Rank-normalised, by fiscal year, core characteristics										
NN	mom1m	mve	mom12m	beta	sp	sp	mom36m	acc	chfeps	lev
NN-W1	mom1m	std_turn	mve	turn	mom12m	mom1m	salecash	sp	grltnoa	dy
NN-W2	mom1m	std_turn	turn	ill	sp	chinv	currat	pchgm_pchsale	std_dolvol	sp
NN-W1W2	mom1m	std_turn	mve	mom12m	turn	turn	cashpr	ep	maxret	roeq
NN-J1	mom1m	mve	std_turn	mom12m	roavol	roeq	sp	invest	chtx	maxret
NN-J2	mom1m	std_turn	mve	mom12m	lev	ep	roeq	acc	ill	cfp
NN-J1J2	mom1m	mom12m	std_turn	mve	chmom	sp	mom1m	mom36m	ill	lgr
NN-J1-m	mom1m	std_turn	mve	mom12m	retvol	pchsale_pchrect	mom1m	maxret	mom36m	retvol
NN-J2-m	mom1m	mve	mom12m	std_turn	indmom	salecash	mom36m	turn	cfp	chinv
NN-J1J2-m	mom1m	chmom	mom12m	sp	beta	roavol	std_turn	mom36m	mom1m	ill
Rank-normalised, by fiscal year, no microcaps, all characteristics										
NN	chmom	mom12m	mom1m	indmom	sp	chmom	mom36m	std_dolvol	maxret	mom12m
NN-W1	mom1m	chmom	mom12m	indmom	ill	mom1m	beta	mom36m	retvol	chinv
NN-W2	mom1m	chmom	mom12m	lev	indmom	mom12m	beta	invest	sp	turn
NN-W1W2	chmom	mom1m	mom12m	lev	indmom	mom12m	mom1m	mom36m	indmom	turn
NN-J1	mom12m	indmom	lev	salecash	dy	lev	retvol	salecash	beta	roaq
NN-J2	mom12m	mom1m	chmom	lev	indmom	lgr	ep	cash	bm	chinv
NN-J1J2	beta	cash	lev	mom12m	indmom	lev	mve	agr	turn	mom36m
NN-J1-m	mom1m	mom12m	turn	indmom	chmom	mom1m	saleinv	roaq	chfeps	maxret
NN-J2-m	mom1m	mom12m	indmom	retvol	chmom	mom12m	beta	acc	sp	lev
NN-J1J2-m	mom1m	chmom	beta	mom12m	sp	lev	roavol	pchsale_pchinvt	salecash	acc
Rank-normalised, by fiscal year, no microcaps, core characteristics										
NN	mom12m	chmom	mom1m	indmom	lev	mom1m	mom36m	cash	maxret	roaq
NN-W1	mom12m	mom1m	chmom	indmom	sp	mom1m	mom36m	ill	salecash	lev
NN-W2	chmom	mom1m	mom12m	indmom	sp	mom1m	chpmia	pchsale_pchxsga	invest	currat
NN-W1W2	mom12m	mom1m	indmom	chmom	turn	mom1m	mom12m	turn	chmom	cfp
NN-J1	beta	lev	sp	mom12m	ep	cashpr	acc	retvol	beta	lev
NN-J2	chmom	mom1m	sp	beta	mom36m	cfp	mom12m	nincr	retvol	cashpr
NN-J1J2	mom12m	mom1m	beta	indmom	bm	cfp	mom36m	grcapx	ep	pchsale_pchxsga
NN-J1-m	mom12m	chmom	mom1m	indmom	retvol	gma	beta	chmom	std_turn	mom36m
NN-J2-m	mom12m	beta	indmom	lev	mom1m	turn	mom12m	std_dolvol	beta	indmom
NN-J1J2-m	mom12m	mom1m	beta	lev	sp	retvol	cfp	cash	std_turn	roavol

Table H.2:
Most important characteristics – rank-normalised input, fiscal year:
The table summarises the most important characteristics measured in absolute median partial derivatives. The left panel reports the most important characteristics over the entire sample, while the right panel only focuses on the most recent five years. The table refers to all neural networks that use rank-normalised input data and that are trained by fiscal year. We consider the case of including and excluding microcaps.

	Most important characteristics: entire sample					Most important characteristics: 2016-2020				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Standardised, by calendar year, all characteristics										
NN	mom1m	mve	beta	mom12m	roavol	gma	indmom	std_dolvol	roavol	cfp
NN-W1	mom1m	mom12m	mve	beta	mom36m	beta	gma	mom12m	cfp	indmom
NN-W2	mom1m	mom12m	retvol	beta	chmom	cfp	turn	gma	saleinv	mom12m
NN-W1W2	mom1m	mve	maxret	mom12m	std_turn	mve	currat	bm	gma	mom12m
NN-J1	mve	cash	beta	mom12m	retvol	indmom	ear	chinv	sp	ill
NN-J2	beta	mom1m	indmom	mom12m	lev	retvol	chpmia	ep	orgcap	beta
NN-J1J2	mve	maxret	std_dolvol	retvol	mom1m	mve	retvol	acc	cash	std_dolvol
NN-J1-m	mom1m	mom12m	beta	indmom	chmom	chpmia	roavol	ill	roaq	acc
NN-J2-m	mom1m	mve	lev	mom12m	roavol	salecash	cashpr	lgr	cfp	mve
NN-J1J2-m	mom12m	mom1m	beta	lev	mve	salecash	depr	mom36m	pchsale_pchrect	grcapx
Standardised, by calendar year, core characteristics										
NN	beta	mve	mom1m	mom12m	ep	roavol	bm	lev	grltnoa	pchsale_pchinvt
NN-W1	mom1m	mve	mom12m	beta	std_turn	ep	lev	grcapx	mve	orgcap
NN-W2	mom12m	beta	mom1m	mve	retvol	beta	mom1m	chatoia	saleinv	chpmia
NN-W1W2	beta	mom12m	mom1m	retvol	sp	agr	indmom	chpmia	cash	nincr
NN-J1	retvol	mom1m	turn	roaq	maxret	maxret	ep	pchgm_pchsale	acc	retvol
NN-J2	mom1m	mom12m	beta	mve	retvol	mom1m	cfp	ill	chpmia	chinv
NN-J1J2	beta	mom1m	mve	lev	sp	mve	maxret	bm	beta	mom12m
NN-J1-m	beta	roaq	mom12m	mve	retvol	bm	currat	invest	roeq	grcapx
NN-J2-m	mom12m	mve	lev	roaq	beta	pchsale_pchinvt	roeq	ear	pchsale_pchrect	agr
NN-J1J2-m	mom12m	maxret	retvol	beta	mom1m	chmom	beta	roaq	chfeps	turn
Standardised, by calendar year, no microcaps, all characteristics										
NN	mom1m	chmom	mom12m	cash	turn	cash	cfp	beta	pchsale_pchxsga	pchgm_pchsale
NN-W1	mom1m	chmom	lev	mom12m	beta	mve	lev	indmom	mom1m	chfeps
NN-W2	mom1m	mom12m	beta	maxret	mve	ill	beta	orgcap	retvol	maxret
NN-W1W2	mom1m	mom12m	lev	chmom	bm	mom1m	beta	acc	indmom	roavol
NN-J1	mom1m	beta	retvol	turn	mom12m	cash	beta	chpmia	ill	depr
NN-J2	mom1m	mom12m	chmom	beta	bm	lev	agr	saleinv	dy	sp
NN-J1J2	turn	beta	mom12m	lev	mom1m	std_dolvol	mve	cash	salecash	lev
NN-J1-m	mom1m	mom12m	indmom	retvol	beta	mve	gma	lev	beta	ill
NN-J2-m	mom1m	beta	chmom	mve	ep	beta	pchsale_pchinvt	bm	indmom	mom1m
NN-J1J2-m	mom1m	beta	mve	mom12m	lev	gma	chatoia	lev	beta	depr
Standardised, by calendar year, no microcaps, core characteristics										
NN	beta	lev	chmom	mom12m	mom1m	beta	maxret	currat	ep	bm
NN-W1	beta	mom12m	chmom	lev	mom1m	beta	lev	mom1m	gma	cash
NN-W2	beta	mom12m	mom1m	lev	mve	mom1m	currat	maxret	cash	retvol
NN-W1W2	mom12m	cash	beta	mom1m	mve	ep	gma	beta	mom1m	mom12m
NN-J1	beta	turn	mom12m	dy	roavol	lev	cash	grcapx	gma	cfp
NN-J2	beta	mom12m	mom1m	mve	retvol	beta	acc	std_turn	mom12m	grcapx
NN-J1J2	mom12m	lev	bm	retvol	turn	depr	ear	roaq	lev	bm
NN-J1-m	beta	retvol	mom12m	mom1m	turn	beta	mom1m	pchgm_pchsale	gma	indmom
NN-J2-m	beta	mom12m	mom1m	chmom	retvol	beta	retvol	orgcap	mom1m	gma
NN-J1J2-m	beta	mom12m	retvol	mom1m	turn	beta	retvol	orgcap	mom1m	pchsale_pchinvt

Table H.3:
Most important characteristics – Standardised, calendar year:
The table summarises the most important characteristics measured in absolute median partial derivatives. The left panel reports the most important characteristics over the entire sample, while the right panel only focuses on the most recent five years. The table refers to all neural networks that use standardised input data and that are trained by calendar year. We consider the case of including and excluding microcaps.

	Most important characteristics: entire sample					**Most important characteristics: 2016-2020**				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Standardised, by fiscal year, all characteristics										
NN	mom1m	std_turn	mve	turn	mom12m	ep	cashpr	beta	std_turn	mom1m
NN-W1	mom1m	mve	turn	std_turn	mom12m	sgr	chatoia	ill	mve	lev
NN-W2	mom1m	turn	indmom	lev	mom12m	lev	chmom	ill	maxret	acc
NN-W1W2	mom1m	mve	std_turn	mom12m	indmom	chmom	dy	chtx	mom36m	ear
NN-J1	mom1m	beta	retvol	turn	indmom	turn	ill	sp	roavol	lgr
NN-J2	mom1m	mve	mom12m	maxret	indmom	mom12m	ill	lev	roeq	mve
NN-J1J2	mom1m	beta	mom12m	bm	mve	indmom	maxret	beta	std_dolvol	roeq
NN-J1-m	mom1m	std_turn	mve	indmom	chmom	mve	roavol	cash	grltnoa	retvol
NN-J2-m	mom1m	mve	mom12m	turn	indmom	depr	mom12m	cfp	lev	roeq
NN-J1J2-m	mom1m	mve	indmom	mom12m	lev	turn	mom1m	cfp	mom12m	roavol
Standardised, by fiscal year, core characteristics										
NN	mom1m	cash	indmom	beta	mve	cashpr	beta	maxret	mom1m	chinv
NN-W1	mom1m	chmom	mve	turn	indmom	chmom	mve	ear	mom36m	nincr
NN-W2	mom1m	mom12m	turn	indmom	beta	roeq	saleinv	roavol	ep	std_turn
NN-W1W2	mom1m	std_turn	mom12m	chmom	mve	maxret	std_turn	mom12m	bm	currat
NN-J1	mom1m	beta	lev	mve	bm	roaq	beta	roavol	turn	retvol
NN-J2	mom1m	mom12m	indmom	turn	std_turn	maxret	cfp	mom36m	cashpr	turn
NN-J1J2	beta	mom1m	lev	maxret	mve	ep	turn	mve	acc	sgr
NN-J1-m	mom1m	mve	mom12m	turn	chmom	salecash	beta	roavol	retvol	ill
NN-J2-m	mom1m	retvol	beta	turn	mom12m	sp	cash	mom1m	bm	std_turn
NN-J1J2-m	mom1m	indmom	mom12m	cash	mve	roeq	std_turn	retvol	acc	maxret
Standardised, by fiscal year, no microcaps, all characteristics										
NN	mom1m	chmom	mom12m	mve	lev	mom36m	cfp	maxret	roaq	mom1m
NN-W1	mom12m	chmom	mom1m	mve	indmom	lev	depr	mom12m	salecash	ear
NN-W2	mom1m	chmom	mom12m	cash	indmom	depr	mom1m	currat	retvol	mom36m
NN-W1W2	mom1m	mom12m	chmom	lev	indmom	mom1m	mom12m	gma	cfp	maxret
NN-J1	beta	cash	mom1m	turn	mom12m	cfp	cash	dy	bm	invest
NN-J2	mom1m	beta	chmom	lev	mom12m	mom36m	cash	turn	mom1m	cfp
NN-J1J2	cash	maxret	agr	sp	bm	cfp	cash	ep	ill	mom36m
NN-J1-m	mom12m	chmom	beta	indmom	mom1m	mve	pchsale_pchinvt	bm	agr	mom1m
NN-J2-m	chmom	mom12m	indmom	mom1m	cash	acc	ill	cfp	nincr	mve
NN-J1J2-m	mom12m	mom1m	chmom	lev	indmom	cfp	roaq	grcapx	pchsale_pchxsga	mom36m
Standardised, by fiscal year, no microcaps, core characteristics										
NN	mom1m	mom12m	chmom	indmom	lev	beta	cashpr	cash	mom12m	turn
NN-W1	mom1m	mom12m	chmom	beta	mom36m	mom1m	retvol	beta	mve	cfp
NN-W2	mom1m	chmom	turn	lev	cash	mom1m	bm	cash	mom36m	beta
NN-W1W2	mom12m	chmom	mom1m	lev	beta	cash	beta	maxret	pchsale_pchxsga	mve
NN-J1	cash	mom1m	beta	mom36m	lev	mom36m	mom12m	grcapx	bm	cash
NN-J2	chmom	beta	mom1m	mom12m	cash	maxret	cash	beta	lev	mom12m
NN-J1J2	mom12m	mom1m	chmom	turn	beta	mom1m	saleinv	pchsale_pchxsga	std_turn	cfp
NN-J1-m	mom12m	beta	mom1m	chmom	turn	mom1m	cash	roavol	pchsale_pchinvt	mve
NN-J2-m	beta	mom12m	mom1m	retvol	lev	retvol	cfp	mom1m	cash	beta
NN-J1J2-m	beta	mom12m	indmom	mom1m	cash	mom36m	cfp	roavol	beta	cash

Table H.4:
Most important characteristics – rank-normalised input, fiscal year:
The table summarises the most important characteristics measured in absolute median partial derivatives. The left panel reports the most important characteristics over the entire sample, while the right panel only focuses on the most recent five years. The table refers to all neural networks that use rank-normalised input data and that are trained by fiscal year. We consider the case of including and excluding microcaps.

In addition, we exemplarily report the time-varying dimensionality reduction analogously to section 4.10 for neural networks trained on all and the core characteristics, using the cross-sectional rank-normalised data, where the models are re-fitted by calendar year. For clarity, we do not report all results for all models, re-fitting regimes or data-preprocessing. Further results can get requested from the authors. Figures H.1 and H.2 show that the dimensionality reduction varies considerably over time, with the element-wise L_1 Jacobian regularisation yielding the strongest dimensionality reduction.

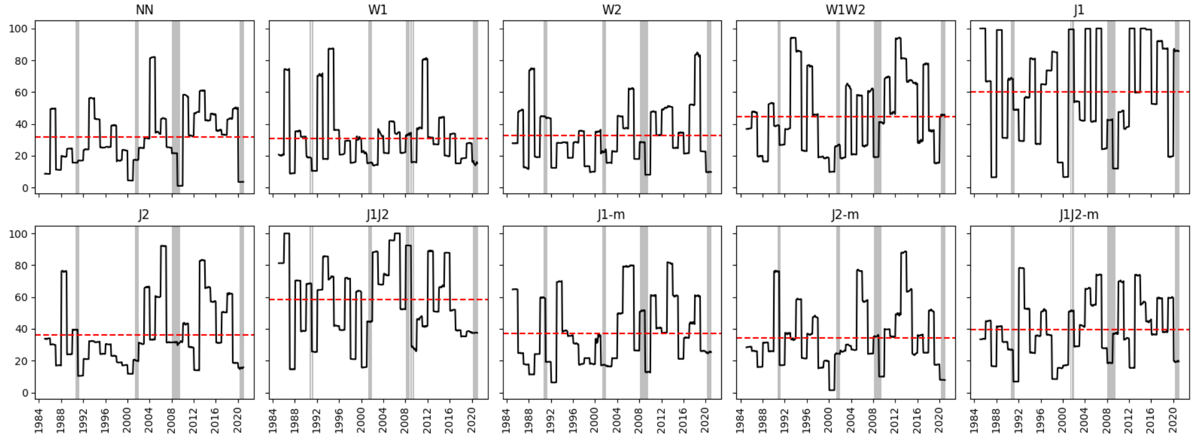


Figure H.1:
Time-varying dimensionality reduction – all characteristics
The graph shows the time-varying dimensionality reduction for all ten neural networks under consideration. The neural networks are trained on all 103 firm characteristics, re-fitted by calendar year, with cross-sectionally rank-normalised data.

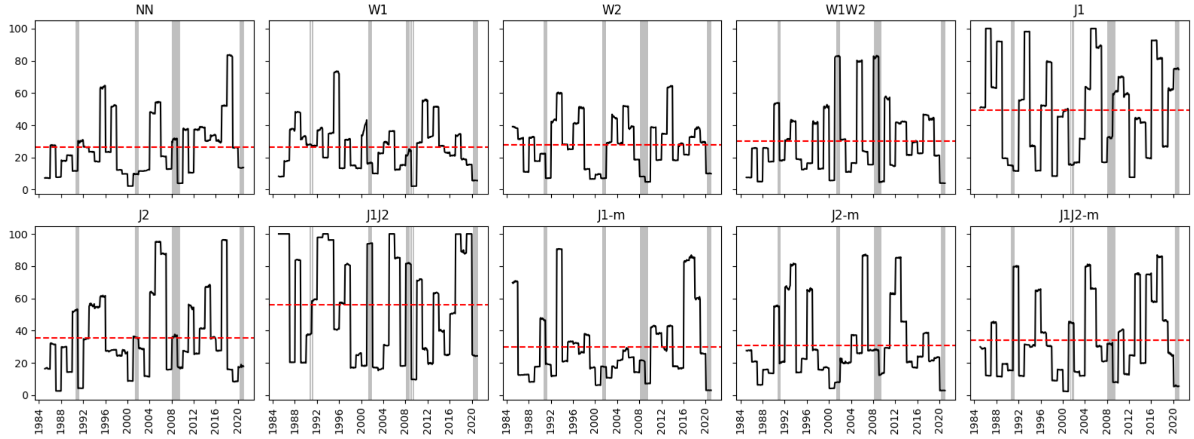


Figure H.2:
Time-varying dimensionality reduction - core characteristics
The graph shows the time-varying dimensionality reduction for all ten neural networks under consideration. The neural networks are trained on the 49 core firm characteristics, re-fitted by calendar year, with cross-sectionally rank-normalised data.

Last but not least, figures H.3 and H.4 exemplarily show the time-varying variable importance displayed as the rank, where a lower rank indicates higher importance. Due to the annual re-fitting, the variable importance changes constantly over time. For clarity, we do not report all time-varying variable importances for all models, data-preprocessing and re-fitting regimes. Further results can be requested from the authors.

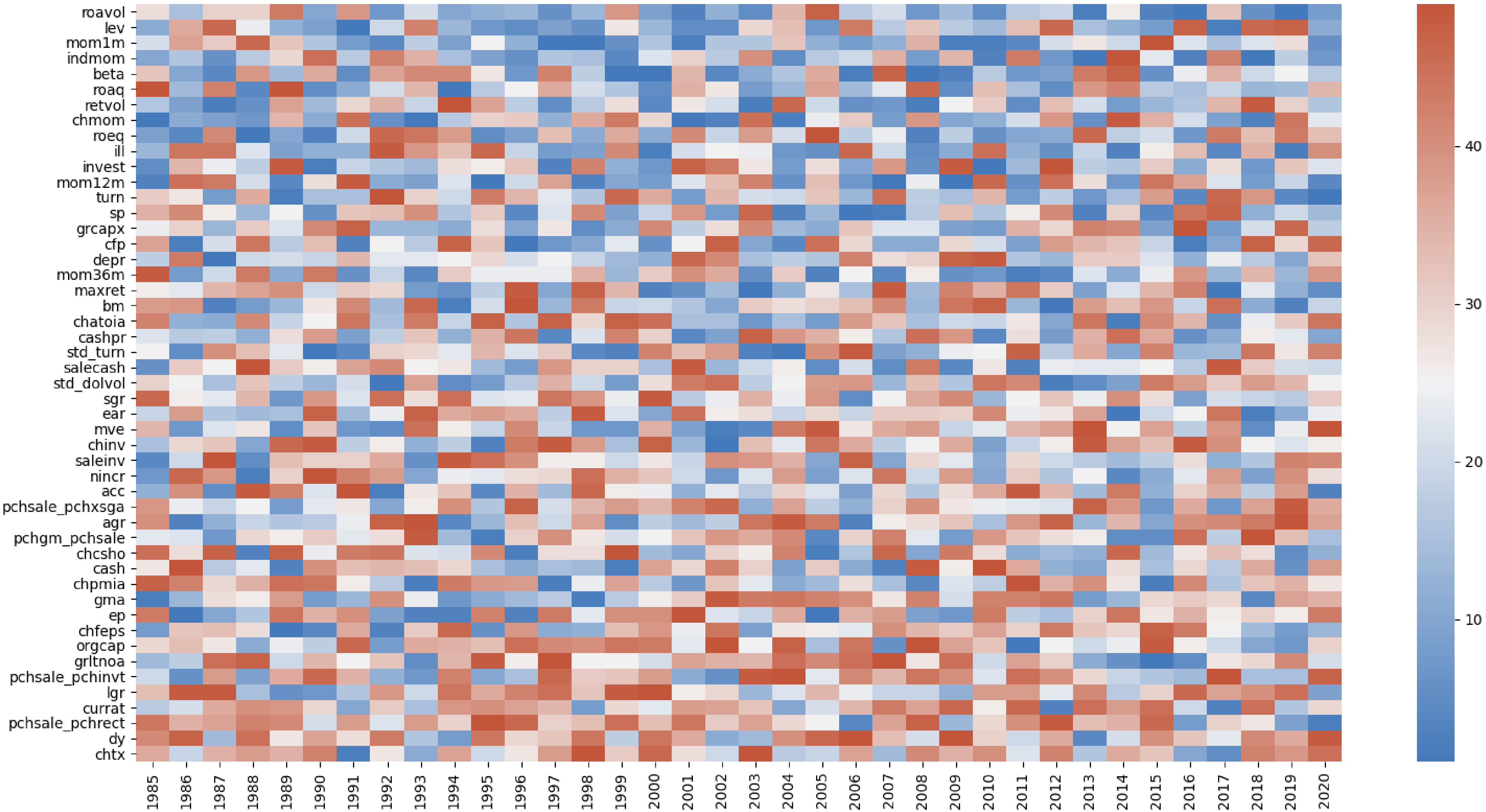


Figure H.3:
Time-varying variable importance – NN-W2
The graph shows the variable importance measured in time-varying rank, estimated by calendar year for the neural networks with L_2 norm weight regularisation in their objective function, and which were trained by calendar year on the core characteristics only. A low rank indicates empirical importance, while a high rank indicates less empirical importance. Due to the annual re-fitting, their architectural freedom and nonlinearity.

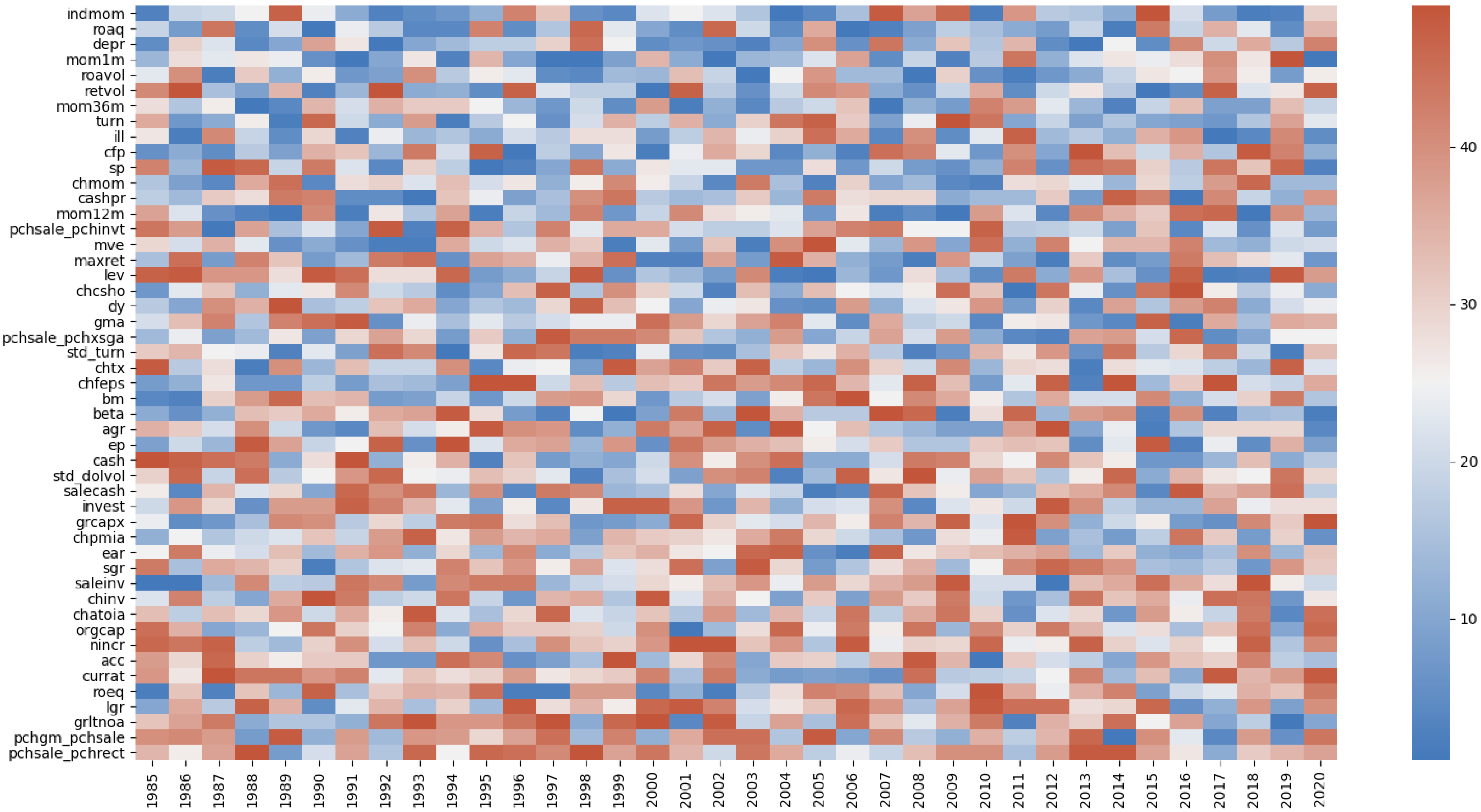


Figure H.4:
Time-varying variable importance – NN-J1-m
The graph shows the variable importance measured in time-varying rank, estimated by calendar year for the neural networks with columns-wise L_1 norm Jacobian regularisation in their objective function, and which were trained by calendar year on the core characteristics only. A low rank indicates empirical importance, while a high rank indicates less empirical importance. Due to the annual re-fitting, their architectural freedom and nonlinearity.

I Risk Prices

This appendix provides further empirical results about the estimated risk prices. Table I.1 reports the linear estimations, analogously to Green et al. (2017). However, the table differs from the original paper, in that the linear models are re-fitted annually to make the results directly comparable to those estimated by the neural networks. Further, we differentiate between the all and core characteristics case. For clarity, table merely reports the linearly estimated risk prices by OLS and WLS for the rank-normalised data (including and excluding microcaps), re-fitted by calendar year, further differentiating between the all and core characteristics case. Further empirical results can be requested from the authors.

In addition to the linear risk price estimates, which serve as a sanity-check or benchmark, figure I.1 visually summarises the nonlinear risk price estimations with empirical tolerance bands for NN-J1 and NN. It can be seen that, as expected, the estimations by NN-J1 are much more restricting than those by NN, as most risk premia are close to zero. We conclude that NN-J1 is potentially too harsh for economically meaningful risk premia estimations. For clarity we refrain from reporting all empirical results. Further empirical details can be requested from the authors.

Table I.1:
Linear risk price estimation

We estimate risk prices linearly, analogously to Green et al. (2017), but *translate* the linear Fama-Macbeth regressions into the setting of this paper to make the results directly comparable to those derived from neural networks. The table report the risk prices as a time-series average from annual re-fitting (scaled by 100), where the models are re-fitted by calendar year. The t-statistics are taken from the time-series of annual coefficient estimates and employ Newey-West adjustments of 12 lags. The sample is cross-sectionally rank-normalised and we consider the inclusion and exclusion of microcaps.

	(A) Rank-normalised, all characteristics				(A) Rank-normalised, core characteristics				(A) Rank-normalised, no micro., all characteristics				(A) Rank-normalised, no micro., core characteristics			
	OLS		WLS		OLS		WLS		OLS		WLS		OLS		WLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
acc	0.03	1.4	−0.04	−0.5	−0.02	−0.7	−0.13	−2.9	0.00	−0.1	0.03	0.6	−0.05	−2.2	−0.05	−1.7
agr	−0.24	−4.5	−0.12	−2.2	−0.33	−3.9	−0.14	−1.6	−0.07	−1.7	−0.16	−3.6	−0.14	−2.1	−0.17	−3.3
beta	0.16	0.7	0.13	0.2	0.08	1.3	0.01	0.1	1.46	1.3	0.89	0.6	0.03	0.6	0.05	0.8
bm	0.11	1.2	0.22	1.5	0.07	1.0	0.03	0.3	0.15	1.5	0.15	1.4	0.05	0.8	−0.01	−0.2
cash	0.23	3.0	0.19	3.5	0.28	3.4	0.22	4.7	0.16	5.1	0.17	3.4	0.22	4.7	0.19	4.6
cashpr	0.00	−0.1	−0.07	−2.3	−0.02	−0.5	−0.09	−2.2	−0.02	−1.0	−0.10	−3.0	−0.02	−1.0	−0.12	−2.7
cfp	−0.11	−1.8	−0.17	−2.7	0.05	1.3	−0.10	−2.3	−0.11	−1.6	−0.05	−0.6	−0.03	−0.5	−0.04	−1.0
chatoia	0.02	1.1	0.05	1.0	0.02	1.1	0.05	1.3	0.02	1.2	0.04	0.8	0.01	0.8	0.04	1.1
chcsho	0.01	0.5	0.02	1.0	−0.02	−1.2	−0.02	−0.6	0.01	0.6	0.01	0.4	−0.02	−1.3	−0.04	−1.2
chfeps	0.23	3.6	−0.04	−0.8	0.23	3.7	−0.04	−0.8	0.12	2.4	−0.04	−0.8	0.09	1.7	−0.04	−0.8
chinv	−0.02	−0.8	−0.01	−0.4	−0.03	−1.1	0.00	0.1	0.00	−0.1	−0.02	−0.5	−0.01	−0.4	0.01	0.5
chmom	0.17	3.6	−0.58	−3.2	−0.17	−4.9	−0.32	−3.6	−0.11	−0.9	−0.29	−2.0	−0.20	−3.8	−0.27	−3.3
chpmia	0.03	0.8	−0.02	−0.7	0.03	0.6	−0.03	−1.0	0.01	0.5	−0.02	−0.9	0.01	0.5	−0.03	−1.2
chtx	0.04	1.7	0.03	0.8	0.06	2.5	0.04	1.2	−0.04	−1.2	0.00	0.1	−0.04	−1.0	0.01	0.4
currat	0.10	1.4	0.03	0.4	0.03	0.5	−0.06	−3.7	−0.02	−0.6	0.09	1.4	−0.01	−0.6	−0.05	−2.4
depr	0.06	1.6	0.09	2.3	0.07	2.1	0.03	0.7	0.04	1.4	0.07	2.5	0.04	1.5	0.01	0.4
dy	−0.08	−2.7	−0.08	−2.3	−0.10	−3.7	−0.20	−4.0	−0.09	−3.4	−0.09	−2.8	−0.13	−3.2	−0.19	−3.9
ear	0.11	12.5	0.12	3.6	0.12	16.6	0.11	3.7	0.09	5.3	0.11	3.8	0.07	5.6	0.10	3.5
ep	0.06	1.6	0.26	4.6	0.00	0.1	0.20	3.3	0.11	3.7	0.19	3.7	0.01	0.3	0.13	2.4
gma	0.02	1.0	0.04	1.9	0.11	5.3	0.16	2.8	0.03	1.0	0.07	2.4	0.11	4.2	0.17	2.7
grcapx	0.00	−0.1	−0.04	−2.0	−0.03	−1.6	−0.05	−1.7	−0.03	−2.3	−0.04	−2.2	−0.06	−3.9	−0.04	−1.9
grltnoa	−0.02	−1.2	−0.05	−1.7	−0.01	−0.5	−0.05	−1.8	−0.04	−1.7	−0.01	−0.3	−0.03	−1.5	−0.01	−0.7

(continued)

Table I.1:
Linear risk price estimation
We estimate risk prices linearly, analogously to Green et al. (2017), but *translate* the linear Fama-Macbeth regressions into the setting of this paper to make the results directly comparable to those derived from neural networks. The table report the risk prices as a time-series average from annual re-fitting (scaled by 100), where the models are re-fitted by calendar year. The t-statistics are taken from the time-series of annual coefficient estimates and employ Newey-West adjustments of 12 lags. The sample is cross-sectionally rank-normalised and we consider the inclusion and exclusion of microcaps.

	(A) Rank-normalised, all characteristics				(A) Rank-normalised, core characteristics				(A) Rank-normalised, no micro., all characteristics				(A) Rank-normalised, no micro., core characteristics			
	OLS		WLS		OLS		WLS		OLS		WLS		OLS		WLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
ill	−0.58	−3.5	0.19	0.7	−0.96	−5.8	0.07	0.3	−0.42	−2.8	0.42	2.0	−0.38	−4.3	0.18	1.7
indmom	0.10	1.6	−0.08	−3.9	0.14	2.1	−0.07	−3.8	−0.07	−2.2	−0.12	−3.0	−0.04	−1.3	−0.12	−3.3
invest	−0.03	−1.0	0.08	1.2	0.01	0.5	0.08	1.8	−0.03	−1.0	0.04	0.8	−0.01	−0.2	0.05	1.3
lev	0.19	2.2	0.24	1.7	0.10	1.7	0.03	0.4	0.25	2.3	0.21	1.7	0.15	2.1	0.06	0.8
lgr	0.03	1.0	0.08	1.8	0.07	2.7	0.10	2.6	0.04	2.0	0.12	2.9	0.06	2.4	0.13	3.9
maxret	0.19	3.1	0.04	0.3	0.15	2.6	0.06	0.5	0.08	1.6	−0.01	−0.1	0.05	1.0	0.01	0.2
mom12m	0.13	2.0	−0.48	−2.7	−0.19	−2.8	0.00	0.0	0.00	0.0	−0.19	−1.5	−0.03	−0.4	0.02	0.3
mom1m	−0.75	−5.3	−0.45	−3.5	−0.66	−4.4	−0.38	−3.4	−0.36	−10.7	−0.45	−3.7	−0.25	−7.2	−0.32	−2.8
mom36m	−0.10	−1.3	−0.14	−1.2	−0.13	−1.4	−0.12	−1.1	−0.03	−0.4	−0.11	−1.0	−0.05	−0.6	−0.09	−0.8
mve	−2.22	−7.1	0.52	3.8	−1.38	−6.4	−0.05	−0.5	−0.32	−3.6	0.38	2.9	−0.40	−3.5	0.05	0.9
nincr	0.02	0.8	0.06	1.8	0.04	1.6	0.06	1.9	0.05	3.1	0.06	1.8	0.06	4.6	0.06	2.1
orgcap	−0.04	−2.3	0.00	−0.2	−0.09	−1.6	−0.01	−0.6	−0.03	−1.7	−0.02	−0.7	0.00	0.0	−0.03	−1.6
pchgm_pchsale	0.03	1.5	0.02	0.5	0.06	3.4	0.03	0.5	0.02	0.9	0.01	0.3	0.03	1.2	0.01	0.2
pchsale_pchinvt	−0.01	−0.3	0.02	0.6	−0.01	−0.2	0.04	1.1	0.00	0.3	0.03	1.1	0.00	0.1	0.05	1.6
pchsale_pchrect	0.00	−0.2	0.01	0.2	0.01	1.5	0.03	1.6	−0.01	−0.7	0.03	1.1	0.00	0.3	0.04	2.0
pchsale_pchxsga	0.00	0.3	−0.04	−1.5	0.00	0.1	−0.06	−2.0	−0.02	−2.9	−0.04	−1.3	−0.03	−3.9	−0.05	−1.8
retvol	−0.49	−5.6	−0.33	−1.4	−0.30	−5.5	−0.46	−3.1	−0.30	−2.1	−0.24	−1.7	−0.18	−3.0	−0.36	−3.7
roaq	0.11	3.0	−0.02	−0.3	0.17	3.4	0.01	0.2	0.17	3.4	0.02	0.3	0.13	1.8	0.02	0.3
roavol	−0.14	−2.4	−0.11	−2.0	−0.18	−4.5	−0.15	−3.6	−0.03	−0.7	−0.07	−1.2	−0.06	−2.0	−0.11	−2.6
roeq	0.10	1.7	0.07	1.3	0.14	2.2	0.07	2.1	0.04	1.6	0.08	2.5	0.04	2.3	0.06	2.1
salecash	0.12	2.0	0.02	0.2	0.11	1.9	0.06	1.1	0.08	2.2	0.02	0.4	0.08	2.2	0.07	1.3
saleinv	0.11	4.7	0.03	0.7	0.09	5.0	0.01	0.3	0.02	1.3	0.03	0.7	0.02	1.5	0.00	0.1
sgr	−0.06	−2.3	−0.06	−1.1	−0.05	−1.8	−0.07	−1.3	0.02	0.5	−0.01	−0.4	0.03	0.6	−0.04	−1.1

(continued)

Table I.1:
Linear risk price estimation
We estimate risk prices linearly, analogously to Green et al. (2017), but *translate* the linear Fama-Macbeth regressions into the setting of this paper to make the results directly comparable to those derived from neural networks. The table report the risk prices as a time-series average from annual re-fitting (scaled by 100), where the models are re-fitted by calendar year. The t-statistics are taken from the time-series of annual coefficient estimates and employ Newey-West adjustments of 12 lags. The sample is cross-sectionally rank-normalised and we consider the inclusion and exclusion of microcaps.

	(A) Rank-normalised, all characteristics				(A) Rank-normalised, core characteristics				(A) Rank-normalised, no micro., all characteristics				(A) Rank-normalised, no micro., core characteristics			
	OLS		WLS		OLS		WLS		OLS		WLS		OLS		WLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
sp	−0.10	−1.9	−0.10	−1.0	−0.07	−0.8	−0.07	−1.0	−0.04	−0.7	−0.13	−1.1	−0.02	−0.5	−0.10	−1.5
std_dolvol	−0.40	−2.5	−0.10	−0.6	−0.48	−2.7	−0.19	−1.6	−0.05	−0.9	−0.15	−1.7	−0.10	−1.2	−0.24	−2.6
std_turn	0.94	3.9	−0.15	−0.7	1.02	3.7	0.25	2.0	0.11	0.9	−0.01	−0.1	0.32	1.6	0.39	3.8
turn	−1.79	−5.2	−0.03	−0.2	−1.20	−6.4	0.04	0.4	−0.40	−1.7	−0.30	−1.5	−0.45	−2.4	−0.10	−1.1
absacc	−0.07	−2.5	−0.02	−0.3					−0.04	−1.7	−0.01	−0.2				
aeavol	0.02	1.7	−0.06	−1.5					−0.03	−2.4	−0.05	−2.3				
age	0.02	0.8	−0.13	−4.2					−0.03	−1.3	−0.13	−4.2				
baspread	0.10	1.3	−0.25	−1.4					0.16	1.6	−0.30	−2.7				
betasq	−0.08	−0.4	−0.21	−0.4					−1.43	−1.3	−0.91	−0.6				
bm_ia	−0.03	−0.8	−0.13	−1.3					−0.07	−1.6	−0.09	−1.3				
cashdebt	0.00	0.0	0.03	0.7					−0.02	−0.7	−0.01	−0.3				
cfp_ia	0.13	3.6	0.05	0.9					0.08	2.6	0.03	0.7				
chempia	0.04	1.0	0.07	1.7					0.05	1.3	0.05	1.5				
chnanalyst	0.02	1.3	0.00	−0.1					0.01	0.7	−0.01	−0.4				
cinvest	−0.01	−0.9	0.00	0.1					−0.04	−2.8	−0.01	−0.2				
convind	−0.07	−2.7	−0.04	−1.0					−0.04	−1.2	−0.05	−1.4				
disp	−0.12	−2.2	−0.03	−0.6					−0.04	−1.5	−0.03	−0.5				
divi	−0.06	−2.4	0.00	0.1					−0.01	−0.4	0.03	0.3				
divo	−0.03	−0.5	−0.07	−1.3					0.01	0.3	−0.06	−1.4				
dolvol	0.98	3.9	−0.58	−3.6					−0.36	−2.3	−0.24	−1.1				
egr	−0.03	−1.3	−0.04	−1.2					−0.06	−1.5	−0.04	−1.1				
fgr5yr	0.00	0.1	0.03	0.6					0.00	0.0	−0.01	−0.1				
herf	−0.05	−2.1	0.01	0.4					−0.01	−0.3	0.01	0.5				

(continued)

Table I.1:
Linear risk price estimation
We estimate risk prices linearly, analogously to Green et al. (2017), but *translate* the linear Fama-Macbeth regressions into the setting of this paper to make the results directly comparable to those derived from neural networks. The table report the risk prices as a time-series average from annual re-fitting (scaled by 100), where the models are re-fitted by calendar year. The t-statistics are taken from the time-series of annual coefficient estimates and employ Newey-West adjustments of 12 lags. The sample is cross-sectionally rank-normalised and we consider the inclusion and exclusion of microcaps.

	(A) Rank-normalised, all characteristics				(A) Rank-normalised, core characteristics				(A) Rank-normalised, no micro., all characteristics				(A) Rank-normalised, no micro., core characteristics			
	OLS		WLS		OLS		WLS		OLS		WLS		OLS		WLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
hire	−0.01	−0.4	−0.07	−1.6					−0.03	−0.8	−0.06	−1.5				
idiovol	−0.03	−0.4	0.10	1.3					−0.01	−0.3	0.16	3.0				
ipo	−0.07	−1.1	−0.14	−2.0					−0.07	−1.3	−0.15	−2.0				
mom6m	−0.52	−7.0	0.35	1.5					−0.13	−1.0	0.03	0.1				
ms	0.06	1.6	0.09	2.0					0.03	0.9	0.07	1.8				
mve_ia	0.09	1.4	0.04	1.6					−0.01	−0.3	0.06	2.9				
nanalyst	0.04	0.6	−0.01	−0.2					0.17	4.7	0.02	0.6				
operprof	0.12	6.0	0.02	0.6					0.05	2.9	0.02	0.6				
pchcapx_ia	−0.04	−2.1	−0.07	−2.4					−0.04	−2.2	−0.08	−4.3				
pchcurrat	0.01	0.2	0.00	0.0					−0.02	−0.6	−0.04	−0.6				
pchdepr	−0.01	−0.6	0.00	−0.1					0.01	0.6	−0.02	−0.8				
pchquick	−0.05	−0.9	0.02	0.3					0.00	−0.1	0.04	0.9				
pchsaleinv	−0.02	−1.1	0.01	0.5					−0.02	−1.8	−0.01	−0.6				
pctacc	−0.11	−3.8	−0.09	−1.3					−0.08	−1.1	−0.05	−0.6				
pricedelay	0.03	1.9	0.01	0.4					0.00	0.3	−0.01	−0.4				
ps	0.00	0.0	0.01	0.8					0.03	1.0	0.01	1.0				
quick	−0.04	−0.8	−0.08	−0.9					0.05	1.1	−0.12	−1.6				
rd	0.02	1.2	0.01	0.6					0.01	0.3	0.00	0.2				
rd_mve	0.27	5.6	0.13	2.2					0.18	4.4	0.04	1.1				
rd_sale	−0.16	−3.1	−0.11	−1.6					−0.10	−2.4	−0.01	−0.4				
realestate	0.04	1.9	0.07	2.6					0.03	1.0	0.05	1.8				
roic	0.00	0.0	0.14	2.1					0.08	1.1	0.10	2.8				
rsup	0.09	4.4	0.05	1.4					0.06	2.7	0.04	1.2				

(continued)

Table I.1:
Linear risk price estimation
We estimate risk prices linearly, analogously to Green et al. (2017), but *translate* the linear Fama-Macbeth regressions into the setting of this paper to make the results directly comparable to those derived from neural networks. The table report the risk prices as a time-series average from annual re-fitting (scaled by 100), where the models are re-fitted by calendar year. The t-statistics are taken from the time-series of annual coefficient estimates and employ Newey-West adjustments of 12 lags. The sample is cross-sectionally rank-normalised and we consider the inclusion and exclusion of microcaps.

	(A) Rank-normalised, all characteristics				(A) Rank-normalised, core characteristics				(A) Rank-normalised, no micro., all characteristics				(A) Rank-normalised, no micro., core characteristics			
	OLS		WLS		OLS		WLS		OLS		WLS		OLS		WLS	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t	β	t
salerec	0.07	2.6	0.04	1.6					0.05	1.4	0.02	1.1				
secured	−0.05	−3.3	0.00	−0.2					−0.01	−0.3	−0.02	−2.4				
securedind	−0.02	−1.2	0.03	1.4					−0.03	−1.8	0.05	3.0				
sfe	−0.16	−1.6	−0.48	−6.9					−0.43	−4.4	−0.44	−8.6				
sgrvol	−0.12	−2.3	−0.05	−1.7					−0.02	−1.0	−0.02	−0.7				
sin	0.19	7.7	0.10	1.8					0.11	2.7	0.10	1.7				
stdacc	−0.13	−3.1	−0.08	−1.6					−0.17	−3.9	−0.09	−2.3				
stdcf	0.05	1.0	0.01	0.2					0.11	3.2	0.03	0.5				
sue	0.15	5.1	0.08	1.0					0.02	0.8	0.01	0.3				
tang	0.05	1.8	0.04	1.1					0.03	1.4	0.01	0.4				
tb	0.05	2.2	0.02	0.6					0.02	1.9	0.01	0.2				
zerotrade	−0.56	−2.5	−0.92	−3.0					−0.29	−1.4	−0.84	−2.6				

(continued)

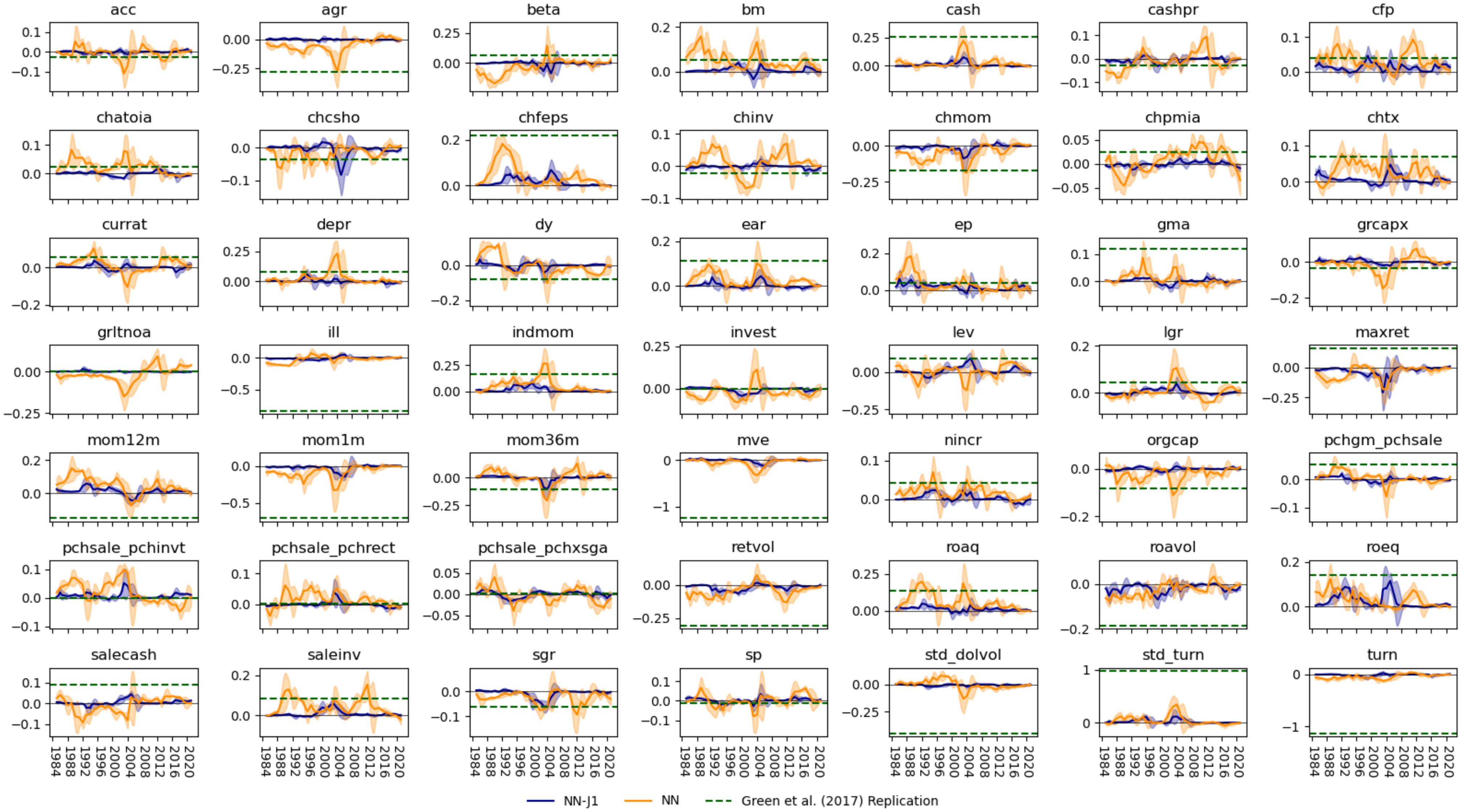


Figure I.1:
Time-varying risk premia, by size class – NN-J1, NN, core characteristics
The graph plots the time-varying risk premia estimations for NN-J1 (blue) and NN (orange), where the dotted green line represents the analogous linear estimation as a benchmark or sanity check.

J Model Insights

In this appendix, we provide further empirical results in addition to the results presented in section 4.12. Figures J.1 and J.2 reveal that the nonlinear sensitivity interactions discussed in section 4.12 not only vary by market capitalisation, but can also be estimated by industry. The figures show that manufacturing constitutes the industry with the largest number of assets. While the sensitivity interactions seem to be relatively homogenous across industry, some differences become apparent, for example, in the case of 1-month momentum and cash holdings. The exemplary display of figures **??** underline the possibility to analyse nonlinear interactions on industry level.

Moreover, figures J.4 to J.10 display the analogous insight to section 4.12, but for all remaining models, namely NN, NN-W1, NN-W1W2, NN-J1, NN-J1J2, NN-J2-m, NN-J1J2-m. It can be seen that the nonlinear interactions in part depend on the objective function.

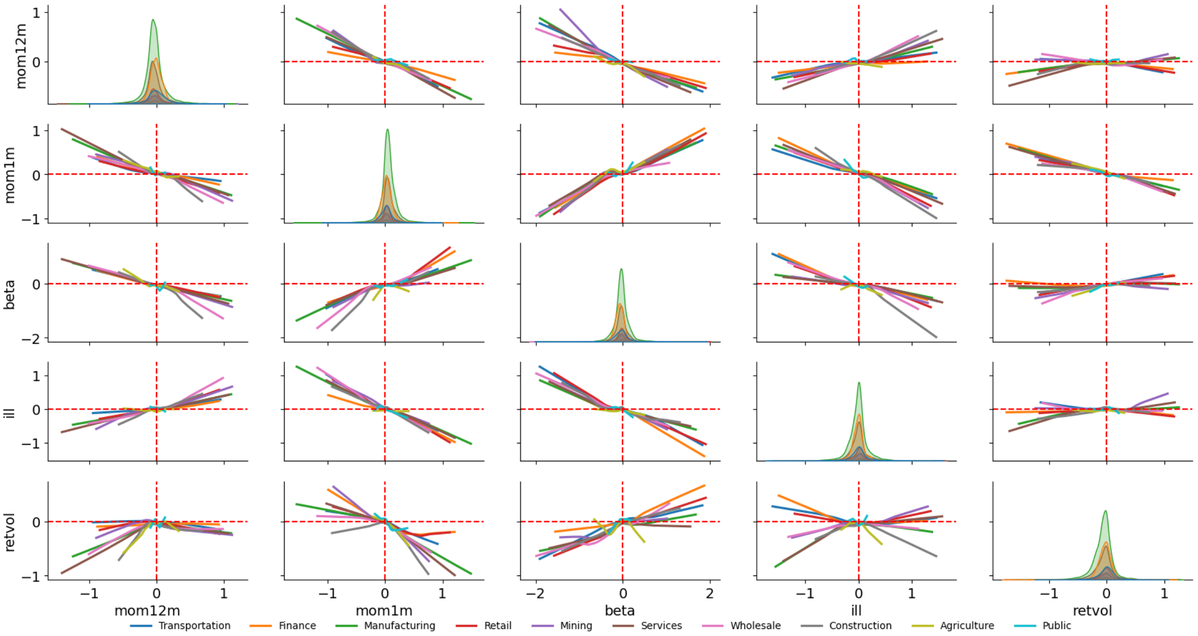


Figure J.1:
Nonlinear sensitivity interactions NN-W2, by industry – exemplary year 2006:
The graph visualises in the off-diagonals how the neural network's return sensitivities to changes in firm characteristics vary nonlinearly across assets, given return sensitivities to changes in other firm characteristics. The nonlinear sensitivity interactions are visualised by industry. The diagonal displays the distribution of sensitivities by industry. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-W2, analogously to table 4. The nonlinear estimations are estimated using data from the year 1985 rather than one month to avoid the curse of dimensionality.

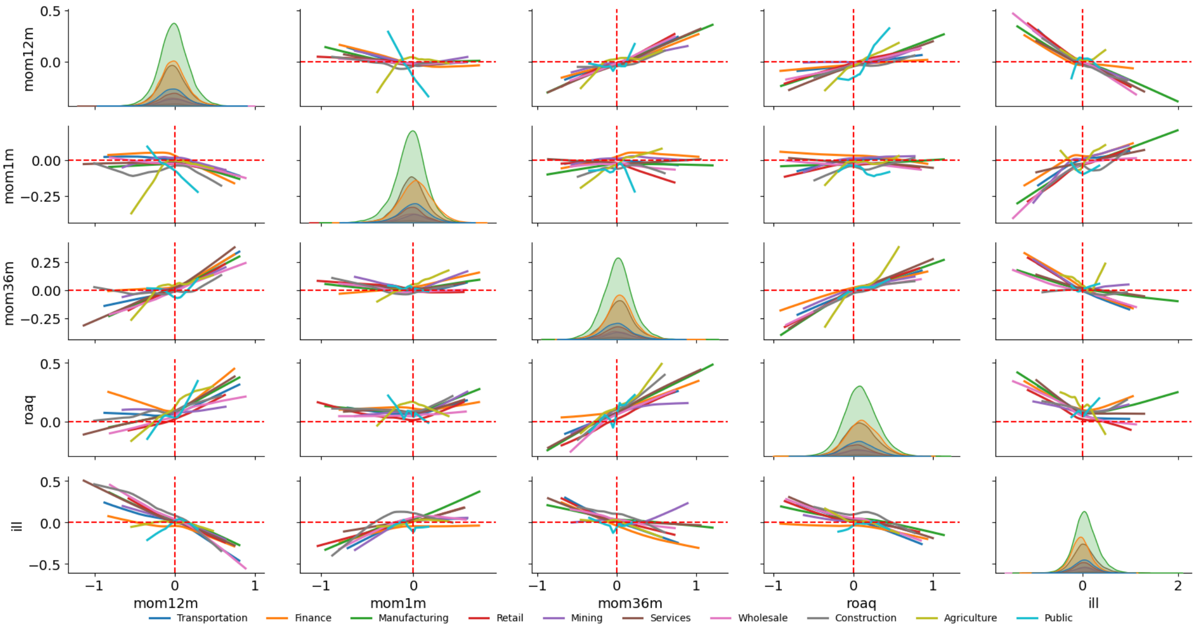


Figure J.2:
Nonlinear sensitivity interactions NN-J1-m, by industry – exemplary year 2006:
The graph visualises in the off-diagonals how the neural network's return sensitivities to changes in firm characteristics vary nonlinearly across assets, given return sensitivities to changes in other firm characteristics. The nonlinear sensitivity interactions are visualised by industry. The diagonal displays the distribution of sensitivities by industry. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J1-m, analogously to table 4. The nonlinear estimations are estimated using data from the year 1985 rather than one month to avoid the curse of dimensionality.

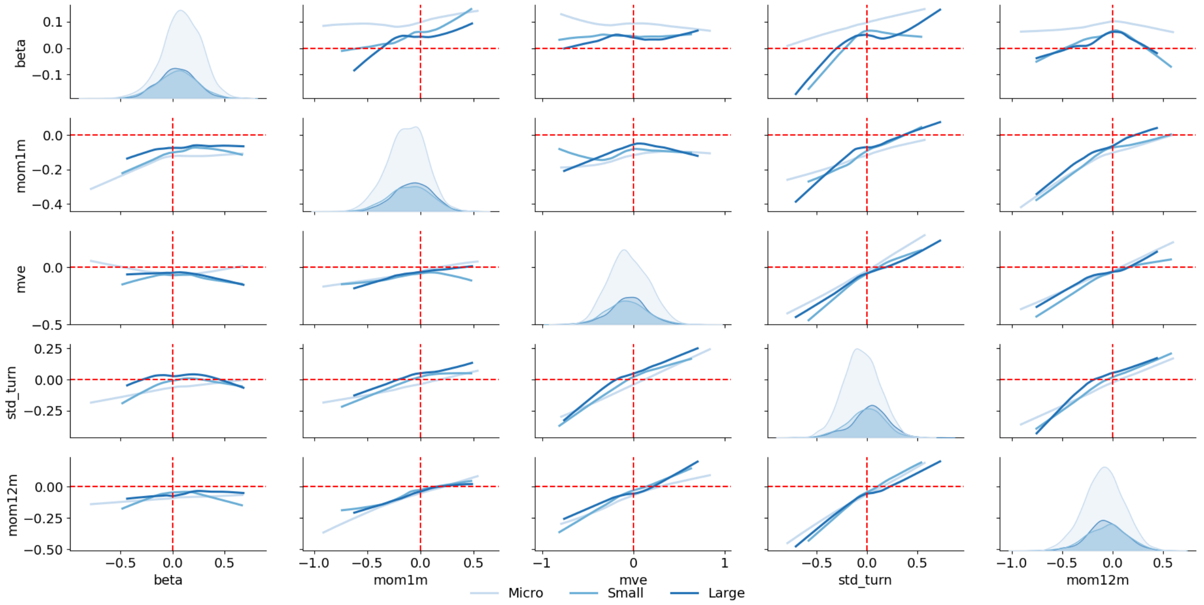


Figure J.3:
Nonlinear sensitivity interactions NN – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN, analogously to table 4.

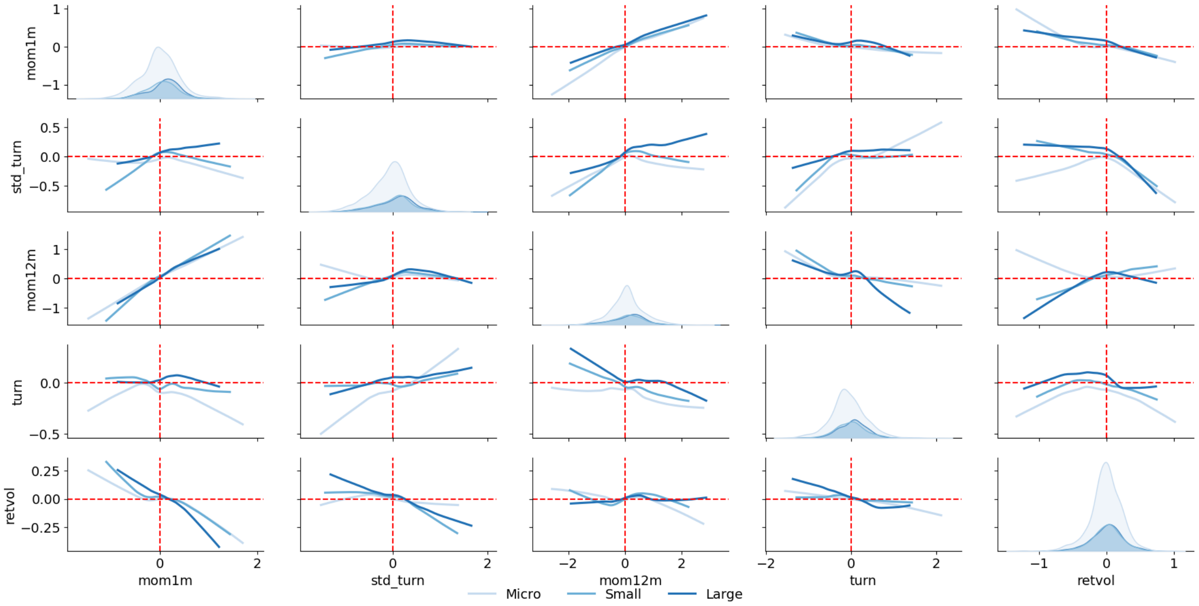


Figure J.4:
Nonlinear sensitivity interactions NN-W1 – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-W1, analogously to table 4.

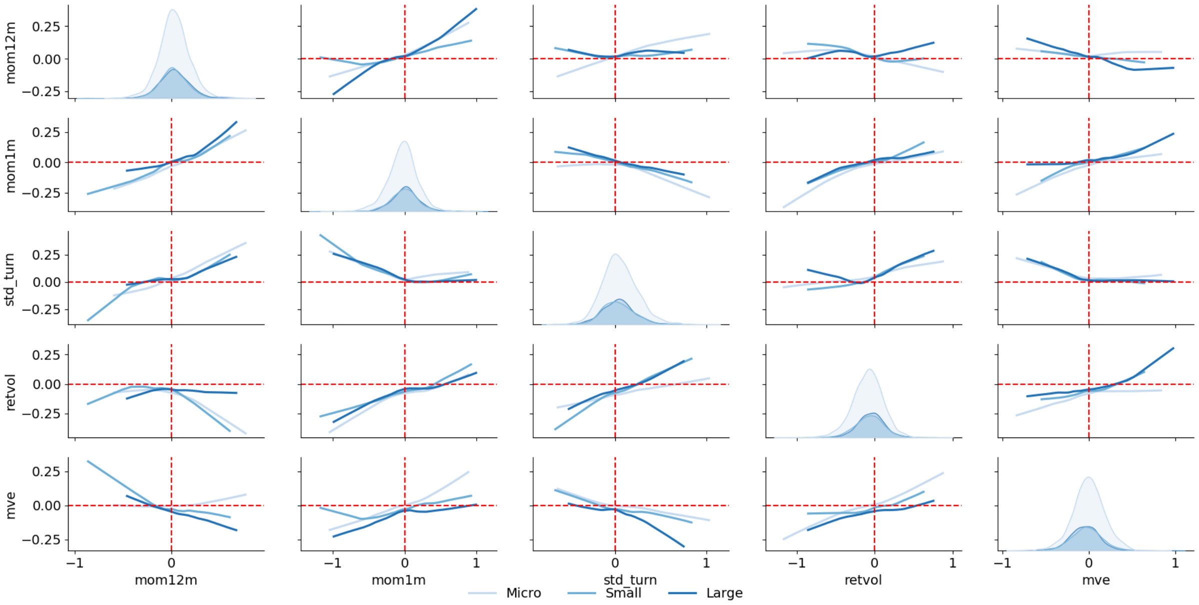


Figure J.5:
Nonlinear sensitivity interactions NN-W1W2 – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-W1W2, analogously to table 4.

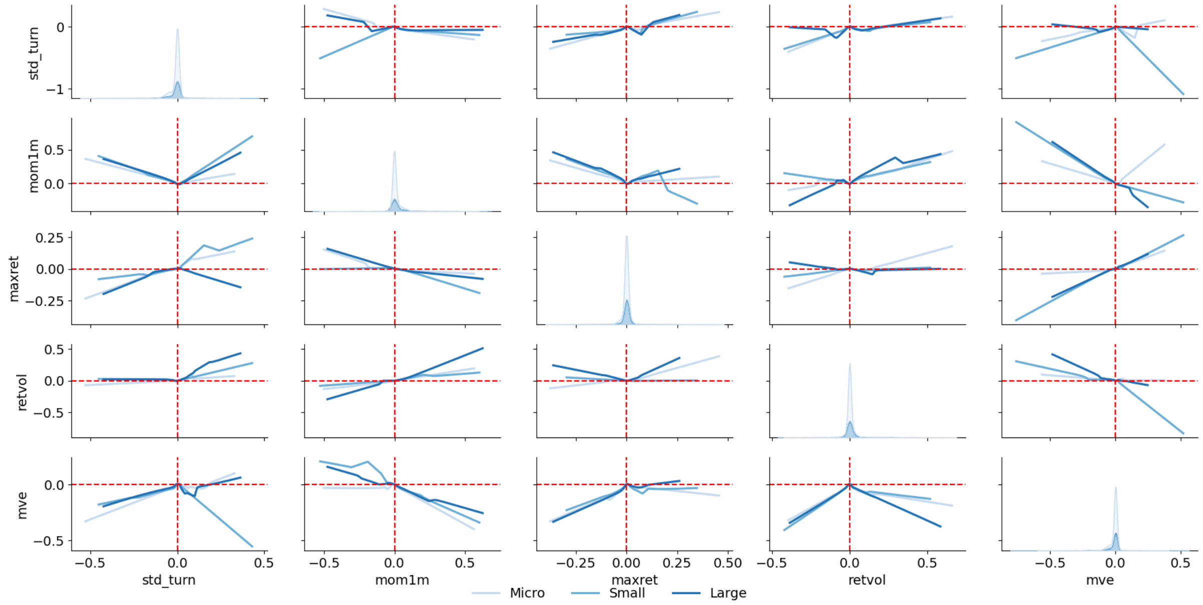


Figure J.6:
Nonlinear sensitivity interactions NN-J1 – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J1, analogously to table 4.

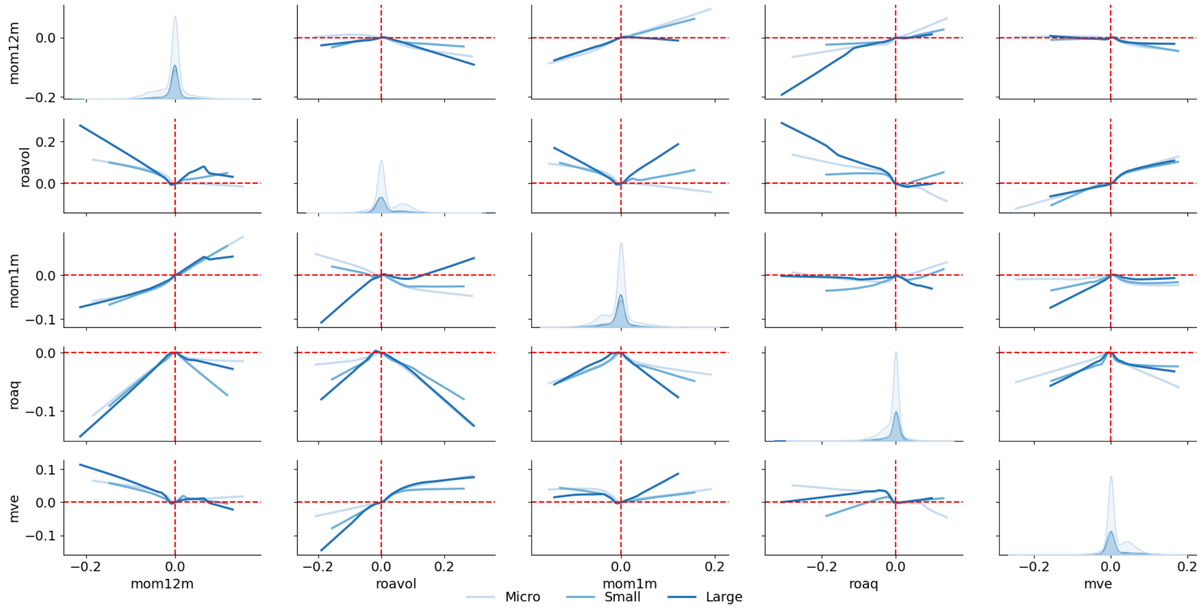


Figure J.7:
Nonlinear sensitivity interactions NN-J2 – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J2, analogously to table 4.

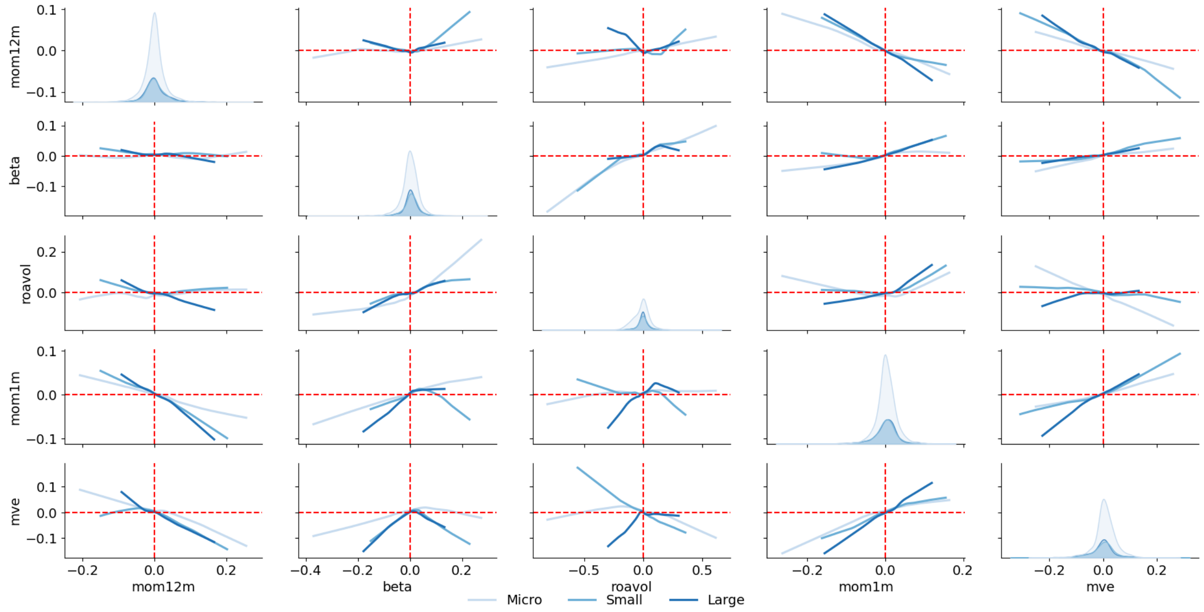


Figure J.8:
Nonlinear sensitivity interactions NN-J1J2 – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J1J2, analogously to table 4.

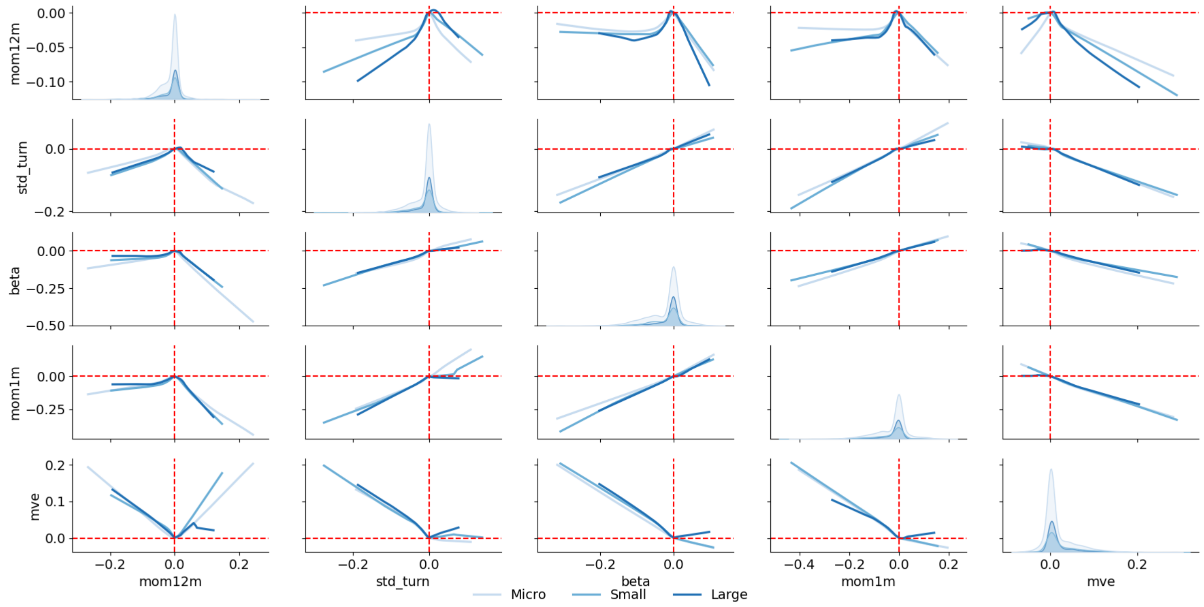


Figure J.9:
Nonlinear sensitivity interactions NN-J2-m – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J2-m, analogously to table 4.

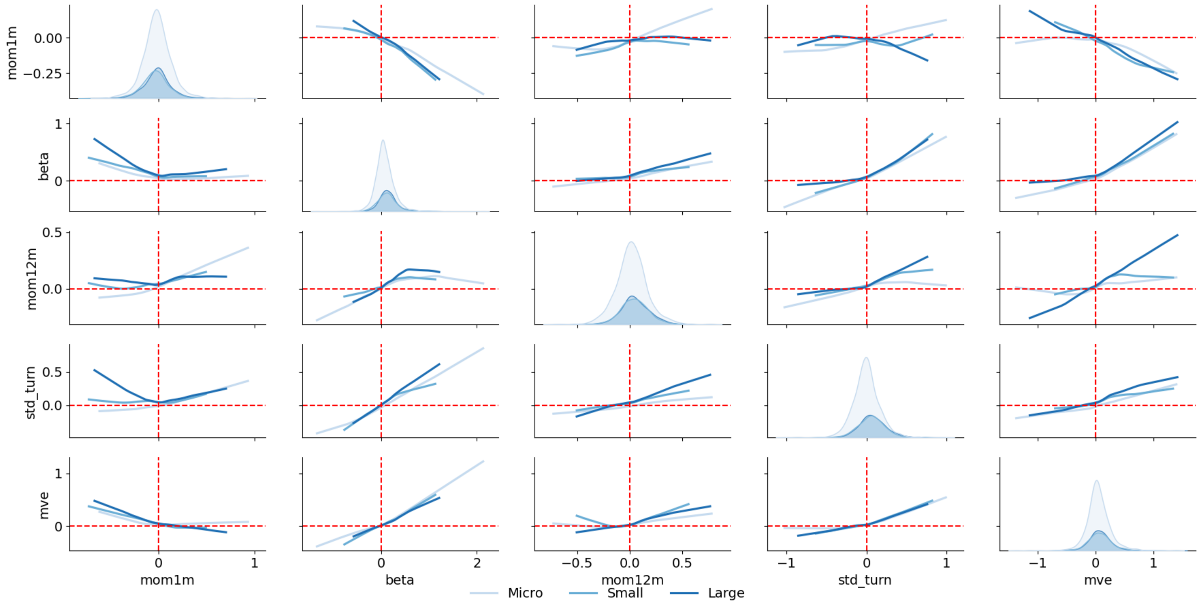


Figure J.10:
Nonlinear sensitivity interactions NN-J1J2-m – exemplary month 2006-09-30:
The graph visualises in the off-diagonals how the sensitivities with respect to the firm characteristics listed on the y-axis are expected to change, given a change in sensitivity with respect to the firm characteristics listed on the x-axis. The nonlinear sensitivity interactions are visualised by size class. The diagonal displays the distribution of sensitivities by size class. The selection of firm characteristics corresponds to the overall most influential firm characteristics for model NN-J1J2-m, analogously to table 4.

K Double-Sorted Portfolios

This appendix provides further empirical details on the double-sorted portfolios. For clarity, we primarily focus on portfolios constructed on return-on-assets and 12-month momentum, both constructed based on out-of-sample sensitivities estimated by NN-W2 and NN-J1-m. Further empirical results can be requested from the authors.

Tables K.1 to K.3 summarise the quintile portfolios, while tables K.4 to K.6 capture the results from common regressions on the Fama-French 3-factor model and a momentum factor. It can be seen that the introduction of a second sort by sensitivity can help improving the Sharpe ratio of the constructed portfolio. This effect seems to be more pronounced for portfolios constructed on sensitivity estimated with neural networks using Jacobian regularisation. For example, table K.4 shows that the annualised Sharpe ratio for equal-weighted portfolios sorted on 12-month momentum sensitivities and characteristics improves from 0.19 to 0.34 from the single to the double sort. This pattern seems to persist across an investment universe excluding microcaps (where the Sharpe ratio improves from 0.40 to 0.49), and value-weighred portfolios. Similar patterns emerge for other characteristics.

Figure K.1 visualises the cumulative returns from portfolios constructed on signals from 12-month momentum.

	Annualised Returns [%]						Annualised Volatility [%]						Annualised Sharpe Ratio					
	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High
Panel A: Value-Weighted Portfolios – NN-W2																		
Low	1.96	2.92	4.89	-1.42	2.17	2.22	30.00	33.94	31.72	31.25	32.03	32.58	0.06	0.09	0.15	-0.05	0.07	0.07
Q2	7.78	11.31	9.23	8.30	5.80	4.80	20.08	20.91	20.52	20.94	21.49	23.66	0.37	0.51	0.43	0.38	0.26	0.20
Q3	7.89	9.03	7.01	8.80	6.51	7.06	15.75	17.12	16.51	16.24	17.28	19.27	0.48	0.51	0.41	0.52	0.37	0.36
Q4	9.79	11.05	10.80	9.56	9.28	7.47	14.67	16.75	15.42	15.66	15.92	18.45	0.64	0.63	0.67	0.59	0.56	0.39
High	12.37	13.94	13.48	11.26	11.31	13.88	18.55	21.25	18.27	20.27	19.92	23.21	0.63	0.62	0.70	0.53	0.54	0.56
Panel B: Equal-Weighted Portfolios – NN-W2																		
Low	10.94	12.24	8.76	10.60	9.89	8.92	32.53	33.12	31.58	32.30	32.38	33.46	0.32	0.35	0.27	0.31	0.29	0.26
Q2	9.12	11.79	9.39	7.59	8.62	7.94	20.44	21.26	20.22	20.56	20.74	21.98	0.43	0.53	0.45	0.36	0.40	0.35
Q3	9.43	12.44	9.56	9.57	7.92	8.39	16.95	17.70	16.59	17.05	17.42	19.70	0.53	0.67	0.55	0.54	0.44	0.41
Q4	12.52	14.75	12.96	12.40	11.13	11.58	16.48	17.43	16.44	16.76	17.06	18.79	0.72	0.79	0.74	0.70	0.62	0.59
High	15.83	17.11	16.50	15.11	15.10	14.44	21.48	22.26	20.91	21.89	21.93	23.74	0.69	0.71	0.74	0.65	0.64	0.57
Panel C: Value-Weighted Portfolios – NN-J1-m																		
Low	1.96	3.17	0.96	3.98	4.02	-1.48	30.00	31.82	32.24	32.39	32.09	33.79	0.06	0.10	0.03	0.12	0.12	-0.04
Q2	7.78	6.46	9.81	8.49	6.69	5.51	20.08	21.92	21.64	21.61	20.89	22.89	0.37	0.29	0.43	0.38	0.31	0.23
Q3	7.89	7.38	7.82	8.45	9.37	5.26	15.75	18.10	17.25	16.06	16.41	17.40	0.48	0.39	0.44	0.51	0.55	0.30
Q4	9.79	10.50	8.42	9.33	9.55	9.42	14.67	16.52	15.78	15.31	15.81	16.54	0.64	0.61	0.51	0.58	0.58	0.55
High	12.37	13.21	12.02	12.24	11.04	13.13	18.55	21.15	20.13	19.32	19.88	20.20	0.63	0.59	0.57	0.60	0.53	0.61
Panel D: Equal-Weighted Portfolios – NN-J1-m																		
Low	10.94	18.95	9.94	12.57	8.13	7.86	32.53	34.33	32.68	34.25	32.63	32.86	0.32	0.51	0.29	0.35	0.24	0.23
Q2	9.12	10.81	9.90	9.26	7.18	8.34	20.44	21.29	20.88	20.79	20.39	21.58	0.43	0.48	0.45	0.43	0.34	0.37
Q3	9.43	10.50	9.66	10.68	8.76	6.88	16.95	17.39	17.37	17.40	17.16	17.81	0.53	0.58	0.53	0.59	0.49	0.37
Q4	12.52	14.02	11.89	13.09	11.41	12.43	16.48	17.86	16.56	16.21	16.50	17.60	0.72	0.74	0.68	0.76	0.66	0.67
High	15.83	16.30	16.13	14.77	16.52	14.74	21.48	23.58	21.73	21.05	21.55	21.84	0.69	0.64	0.69	0.66	0.71	0.63

Table K.1:
Double-sorted portfolios – 12-month momentum:
The table summarises the annualised average monthly returns, volatilities and Sharpe ratios of double-sorted quintile portfolios. The portfolios are benchmarked against single-sorted quintile portfolios. Double-sorted portfolios are sorted on the characteristic first and by the out-of-sample sensitivity with respect to that sensitivity second.

	Annualised Returns [%]						Annualised Volatility [%]						Annualised Sharpe Ratio					
	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High
Panel A: Value-Weighted Portfolios (No Microcaps) – NN-W2																		
Low	6.35	10.29	4.90	5.59	4.17	4.26	23.62	25.40	22.87	24.87	24.68	27.88	0.26	0.39	0.21	0.22	0.17	0.15
Q2	9.43	10.20	11.64	10.77	8.68	5.57	16.60	17.62	17.62	17.89	17.67	20.06	0.54	0.55	0.63	0.57	0.47	0.27
Q3	8.01	9.98	8.60	8.00	6.53	6.09	14.68	15.83	15.24	16.03	16.41	18.16	0.53	0.60	0.54	0.48	0.39	0.33
Q4	9.92	12.70	11.68	9.63	10.04	8.09	14.90	16.30	15.50	16.06	16.57	18.04	0.64	0.74	0.72	0.57	0.58	0.43
High	13.42	14.05	14.26	12.99	15.03	13.22	19.87	19.43	19.70	21.20	22.64	23.48	0.64	0.68	0.68	0.58	0.62	0.53
Panel B: Equal-Weighted Portfolios (No Microcaps) – NN-W2																		
Low	6.09	8.86	8.12	5.29	4.29	3.86	26.18	25.54	24.83	25.41	26.86	28.50	0.23	0.33	0.32	0.20	0.16	0.13
Q2	10.16	10.08	11.38	10.23	9.69	7.63	18.30	18.20	18.24	18.18	18.76	20.46	0.53	0.53	0.59	0.54	0.49	0.36
Q3	9.73	10.75	10.41	9.47	9.39	7.39	16.47	16.47	16.08	16.77	17.42	19.40	0.57	0.62	0.62	0.54	0.52	0.37
Q4	11.14	12.63	11.64	11.71	10.77	10.87	16.80	16.64	16.10	16.94	18.36	19.43	0.63	0.72	0.69	0.66	0.56	0.53
High	14.90	16.39	15.02	15.71	14.59	16.42	23.10	21.93	22.21	22.94	24.32	25.79	0.60	0.70	0.63	0.64	0.56	0.59
Panel C: Value-Weighted Portfolios (No Microcaps) – NN-J1-m																		
Low	6.35	7.36	6.72	8.30	3.72	8.95	23.62	23.33	26.21	24.19	25.87	25.75	0.26	0.31	0.25	0.33	0.14	0.33
Q2	9.43	7.64	9.01	9.92	8.73	11.70	16.60	17.05	18.10	18.24	18.31	18.76	0.54	0.43	0.48	0.52	0.46	0.59
Q3	8.01	7.82	8.14	7.54	7.61	8.95	14.68	15.69	15.88	15.82	16.16	17.70	0.53	0.48	0.49	0.46	0.46	0.49
Q4	9.92	8.71	10.51	9.23	11.38	10.57	14.90	16.00	16.46	16.48	15.96	17.60	0.64	0.52	0.61	0.54	0.68	0.57
High	13.42	11.77	14.30	13.23	15.61	14.37	19.87	21.50	21.41	21.16	21.88	21.42	0.64	0.52	0.63	0.59	0.67	0.63
Panel D: Equal-Weighted Portfolios (No Microcaps) – NN-J1-m																		
Low	6.09	5.92	6.91	6.99	5.88	6.19	26.18	25.71	27.19	26.02	26.15	27.02	0.23	0.22	0.25	0.26	0.22	0.22
Q2	10.16	9.60	9.48	10.72	9.29	11.86	18.30	17.60	18.91	18.59	18.93	19.79	0.53	0.52	0.48	0.55	0.47	0.57
Q3	9.73	9.12	9.49	9.16	10.30	10.24	16.47	16.81	16.46	16.97	16.88	18.09	0.57	0.52	0.55	0.52	0.58	0.54
Q4	11.14	9.65	10.44	10.35	12.81	12.98	16.80	17.29	17.07	17.34	17.53	18.19	0.63	0.54	0.58	0.57	0.69	0.67
High	14.90	13.47	16.30	14.06	16.44	14.62	23.10	22.87	23.78	23.37	23.19	24.75	0.60	0.56	0.64	0.57	0.66	0.55

Table K.2:
Double-sorted portfolios (no microcaps) – 12-month momentum:
The table summarises the annualised average monthly returns, volatilities and Sharpe ratios of double-sorted quintile portfolios. The portfolios are benchmarked against single-sorted quintile portfolios. Double-sorted portfolios are sorted on the characteristic first and by the out-of-sample sensitivity with respect to that sensitivity second. Microcaps are excluded.

	Annualised Returns [%]						Annualised Volatility [%]						Annualised Sharpe Ratio					
	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High	Single Sort	Sens. – Low	Q2	Q3	Q4	Sens. – High
Panel A: Value-Weighted Portfolios (No Microcaps) – NN-W2																		
Low	6.48	7.68	9.19	4.81	6.76	8.01	23.28	24.10	25.53	24.08	25.04	23.90	0.27	0.31	0.35	0.20	0.26	0.32
Q2	8.74	10.29	9.55	6.91	8.34	9.25	17.59	19.14	18.73	18.24	19.59	21.17	0.48	0.51	0.49	0.37	0.41	0.42
Q3	9.15	9.60	8.37	7.75	9.25	10.51	14.74	16.23	15.82	16.18	15.80	18.92	0.60	0.57	0.51	0.46	0.56	0.53
Q4	9.69	11.31	9.73	8.59	8.31	11.23	14.62	16.63	16.80	15.98	18.14	18.27	0.63	0.65	0.55	0.52	0.44	0.59
High	11.42	10.33	13.34	11.07	12.00	12.46	16.40	18.23	17.44	17.84	18.83	20.54	0.66	0.54	0.72	0.59	0.60	0.57
Panel B: Equal-Weighted Portfolios (No Microcaps) – NN-W2																		
Low	6.79	7.34	8.91	5.39	5.98	7.15	26.46	24.78	25.61	26.26	27.57	27.11	0.25	0.29	0.33	0.20	0.21	0.26
Q2	10.08	8.91	10.97	10.33	9.66	9.43	17.50	18.56	17.99	17.56	18.67	20.23	0.55	0.46	0.58	0.56	0.50	0.45
Q3	10.92	10.20	10.06	11.44	11.19	10.62	17.20	18.05	17.65	17.49	17.66	19.42	0.60	0.54	0.55	0.62	0.60	0.52
Q4	11.45	11.47	11.03	11.47	11.68	11.20	17.78	18.73	17.82	17.91	18.86	19.75	0.61	0.58	0.59	0.61	0.59	0.54
High	12.70	12.12	12.25	13.42	13.26	14.51	19.48	19.79	19.20	19.74	20.07	21.91	0.62	0.58	0.60	0.64	0.62	0.62
Panel C: Value-Weighted Portfolios (No Microcaps) – NN-J1-m																		
Low	6.48	7.86	6.85	5.56	6.59	6.47	23.28	24.69	24.59	25.50	25.18	25.24	0.27	0.31	0.27	0.21	0.25	0.25
Q2	8.74	9.78	7.39	8.51	8.38	9.06	17.59	18.20	17.16	18.70	19.11	19.87	0.48	0.51	0.42	0.44	0.42	0.44
Q3	9.15	10.53	7.48	8.86	9.60	11.25	14.74	16.26	15.99	16.14	16.31	16.65	0.60	0.62	0.45	0.53	0.56	0.64
Q4	9.69	10.08	10.38	8.33	10.61	9.30	14.62	16.30	16.01	16.41	16.30	15.96	0.63	0.59	0.62	0.49	0.62	0.56
High	11.42	9.65	10.79	12.69	12.28	11.57	16.40	19.10	18.47	17.77	17.48	19.26	0.66	0.48	0.56	0.68	0.67	0.57
Panel D: Equal-Weighted Portfolios (No Microcaps) – NN-J1-m																		
Low	6.79	7.65	7.22	6.51	6.25	6.45	26.46	26.67	26.79	27.43	26.84	27.15	0.25	0.28	0.26	0.23	0.23	0.23
Q2	10.08	10.92	10.11	9.83	9.34	9.87	17.50	17.79	17.71	18.02	18.00	18.91	0.55	0.59	0.55	0.52	0.50	0.50
Q3	10.92	11.65	10.89	10.48	10.63	11.17	17.20	18.16	17.48	17.41	17.85	18.68	0.60	0.61	0.59	0.57	0.57	0.57
Q4	11.45	12.24	11.77	10.56	11.66	11.05	17.78	18.56	17.96	18.19	18.01	18.76	0.61	0.63	0.62	0.55	0.62	0.56
High	12.70	12.61	12.22	13.54	12.41	13.12	19.48	20.09	19.59	20.41	19.39	19.57	0.62	0.59	0.59	0.63	0.61	0.63

Table K.3:
Double-sorted portfolios (no microcaps) – return-on-assets:
The table summarises the annualised average monthly returns, volatilities and Sharpe ratios of double-sorted quintile portfolios. The portfolios are benchmarked against single-sorted quintile portfolios. Double-sorted portfolios are sorted on the characteristic first and by the out-of-sample sensitivity with respect to that sensitivity second. Microcaps are excluded.

	All stocks				No microcaps			
	NN-J1-m		NN-J1-m		NN-J1-m		NN-J1-m	
	equal-weighed		value-weighed		equal-weighed		value-weighed	
	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort
intercept (α)	0.000	−0.003	0.006***	0.002	0.002	0.000	0.003	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
Mktrf (β_1)	0.018	0.010	−0.142*	−0.166***	0.049	0.029	0.023	0.061**
	(0.058)	(0.054)	(0.079)	(0.044)	(0.048)	(0.031)	(0.066)	(0.026)
HML (β_2)	0.193*	0.231**	−0.057	−0.021	−0.079	−0.067	−0.240**	−0.113***
	(0.101)	(0.111)	(0.067)	(0.064)	(0.073)	(0.063)	(0.109)	(0.041)
SMB (β_3)	−0.206***	−0.199***	−0.317***	−0.156**	0.190***	0.246***	0.131*	0.203***
	(0.076)	(0.070)	(0.080)	(0.067)	(0.071)	(0.093)	(0.079)	(0.057)
UMD (β_4)	1.135***	1.242***	1.398***	1.437***	1.201***	1.208***	1.263***	1.302***
	(0.082)	(0.116)	(0.070)	(0.073)	(0.085)	(0.063)	(0.105)	(0.064)
Observations	432		432		432		432	
R^2	0.633	0.706	0.649	0.875	0.786	0.904	0.710	0.913
Adjusted R^2	0.630	0.703	0.646	0.874	0.784	0.903	0.707	0.912
Annualised Return [%]	7.85	4.45	14.99	10.24	10.98	8.35	11.38	6.64
Annualised Volatility [%]	22.23	23.01	28.63	25.13	21.55	20.31	24.46	21.75
Annualised Sharpe Ratio	0.34	0.19	0.49	0.39	0.49	0.40	0.44	0.30

Note: *p<0.1; **p<0.05; ***p<0.01

Table K.4:
Portfolio summary – J1-m, 12-month momentum:
The table summarises standard Fama-French regressions, following

$$R_{p,t} = \alpha + \beta_1 \text{Mktrf}_t + \beta_2 \text{HML}_t + \beta_3 \text{SMB}_t + \beta_4 \text{UMD}_t + \epsilon_t,$$

where we regress the respective portfolio returns on the Fama-French 3 Factor model plus a market factor. The market and Fama-French portfolios are sourced directly from Fama's website through WRDS. The constructed portfolio returns are in excess of the risk-free rate.

	All stocks				No microcaps			
	NN-W2		NN-W2		NN-W2		NN-W2	
	equal-weighed		value-weighed		equal-weighed		value-weighed	
	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort
intercept (α)	0.004	0.002	0.004	0.005^{**}	0.005^{**}	0.006^{***}	0.003	0.006^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Mktrf (β_1)	-0.075	-0.070	-0.233^{***}	-0.297^{***}	-0.202^{***}	-0.199^{***}	-0.222^{***}	-0.282^{***}
	(0.057)	(0.055)	(0.066)	(0.062)	(0.050)	(0.053)	(0.070)	(0.059)
HML (β_2)	0.425^{**}	0.405^{**}	0.258	0.196	0.084	-0.020	-0.170	-0.290^{**}
	(0.184)	(0.188)	(0.218)	(0.164)	(0.179)	(0.150)	(0.193)	(0.134)
SMB (β_3)	-0.695^{***}	-0.788^{***}	-0.684^{***}	-0.889^{***}	-0.241^{**}	-0.533^{***}	-0.092	-0.532^{***}
	(0.135)	(0.170)	(0.102)	(0.098)	(0.094)	(0.179)	(0.069)	(0.134)
UMD (β_4)	0.622^{***}	0.411^{***}	0.509^{***}	0.229^{***}	0.383^{***}	0.152	0.475^{***}	0.182^{**}
	(0.147)	(0.120)	(0.114)	(0.086)	(0.137)	(0.095)	(0.149)	(0.082)
Observations	432		432		432		432	
R^2	0.324	0.357	0.387	0.512	0.332	0.331	0.273	0.415
Adjusted R^2	0.383	0.351	0.346	0.508	0.235	0.325	0.209	0.410
Annualised Return [%]	7.54	4.28	5.05	4.26	6.74	5.57	4.77	4.73
Annualised Volatility [%]	21.44	20.1	21.7	17.91	16.44	13.3	20.15	13.68
Annualised Sharpe Ratio	0.34	0.21	0.23	0.23	0.4	0.41	0.23	0.34

Note: *p<0.1; **p<0.05; ***p<0.01

Table K.5:
Portfolio summary – W2, return-on-assets:
The table summarises standard Fama-French regressions, following

$$R_{p,t} = \alpha + \beta_1 \mathrm{Mktrf}_t + \beta_2 \mathrm{HML}_t + \beta_3 \mathrm{SMB}_t + \beta_4 \mathrm{UMD}_t + \epsilon_t,$$

where we regress the respective portfolio returns on the Fama-French 3 Factor model plus a market factor. The market and Fama-French portfolios are sourced directly from Fama's website through WRDS. The constructed portfolio returns are in excess of the risk-free rate.

	All stocks				No microcaps			
	NN-W2 equal-weighed		NN-W2 value-weighed		NN-W2 equal-weighed		NN-W2 value-weighed	
	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort	Double-Sort	Single-Sort
intercept (α)	−0.001	−0.003	0.002	0.002	0.002	0.000	−0.002	−0.002
	0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.003)	(0.001)
Mktrf (β_1)	0.048	0.010	−0.066	−0.166***	0.081	0.029	0.049	0.061**
	(0.080)	(0.054)	(0.071)	(0.044)	(0.050)	(0.031)	(0.074)	(0.026)
HML (β_2)	0.331**	0.231**	0.068	−0.021	0.072	−0.067	0.063	−0.113***
	(0.149)	(0.111)	(0.079)	(0.064)	(0.099)	(0.063)	(0.129)	(0.041)
SMB (β_3)	−0.230**	−0.199***	−0.172**	−0.156**	0.146**	0.246***	0.110	0.203***
	(0.099)	(0.070)	(0.070)	(0.067)	(0.067)	(0.093)	(0.080)	(0.057)
UMD (β_4)	1.338***	1.242***	1.571***	1.437***	1.294***	1.208***	1.403***	1.302***
	(0.169)	(0.116)	(0.099)	(0.073)	(0.090)	(0.063)	(0.113)	(0.064)
Observations	432		432		432		432	
R^2	0.606	0.706	0.718	0.875	0.779	0.904	0.708	0.913
Adjusted R^2	0.603	0.703	0.716	0.874	0.777	0.903	0.706	0.912
Annualised Return [%]	7.56	4.45	11.54	10.24	11.39	8.35	6.87	6.64
Annualised Volatility [%]	26.55	23.01	29.46	25.13	22.72	20.31	25.95	21.75
Annualised Sharpe Ratio	0.28	0.19	0.37	0.39	0.48	0.40	0.26	0.30
Note:								*p<0.1; **p<0.05; ***p<0.01

Table K.6:
Portfolio summary – W2, 12-month momentum:
The table summarises standard Fama-French regressions, following

$$R_{p,t} = \alpha + \beta_1 \mathrm{Mktrf}_t + \beta_2 \mathrm{HML}_t + \beta_3 \mathrm{SMB}_t + \beta_4 \mathrm{UMD}_t + \epsilon_t,$$

where we regress the respective portfolio returns on the Fama-French 3 Factor model plus a market factor. The market and Fama-French portfolios are sourced directly from Fama's website through WRDS. The constructed portfolio returns are in excess of the risk-free rate.

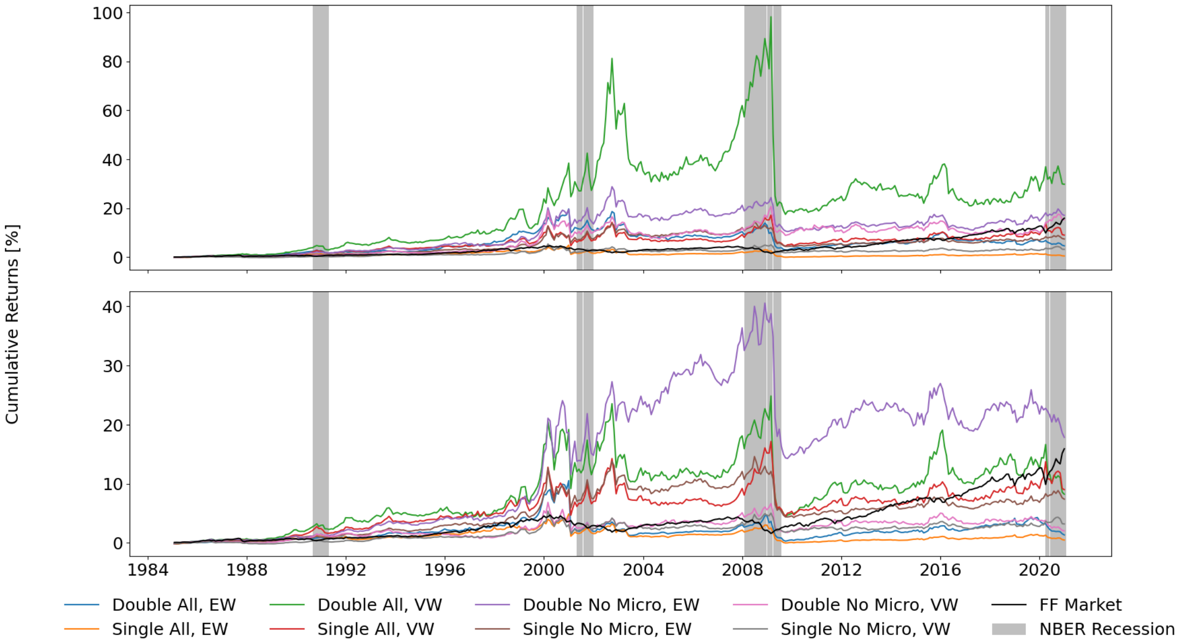


Figure K.1:
Cumulative portfolio returns – 12-month momentum)
The graphs shows the cumulative portfolio returns of single and double-sorted portfolio returns, where the out-of-sample sensitivities of the double-sorted portfolios are estimated by NN-J1-m (top) and NN-W2 (bottom). The cumulative portfolio returns are benchmarked against the cumulative market return, where the market portfolio is sourced from Kenneth French's website.

The Data Analytics for Finance and Macro (DAFM) Research Centre
Forecasting trends is more important now than it has ever been. Our quantitative financial and macroeconomic research helps central banks and statistical agencies better understand their markets. Having our base in central London means we can focus on issues that really matter to the City. Our emphasis on digitalisation and the use of big data analytics can help people understand economic behaviour and adapt to changes in the economy.

Find out more at kcl.ac.uk/dafm

KING'S BUSINESS SCHOOL

Data Analytics for Finance & Macro Research Centre