Quantifying Equity Risk Premia
Financial Economic Theory and High-Dimensional Statistical Methods

Dissertation
zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
Constantin Hanenberg

Tübingen
2023

1. Betreuer:	Prof. Dr. Joachim Grammig
2. Betreuer:	Prof. Dr. Christian Koziol
Tag der mündlichen Prüfung:	13.12.2023
Dekan:	Prof. Dr. Ansgar Thiel
1. Gutachter:	Prof. Dr. Joachim Grammig
2. Gutachter:	Prof. Dr. Christian Koziol

Contents

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals who have played a pivotal role in the completion of this dissertation: First and foremost, I want to thank my wonderful wife for her unwavering patience and support throughout these years. Her encouragement has been a constant source of motivation, and I am truly grateful to have her by my side. In addition, I am deeply indebted to my supervisor, Prof. Dr. Joachim Grammig, for providing the best possible environment for me to thrive in my studies. His tireless guidance and support have been crucial to my academic career. A special word of thanks goes to my coauthors, Dr. Jantje Sönksen and Prof. Dr. Christian Schlag, for outstanding collaboration and stimulating discussions, which have been essential to the quality of my work. I must also acknowledge the support of my parents, without whom I would not be where I am today. Their belief in my potential and their sacrifices have been the driving force behind everything I have accomplished so far. Last but not least, I extend my gratitude to my coworkers, whose camaraderie and humor have made these four years so memorable. Their support and friendship have been a valuable part of this academic journey. Lastly, I acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grants INST 35/1597-1 FUGG, GR 2288/7-1, and SCHL 558/7-1.

Remarks

This doctoral thesis is part of a joint research project of the Chair of Finance (Prof. Christian Schlag) at Goethe University Frankfurt and the Chair of Statistics and Econometrics (Prof. Joachim Grammig) at Eberhard Karls University Tübingen. The research project is funded by the German Research Foundation (DFG) through grants GR 2288/7-1 and SCHL 558/7-1. The use of the WRDS data for the present thesis is possible because of the access and license available to the Frankfurt-based cooperation partner. The results presented herein are part of this joint research project, and publication of the reported results beyond this thesis must only occur via joint research papers coauthored with members of the group who are licensed to use the WRDS data.

Chapter 1
Introduction

The conditional stock risk premium is a quantity of central interest in financial economics. It is the expected return that investors demand in excess of the risk-free rate for holding equity in the issuing company. As such, it contains valuable information about the risks traded in financial markets. However, because stock risk premia are not directly observable, one has to rely on approximations to harness the information they contain. One possibility is to interpret this as a statistical exercise in which approximations are obtained from predictive regressions of excess returns onto firm characteristics that are observed prior to the investment period. While early proponents of this approach mainly considered linear specifications of these regressions (e.g., Moskowitz and Grinblatt, 1999; Amihud, 2002), recent years have seen a surge in applications of machine learning methods that promise greater flexibility in describing the risk-return relationship (e.g., Gu et al., 2021; Chen et al., 2023). The added flexibility, however, raises concerns about the generalization abilities of these models, their interpretability, and their compatibility with classical methods of statistical inference.

Another approach to measuring stock risk premia avoids any statistical estimation and attempts to derive the investors' risk perception from theoretical considerations alone (e.g., Martin and Wagner, 2019; Chabi-Yo et al., 2023). This is achieved by establishing a relationship between unobserved stock risk premia and risk-neutral moments of returns, which can be inferred from European option prices. A key advantage of this approach is that sudden changes in investor sentiment are immediately reflected in changing conditional expectations. Statistical models are much more rigid in this respect, as their parameterization is chosen on the basis of past investor behavior. One drawback of the theory-based approach, however, is that structural assumptions are needed to explain the change of measure associated with establishing the aforementioned relationship. The purpose of this dissertation is to compare these two contrasting philosophies and explore ways in which they can be combined to improve our understanding of conditional risk premia.

The first chapter, titled "Theory-based versus machine learning-implied stock risk premia", starts with a comparison of two prominent representatives of these two competing strands of the literature. For the machine learning approach, we adopt the methodology by Gu et al. (2020), who employ various statistical models to account for nonlinearities and interaction effects in the approximation of conditional stock risk premia. While most of these models require large amounts of historical data to unfold their potential, the theory-based approach by Martin and Wagner (2019) offers a parsimonious alternative. They propose a formula for the expected excess return

that is free of unknown parameters and can be implemented using a panel of option prices. The goal of this chapter is to evaluate which of the two approaches provides a better approximation of conditional stock risk premia, as measured by their ability to predict future realized excess returns. Beyond this comparison, we present an alternative hybrid approach that employs machine learning to approximate what is left unexplained by the option-based formula. In this way, we seek to identify the strengths and weaknesses of the two competing approaches.

Our conclusions are as follows: At a monthly investment horizon, the theory-based approach provides better out-of-sample forecasts than any of the machine learning models, both in terms of the predictive R-squared and the Sharpe ratio of an investment strategy that exploits the differences between the highest and lowest deciles of expected excess returns. At an annual investment horizon, however, the ranking is reversed in favor of the machine learning models, which may be due to the fact that options with maturies of one year are much less frequently traded than their monthly counterparts. This suggests that the performance of the theory-based approach is highly dependent on the quality of the options data that is used for its implementation. One lesson we draw from the hybrid approach is that the formula by Martin and Wagner (2019) can be further improved, as its approximation errors are absorbed by other stock characteristics.

The second chapter of this thesis, titled "The uncertainty principle in asset pricing", builds on the findings of the previous chapter and introduces an alternative set of assumptions to link physical and risk-neutral return distributions. We employ these assumptions to derive a fully-implied representation of the conditional capital asset pricing model (CAPM) in which both the betas and the equity premium are jointly characterized by the information embedded in option prices. The novelty of this approach is that the implied beta and the equity premium represent valid measurements of their physical counterparts without the need for further risk adjustment. Moreover, because we do not need to estimate any of the model's time-varying parameters, we are able to test its unconditional implications directly.

Leveraging these advantages, we study a phenomenon that is synonymous with the failure of the CAPM – the flat relationship between average predicted and realized excess returns of beta-sorted portfolios. To provide a coherent explanation for the persistence of this phenomenon across investment horizons, we decompose the model's testable restrictions in a way that allows us to distinguish between asset-specific and aggregate components of market risk. Our results indicate that, at shorter investment horizons, the CAPM's failure is due to the inherently unpredictable component of the market excess return, while at longer horizons, it is due to the limited cross-sectional explanatory power of the betas. In analogy to the uncertainty principle in quantum mechanics, we refer to this observation as the uncertainty principle in asset pricing.

The third chapter, titled "Multi-task learning in cross-sectional regressions", continues where the second chapter left off and evaluates the short-term explanatory power of the implied betas. To this end, we derive testable restrictions from period-by-period

cross-sectional regressions that include the implied beta and other stock characteristics as regressors. According to these restrictions, the beta should be the only characteristic relevant for explaining the cross-section of returns, i.e., it should drive out any of the other characteristics.

One of the challenges associated with this test lies in determining which of the myriad candidate characteristics should be used as competitors. In contrast to previous literature, we address this issue systematically using a combination of ℓ_1- and ℓ_2-regularization, known as the multi-task Lasso. The appeal of our procedure is that it leverages the entire panel of returns and characteristics to select a common set of covariates, while taking into account that each of the regressions is subject to its own parameterization. Moreover, by combining the multi-task Lasso with standard ℓ_1-regularization, we are able to distinguish between stable and anomalous return predictive signals, which allows us to examine the robustness of prior research on stock return predictability.

One problem that inevitably arises when we systematically select characteristics based on in-sample information is that classical methods of statistical inference are no longer valid. We address this problem by performing the selection and estimation steps separately on multiple random subsamples of the data, thus accounting for both the bias that is induced by selection and the uncertainty that is due to splitting the data.

In the empirical part of this chapter, we implement this testing strategy using different sets of test assets and an extensive set of stock characteristics. We find that while the implied beta is by far the most important predictor of return variation, there are still some characteristics that provide incremental information for the cross-section of returns. Depending on the chosen set of test assets, the few remaining characteristics are either too important to be ignored, as in the case of individual stocks, or not important enough for the conditional CAPM to be rejected, as in the case of characteristic-sorted portfolios. Regarding the stability of the return predictive signals, we find that most of the characteristics proposed in the previous literature are stable rather than anomalous predictors. This finding casts a positive light on empirical finance research, which is often criticized for its lack of replicability.

Chapter 2

Theory-based versus machine learning-implied stock risk premia*

2.1 Motivation

When it comes to measuring stock risk premia, two roads diverge in the finance world – or at least, so it may seem to an observer of recent literature on empirical asset pricing. Two prominent studies exemplify this impression: Martin and Wagner (2019) quantify the conditional expected return of a stock by exploiting the information contained in option prices, as implied by financial economic theory.[1] Gu et al. (2020) pursue the same end but along a completely different path, leveraging the surge of machine learning applications in economics and finance, together with advances in computer technology.[2] Approaches similar to the one adopted by Martin and Wagner (2019) derive results from asset pricing paradigms and have no need of historical data to quantify stock risk premia; Gu et al. (2020) and related papers instead do not refer substantially to financial economic theory and prefer to "let the data speak for themselves."

These radically different ways to address the same issue motivate us to conduct a fair, comprehensive performance comparison of theory-based and machine learning approaches to measuring stock risk premia and to explore the potential of hybrid strategies. The comparison is based on the fact that the risk premium is the conditional expected value of an excess return and that, in the present context, the machine

*This chapter is based on Grammig et al. (2022), available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3536835. Earlier versions of this paper were presented at the 48th Annual Meeting of the European Finance Association, the 12th Econometric Society World Congress, the 13th Annual Conference of the Society for Financial Econometrics, and several other conferences and seminars. We thank the participants, and in particular, Michael Bauer, Svetlana Bryzgalova, Emanuele Guidotti, Christoph Hank, Jens Jackwerth, Alexander Kempf, Michael Kirchler, Christian Koziol, Michael Lechner, Marcel Müller, Elisabeth Nevins, Yarema Okhrin, Olaf Posch, Éric Renault, Olivier Scaillet, Julie Schnaitmann, and Grigory Vilkov for helpful comments.

[1]Their strategy to quantify the risk premia of financial assets draws on Martin's (2017) derivation of a lower bound for the conditional expected return of the market, which in turn is based on concepts outlined by Martin (2011). Kadan and Tang (2020) take up Martin's (2017) idea and argue that it can be applied to quantify risk premia for a certain type of stocks. Bakshi et al. (2020) propose an exact formula for the expected return of the market that relies on all risk-neutral moments of returns. In a similar vein, Chabi-Yo et al. (2023) consider bounds for expected excess stock returns that take into account higher risk-neutral moments using calibrated preference parameters.

[2]Recent studies in a similar vein include those by Light et al. (2017), Martin and Nagel (2022), and Freyberger et al. (2020).

learning objective is to minimize the mean squared forecast error (MSE). Because the conditional expectation is the best predictor in terms of MSE, it seems natural to compare the opposing philosophies by gauging the quality of their excess return forecasts: A superior forecast indicates a better approximation of the risk premium. Such a comparative analysis can reveal whether the use of the information theoretically embedded in current option prices is preferable to sophisticated statistical analyses of historical data, or vice versa.

Beyond this direct comparison, we also investigate the potential of hybrid strategies that combine the theory-based and machine learning paradigms. In particular, we rely on machine learning to address the approximation errors of the theory-based approach. These residuals are functions of moments conditional on time t information, and machine learning is employed to approximate the conditional moments using time t stock- and macro-level variables. We refer to this strategy as *theory assisted by machine learning*. We also consider a machine learning approach that includes theory-implied risk premium measures computed from current option data, along with historical stock- and macro-level feature data. To ensure a fair comparison we deliberately adhere to the model specifications used in the base papers, for example regarding the features considered and the training and validation strategy adopted for machine learning.

To level the playing field, we need data for which both theory-based and machine learning approaches are applicable. For our large-scale empirical study, we use data on the S&P 500 constituents from 1964 to 2018, including firm- and macro-level variables, as well as return and option data. The analysis centers on theory-based and machine-learning-implied estimates of stock risk premia, computed at one-month and one-year investment horizons. We focus on the machine learning methods that Gu et al. (2020) identify as most promising, namely, an ensemble of artificial neural networks (ANN), gradient boosted regression trees (GBRT), and random forests (RF). We also include the elastic net (ENet), as a computationally less demanding benchmark. We consider two training and validation strategies, starting in 1974 (*long training*) and 1996 (*short training*), respectively. Using the short training scheme is necessary for all hybrid approaches, because the option data are not available earlier.

The main results are as follows: Of the two theory-based approaches that we consider, the one proposed by Martin and Wagner (2019) (henceforth, MW) is preferable to Kadan and Tang's (2020) approach (henceforth, KT). At the one-month horizon, MW is also superior to three of the four machine learning methods. Only MW and the ANN deliver a positive predictive R^2 of comparable size, according to the analyses that use forecasts issued at the end of each month. When using risk premium estimates at a daily frequency, the predictive R^2 by MW increases from 0.2% to 0.9%. Adapting the machine learning models to deliver daily risk premium estimates improves their performance, but it does not match that of MW; the best machine learning result is achieved by the ANN, with a predictive R^2 of 0.5%. We note that among all the machine learning approaches and stock universes considered by Gu et al. (2020), the highest reported predictive R^2 is 0.7%; the one-month horizon is a low signal-to-noise

environment. Constructing prediction-sorted portfolios, we find that the alignment of predicted and realized mean excess returns works better and the cross-sectional variation of mean realized returns across prediction-sorted portfolios is highest when using MW.[3]

The signal-to-noise ratio increases at the one-year horizon. ANN and GBRT achieve predictive R^2 around 9%, very similar to MW. While ENet and KT are less successful, the RF delivers the highest annual predictive R^2 of about 19%. The analysis of the alignment and cross-sectional variation of prediction-sorted portfolios also provides corroborative evidence. To achieve this performance, the RF relies on the long training scheme. Generally, the performance of machine learning approaches is attenuated when using a short training scheme, but hybrid strategies can compensate for this drawback. A theory assisted by machine learning strategy that takes MW as a basis and trains an RF or an ANN to deal with the approximation errors implied by the theory-based formula is particularly successful. The assistance by the RF increases the predictive R^2 delivered by MW from 9% to 16%. The analysis of prediction-sorted portfolios further establishes the expediency of this hybrid approach: It produces the best alignment and highest variation of the mean realized excess returns across the prediction-sorted portfolios. The MW+RF and MW+ANN combinations answer critiques of machine learning as measurement without theory, because they reflect financial economic paradigms and employ statistical assistance only for the components that remain unaccounted for by theory.

When risk premia need to be estimated at a daily frequency, the theory-based methods offer a natural advantage. The required option data are available at a daily frequency, whereas many stock- and all macro-level features are updated monthly at best. However, we find that a modified hybrid strategy that uses daily updated theory-based features for an RF, trained using end-of-month data, does a good job providing daily risk premium estimates. The annual predictive R^2 of the RF without theory-based features and evaluated at a daily frequency is 9%. Including theory-based features doubles this value.

Further analysis reveals that the importance of firm- and macro-level features does not differ markedly across the two applications of the RF, that is, its pure usage or when assisting the theory-based approach. At the one-year horizon, the familiar firm-level return predictive signals are most important in both applications: the book-to-market ratio, liquidity-related indicators, and momentum variables (in that order). The dominance of the short-run price reversal at the one-month horizon vanishes at the one-year horizon. The importance of the Treasury bill rate (a macro-level predictor) in both applications supports the use of short-term interest rates as state variables in variants of the intertemporal capital asset pricing model. The benefits of theory assistance by machine learning are also corroborated by disaggregated analyses, for which we create portfolios by sorting stocks according to valuation ratios, liquidity

[3]The advantage of the theory-based paradigm at the one-month horizon is confirmed by a complementary analysis in which we apply Chabi-Yo et al.'s (2023) option-based method to approximate stock risk premia.

variables, momentum indicators, and industry affiliation.

Overall, these results indicate the usefulness of hybrid strategies that combine theory-based and machine learning methods for quantifying stock risk premia. In this respect, the present study complements recent literature that links machine learning with theory-based empirical asset pricing and for which Giglio et al. (2022) provide a comprehensive survey and guide. For example, Gu et al. (2021) note that a focus of machine learning on prediction aspects does not constitute a genuine asset pricing framework, so they propose using a machine learning method (autoencoder) that takes account of the risk-return trade-off directly. Chen et al. (2023) use the results reported by Gu et al. (2020) as a benchmark and find that the inclusion of no-arbitrage considerations improves the empirical performance. In another combination of theory and data science methods, Wang (2018) employs partial least squares to account for higher risk-neutral cumulants when modeling stock risk premia. Kelly et al. (2019) use an instrumented principle components analysis to construct a five-factor model that spans the cross-section of average returns, and Kozak et al. (2020) use penalized regressions to shrink the coefficients on risk factors in the pricing kernel. Bryzgalova et al. (2021) generalize this idea and use decision trees to construct a set of base assets that span the efficient frontier. In their attempt to address the plethora of factors described in recent asset pricing literature, Feng et al. (2020) combine two-pass regression with regularization methods. In what might be considered a broad reality check, Avramov et al. (2023) take a practitioner's perspective and assess the advantages and limitations of the aforementioned approaches.[4]

The remainder of this chapter is structured as follows: Section 2.2 contrasts theory-based and machine learning methodologies for measuring stock risk premia, then outlines ideas to combine them. Section 2.3 explains the construction of the database and the implementation of the respective strategies. Section 2.4 contains a performance comparison between theory-based and machine learning methods at varying horizons and the assessment of the potential of hybrid strategies. Section 2.5 concludes. Appendix A provides details on methodologies, data, and implementation.

2.2 Methodological considerations

2.2.1 Two diverging roads

This section outlines the concepts and key equations associated with the theory-based and machine learning approaches that are the focus of our study. We explain how, from a common starting point, the methodologies to measure stock risk premia diverge. For

[4]Although our study is related to this strand of literature in the general sense of combining financial economic theory with machine learning, our focus is on using this framework for approximating conditional stock risk premia. We do not aim at providing hybrid approaches for the explicit recovery of the stochastic discount factor. Rather, our strategy of using machine learning to deal with the approximation errors inherent to the theory-based approach could be viewed as an exercise in predicting risk-adjusted returns or being related to the notion of boosting.

conciseness, the details of the respective approaches are presented in Appendix A.

The theory-based approach (explicitly) and the machine learning approach (implicitly) take as a point of reference the basic asset pricing equation applied to a gross return of asset i from time t to T ($R^i_{t,T}$) in excess of the gross risk-free rate ($R^f_{t,T}$),

$$\mathbb{E}_t(R^{ei}_{t,T}) = \mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = -R^f_{t,T} \cdot \text{cov}_t(m_{t,T}, R^i_{t,T}), \tag{2.1}$$

where expected values are conditional on time t information. In preference-based asset pricing, the stochastic discount factor (SDF) $m_{t,T}$ represents the marginal rate of substitution between consumption in t and T. In the absence of arbitrage, a positive SDF exists, such that $R^f_{t,T} = \mathbb{E}_t(m_{t,T})^{-1} > 0$. The sign and size of the risk premium, reflected in the conditional expected excess return on asset i, are determined by the conditional covariance on the right-hand side of Equation (2.1).

Theory-/option-based approach

We first take a look down the theory-based route. Using Equation (2.1) as a starting point, we delineate in Section A.1 of the appendix how Martin and Wagner (2019) derive the following reformulation:

$$\mathbb{E}_t(R^{ei}_{t,T}) = R^f_{t,T} \cdot \left\{ \text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2} \cdot \left[\text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right) \right] \right\} + a^i_{t,T}, \tag{2.2}$$

where $R^m_{t,T}$ denotes the return of a market index proxy, w^j_t is the time-varying value weight of index constituent j, var^*_t denotes a conditional variance under the risk-neutral measure, and $a^i_{t,T}$ is a time-varying, asset-specific component that, as shown in Section A.1 of the appendix, is a function of conditional moments either under the risk-neutral or the physical measure. In a similar vein, Kadan and Tang (2020) advocate an even more succinct formula:

$$\mathbb{E}_t(R^{ei}_{t,T}) = \frac{1}{R^f_{t,T}} \cdot \text{var}^*_t(R^i_{t,T}) - \xi^i_{t,T}, \tag{2.3}$$

where $\xi^i_{t,T} = \text{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T})$. In Section A.1 of the appendix, we show how Kadan and Tang (2020) draw on Martin's (2017) derivation of a lower bound for the market equity premium. They argue that, depending on the acceptable level of risk aversion, $\xi^i_{t,T} < 0$ holds for a large fraction of stocks, such that $1/R^f_{t,T} \cdot \text{var}^*_t(R^i_{t,T})$ represents a lower bound for the risk premium.

According to Martin (2017), the risk-neutral variances in Equations (2.2) and (2.3) can be obtained as follows (suppressing the asset index i for notational brevity):

$$\text{var}^*_t\left(\frac{R_{t,T}}{R^f_{t,T}}\right) = \frac{\int_0^{F_{t,T}} \text{put}_{t,T}(K)dK + \int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)dK}{0.5 \cdot S^2_t \cdot R^f_{t,T}}, \tag{2.4}$$

where $\text{call}_{t,T}(K)$ and $\text{put}_{t,T}(K)$ denote the time t prices of European call and put

options, respectively, with strike price K and time to maturity T. Furthermore, S_t is the spot price, and $F_{t,T}$ is the forward price of the underlying asset. The components of the right-hand sides of Equations (2.2) and (2.3), except for the residuals $a^i_{t,T}$ and $\xi^i_{t,T}$, can thus be approximated using a range of option prices at different strikes.[5] For Equation (2.3), these data are only required for asset i. Equation (2.2) is more demanding, in that the option data must be provided for both the market index proxy and its constituents, along with the time-varying index weights. Martin and Wagner (2019) argue that the consequences of setting $a^i_{t,T} = 0$ should be benign, such that stock risk premia can be quantified without the need to estimate any unknown parameters, by using:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx R^f_{t,T}\left\{\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\cdot\left[\text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right]\right\}. \tag{2.5}$$

Similarly, assuming that the negative correlation condition holds and that the lower bound in Equation (2.3) is binding, Kadan and Tang's (2020) approximative formula for the risk premium on stock i is given by:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \frac{1}{R^f_{t,T}} \cdot \text{var}^*_t(R^i_{t,T}). \tag{2.6}$$

Machine learning approach

Recalling that the conditional expectation is the best predictor in terms of the MSE, Equation (2.1) states that the optimal forecast of $R^{ei}_{t,T}$ is given by $-R^f_{t,T}\cdot\text{cov}_t(m_{t,T}, R^i_{t,T})$. Because the functional form of the conditional covariance is not known, one can treat $-R^f_{t,T}\cdot\text{cov}_t(m_{t,T}, R^i_{t,T})$ as a function that depends on state variables $z^i_t \in \mathcal{F}_t$, such that

$$\mathbb{E}_t(R^{ei}_{t,T}) = g^0_T(z^i_t), \tag{2.7}$$

where the subindex T indicates dependence on the horizon of interest. The machine learning approach then proceeds to approximate $g^0_T(z^i_t)$ by $g_T(z^i_t, \theta_T)$, a parametric function implied by some statistical model with a parameter vector θ_T to be estimated. The estimation of θ_T using machine learning procedures (MLPs) instead of standard econometric methods may be advocated for the following reasons.

First, there are a lot of candidates for the state variables z^i_t. A myriad of stock- and macro-level return predictive signals (*features* in machine learning terms) appear in empirical finance literature, and dimension reduction and feature selection are the very domain of MLPs. Second, the suite of statistical models employed for MLPs trade analytical tractability and rigorous statistical inference for flexible functional forms and predictive performance. The prediction implications of the basic asset pricing equation (2.1) naturally establish a learning objective, that is, minimization of the forecast MSE. However, the combination of these two issues – many features and

[5] Details on the approximation can be found in Section A.3 of the appendix.

a desire for flexibility – creates a vast risk of overfitting. To deal with this concern, MLPs divide the data into a training, a validation, and a test sample and introduce regularization in the estimation process. Regularization is controlled by the tuning of hyperparameters, which might take the form of a penalty applied to the learning objective, early stopping rules applied to its optimization, or, more generally, coefficients that determine the complexity of the statistical model (e.g., number of layers in an ANN). Using a given combination of hyperparameters, the parameter vector θ_T is estimated on the training sample, and the model performance gets evaluated, in terms of the MSE, on the validation sample. A search across hyperparameter combinations ultimately points to the specification that delivers the best performance. Using the hyperparameter combination thus selected, θ_T is re-estimated on the merged training/validation sample. The result is the final estimated model, $g_T(z_t^i, \hat{\theta}_T)$, which is used as a machine learning-implied approximative risk premium,

$$\mathbb{E}_t(R_{t,T}^{ei}) \approx g_T(z_t^i, \hat{\theta}_T). \tag{2.8}$$

Machine learning encompasses a variety of statistical models that offer flexible approximations of $g_T^0(z_t^i)$. In this study, we consider an ENet, GBRT, RF, and ANN. We discuss the associated hyperparameter configurations in Section 2.3.2.

2.2.2 Pros and cons

As far as the empirical implementation is concerned, the theory-based and data science approaches have their own unique pros and cons.

Parameter estimation and approximation errors

Using the theory-based formulas in Equation (2.5) or (2.6) and working under the risk-neutral measure, one can dispense with the estimation of unknown model parameters altogether. However, this parsimony of the theory-based approach comes at the cost of approximation errors, the practical consequences of which are not quite clear. In contrast, the machine learning approach deals with a huge number of parameters, which must be estimated without the risk of overfitting.

Time-varying parameters

A conspicuous feature of the theory-based approach is that it can deal naturally with changing conditional distributions and even non-stationary data. The machine learning approach, like any statistical/econometric method, struggles more with ensuing problems like an incidental parameter problem that would occur if the parameters in θ_T were time-varying. This caveat can be accounted for by employing a dynamic procedure, in which the training sample is gradually extended and the validation and test sample are shifted forward in time. (Hyper-)parameter estimation is performed for each of these "sample splits." Compared with Equation (2.8), it is thus notationally

more precise, albeit more cluttered, to write

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx g_{s,T}(z^i_t, \hat{\theta}_{s,T}), \tag{2.9}$$

indicating the dependence of the functional form and estimates on the sample split s and investment horizon T.

Data quality and computational resource demands

The demands for data quality and quantity in both the theory-based and machine learning strategies are considerable, distinct, and complementary. The machine learning approach needs historical data on stock-level predictors for every asset of interest. A critical aspect is that these data suffer from a missing value problem that is most severe in the more distant past. As pointed out by Freyberger et al. (2022), the imputation of those observations is not innocuous and may hamper the application of data-intensive machine learning methods. This issue is mitigated using theory-based approaches. However, both MW and KT require high quality option data. In particular, for the option prices, the times-to-maturity must match (at least approximately) the horizons of interest, and only a sufficiently large number of strike prices K can provide a good approximation of the integrals in Equation (2.4). Moreover, Equation (2.5) reveals that these data are required for not only the stocks of interest but also every member of the market index, as well as the index itself.

An advantage of the option-based approaches is that the computational resources needed to provide quantifications of stock risk premia are moderate. Machine learning approaches instead mandate ready access to considerable computing power. Training and hyperparameter tuning are required for each statistical model, for each horizon of interest, and for every new test sample.

2.2.3 Hybrid approaches

Because of the diversity of their respective pros and cons, it is intriguing to combine the theory-based and machine learning philosophies. Our primary hybrid approach is based on MW; it starts from Equation (2.2) and the approximative formula in Equation (2.5) and then employs machine learning to account for the approximation residuals $a^i_{t,T}$.[6] Let us use $\widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ to denote the right-hand side of Equation (2.5). Then $\widetilde{R}^{ei}_{t,T} = R^{ei}_{t,T} - \widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ gives the component of the excess return left unexplained by MW. Provided that the aforementioned data requirements are met, $\widetilde{R}^{ei}_{t,T}$ can be computed for every i, t, and T. Emphasizing the prediction aspect of the basic asset pricing equation, we consider the following decomposition:

$$\widetilde{R}^{ei}_{t,T} = a^i_{t,T} + \varepsilon^i_{t,T}, \tag{2.10}$$

[6] Alternatively, we could also use KT as a starting point, but MW is arguably more appropriate for a larger number of stocks.

where $\varepsilon^i_{t,T} = R^{ei}_{t,T} - \mathbb{E}_t(R^{ei}_{t,T})$ can be conceived of as the irreducible idiosyncratic forecast error. We can now apply the MLPs not to $R^{ei}_{t,T}$ and $\mathbb{E}_t(R^{ei}_{t,T})$ but rather to $\widetilde{R}^{ei}_{t,T}$ and $a^i_{t,T}$. This is a sensible approach because the approximation residual $a^i_{t,T}$ is a function of time t conditional moments, as is shown in Section A.1 of the appendix. Similar to the treatment of $g^0_T(z^i_t)$ in Equation (2.7), we can represent $a^i_{t,T}$ as a function of the time t state variables z^i_t, such that $a^i_{t,T} = h^0_T(z^i_t)$, and use a statistical model with parameters ϑ_T to approximate $h^0_T(z^i_t) \approx h_T(z^i_t, \vartheta_T)$.

The machine learning-style estimation of the parameters ϑ_T entails minimizing the MSE associated with the forecast error $\widetilde{R}^{ei}_{t,T} - h_T(z^i_t, \vartheta_T)$ instead of $R^{ei}_{t,T} - g_T(z^i_t, \theta_T)$. The hybrid risk premium quantification is then given by:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_T(z^i_t, \hat{\vartheta}_T), \tag{2.11}$$

which yields the familiar decomposition:

$$R^{ei}_{t,T} - \underbrace{(\widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_T(z^i_t, \hat{\vartheta}_T))}_{\text{hybrid forecast}} = \underbrace{(a^i_{t,T} - h_T(z^i_t, \vartheta_T))}_{\text{approximation error}} + \underbrace{(h_T(z^i_t, \vartheta_T) - h_T(z^i_t, \hat{\vartheta}_T))}_{\text{estimation error}} + \varepsilon^i_{t,T}. \tag{2.12}$$

To account for time-varying model parameters, the dynamic hyperparameter tuning described in Section 2.2.3 can be applied in the same way, which yields the following hybrid approximative formula for the stock risk premium:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_{s,T}(z^i_t, \hat{\vartheta}_{s,T}). \tag{2.13}$$

Neither the theory-based ("Econ") nor the machine learning ("Metrics") approach would be described as econometrics, the discipline founded to connect economic theory and statistics. Yet, the formula in Equation (2.13) may be seen as a novel way to combine Econ and Metrics in the modern age of data science. We refer to this hybrid strategy as *theory assisted by machine learning*.

An obvious alternative hybrid strategy is motivated by the observation that though GKX include a plethora of stock-level and macro features, they do not use the information provided by the theory-based risk premium measures, or any other conditional time t moment computed under the risk-neutral measure. By augmenting the set of features accordingly, we can assess whether the theory-based measurements enhance the explanatory power of the data science approach. We refer to this hybrid approach as *machine learning with theory features*.

A central tenet of financial economics, derived from Equation (2.1), states that marginal utility-weighted prices follow martingales. This tenet implies that return predictability should be a longer-horizon phenomenon. High frequency price processes are expected to behave like martingales, such that the MSE-optimal return prediction at very short horizons should be close to the zero forecast (cf. Cochrane (2005), Section 2.4). The signal-to-noise ratio – $\mathbb{E}_t(R^{ei}_{t,T})$ to $\varepsilon^i_{t,T}$ – is expected to increase at longer forecast horizons. So, the empirical question that we seek to address refers to which

of the approaches – theory-based, machine learning or hybrid – delivers a better approximation of $\mathbb{E}_t(R^{ei}_{t,T})$, i.e. a superior out-of-sample performance, at given horizons. To answer this question we need a comprehensive database.

2.3 Data, implementation, and performance assessments

2.3.1 Assembling the database

Selection of stocks and linking databases

The universe of stocks for which we compare the alternative risk premium measures is defined by a firm's membership in the S&P 500 index.[7] One reason to choose this criterion is that if we want to compute theory/option-based risk premia according to Equation (2.5), we have to provide information about the constituents of the market index proxy. Because the S&P 500 is used for that purpose, index membership is the obvious criterion to select the cross-section of stocks considered for our analysis. For the identification of historical S&P 500 constituents (HSPC) across databases, we start by extracting information about a firm's S&P 500 membership status from Compustat. We thereby obtain, for every month from March 1964 to December 2018, a list of HSPC. In total, we find 1,675 firms that have been in the S&P 500 for at least one month. For the HSPC identified in Compustat, we retrieve price and return data from CRSP. Compustat and CRSP also supply the data used for the machine learning approaches. The option data, which are required to compute the theory-based measures, come from OptionMetrics. Section A.2 in the appendix explains in detail how we link the three databases and documents the quality of the matching procedure.

Stock-level and macro features

Following GKX, we retrieve from Compustat and CRSP 93 firm-level variables that have been identified as predictors for stock returns in previous literature. We also construct 72 binary variables that identify a firm's industry (see Table A.1 in the appendix).[8] A cross-sectional median-based imputation is applied to deal with missing observations.[9]

[7]Each company in the S&P 500 may be associated with multiple securities. An S&P 500 constituent is a specific company-security combination, but we refer to them, as is common in the literature, interchangeably as "securities," "stocks" or "firms."

[8]For that purpose, we adapt the SAS program from Jeremiah Green's website, https://sites.google.com/site/jeremiahrgreenacctg/home, accessed January 20, 2020. The industry indicators are based on the first two digits of the standard industrial classification (SIC) code.

[9]Median-based imputation is frequently applied in related literature. However, Bryzgalova et al. (2022) point out that firm characteristics are typically not missing at random, rendering median-based imputation problematic. They propose an alternative approach that exploits cross-sectional and time series dependencies between characteristics to impute missing values. For their empirical analysis Bryzgalova et al. (2022) use a sample that comprises more than 22,000 stocks (including penny stocks) and starts in 1967. Missing data occur particularly often at the beginning of the sample and for small firms. Being aware of the missing value issue, we do not follow GKX, who use

We consider two types of transformation for firm-level fetures: standard mean-variance and median-interquartile range scaling, the latter being more robust in the presence of outliers. The choice of the scaling procedure (standard or robust) is treated as a hyperparameter.[10] In either case, we make sure that no information from the future enters the validation or tests sets in order to prevent a look-ahead bias. The stock-level features are augmented by macro-level variables, obtained from Amit Goyal's website.[11] These variables are the market-wide dividend-price ratio, earnings-price ratio, book-to-market ratio, net equity expansion, stock variance, the Treasury bill rate, term spread, and default spread. Their detailed definitions can be found in Welch and Goyal (2008).

The variables retrieved have a mixed frequency: monthly (20 stock-level + 8 macro-level variables), quarterly (13 stock-level variables), or annual (60 stock-level variables). Using the date of the last trading day of each month as a point of reference, they are aligned according to Green et al.'s (2017) assumptions about delayed availability to avoid any forward-looking bias. Features at the monthly frequency are delayed at most one month, quarterly variables by at least a four-month lag, and annual variables by at least a six-month lag. Moreover, we match CRSP returns at horizons of one month (30 calendar days) and one year (365 calendar days), such that they are forward-looking from the vantage point of the end-of-month alignment day.

A considerable number of missing values for stock-level features arise, if we go further back in time than the mid-1970s. To mitigate the aforementioned negative consequences associated with massively imputing missing values, we start using the data in October 1974, when the problem is alleviated. Moreover, two of the originally 93 stock-level features retrieved are excluded, because they contain an excessive amount of missing values. Figure 2.1 shows a heatmap that illustrates how the share of missing values of stock-level features changes over time.

The out-of-sample analysis is performed for the period from January 1996, the starting date of OptionMetrics, until December 2018. Proceeding as described, we obtain an unbalanced panel data set at a monthly frequency that ranges from October 1974 until December 2018. The number of HSPC during that period is 1,145, with a varying number of observations per stock. In total, there are 362,306 stock/month observations.

Option data

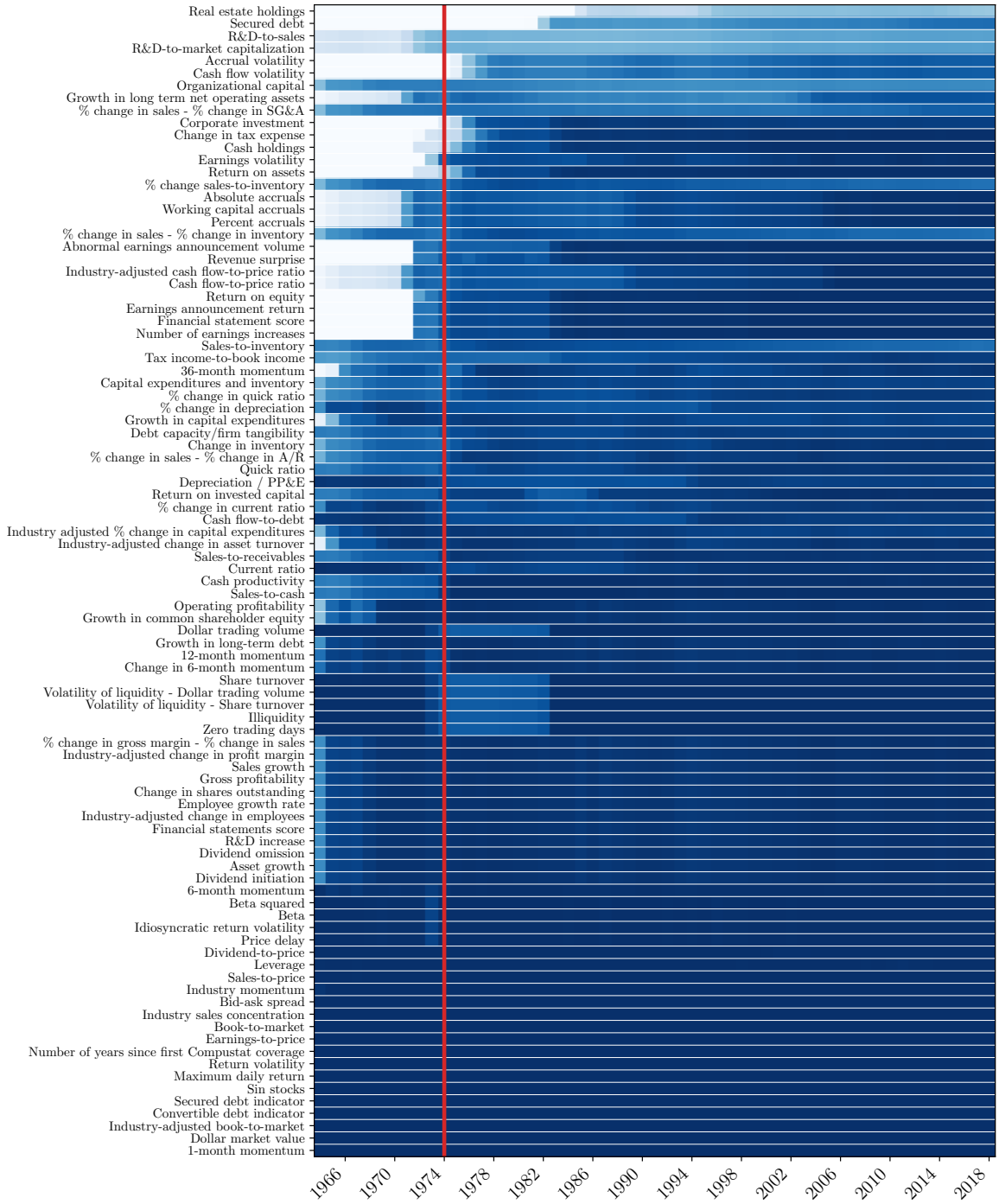
The data to implement the option-based risk premium formulas in Equations (2.5) and (2.6) are retrieved from OptionMetrics. Two issues must be resolved in the process.

data from the late 1950s, but instead commence the training process in 1974. Focusing on HSPC, which are large firms by constructions, further mitigates the problem of missing values.

[10] Here we deviate from GKX, who achieve outlier robustness by applying a cross-sectional rank transformation and re-scaling the stock-level features to the interval -1 to 1. Various studies (e.g., Da et al., 2022 and Kelly et al., 2019) report that their results do not critically depend on the choice of scaling. To assess whether this conclusion also holds true in our setting, Section A.7 of the appendix reports the results of a robustness check, in which the empirical analysis is conducted with rank-transformed features.

[11] See http://www.hec.unil.ch/agoyal, accessed January 20, 2020.

Figure 2.1: Proportion of non-missing observations for each stock-level feature and year. This figure illustrates, for each of the stock-level features used in the machine learning approaches, the proportion of non-missing firm-date observations per year. The sample period ranges from 1964 to 2018, and the features are sorted from top to bottom in ascending order, according to their average proportion of non-missing observations. The darker the color, the more observations are available. The lighter the color, the less observations are available. All white indicates 100% missing values, the darkest blue means no missing values. The red vertical line indicates the year 1974, which is the first year that we use in the long training scheme described in Figure 2.2. Because of the excessive amount of missing values, we exclude the variables *real estate holdings* and *secured debt* from the empirical analysis.



First, options on S&P 500 stocks are American options, yet the computation of risk-neutral variances according to Equation (2.4) relies on European options. Second, a continuum of strike prices is not available, so the integrals in Equation (2.4) must be approximated using a grid of discrete strikes. As pointed out by Martin (2017), a lack of a sufficient number of strikes may severely downward bias the computation of risk-neutral variances. Martin and Wagner (2019) advocate for the use of the OptionMetrics volatility surface to address these issues and compute risk-neutral variances according to Equation (2.4). Although European options are traded on the S&P 500 index, and their prices are available in OptionMetrics, we also rely on the volatility surface to compute risk-neutral index variances. Using the OptionMetrics volatility surface, we compute the theory-based risk premium measures for the selected stocks and the two horizons of interest. These data are matched, by their security identifier and end-of-month date, with the aforementioned unbalanced panel. A detailed explanation of our use of the volatility surface is provided in Section A.3 of the appendix.

Risk-free rate proxies

To compute excess returns and all of the option-based measures, we need a risk-free rate proxy that matches the investment horizon. It can be computed for different horizons at a daily frequency using the zero curve provided by OptionMetrics. However, like any data supplied by OptionMetrics, the zero curve is not available before January 1996. We therefore employ the Treasury bill rate as a risk-free rate proxy for earlier periods.

2.3.2 Empirical implementation

In the following we provide information about the hyperparameter configurations of the statistical models, the construction of the vector of state variables z_t^i, and the long and short training schemes.

As mentioned previously, our machine learning approaches employ four popular statistical models: the ANN, RF, GBRT, and ENet. The first three were identified by GKX as the most appropriate for the task at hand. The ENet is included as an instance of penalized regression because of the less demanding hyperparameter tuning.[12] The hyperparameter configurations for these models are listed in Table 2.1.

The selection of features collected in the vector z_t^i follows GKX, such that we use the 91 stock-level variables (included in the vector c_t^i) and their interactions with the eight macro predictors (included in the vector x_t). Formally, z_t^i is comprised of the vector $(1, x_t')' \otimes c_t^i$, augmented with industry dummies, such that altogether we have $91 \times 9 + 72 = 891$ features.[13]

The implementation of the sequential validation procedure mentioned in Section

[12] We assume that the reader has some familiarity with these approaches, which are covered by Hastie et al. (2017).

[13] In principle, it would also be possible to explicitly consider the time series of macroeconomic variables, as proposed by Chen et al. (2023). In line with GKX, we choose to focus on the last observation of these series instead.

Table 2.1: Hyperparameter search space. This table shows the hyperparameter search space and the Python packages used for both long and short training. Parameter configurations not listed here correspond to the respective default settings.

Panel A: ENet	Panel B: RF
Package: Scikit-learn (SGDRegressor)	*Package:* Scikit-learn (RandomForestRegressor)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: ℓ_1-ℓ_2-penalty: $\{x \in \mathbb{R} : 10^{-5} \leq x \leq 10^{-1}\}$ ℓ_1-ratio: $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$	*Model parameters:* Number of trees: 300 Max. depth: $\{x \in \mathbb{N} : 2 \leq x \leq 30\}$ Max. features: $\{x \in \mathbb{N} : 2 \leq x \leq 150\}$
Optimization: Stochastic gradient descent Tolerance: 10^{-4} Max. epochs: $1,000$ Learning rate: $10^{-4}/t^{0.1}$	
Random search: Number of combinations: $1,000$	*Random search:* Number of combinations: 500

Panel C: GBRT	Panel D: ANN
Package: Scikit-learn (GradientBoostingRegressor)	*Package:* Tensorflow/Keras (Sequential)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: Number of trees: $\{x \in \mathbb{N} : 2 \leq x \leq 100\}$ Max. depth: $\{x \in \mathbb{N} : 1 \leq x \leq 3\}$ Max. features: $\{20, 50, \text{All}\}$ Learning rate: $\{x \in \mathbb{R} : 5 \times 10^{-3} \leq x \leq 1.2 \times 10^{-1}\}$	*Model parameters:* Activation: TanH (Glorot), ReLU (He) Hidden layers: $\{1, 2, 3, 4, 5\}$ First hidden layer nodes: $\{32, 64, 128\}$ Network architecture: Pyramid Max. weight norm: 4 Dropout rate: $\{x \in \mathbb{R} : 0 \leq x \leq 0.5\}$ ℓ_1-penalty: $\{x \in \mathbb{R} : 10^{-7} \leq x \leq 10^{-2}\}$
	Optimization: Adaptive moment estimation Batch size: $\{100, 200, 500, 1,000\}$ Learning rate: $\{x \in \mathbb{R} : 10^{-4} \leq x \leq 10^{-2}\}$ Early stopping patience: 6 Max. epochs: 50 Batch normalization before activ. Number of networks in ensemble: 10
Random search: Number of combinations: 300	*Random search:* Number of combinations: $1,000$

2.2.1 is illustrated in Figure 2.2 (long training scheme). It shows that the length of the training period increases from 10 years initially to 31 years; the 12-year validation period shifts forward by one year with every new test sample. There are S=22 out-of-sample years with the final one-year predictions made in December 2017 for December 2018. For every sample and statistical model, hyperparameter tuning is performed at the one-month and one-year forecast horizon. When considering the one-month horizon, the number of test samples increases to S=23, because monthly forecasts are possible during the year 2018. Details on the hyperparameter tuning are provided in Section A.5 of the appendix.[14]

The basic setup remains the same when considering the hybrid approaches. However, the training and validation procedure changes because of the delayed availability of the OptionMetrics data beginning January 1996. We therefore consider the alternative, short training scheme illustrated in Figure 2.3; it is used for the *theory assisted by ML* and *ML with theory features* strategies. The short training scheme reduces the initial training period to one year and the validation period comprises 1 year instead of 12. With this configuration, we can retain a sufficiently large number of out-of-sample years, comparable to the long training scheme.

To establish a benchmark for the performance of the hybrid approaches, we also train the models using the original feature set and the short training scheme. A comparison with the long training results is interesting for another reason too: It allows us to study how important the length of the training period is and to assess the effect of the length of the validation period.

2.3.3 Performance assessments

We compare the alternative approaches to measure stock risk premia by assessing their out-of-sample forecast performance. This represents a useful criterion, because the different methodologies provide approximations of the conditional expected excess return, which is the MSE-optimal prediction. The smaller the MSE, the better the approximation of the stock risk premium. We consider forecasts with horizons of one month (30 calendar days) and one year (365 calendar days), issued at an end-of-month and daily frequency, respectively.

Following Welch and Goyal (2008), we rely on a performance measure that relates the MSE of a model's out-of-sample forecast to that of a benchmark. We use the zero forecast for that purpose, which has the appeal of providing a parameter-free alternative and comparability across studies. More specifically, the performance criterion is the pooled predictive R^2 given by:

$$R^2_{oos} = 1 - \frac{\sum_t \sum_i \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2}{\sum_t \sum_i \left(R^{ei}_{t,T}\right)^2}, \tag{2.14}$$

[14] While our implementation of the machine learning approaches draws on GKX, it deviates in some respects. Section A.6 in the appendix provides a detailed juxtaposition.

Figure 2.2: Long training scheme. The figure depicts the annual horizon variant of the long training scheme. The data range from October 1974 to December 2017. The training period (red/dark grey) initially spans 10 years and increases by one year after each validation step. Each of the 22 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers 12 years and is rolled forward with a fixed width, followed by one year of out-of-sample testing (checkered blue).

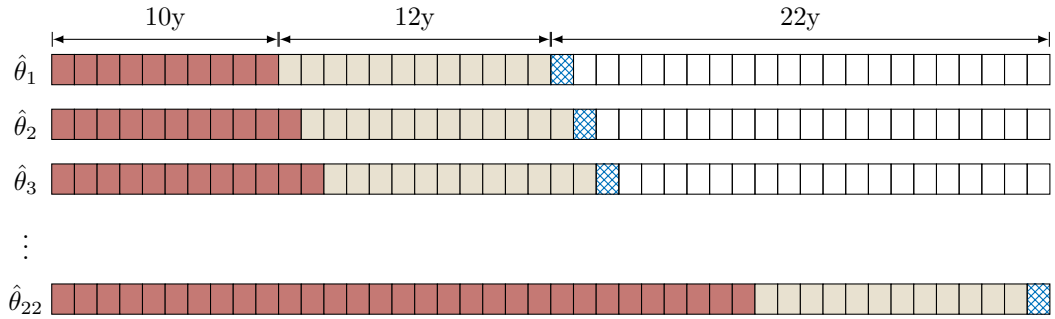
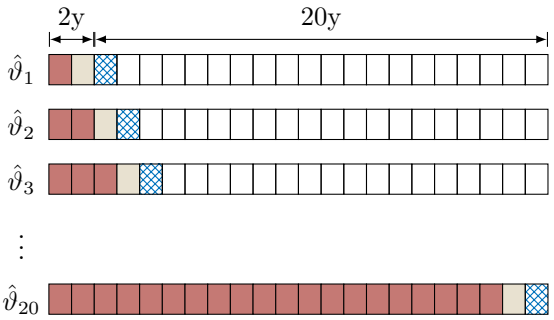


Figure 2.3: Short training scheme. The figure depicts the annual horizon variant of the short training scheme. The data range from January 1996 to December 2017. The training period (red/dark grey) initially spans one year and increases by one year after each validation step. Each of the 20 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers one year, followed by one year of out-of-sample testing (checkered blue).



where $\hat{R}^{ei}_{t,T}$ denotes the respective forecast/risk premium estimate. The calculation is based solely on observations included in the S test sample years that were not used for training or validation.

To study performance over time, we also compute the predictive R^2 for each of the test samples separately:

$$R^2_{oos,s} = 1 - \frac{\sum_i \sum_t \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]}{\sum_i \sum_t \left(R^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]} \qquad s = 1, 2, \ldots, S, \tag{2.15}$$

where $\mathcal{S}(s)$ denotes the set of time indices of forecast sample s, such that $\mathbb{1}[t \in \mathcal{S}(s)]$ is equal to 1 if the observation in t belongs to the sample year s, and 0 otherwise. For the assessment of statistical significance, we report the p-values associated with a test whether a model has no explanatory power over the zero forecast; formally, the null hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$. To construct a convenient test statistic, we take the mean of the $R^2_{oos,s}$ across the test samples, $\overline{R^2_{oos}} = \frac{1}{S}\sum_{s=1}^{S} R^2_{oos,s}$, and compute its standard error $\hat{\sigma}(\overline{R^2_{oos}})$, using a Newey-West correction to account for serial correlation. Provided that a central limit theorem applies, and assuming that $\mathbb{E}(R^2_{oos,s}) = 0$, the t-statistic $\overline{R^2_{oos}}/\hat{\sigma}(R^2_{oos})$ is approximately standard normally distributed, such that a one-sided p-value can be provided.[15]

As an alternative to the R^2_{oos} in Equation (2.14), we also consider the time-series R^2 used by Chen et al. (2023), which accounts for the fact that the number of stocks in period t (N_t) can change over time:

$$EV_{oos} = 1 - \frac{\sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2}{\sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} \left(R^{ei}_{t,T}\right)^2}. \tag{2.16}$$

As this study is ultimately concerned with approximating stock risk premia, both the level and cross-sectional properties of the excess return predictions should be taken into account for performance assessment. However, the R^2_{oos} can be dominated by the forecast error in levels, potentially masking the cross-sectional explanatory power of a model. To explicitly account for this dimension of return predictability, we use the following measures: First, we compute a cross-sectional out-of-sample R^2 similar to those advocated by Maio and Santa-Clara (2012) and Bryzgalova et al. (2021):

$$XS_{oos} = 1 - \frac{\mathrm{Var}_N(\overline{\hat{\varepsilon}^i_T})}{\mathrm{Var}_N(\overline{R^{ei}_T})}, \tag{2.17}$$

where $\mathrm{Var}_N(\cdot)$ stands for the cross-sectional variance across the N sample stocks; $\overline{\hat{\varepsilon}^i_T}$ and $\overline{R^{ei}_T}$ are the stock-specific time-series averages of $R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}$ and $R^{ei}_{t,T}$, respectively. Second, we assess cross-sectional performance by forming decile portfolios based on

[15] The Diebold-Mariano test employed by GKX to gauge differences in forecast performances is constructed in a similar vein. We provide p-values associated with this test in Section A.8 of the appendix.

the respective model's excess return predictions and comparing predicted and realized mean excess returns across approaches. If an approach delivers sensible risk premium estimates then a) the mean predicted excess returns and mean realized excess returns of the prediction-sorted portfolios should align, and b) there should be sizable variation in the mean realized excess returns across these portfolios. Besides graphical assessments and rank correlations, we also compare the annualized Sharpe ratios of zero-investment portfolios long in the decile portfolio of stocks with the highest excess return prediction and short in that with the lowest. The Sharpe ratio accounts for the desideratum that the cross-sectional differentiation of the mean realized excess returns should be achieved by a small variation over the years of the test sample.

The machine learning models are trained on data at a monthly frequency. Accordingly, the respective excess return forecasts are updated once at the end of each month. Forecasts at these same dates are also available using the option-based approaches, which additionally can provide risk premium estimates at higher frequencies, up to daily. To facilitate comparisons at a daily frequency, we retain the most recent ML-based risk premium estimate until an update becomes available by the end of the next month. For example, the estimate of an annual horizon stock risk premium in mid-April 2015 corresponds to the last available estimate calculated at the end of March 2015. For the *ML with theory features* strategy, the hybrid model's daily estimate employs the statistical model (trained on monthly data) endowed with the prevailing end-of-month firm- and macro-level features and daily updated theory-based measures. Similarly, the adaption of the *theory assisted by ML* approach combines the theory-based daily risk premium estimate with the prevailing end-of-month ML-based residual approximation.

2.4 Empirical results

2.4.1 Comparison at monthly and annual horizons

One-month horizon

Table 2.2 contains the results for the one-month horizon; in Panel A, the forecasts are issued at a daily frequency, whereas in Panel B, they are issued monthly (end-of-month). Among the machine learning approaches in Panel B, only the ANN achieves a positive predictive R^2 (0.2%); the same R^2_{oos} is delivered by the theory-based MW.[16] Evaluating the daily MW forecasts, we find that the predictive R^2 increases to 0.9%, which represents the only instance in which we can reject the hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$ at significance levels below 5%. For a daily forecast frequency, the ANN

[16]To avoid a cluttered exposition, we focus in the main text on reporting and interpreting the R^2_{oos} results. Section A.8 in the appendix includes extended tables that also report XS_{oos} and EV_{oos}. It can be seen that R^2_{oos} and EV_{oos} take on very similar values, and while the level of XS_{oos} is somewhat smaller, its pattern across approaches corresponds to that of R^2_{oos}. Accordingly, the conclusions obtained by using the alternative performance measures remain the same.

achieves an R^2_{oos} of 0.5%, the highest among the machine learning approaches.[17]

The comparatively good performance of the theory-based approach is corroborated by a complementary analysis based on the data that Chabi-Yo et al. (2023) used to introduce their alternative option-based risk premium estimate, and which contain their estimates at the one-month and one-year horizons.[18] Although the universe of stocks is different, there is an overlap with our study. When we conduct an analysis at the intersection of firms and dates, it yields a monthly R^2_{oos} of 1% implied by Chabi-Yo et al.'s (2023) method (daily forecast frequency). For this merged sample, the predictive R^2 produced by MW remains unchanged (0.9%); the R^2_{oos} of the machine learning approaches do not improve.

The relative advantages of the theory-based paradigm are also evident in Figure 2.4. Panel A (monthly forecast frequency) and conspicuously Panel B (daily) both show that MW yields a better alignment of the prediction-sorted portfolios. The rank correlation between mean predicted and mean realized excess returns is 0.96, whereas that implied by the ANN is 0.56 (monthly forecast frequency). Figure 2.4 also shows that the variation of the mean realized excess returns across prediction-sorted portfolios is favorably wider using MW than the variation implied by the ANN. This result is reflected in the Sharpe ratios of the zero investment portfolios (cf. Table 2.2), which are 0.30 (monthly forecast frequency) and 0.37 (daily) for MW, compared with 0.28 (monthly) and 0.26 (daily) for the ANN.[19] Overall, these findings indicate that at the one-month horizon, care is needed when investing in machine learning-based methods; their superiority over the theory-based paradigm is by no means a given.

An alternative conclusion might refer to the sample period and universe of stocks, for which the task at hand might be more difficult for machine learning. Compared with GKX, we consider fewer stocks for training and validation, and the training begins in a later year, both of which are factors that could prevent the machine learning approaches from reaching their full potential.

One-year horizon

Most of these concerns can be alleviated by a review of Table 2.3, which shows the results for the one-year horizon. Contrasting Panels A and B, we observe that it matters little whether we use daily or monthly forecasts, so we simply focus on the latter in the following discussion.

Compared with the one-month horizon results, the annual predictive R^2 increase by an order of magnitude; the R^2_{oos} delivered by MW is about 9%. The results in Table 2.3 mitigate any concerns that the present selection of stocks constitutes a more

[17] A monthly predictive R^2 of about 1% may appear small, but it is actually higher than any reported by GKX. Their ANNs yield monthly predictive R^2 between 0.3% and 0.7%, depending on the universe of stocks and ANN architecture.

[18] We are grateful to Grigory Vilkov for providing access to these data.

[19] Tables 2.2 and 2.3 also show that, in terms of predictive R^2, KT is less successful. Yet, regarding prediction-sorted portfolios, KT and MW are equivalent. Both achieve cross-sectional differentiation through risk-neutral variances $\text{var}^*_t(R^i_{t,T})$. Thus, the prediction-sorted portfolios include the same stocks and yield the same mean realized excess returns and Sharpe ratios.

Table 2.2: Performance comparison, one-month horizon: long training. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-month horizon forecasts are issued at a daily frequency. For Panel B, the one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in November 2018. The machine learning results are obtained using the long training scheme depicted in Figure 2.2.

Panel A: daily forecast frequency

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.9	2.3	0.008	0.37
	KT	−0.5	5.3	0.530	0.37
Machine Learning	ENet	0.0	2.9	0.072	0.07
	ANN	0.5	3.1	0.038	0.26
	GBRT	0.3	2.9	0.036	0.29
	RF	−0.5	3.8	0.215	0.15

Panel B: monthly forecast frequency

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.2	3.2	0.154	0.30
	KT	−1.8	6.9	0.704	0.30
Machine Learning	ENet	−0.3	3.5	0.161	0.00
	ANN	0.2	3.5	0.096	0.28
	GBRT	−0.6	4.2	0.248	0.20
	RF	−1.6	5.2	0.435	0.13

Figure 2.4: Prediction-sorted portfolios, one-month horizon: long training. The stocks are sorted into deciles according to the one-month horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot the predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to November 2018. Machine learning results are based on the long training scheme depicted in Figure 2.2.

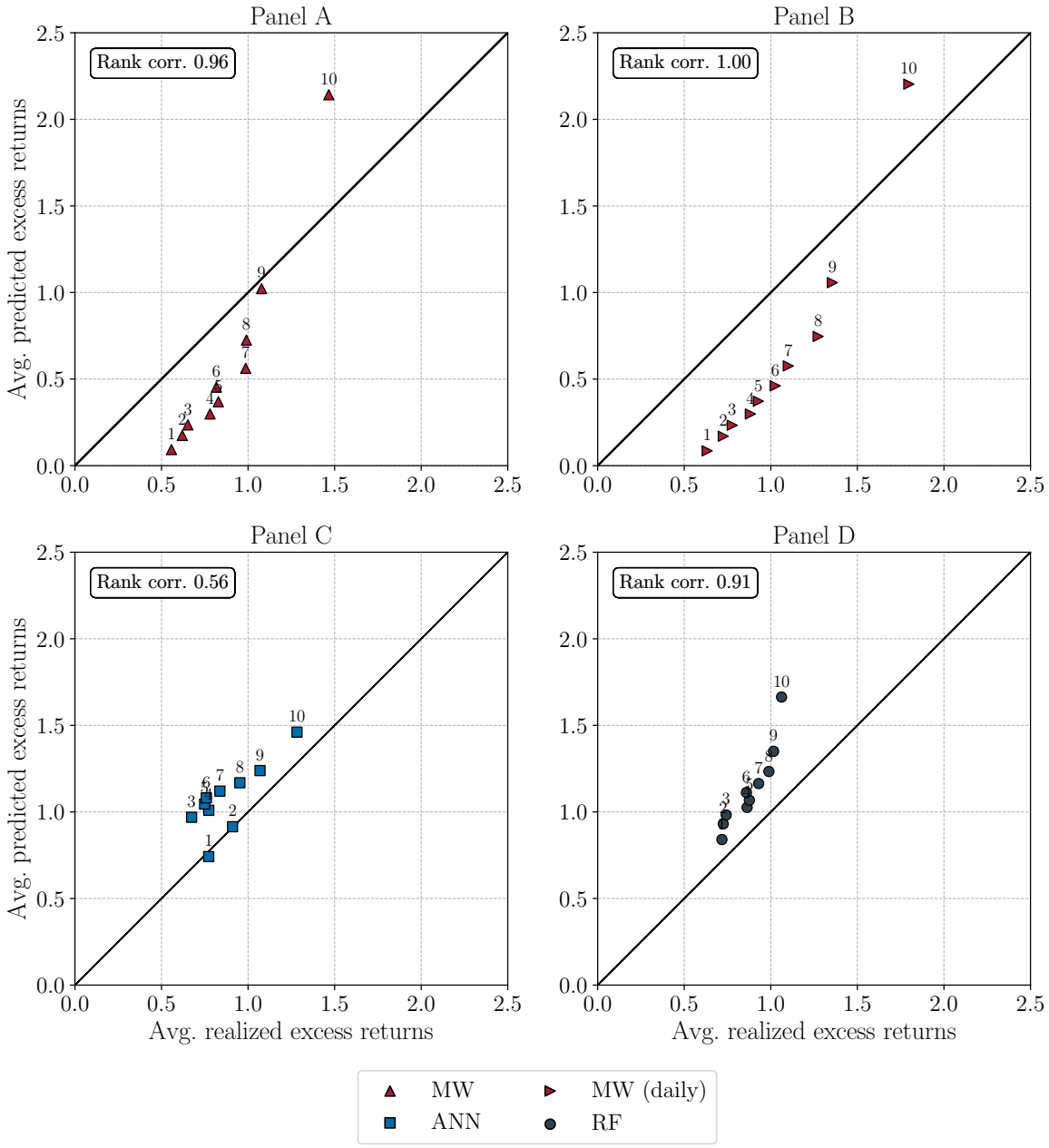


Table 2.3: Performance comparison, one-year horizon: long training. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized SR (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-year horizon forecasts are issued at a daily frequency. For Panel B, the one-year horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in December 2017. The machine learning results are obtained using the long training scheme depicted in Figure 2.2.

Panel A: daily forecast frequency

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.1	16.0	0.040	0.38
	KT	3.5	47.5	0.675	0.38
Machine Learning	ENet	4.0	19.5	0.201	0.35
	ANN	8.2	17.6	0.029	0.49
	GBRT	9.9	19.9	0.039	0.36
	RF	18.2	22.6	0.003	0.56

Panel B: monthly forecast frequency

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	8.8	16.3	0.051	0.37
	KT	3.1	47.6	0.694	0.37
Machine Learning	ENet	5.5	18.5	0.125	0.36
	ANN	9.0	19.0	0.028	0.50
	GBRT	10.6	20.5	0.035	0.36
	RF	19.5	23.6	0.002	0.58

difficult environment for machine learning approaches or that their training is flawed. For example, the ANN achieves an annual R^2_{oos} notably higher than those reported by GKX.[20] Furthermore, MW, GBRT, and the ANN perform comparably well, with R^2_{oos} ranging between 8.8% and 10.6% and p-values for the hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$ ranging from 3.5% to 5.1%.[21] Notably smaller predictive R^2 and higher p-values are implied by the ENet and KT; that is, not all option-based and machine learning approaches perform equally well.

In terms of predictive R^2, the RF stands out, delivering an annual R^2_{oos} of 19.5% with a p-value of 0.2%. The good RF results are confirmed by the favorable alignment and cross-sectional variation in realized mean excess returns of the prediction decile portfolios (cf. Panel D of Figure 2.5), and the highest Sharpe ratio of the long-short portfolio among the approaches considered. We thus conclude that at the one-year horizon, there exists a machine learning method that offers a comparative advantage over the theory-based approach.[22]

Time-series variation

The time-series variation of the predictive R^2 is illustrated in Figure 2.6. In Panel A, we present a comparison of MW with the random forest, the best-performing machine learning method; the other approaches are in Panel B. The $R^2_{oos,s}$ values depicted in Figure 2.6 refer to the year the forecast was issued. For example, the annual predictive R^2 associated with the year 2008 is based on forecasts issued from January to December 2007.

The volatility of the $R^2_{oos,s}$ values indicated by Figure 2.6 is not surprising; the years 1996-2018 represent a period rife with crises and crashes. These events have a notable effect on the standard deviations of the predictive R^2 in Tables 2.2 and 2.3. We observe that at the one-year horizon, the impact of the build-up and burst of the so-called dot-com bubble is more pronounced than that of the 2008 financial crisis. Both theory-based and machine learning approaches yield large negative annual $R^2_{oos,s}$ values associated with forecasts issued during 2000 and 2001. Panel A in Figure 2.6 also illustrates how the RF achieves its improvement over MW at the one-year horizon.

2.4.2 Hybrid approaches and short training

Next, we assess the potential of hybrid strategies that combine the theory-based and machine learning paradigms. Table 2.4 indicates the promise of this idea: Although theory-based and machine learning forecasts covary positively, the correlations are not strong, so the two approaches seem to account for different components of the stock

[20] Depending on the selection of stocks, they report annual predictive R^2 for ANNs that range from 3.4% to 5.2%.

[21] A complementary analysis using data provided by Grigory Vilkov yields very similar annual predictive R^2 values for MW and Chabi-Yo et al.'s (2023) alternative approach.

[22] As mentioned in Section 2.3.3, the R^2_{oos} can be dominated by the forecast error in levels, whereas the Sharpe ratio captures purely cross-sectional aspects. Hence, it is not necessary for R^2_{oos} and the Sharpe ratio to point into the same direction in terms of favored approaches.

Figure 2.5: Prediction-sorted portfolios, one-year horizon: long training. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to December 2017. Machine learning results are based on the long training scheme depicted in Figure 2.2.

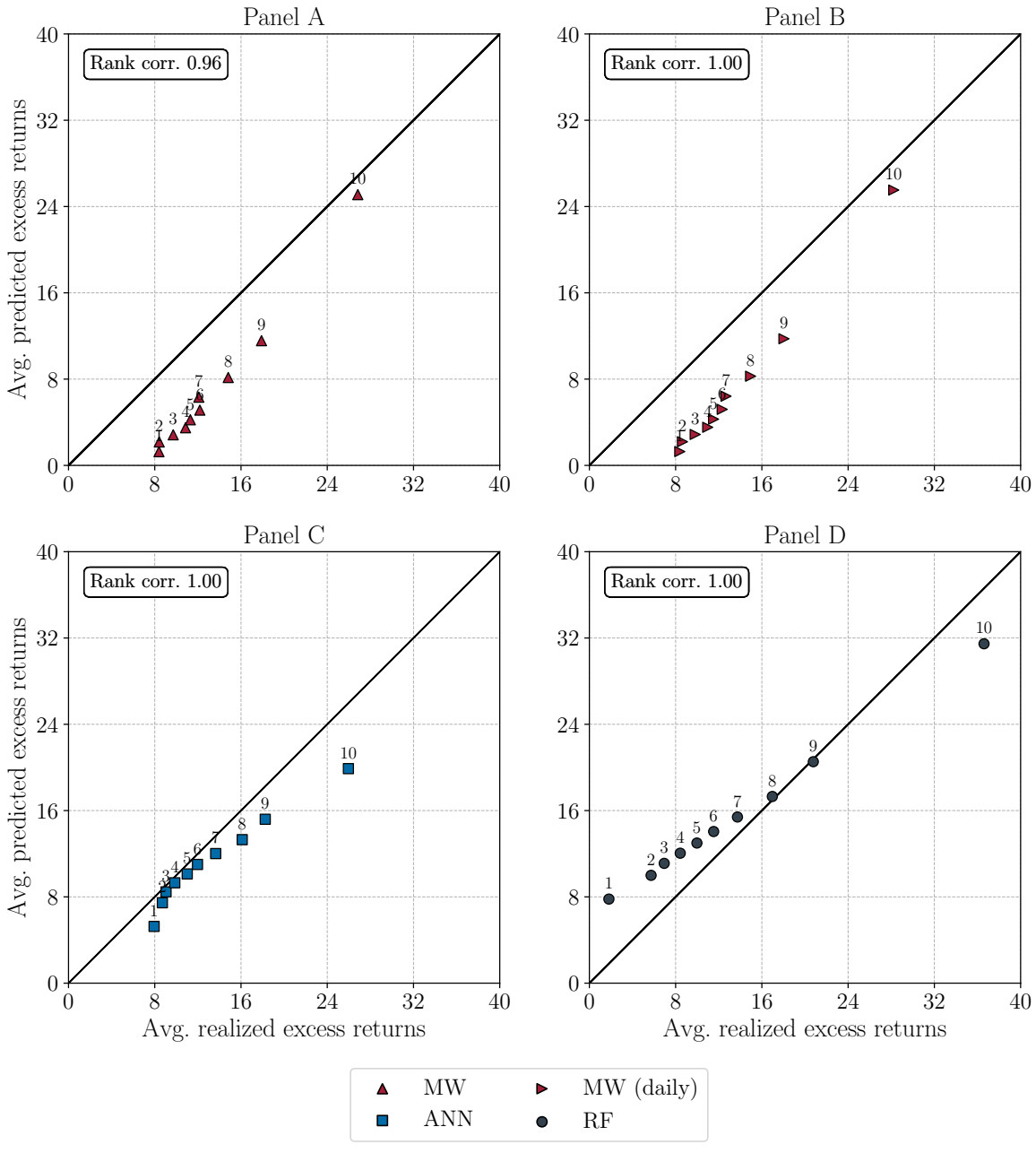
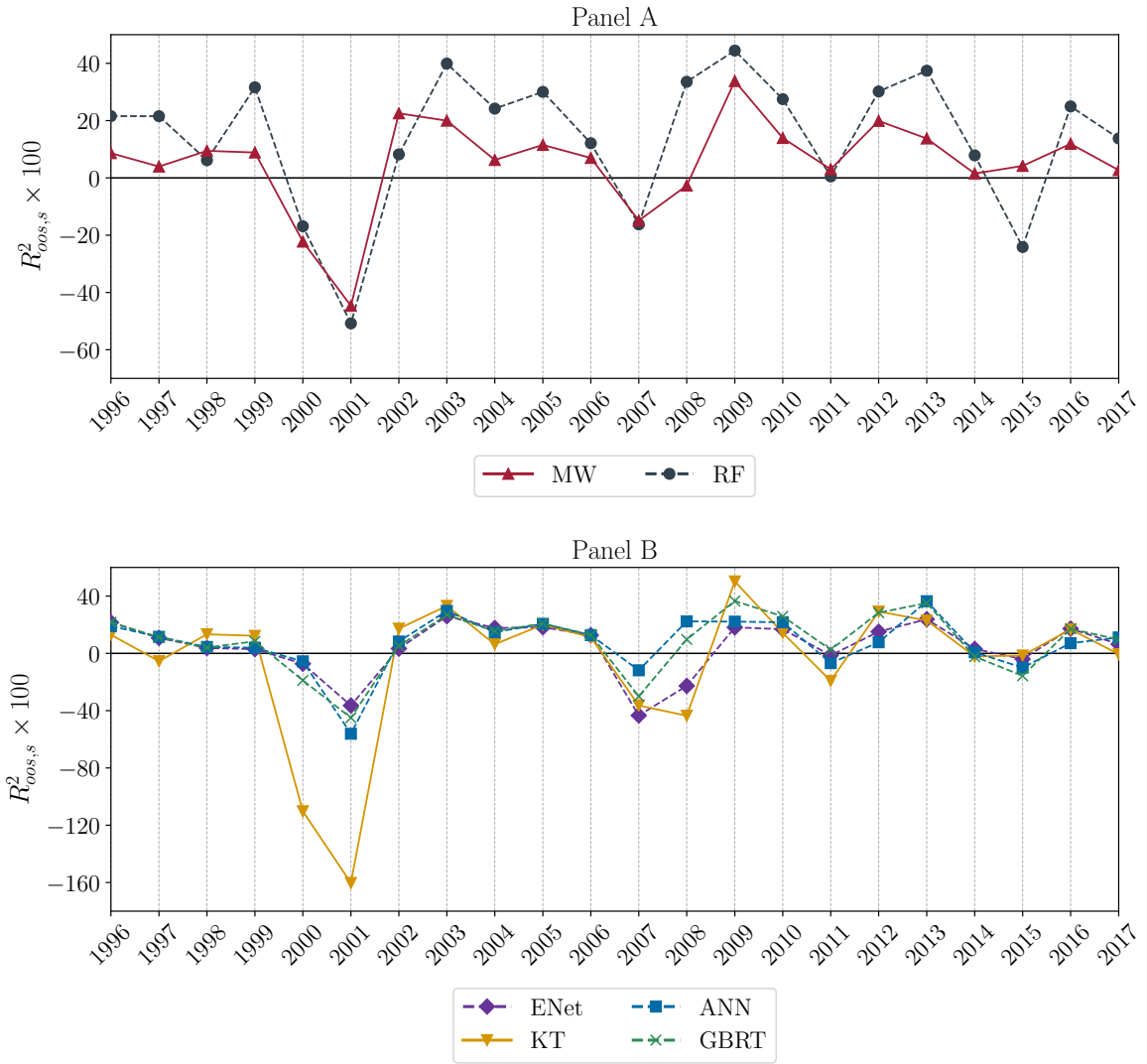


Figure 2.6: Time series of predictive R^2, one-year horizon: long training. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples. The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1996 to December 2017. Panel A contrasts the MW results with the RF, which in terms of R^2_{oos} is the best among the machine learning approaches. Panel B shows the $R^2_{oos,s}$ time series of the remaining approaches. The machine learning results are obtained using the long training scheme depicted in Figure 2.2.



risk premium.

Short-training effects and ML with theory features

Any hybrid methodology must accommodate the late availability of the OptionMetrics data. As discussed previously, we deal with this issue by applying the short-training scheme in Figure 2.3. Tables 2.5 (one-month horizon) and 2.6 (one-year horizon) present two sets of machine learning results obtained by short training. The first uses the same 891 features as selected for long training. The second, referred to as *ML with theory features*, results from adding the two option-based stock risk premium measures (according to MW and KT) and Martin's (2017) lower bound of the expected market return. The following discussion contains an assessment of the incremental effects of applying the short-training scheme and including the theory-based features.[23]

We have already seen that at the one-month horizon, most of the machine learning approaches do not perform well. Table 2.5 shows that the results worsen when applying the short-training scheme. All MLPs, including the ANN, now yield a negative predictive R^2. Their standard deviations increase, and the Sharpe ratios of the long-short portfolios decline. The segments labeled *ML with theory features* in Table 2.5 reveal that this deterioration is not mitigated by the inclusion of theory-based features. Using MW to obtain risk premium estimates remains the preferred strategy at the one-month horizon.

Table 2.6 shows that the short-training effects are more ambiguous with regard to end-of-month issued forecasts with a one-year horizon. While the ENet now performs poorly, the ANN benefits from short training: Its R^2_{oos} increases from 9% (long training) to 14%, with a p-value of 0.4%. In contrast, short training reduces the RF's predictive R^2 from 19.1% (long training) to 12.4%, accompanied by increases of the standard deviation and p-value. However, Panel A of Figure 2.7, which depicts the time-series variation of the predictive R^2, shows that the adverse effects of short training on the RF are mitigated as the training sample grows. At the start of the sequential validation procedure, there are only a few years of observations available for training. When the dot-com crisis confronts such an RF, it results in a sharp decline of the $R^2_{oos,s}$ associated with the one-year forecasts issued in the year 2000. This drop causes the increase of the time-series standard deviation and p-value compared with the long-trained RF.[24] As the training sample grows, the performance of the short-trained RF improves and reaches, near the end of the sample period, the level of its long-trained counterpart. Table 2.6 also shows that the *machine learning with theory features* strategy yields a positive effect only when using the RF. Though the improvement is moderate for end-of-month-issued forecasts – the R^2_{oos} increases from 12.4% to 14.6%, and the Sharpe ratio increases from 0.59 to 0.62 – we note that the augmentation with theory features helps the short-trained RF improve the 2008 crisis year forecasts (cf. Figure 2.7).

[23] Comparing Table 2.5 with Table 2.2, we note that the theory-based results only change because the out-of-sample evaluation period is shorter. The years 1996 and 1997 are excluded to ensure comparability with the short-trained MLPs.

[24] Figure 2.7 shows that this drop is much less pronounced for the short-trained ANN, which explains the smaller standard deviation and p-value in Table 2.6.

Table 2.4: Forecast correlations. The table reports Pearson correlation coefficients for the out-of-sample forecasts of the theory-based approaches (Martin and Wagner, 2019; Kadan and Tang, 2020) and the four machine learning models with the long training scheme depicted in Figure 2.2. Panel A refers to a forecast horizon of one month with a testing period from January 1996 to November 2018. Panel B refers to a forecast horizon of one year and a testing period from January 1996 to December 2017. All forecasts are issued at the end of each month.

Panel A: One-month horizon

	ANN	RF	GBRT	ENet	KT
MW	0.01	0.25	0.32	−0.06	0.98
KT	0.02	0.25	0.31	−0.04	
ENet	0.32	0.70	0.45		
GBRT	0.11	0.82			
RF	0.22				

Panel B: One-year horizon

	ANN	RF	GBRT	ENet	KT
MW	0.19	0.33	0.34	0.00	0.98
KT	0.20	0.32	0.35	0.02	
ENet	0.69	0.49	0.57		
GBRT	0.70	0.72			
RF	0.59				

Figure 2.7: Time series of predictive R^2, one-year horizon: theory-based vs. machine learning with and without theory features. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples. The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1998 to December 2017. The machine learning results are obtained using the short training scheme depicted in Figure 2.3. For a comparison, we also display the $R^2_{oos,s}$ for MW and the long-trained RF from Panel A of Figure 2.6.

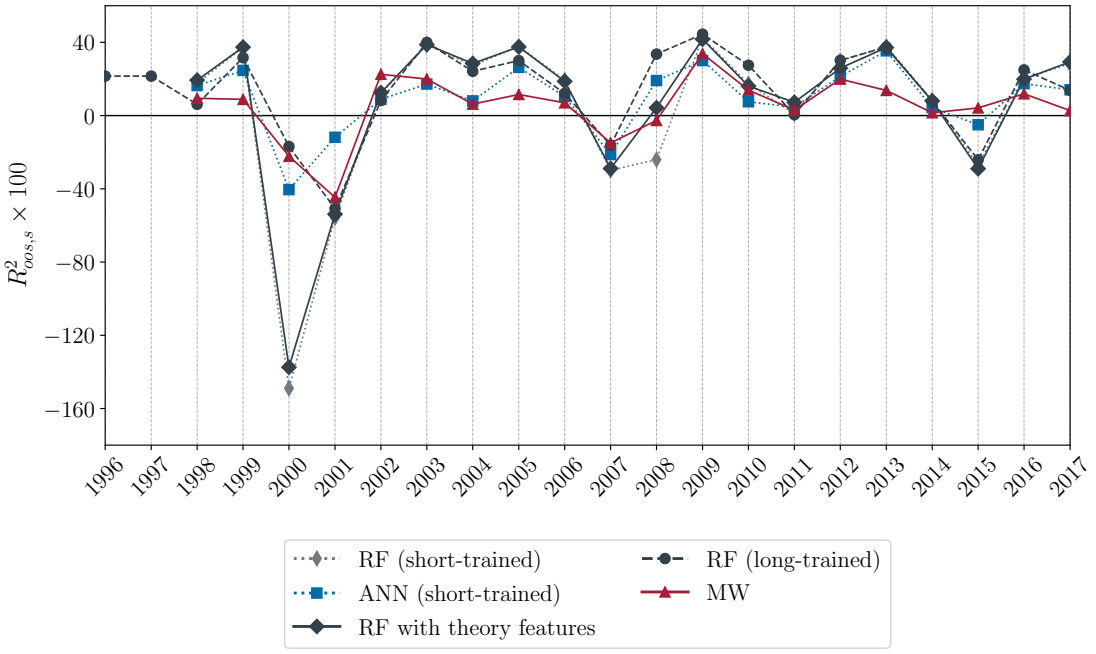


Table 2.5: Performance comparison, one-month horizon: theory-based vs. machine learning approaches vs. hybrid approach. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, the four machine learning models, and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-month horizon forecasts are issued at a daily frequency, and for Panel B, the one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1998 and ends in November 2018. The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

		Panel A: daily forecast frequency			
		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.8	2.4	0.017	0.37
	KT	−0.7	5.5	0.590	0.37
Machine Learning	ENet	−4.0	8.1	0.844	0.33
	ANN	−2.7	5.0	0.864	0.22
	GBRT	−22.6	30.7	0.884	0.12
	RF	−5.4	7.8	0.924	−0.04
ML with theory features	ENet	−3.0	6.4	0.870	0.46
	ANN	−30.7	68.7	0.853	0.20
	GBRT	−10.7	21.5	0.844	0.37
	RF	−3.0	5.8	0.868	0.17
		Panel B: monthly forecast frequency			
		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.1	3.4	0.206	0.32
	KT	−2.0	7.2	0.739	0.32
Machine Learning	ENet	−4.0	8.6	0.840	0.21
	ANN	−3.1	5.0	0.853	0.13
	GBRT	−29.5	57.7	0.860	0.15
	RF	−8.4	15.1	0.869	0.00
ML with theory features	ENet	−3.2	7.1	0.790	0.29
	ANN	−36.0	69.5	0.859	0.07
	GBRT	−25.6	53.1	0.855	0.20
	RF	−7.6	13.3	0.871	0.01

Table 2.6: Performance comparison, one-year horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which the machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued monthly (end-of-month). The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.1	17.1	0.072	0.37
	KT	3.1	49.9	0.706	0.37
Machine Learning	ENet	−31.6	153.6	0.873	0.36
	ANN	14.1	18.1	0.004	0.47
	GBRT	10.3	36.6	0.308	0.45
	RF	12.4	45.1	0.329	0.59
ML with theory features	ENet	−32.6	160.3	0.868	0.36
	ANN	14.1	19.7	0.013	0.57
	GBRT	9.7	39.7	0.356	0.42
	RF	14.6	42.3	0.244	0.62
Theory assisted by ML	MW+ENet	−38.2	192.9	0.885	0.45
	MW+ANN	14.2	25.8	0.073	0.51
	MW+GBRT	9.2	45.2	0.440	0.40
	MW+RF	16.1	50.6	0.259	0.65

Table 2.7: Performance comparison, one-year horizon, daily forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued daily. The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.5	16.8	0.057	0.37
	KT	3.4	49.8	0.689	0.37
Machine Learning	ENet	−35.5	140.9	0.898	0.36
	ANN	12.0	18.7	0.032	0.45
	GBRT	8.8	36.9	0.394	0.44
	RF	9.0	46.1	0.462	0.56
ML with theory features	ENet	−27.4	138.6	0.861	0.38
	ANN	16.1	20.0	0.005	0.58
	GBRT	11.6	38.5	0.308	0.44
	RF	18.6	39.9	0.126	0.67
Theory assisted by ML	MW+ENet	−41.2	176.6	0.902	0.45
	MW+ANN	12.8	26.3	0.154	0.50
	MW+GBRT	8.2	47.1	0.522	0.40
	MW+RF	14.1	51.9	0.355	0.62

Table 2.7 suggests that the *ML with theory features* strategy is more rewarding for forecasts at a daily frequency, and in particular when using the RF. Augmented with daily theory-based features, the RF's predictive R^2 increases from 9.0% to 18.6%, while also reducing the time-series variation across test samples. Considering that the pure theory-based (MW) R^2_{oos} amounts to 9.5%, this hybrid approach makes particularly good use of the additional data. The highest Sharpe ratio of the long-short portfolio in the field of competitors corroborates this conclusion.

Theory assisted by machine learning

For our implementation of the *theory assisted by machine learning* strategy we rely on Martin and Wagner's (2019) approach to measuring stock risk premia (MW for short), which explicitly starts from the basic asset pricing equation, the keystone of financial economics. MW is empirically not unsuccessful, and we propose building on it, as a basis, to model only that which theory cannot account for – the approximation errors – by applying machine learning techniques.

The segment labeled *theory assisted by ML* in Table 2.6 contains the results obtained from applying this idea.[25] We observe that not all machine learning assistance improves the performance of the theory-based approach; the ENet even drives the R^2_{oos} into a negative domain. GBRT yield a moderate improvement, whereas the ANN and RF are more successful. Their support increases the baseline MW R^2_{oos} by 5.1 percentage points (MW+ANN) and 7 percentage points (MW+RF), respectively. The standard deviations of the predictive R^2 grow, but Figure 2.8 shows that this increase is mainly due to the short-training effect, which in turn is reflected in the harsh drop of the $R^2_{oos,s}$ associated with the year 2000 forecasts, which we also identified for the short-trained RF. By zooming in on more recent forecast samples, we observe that with an increasing training sample size, the performance of the MW+RF hybrid matches that of the long-trained RF.

The prediction decile plots in Figure 2.9 show that the alignment of mean predicted and realized excess returns of the prediction-sorted portfolios is particularly good for the MW+RF approach and that the variation of the mean realized excess returns across the prediction-sorted portfolios is favorably high. Consistently, RF assistance increases the Sharpe ratio for the long-short portfolio from 0.37 (pure MW) to 0.65, as reported in Table 2.6. For daily forecasts rather than forecasts issued at the end of the month, these conclusions remain the same (cf. Table 2.7).

These results lead to the conclusion that at the one-year horizon, the MW+RF approach qualifies as a promising alternative for the task of quantifying stock risk premia. This hybrid strategy also has the appeal of effectively combining theory with measurement.

[25] Short-trained MLPs do not perform well at the one-month horizon, and when using them to account for the approximation errors of MW, we find no improvement. We therefore discuss in detail only the one-year horizon results.

2.4.3 Feature importance and disaggregated analyses

We also investigate how the importance of features with respect to stock risk premia might differ between pure machine learning and theory assisted by machine learning. We consider both pure RF and the MW+RF hybrid and focus on the one-year horizon with end-of-month issued forecasts. To gauge a feature's importance by the reduction of the predictive R^2 induced, we use a disruption of the temporal and cross-sectional alignment of the feature with the prediction target. This disruption is implemented by replacing the feature's observed values by 0 when computing the predictive R^2. We compute the importance measure on the test samples, and report the size of the induced R^2_{oos} reduction.[26] Figures 2.10 (RF) and 2.11 (MW+RF) illustrate the results.

A comparison of Figures 2.10 and 2.11 reveals that the conclusions regarding the relative importance of features remain the same, regardless of whether the RF serves to assist the theory-based approach or is applied for its original use. The pattern is similar in both applications. With respect to stock-level variables, the established return predictive signals (RPS) are most important: The book-to-market ratio ranks first (along with other valuation ratios), followed by variables associated with liquidity (dollar trading volume, Amihud illiquidity), and then momentum indicators (industry momentum and 12-month momentum). None of the other more than 80 stock level features is among the top four. The revival of the classic RPS, and in particular the conspicuous role of the book-to-market ratio, is noteworthy. In GKX's study, the short-term price reversal dominated the feature importance at the one-month horizon, whereas the book-to-market ratio remained nondescript. The consistent feature importance in both applications – RF and MW+RF – may seem surprising, because MW already accounts for a considerable part of the excess return variation. We might have expected that modeling the approximation error of the theory-based approach would reveal other important features. But it is the familiar triad – valuation ratio, liquidity, and momentum – that dominates in both applications.

A corresponding conclusion arises from an analysis of the importance of the market-wide variables (Panels B in Figures 2.10 and 2.11). In both uses of the RF, the Treasury bill rate is the most important variable. Its conspicuous role highlights the relevance of asset pricing approaches that adopt Merton's (1973) suggestion to use short-term interest rates as state variables in variants of the intertemporal CAPM (e.g., Brennan et al. (2004), Petkova (2006), Maio and Santa-Clara (2017)), as well as preference-based asset pricing models that motivate a short-term interest rate-related risk factor,

[26] Alternatively, it is possible to compute the importance measure on the training samples and provide a relative measure of feature importance, as done by GKX. Moreover, feature importance could be assessed by randomly drawing a feature from the empirical distribution instead of replacing it by 0. We prefer the present approach for its straightforward interpretability. Another approach to assess the importance of features is based on the absolute gradient of the loss function with respect to each feature respectively, which is very convenient in the context of neural networks (cf. Chen et al., 2023), but not suitable for all machine learning techniques. Shapley additive explanations (cf. Lundberg and Lee, 2017) would be well suited to account for dependencies between features, but are computationally infeasible given our number of characteristics.

as in Lioui and Maio (2014).[27]

The feature importance results provide the foundation for disaggregated analyses, for which we form portfolios by sorting stocks into quintiles according to key characteristics associated with valuation ratios, liquidity, and momentum. As suggested by the previous results, we choose book-to-market and earnings-to-price as valuation ratios; for liquidity, we use dollar trading volume and Amihud's illiquidity measure. Momentum portfolios are based on 12-month and industry momentum. The sorting of stocks into quintile portfolios on the basis of the respective characteristic gets renewed each month. We also form 10 industry portfolios based on one-digit SIC codes. For each quintile and industry portfolio and each approach of interest – MW, pure machine learning (ANN and RF), and theory assisted by machine learning (MW+RF and MW+ANN) – we compute the annual R^2_{oos} according to Eq. (2.15).

The results in Table 2.8 generally corroborate the conclusions of the aggregated analysis and also reveal the following detailed insights: For all portfolios based on valuation ratios, we observe an improvement of the theory-based method by machine learning assistance. Moreover, the hybrid approaches are preferred across all quintile portfolios. MW+RF is particularly successful in quintiles 2 to 5, and MW+ANN is optimal in quintile 1. For all momentum portfolios, machine learning assistance improves the performance of the theory-based approach. For momentum quintiles 1 to 4, MW+RF is the preferred strategy. For momentum quintile 1, pure ANN and MW+ANN perform better. Regarding the liquidity-sorted portfolios, machine learning assistance again improves the theory-based results, but we note that MW+RF does not perform well on the high liquidity portfolios. The explanation is that the short training effect that we discussed previously has the strongest effect on the performance of both RF and MW+RF in the high liquidity portfolios.[28] The pure ANN, less affected by short training, delivers more consistent performance across liquidity portfolios. Nevertheless, a hybrid strategy is preferred over pure machine learning for four (dollar trading volume), respectively three (Amihud illiquidity) quintile portfolios.

Panel B of Table 2.8 shows that for all industry portfolios, RF assistance improves the performance of MW; the ANN assistance does so in seven out of ten cases. With the exception of one of the sector portfolios for which the pure ANN is preferred, the hybrid strategies yield the highest predictive R^2. In addition, MW+RF is preferred in seven of ten sector portfolios, and MW+ANN is preferred in two. The complementary advantage of the two hybrid approaches is thus a recurring result.

[27] We also check whether feature importance differs when we measure the effect of an exclusion of a feature on the cross-sectional performance, measured by the Sharpe ratio of the long-short portfolio. The conclusions remain qualitatively the same as when we use the predictive R^2. Details of this analysis are available in Section A.8 of the appendix.

[28] For more details, refer to Section A.8 of the appendix, which contains time series plots of the predictive R^2 corresponding to Figure 2.6. They illustrate the short training effect broken down by quintile portfolios based on Amihud illiquidity.

Table 2.8: Disaggregated performance comparison, one-year horizon, monthly forecast frequency. To obtain the results in Panel A, we sort the sample stocks into quintiles, according to the size of stock-specific valuation ratios (book-to-market and earnings-to-price), liquidity (Amihud illiquidity and dollar trading volume), and momentum (industry and 12-month). The sorting is renewed each month, taking into account the availability conditions outlined in Section 2.3. The pooled $R^2_{oos} \times 100$ according to Equation (2.15) is reported for each quintile portfolio and the approaches of interest, namely, MW, pure ML (ANN and RF), and theory assisted by machine learning (MW+RF and MW+ANN). Panel B shows the pooled $R^2_{oos} \times 100$ for each of the 10 industry portfolios based on the one-digit SIC code. The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

Panel A: $R^2_{oos} \times 100$ for quintile portfolios

		Book-to-market					Earnings-to-price				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Valuation ratios	MW	8.1	7.1	8.7	9.1	12.6	8.9	7.3	8.8	10.1	11.6
	ANN	14.7	17.1	11.9	14.0	12.1	13.1	14.6	16.8	13.4	14.1
	RF	6.7	16.2	9.4	17.8	15.4	8.0	13.0	17.7	16.1	16.7
	MW+ANN	14.9	15.7	10.8	13.4	14.9	13.1	13.8	16.7	14.5	15.5
	MW+RF	8.9	19.0	13.4	21.8	21.4	10.1	17.0	22.4	20.4	22.5
		Dollar trading volume					Amihud illiquidity				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Liquidity	MW	15.7	10.5	10.2	6.2	-0.9	-1.0	4.1	7.3	10.7	14.9
	ANN	17.2	13.1	14.5	15.8	8.0	8.2	12.4	12.8	16.0	16.5
	RF	21.8	16.0	16.8	14.0	-11.3	-8.9	4.8	12.4	19.4	20.1
	MW+ANN	19.6	13.9	15.7	16.2	2.9	4.1	10.2	12.7	17.3	18.7
	MW+RF	27.5	20.0	20.7	17.1	-11.2	-7.5	8.1	15.1	23.3	25.0
		12-month momentum					Industry momentum				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Momentum	MW	13.9	9.4	7.5	5.9	5.8	7.7	11.1	10.3	10.3	6.4
	ANN	13.9	11.1	14.8	13.2	15.9	13.0	17.4	15.3	14.0	10.8
	RF	15.2	12.7	13.1	15.3	7.2	13.1	18.9	19.6	11.1	0.5
	MW+ANN	17.0	10.8	13.1	12.2	14.3	11.9	17.8	16.1	16.2	9.5
	MW+RF	21.4	18.1	16.3	18.4	7.4	15.6	23.4	23.6	17.5	1.8

Panel B: $R^2_{oos} \times 100$ for industry portfolios (one digit SIC code)

	0	1	2	3	4	5	6	7	8	9
MW	6.6	5.4	11.9	8.0	9.0	8.7	12.0	8.0	16.9	2.1
ANN	23.9	12.7	12.2	15.8	16.6	8.1	12.0	17.3	3.6	12.9
RF	29.3	15.6	10.8	13.2	16.5	7.7	11.9	11.4	9.5	15.2
MW+ANN	22.7	8.3	13.4	14.3	19.2	8.6	15.6	16.5	11.0	18.4
MW+RF	31.6	18.1	14.6	16.0	22.5	12.5	18.1	12.4	21.5	12.6

Figure 2.8: Time series of predictive R^2, one-year horizon: MW+RF vs. pure RF (long-training) vs. MW. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples for the MW+RF hybrid (theory assisted by machine learning). The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1998 to December 2017. The MW+RF results are based on the short training scheme depicted in Figure 2.3. For a comparison, we also display the $R^2_{oos,s}$ for MW and the long-trained RF from Panel A of Figure 2.6.

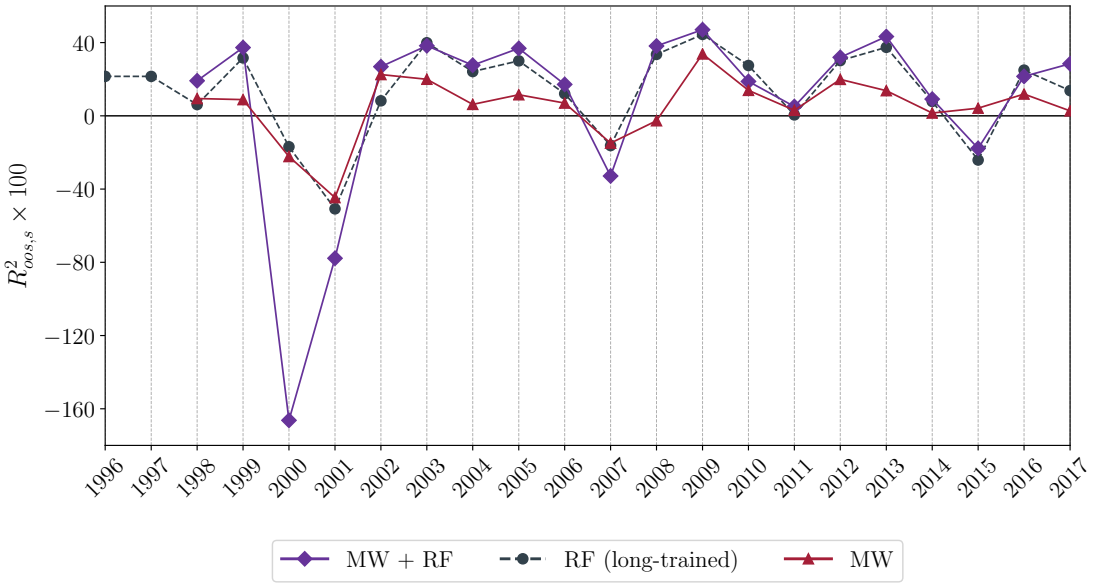


Figure 2.9: Prediction-sorted portfolios, one-year horizon: theory assisted by machine learning approaches. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed at the end of each month. The two panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW assisted by an ANN (MW + ANN, Panel A) and MW assisted by RF (MW+RF, Panel B). The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 2.3.

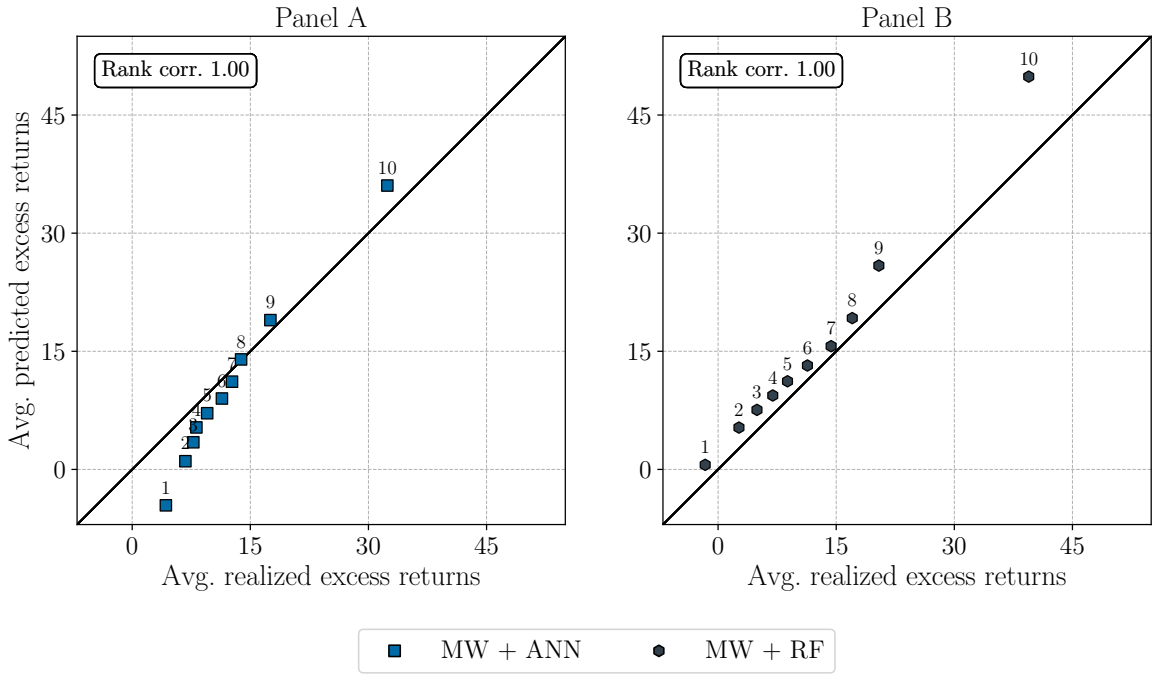


Figure 2.10: Feature importance, one-year horizon: random forest (short training). The figure depicts feature importance (Panel A: firm-level features, Panel B: macro-level features) for the RF. The forecast horizon is one year; the prediction frequency is end-of-month. A feature's importance is measured by the reduction of the predictive R^2 that is induced by setting the feature's values in the test samples to 0. In both panels, the features are sorted in descending order of importance. Panel A focuses on the ten most important firm-level features. The dashed vertical line, included for reference, represents the R^2_{oos} that is obtained without setting any feature's values to 0. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 2.3.

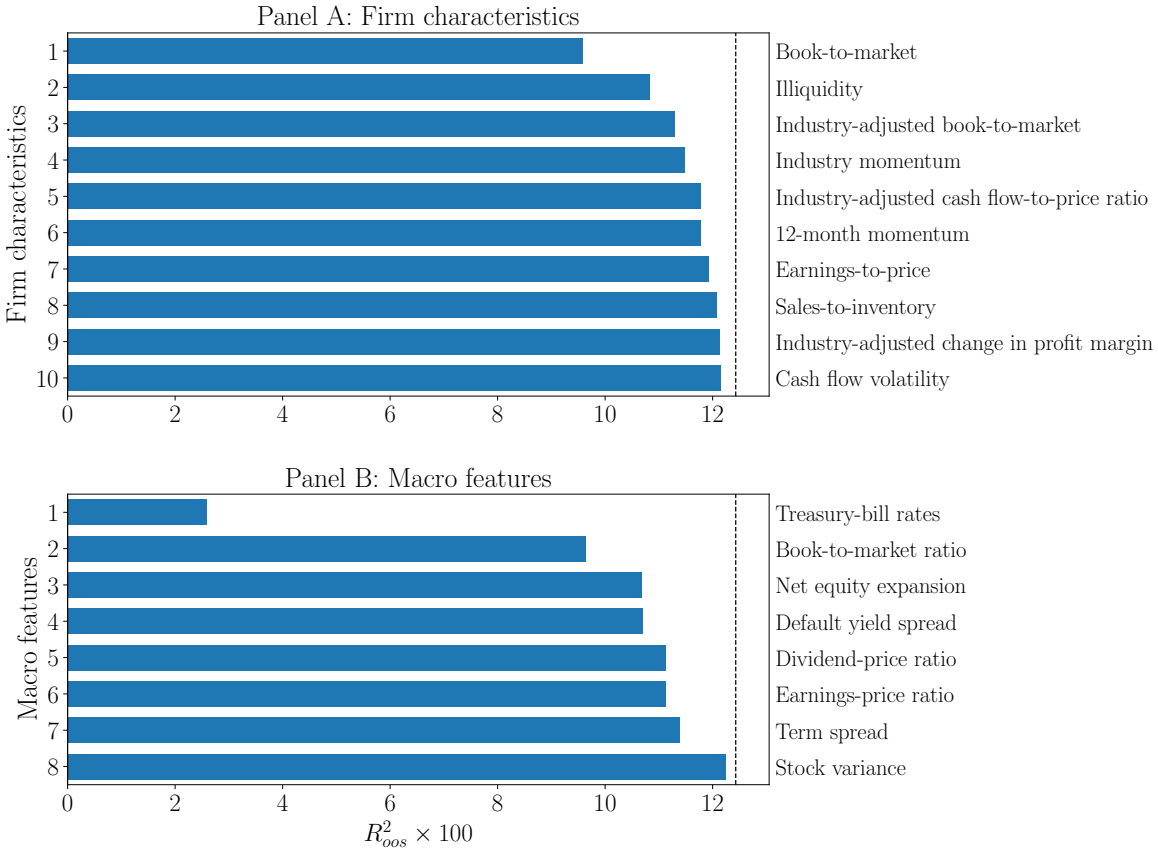
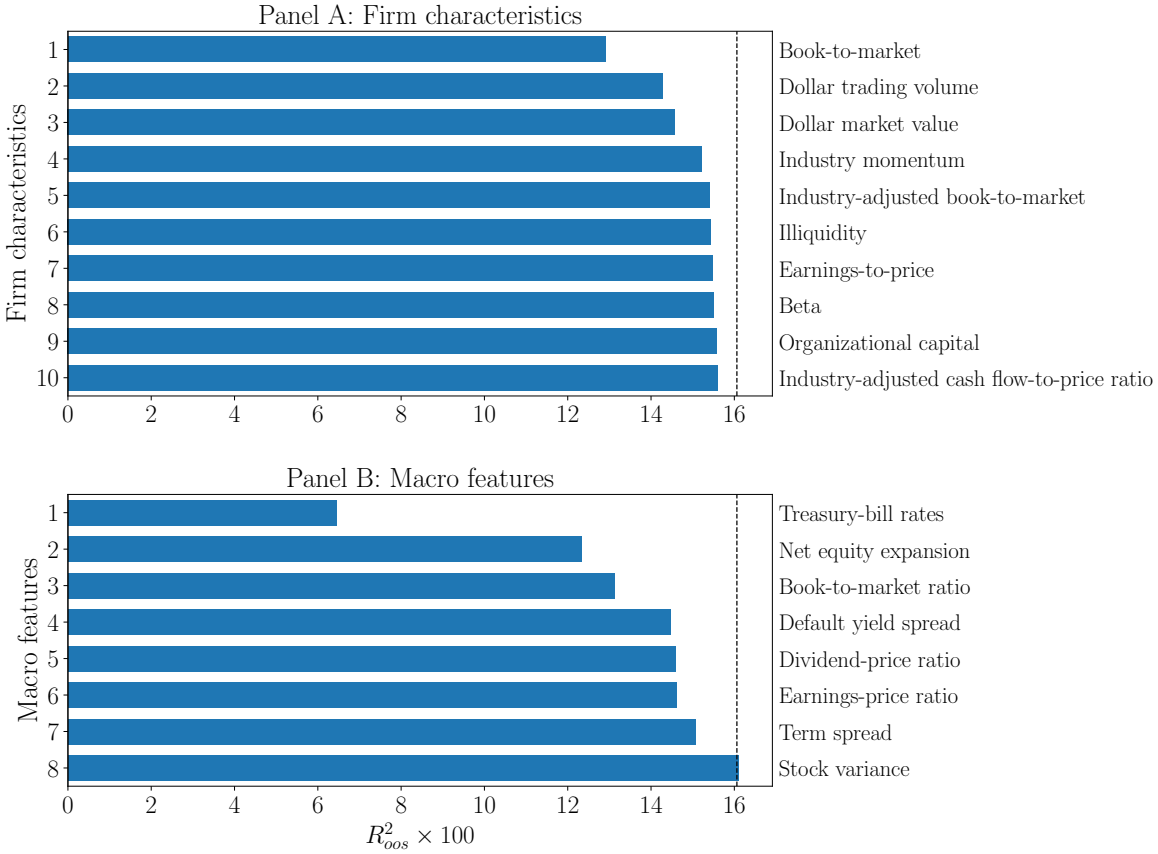


Figure 2.11: Feature importance, one-year horizon: MW+RF. The figure depicts feature importance (Panel A: firm-level features, Panel B: macro-level features) for the MW assisted by RF strategy. The forecast horizon is one year; the prediction frequency is end-of-month. A feature's importance is measured by the reduction in R^2 that is induced by setting the feature's values in the test samples to 0. In both panels, the features are sorted in descending order of importance. Panel A focuses on the ten most important firm-level features. The dashed vertical line, included for reference, represents the R^2_{oos} that is obtained without setting any feature's values to 0. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 2.3.



2.5 Conclusions

In this study, we took two diverging paths to measure stock risk premia in an attempt to assess and reconcile the opposing philosophies that underlie them. The comparison, at one-month and one-year investment horizons, reveals that the theory/option-based method offers an advantage at the shorter horizon, especially if stock risk premium estimates are to be delivered at higher frequencies. At the one-year horizon, the picture is more complex. Of the four machine learning methods considered in this study, one delivers weaker performance than the theory-based strategy (elastic net), two are comparable (gradient boosted regression trees and artificial neural networks), and one (random forest) offers the best results. To achieve this performance, a sufficiently long training period is required though.

Noting the concerns regarding the use of agnostic machine learning procedures in a theoretically well-developed discipline like finance, we put forth a methodology that takes Martin and Wagner's (2019) theory-based approximate formula for the stock risk premium as its basis and then applies machine learning to account for the approximation error. Although a pure theory-based method remains the preferred choice at the one-month horizon, the empirical performance of this *theory assisted by machine learning* approach at the one-year horizon is encouraging. Using a random forest, the theory-based component provides 57% of the hybrid model's explanatory power in terms of the predictive R^2; 43% is attributable to machine learning assistance. The conclusion that such a supportive use of machine learning captures fundamental components of stock risk premia is supported by the conspicuous role of valuation ratios and liquidity indicators in an analysis of feature importance. Disaggregated analyses based on stock portfolios sorted according to these characteristics corroborate the expediency of the proposed hybrid approach. We view it as a promising alternative for bringing together the diverging paths in finance.

A Appendix

A.1 Theory-based stock risk premium formulas

This section provides details on the stock risk premium formulas in Equations (2.2) and (2.3) and the nature of the approximation residuals $a^i_{t,T}$ and $\xi^i_{t,T}$. We delineate the assumptions and rationales behind their omission, which provide the theory-based approximation formulas in Equations (2.5) and (2.6).

Martin and Wagner's (2019) derivations originate from the basic asset pricing equation, with a focus on the gross return of a portfolio with maximal expected log return ($R^g_{t,T}$). This growth-optimal return has the unique property among gross returns that its reciprocal is an SDF, such that $m_{t,T} = 1/R^g_{t,T}$. Using this SDF to price the payoff $X^i_{t,T} = R^i_{t,T} \cdot R^g_{t,T}$ gives:

$$\mathbb{E}_t\big(m_{t,T} \cdot X^i_{t,T}\big) = \mathbb{E}_t\big(R^i_{t,T}\big) = \frac{1}{R^f_{t,T}} \mathbb{E}^*_t(R^i_{t,T} \cdot R^g_{t,T}), \tag{A-1}$$

where the $*$ notation indicates that the expected value is computed with respect to the risk-neutral measure. Division by $R^f_{t,T}$ and subtracting $\mathbb{E}^*_t\big(R^i_{t,T}/R^f_{t,T}\big) \cdot \mathbb{E}^*_t\big(R^g_{t,T}/R^f_{t,T}\big) = 1$ (the price of any gross return is 1) yields:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = 1 + \mathrm{cov}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}, \frac{R^g_{t,T}}{R^f_{t,T}}\right). \tag{A-2}$$

An orthogonal projection under the risk-neutral measure of $R^i_{t,T}/R^f_{t,T}$ on $R^g_{t,T}/R^f_{t,T}$ and a constant gives:

$$\frac{R^i_{t,T}}{R^f_{t,T}} = \alpha^i_{t,T} + \beta^i_{t,T} \cdot \frac{R^g_{t,T}}{R^f_{t,T}} + u^i_{t,T}, \tag{A-3}$$

where the moment conditions $\mathbb{E}^*_t(u^i_{t,T}) = 0$ and $\mathbb{E}^*_t(u^i_{t,T} \cdot R^g_{t,T}) = 0$ define the projection coefficients

$$\beta^i_{t,T} = \frac{\mathrm{cov}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}, \frac{R^g_{t,T}}{R^f_{t,T}}\right)}{\mathrm{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right)},$$

and $\alpha^i_{t,T} = 1 - \beta^i_{t,T}$. Inserting these insights into Equation (A-2) produces:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = 1 + \beta^i_{t,T} \cdot \mathrm{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right). \tag{A-4}$$

Moreover, Equation (A-3) implies:

$$\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) = (\beta_{t,T}^i)^2 \cdot \text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) + \text{var}_t^*(u_{t,T}^i). \tag{A-5}$$

To make these results practically usable, Martin and Wagner (2019) propose to linearize $(\beta_{t,T}^i)^2 \approx 2\beta_{t,T}^i - k$, which for $k = 1$ amounts to a first-order Taylor approximation at $\beta_{t,T}^i = 1$. Using this approximation and inserting it into Equation (A-4) (for $k = 1$) removes the dependence on $\beta_{t,T}^i$,

$$\mathbb{E}_t\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) \approx 1 + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) - \frac{1}{2}\text{var}_t^*(u_{t,T}^i). \tag{A-6}$$

The term neglected on the right-hand side of Equation (A-6) due to the linearization is $-\text{var}_t^*(R_{t,T}^g/R_{t,T}^f)(\beta_{t,T}^i - 1)^2$. The approximation thus should be reasonable for stocks whose $\beta_{t,T}^i$ is close to 1.

Using w_t^j, the weight of stock j in a market index with gross return $R_{t,T}^m$, Martin and Wagner (2019) perform a value-weighting of Equation (A-6) to obtain:

$$\mathbb{E}_t\left(\frac{R_{t,T}^m}{R_{t,T}^f}\right) \approx 1 + \frac{1}{2}\sum_j w_t^j \text{var}_t^*\left(\frac{R_{t,T}^j}{R_{t,T}^f}\right) + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) - \frac{1}{2}\sum_j w_t^j \cdot \text{var}_t^*(u_{t,T}^i). \tag{A-7}$$

Subtracting Equation (A-7) from (A-6) removes the dependence on the unobservable optimal growth portfolio, such that

$$\begin{aligned}\mathbb{E}_t\big(R_{t,T}^i\big) \approx \mathbb{E}_t\big(R_{t,T}^m\big) &+ \frac{R_{t,T}^f}{2}\left[\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) - \sum_j w_t^j \cdot \text{var}_t^*\left(\frac{R_{t,T}^j}{R_{t,T}^f}\right)\right] \\ &- \frac{R_{t,T}^f}{2}\left(\text{var}_t^*(u_{t,T}^i) - \sum_j w_t^j \cdot \text{var}_t^*(u_{t,T}^j)\right).\end{aligned} \tag{A-8}$$

Keeping track of the approximation error due to the linearization, we note that the term that is omitted on the right-hand side of Equation (A-8) is

$$\kappa_{t,T}^i = -\frac{1}{2R_{t,T}^f}\text{var}_t^*\big(R_{t,T}^g\big) \cdot \left[(\beta_{t,T}^i - 1)^2 - \sum_j w_t^j \cdot (\beta_{t,T}^i - 1)^2\right].$$

To account for the first term on the right-hand side of Equation (A-8), Martin and Wagner (2019) draw on a result by Martin (2017), who derives a lower bound for the expected return of a market index. His starting point is again the basic asset pricing Equation (2.1), which can be written in terms of the price of the payoff $(R_{t,T}^i)^2$ using an add-and-subtract strategy:

$$\mathbb{E}_t(R_{t,T}^i) - R_{t,T}^f = \big(\mathbb{E}_t[m_{t,T} \cdot (R_{t,T}^i)^2] - R_{t,T}^f\big) - \big(\mathbb{E}_t[m_{t,T} \cdot (R_{t,T}^i)^2] - \mathbb{E}_t(R_{t,T}^i)\big). \tag{A-9}$$

The first term on the right-hand side of Equation (A-9) can be related to a risk-neutral variance, and the second term to a covariance under the physical measure, such that

$$\mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = \frac{1}{R^f_{t,T}} \text{var}^*_t(R^i_{t,T}) - \text{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T}). \tag{A-10}$$

As noted in the main text, Kadan and Tang (2020) use Equation (A-10) for their quantification and approximation of stock risk premia.

Martin (2017) argues that for an asset return that qualifies as a market return proxy (denoted $R^m_{t,T}$), it should be the case that

$$\xi_{t,T} = \text{cov}_t(m_{t,T} \cdot R^m_{t,T}, R^m_{t,T}) < 0. \tag{A-11}$$

Intuitively, an investor's marginal rate of intertemporal substitution should be negatively correlated with any portfolio that qualifies as a market index. Accordingly,

$$\mathbb{E}_t(R^m_{t,T}) - R^f_{t,T} \geq \frac{1}{R^f_{t,T}} \text{var}^*_t(R^m_{t,T}). \tag{A-12}$$

Assuming that the inequality (A-12) is binding, we can use it with Equation (A-8), which yields:

$$\begin{aligned} \mathbb{E}_t\big(R^i_{t,T}\big) - R^f_{t,T} \approx R^f_{t,T} \cdot \left[\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\left\{ \text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right) \right\} \right] \\ - \frac{R^f_{t,T}}{2} \cdot \left[\text{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \text{var}^*_t(u^j_{t,T}) \right], \end{aligned} \tag{A-13}$$

where the approximative formula in Equation (A-13) omits the term $\kappa^i_{t,T} - \xi_{t,T}$ on the right-hand side. Equation (2.2) thus results from

$$a^i_{t,T} = \kappa^i_{t,T} - \xi_{t,T} - \zeta^i_{t,T}, \tag{A-14}$$

where

$$\zeta^i_{t,T} = \frac{1}{2} R^f_{t,T} \cdot \big[\text{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \text{var}^*_t(u^j_{t,T}) \big]. \tag{A-15}$$

Working with the abbreviated formula in Equation (2.5) thus entails three approximations: (1) the linearization of $(\beta^i_{t,T})^2$, (2) the assumption that Martin's (2017) lower bound for the expected return of the market is binding, and (3) the assumption that the residual variances $\text{var}^*_t(u^i_{t,T})$ are very similar across stocks, such that $\zeta^i_{t,T}$ is negligibly small in absolute terms.

A.2 Construction of the database

As outlined in the main text, this study focuses on a universe of firms that appear at least once as an S&P 500 constituent during October 1974 to December 2018. For that purpose, we must identify the set of historical S&P 500 constituents (HSPC) for each date of the sample period.[29] Our strategy to identify HSPC is based on a monthly security query from Compustat's *SECM* file, in which the variable *SPMIM* (S&P Major Index Code - Historical) identifies S&P 500 members. As recommended by WRDS, until November 1994, we select $SPMIM \in \{10, 40, 49, 60\}$ to identify S&P 500 constituents.[30] After December 1994, WRDS recommends selecting $SPMIM=10$ instead. The data table resulting from the query contains the variables *GVKEY* and *IID*, which together constitute Compustat's permanent security identifier, and the dates when the firm thus identified has been a S&P 500 member.[31] The table also contains the security identifier *CUSIP*, which, like the ticker symbol, can change over the lifetime of a firm.[32]

Establishing a connection between Compustat and CRSP, and merging the respective data for a given security, is a common problem in empirical finance. To facilitate such a merge, WRDS provides a linkage table that enables the cross-database identification of securities using each database's permanent identifier. For Compustat this is the aforementioned combination of *GVKEY* and *IID*, while CRSP uses the security identifier *PERMNO*. The linkage table provides (via the variables *LINKDT* and *LINKENDDT*) information about the validity of a connection of the permanent CRSP and Compustat identifiers at a certain point in time. Applying the linkage table to the Compustat identified HSPC, the connection of the permanent identifiers in CRSP and Compustat is one-to-one at all dates of the period considered for the analysis. Using the list of S&P 500 constituents obtained from Compustat and the matched *PERMNO* identifiers, security-level information can be extracted from CRSP. In particular, the *crspa* library provides price information and the number of outstanding shares on a daily frequency for each index constituent. The CRSP index price data are obtained from the library *crsp* with the table *dsi*.

[29] WRDS suggests several ways to perform this task. Due to license-specific data access constraints, not all of them may be feasible, though. For example, we cannot access the CRSP table *dsp500list*, which contains the starting and ending dates of S&P 500 membership for each security identified by CRSPs permanent security identifier *PERMNO*.

[30] According to WRDS, S&P 500 constituents represent the union of S&P Transportation (*SPMIM*=40), Utilities (*SPMIM*=49), Financial (*SPMIM*=60), and Industrial (*SPMIM* 10). This information is obtained from https://wrds-www.wharton.upenn.edu/pages/support/applications/programming-examples-and-other-topics/sp-500-datasets-and-constituents/

[31] An alternative way to generate the HSPC list is to use the Compustat table *IDXCST_HIS*, which collects securities identified by the variable *GVKEYX*, indicating membership of a company in the S&P 500. We have implemented both methods, and the HSPC lists resulting from these two approaches differ only slightly. We choose the first approach because it provides a more consistent coverage of HSPC during the 1970s

[32] As noted by WRDS,"[a] change in *CUSIP* [...] could be triggered by any change in the security, including non-fundamental events such as splits and company name changes." For a detailed description of the cross-database identification problem see https://wrds-www.wharton.upenn.edu/pages/support/applications/linking-databases/linking-crsp-and-compustat/

The OptionMetrics library *optionm* contains a separate volatility surface table for each available year, which we connect to the merged Compustat/CRSP data. Putting aside the aformentioned shortcomings associated with *CUSIP* identification, we search OptionMetrics for the HSPC detected in Compustat, using the *CUSIP* identifier. Although this approach does not yield a 100% coverage of S&P 500 constituents in the OptionMetrics data, it is still very close to the one reported by Martin and Wagner (2019).

Panel A in Figure A.1 shows the number of HSPC that we are able to identify in Compustat, CRSP, and OptionMetrics for the period from March 1964 to December 2018. The coverage rate that we achieve with our procedure is higher than that reported by Martin and Wagner (2019). Averaged over the respective sample periods, we manage to recover 483/500 HSPC; Martin and Wagner's (2019) coverage ratio is 451/500. Panel B of Figure A.1 shows that the actual S&P 500 market capitalization is closely tracked by that of the HSPC identified in Compustat, CRSP, and OptionMetrics.

A.3 Approximating risk-neutral variances

In the following, we describe how we approximate the risk-neutral variances in Equation (2.4) using the volatility surface provided by OptionMetrics. The ingredients we need for this approximation are the price of the underlying, a proxy for the risk-free rate, the price of the forward contract, and the prices of European call and put options for various strikes, each expiring in T. However, the prices of European options are not directly observable because options on constituents of the S&P 500 are exclusively traded American-style. Therefore, we must first determine the prices of equivalent European options with identical maturities and strike prices. We follow the approach by Martin and Wagner (2019) and assume that the implied volatility serves as a conversion factor between the two exercise styles. That is, we determine the prices of European options by using standardized grid points of the volatility surface derived from American options as inputs to the Black-Scholes-Merton formula.

Moreover, we need to approximate the integrals in Equation (B-1), as we do not observe option contracts for a continuum of strike prices. We employ the conservative approach by Martin (2017), where the numerator on the right-hand side of Equation (2.4) is replaced by[33]

$$\int_0^\infty \Omega(K)dK \approx \sum_j \Omega_j(K_j)\cdot\Delta K_j, \tag{A-16}$$

and

$$\Omega_j(K_j) = \begin{cases} \text{put}_j(K_j) & \text{if} \quad K_j < F_j \\ \text{call}_j(K_j) & \text{if} \quad K_j \geq F_j \end{cases} \tag{A-17}$$

[33]For notational convenience, we drop the security, time, and maturity indices.

Figure A.1: Identification of S&P 500 constituents. The figure illustrates the reliability of the procedure with which we identify historical S&P 500 constituents. Panel A presents the coverage of HSPC achieved at different stages of the data processing. The line in light grey refers to the HSPC found in Compustat. The blue line shows for how many of these constituents it is possible to find stock price information in CRSP. The red line starting in 1996 illustrates for how many HSPC it is also possible to find information in OptionMetrics. Panel B depicts the aggregate market capitalization for each of these three groups of HSPC.

denotes the price of an out-of-the-money option with strike price K_j. The spacing of the grid with terminal value K_n is determined according to

$$\begin{aligned}\Delta K_j &= \frac{K_{j+1}-K_{j-1}}{2} \quad j = 2,\ldots,n-1,\\ \Delta K_1 &= K_2-K_1,\\ \Delta K_n &= K_n-K_{n-1}.\end{aligned}$$

A.4 Theory-based, stock-level, and macro-level variables

Table A.1 gives a description of the variables used in this study. The content displayed in Panel B1 is obtained from Table A.6 in GKX. The stock-level features are retrieved using the SAS program provided by Jeremiah Green that we update and modify for our purposes. The features are originally used in the study by Green et al. (2017).

A.5 Hyperparameter tuning

We adapt the search space for the hyperparameters of each machine learning model to the requirements of our restricted sample. In particular, GKX set the maximum depth of each tree in their random forest to 6. We increase this upper boundary to 30, which improves the validation results, especially at the one-year horizon. We also extend the search space for the elastic net's ℓ_1-ratio, which in GKX is fixed at 0.5, to allow for a more flexible combination of ℓ_1- and ℓ_2-penalization. For the gradient boosted regression trees, we limit the number of trees to the interval $[2, 100]$, increase the maximum tree depth to 3, and extend the interval for the learning rate to $[0.005, 0.12]$. In the case of the neural networks, we switch from the seed value-based ensemble approach advocated by GKX to dropout regularization, in combination with a structural ensemble approach, such that each neural network in the ensemble can have a different architecture. Ensemble methods have proven to be the gold standard in many machine learning applications, because they can subsume the different aspects learned by each individual model within a single prediction. However, creating ensembles can become prohibitively expensive if the number of sample observations is large and/or each individual model is highly complex. Srivastava et al. (2014) address this issue by proposing dropout regularization, which retains the capability of neural networks to learn different aspects of the data while also being computationally more efficient than the standard ensemble approach. We also introduce a maximum weight norm for each hidden layer. By applying both dropout regularization and a structural ensemble approach with ten different neural networks per ensemble, we seek to combine the best of both worlds. Compared to GKX, we also reduce the batch size; a smaller batch size typically improves the generalization capabilities of a model that is trained with stochastic gradient descent (cf. Keskar et al., 2016). For a detailed comparison of the hyperparameter search spaces, please refer to Table 2.1 in the main text and Table A.5 in GKX.

Table A.1: Variable description. The table contains information on the variables used for the empirical analysis. Panel A covers the theory/option-based risk premium measures proposed by Martin and Wagner (2019), Kadan and Tang (2020), and Martin (2017). The information in Panels B1 and B2 is taken from Table A.6 in Gu et al. (2020). For each variable, the table reports its debut in finance literature (author(s), year, journal), from which database it can be constructed (source), and at which frequency it is reported (freq.). For the stock-level features, we also supply the name of the respective variable used in the SAS program supplied by Jeremiah Green. The names of the macro-level variables come from Amit Goyal's original data files.

Panel A: Theory-based variables	Source	Freq.	Author(s)	Year	Jnl.
MW	Compustat, CRSP, OptionMetrics	Daily	Martin & Wagner	2019	JF
KT	Compustat, CRSP, OptionMetrics	Daily	Kadan & Tang	2019	RFS
Lower bound market equity premium	Compustat, CRSP, OptionMetrics	Daily	Martin	2017	QJE

Panel B1: Stock-level variables	Code name	Source	Freq.	Author(s)	Year	Jnl.
1-month momentum	mom1m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
6-month momentum	mom6m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
12-month momentum	mom12m	CRSP	Monthly	Jegadeesh	1990	JF
36-month momentum	mom36m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
Abnormal earnings announcement volume	aeavol	Compustat, CRSP	Quarterly	Lerman, Livnat & Mendenhall	2007	WP
Absolute accruals	absacc	Compustat	Annual	Bandyopadhyay, Huang & Wirjanto	2010	WP
Accrual volatility	stdacc	Compustat	Quarterly	Bandyopadhyay, Huang & Wirjanto	2010	WP
Asset growth	agr	Compustat	Annual	Cooper, Gulen & Schill	2008	JF
Beta	beta	CRSP	Monthly	Fama & MacBeth	1973	JPE
Beta squared	betasq	CRSP	Monthly	Fama & MacBeth	1973	JPE
Bid-ask spread	baspread	CRSP	Monthly	Amihud & Mendelson	1989	JF
Book-to-market	bm	Compustat, CRSP	Annual	Rosenberg, Reid & Lanstein	1985	JPM
Capital expenditures and inventory	invest	Compustat	Annual	Chen & Zhang	2010	JF
Cash flow-to-debt	cashdebt	Compustat	Annual	Ou & Penman	1989	JAE
Cash flow-to-price	cfp	Compustat	Annual	Desai, Rajgopal & Venkatachalam	2004	TAR
Cash flow volatility	stdcf	Compustat	Quarterly	Huang	2009	JEF
Cash holdings	cash	Compustat	Quarterly	Palazzo	2012	JFE
Cash productivity	cashpr	Compustat	Annual	Chandrashekar & Rao	2009	WP
Change in 6-month momentum	chmom	CRSP	Monthly	Gettleman & Marks	2006	WP
Change in inventory	chinv	Compustat	Annual	Thomas & Zhang	2002	RAS
Change in shares outstanding	chcsho	Compustat	Annual	Pontiff & Woodgate	2008	JF
Change in tax expense	chtx	Compustat	Quarterly	Thomas & Zhang	2011	JAR
Convertible debt indicator	convind	Compustat	Annual	Valta	2016	JFQA
Corporate investment	cinvest	Compustat	Quarterly	Titman, Wei & Xie	2004	JFQA
Current ratio	currat	Compustat	Annual	Ou & Penman	1989	JAE

Table A.1 continued . . .

Table A.1 continued . . .

. . .	Code name	Source	Freq.	Author(s)	Year	Jnl.
Debt capacity/firm tangibility	tang	Compustat	Annual	Almeida & Campello	2007	RFS
Depreciation/PP&E	depr	Compustat	Annual	Holthausen & Larcker	1992	JAE
Dividend initiation	divi	Compustat	Annual	Michaely, Thaler & Womack	1995	JF
Dividend omission	divo	Compustat	Annual	Michaely, Thaler & Womack	1995	JF
Dividend-to-price	dy	Compustat	Annual	Litzenberger & Ramaswamy	1982	JF
Dollar market value	mve	CRSP	Monthly	Banz	1981	JFE
Dollar trading volume	dolvol	CRSP	Monthly	Chordia, Subrahmanyam & Anshuman	2001	JFE
Earnings announcement return	ear	Compustat, CRSP	Quarterly	Kishore, Brandt, Santa-Clara & Venkatachalam	2008	WP
Earnings-to-price	ep	Compustat	Annual	Basu	1977	JF
Earnings volatility	roavol	Compustat	Quarterly	Francis, LaFond, Olsson & Schipper	2004	TAR
Employee growth rate	hire	Compustat	Annual	Bazdresch, Belo & Lin	2014	JPE
Financial statement score (q)	ms	Compustat	Quarterly	Mohanram	2005	RAS
Financial statements score (a)	ps	Compustat	Annual	Piotroski	2000	JAR
Gross profitability	gma	Compustat	Annual	Novy-Marx	2013	JFE
Growth in capital expenditures	grcapx	Compustat	Annual	Anderson & Garcia-Feijoo	2006	JF
Growth in common shareholder equity	egr	Compustat	Annual	Richardson, Sloan, Soliman & Tuna	2005	JAE
Growth in long term net operating assets	grltnoa	Compustat	Annual	Fairfield, Whisenant & Yohn	2003	TAR
Growth in long-term debt	lgr	Compustat	Annual	Richardson, Sloan, Soliman & Tuna	2005	JAE
Idiosyncratic return volatility	idiovol	CRSP	Monthly	Ali, Hwang & Trombley	2003	JFE
(Amihud) Illiquidity	ill	CRSP	Monthly	Amihud	2002	JFM
Industry momentum	indmom	CRSP	Monthly	Moskowitz & Grinblatt	1999	JF
Industry sales concentration	herf	Compustat	Annual	Hou & Robinson	2006	JF
Industry-adjusted book-to-market	bm_ia	Compustat, CRSP	Annual	Asness, Porter & Stevens	2000	WP
Industry-adjusted cash flow-to-price ratio	cfp_ia	Compustat	Annual	Asness, Porter & Stevens	2000	WP
Industry-adjusted change in asset turnover	chatoia	Compustat	Annual	Soliman	2008	TAR
Industry-adjusted change in employees	chempia	Compustat	Annual	Asness, Porter & Stevens	1994	WP
Industry-adjusted change in profit margin	chpmia	Compustat	Annual	Soliman	2008	TAR
Industry-adjusted % change in capital exp.	pchcapx_ia	Compustat	Annual	Abarbanell & Bushee	1998	TAR
Leverage	lev	Compustat	Annual	Bhandari	1988	JF
Maximum daily return	maxret	CRSP	Monthly	Bali, Cakici & Whitelaw	2011	JFE
Number of earnings increases	nincr	Compustat	Quarterly	Barth, Elliott & Finn	1999	JAR
Number of years since first Compustat coverage	age	Compustat	Annual	Jiang, Lee & Zhang	2005	RAS
Operating profitability	operprof	Compustat	Annual	Fama & French	2015	JFE
Organizational capital	orgcap	Compustat	Annual	Eisfeldt & Papanikolaou	2013	JF
% change in current ratio	pchcurrat	Compustat	Annual	Ou & Penman	1989	JAE
% change in depreciation	pchdepr	Compustat	Annual	Holthausen & Larcker	1992	JAE
% change in gross margin - % change in sales	pchgm_pchsale	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in quick ratio	pchquick	Compustat	Annual	Ou & Penman	1989	JAE
% change in sales - % change in A/R	pchsale_pchrect	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in sales - % change in inventory	pchsale_pchinvt	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in sales - % change in SG&A	pchsale_pchxsga	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change sales-to-inventory	pchsaleinv	Compustat	Annual	Ou & Penman	1989	JAE

Table A.1 continued . . .

Table A.1 continued . . .

. . .	Code name	Source	Freq.	Author(s)	Year	Jnl.
Percent accruals	pctacc	Compustat	Annual	Hafzalla, Lundholm & Van Winkle	2011	TAR
Price delay	pricedelay	CRSP	Monthly	Hou & Moskowitz	2005	RFS
Quick ratio	quick	Compustat	Annual	Ou & Penman	1989	JAE
R&D increase	rd	Compustat	Annual	Eberhart, Maxwell & Siddique	2004	JF
R&D-to-market capitalization	rde_mve	Compustat	Annual	Guo, Lev & Shi	2006	JBFA
R&D-to-sales	rd_sale	Compustat	Annual	Guo, Lev & Shi	2006	JBFA
Real estate holdings	realestate	Compustat	Annual	Tuzel	2010	RFS
Return on assets	roaq	Compustat	Quarterly	Balakrishnan, Bartov & Faurel	2010	JAE
Return on equity	roeq	Compustat	Quarterly	Hou, Xue & Zhang	2015	RFS
Return on invested capital	roic	Compustat	Annual	Brown & Rowe	2007	WP
Return volatility	retvol	CRSP	Monthly	Ang, Hodrick, Xing & Zhang	2006	JF
Revenue surprise	rsup	Compustat	Quarterly	Kama	2009	JBFA
Sales growth	sgr	Compustat	Annual	Lakonishok, Shleifer & Vishny	1994	JF
Sales-to-cash	salecash	Compustat	Annual	Ou & Penman	1989	JAE
Sales-to-inventory	saleinv	Compustat	Annual	Ou & Penman	1989	JAE
Sales-to-price	sp	Compustat	Annual	Barbee, Mukherji, & Raines	1996	FAJ
Sales-to-receivables	salerec	Compustat	Annual	Ou & Penman	1989	JAE
Secured debt indicator	securedind	Compustat	Annual	Valta	2016	JFQA
Share turnover	turn	CRSP	Monthly	Datar, Naik & Radcliffe	1998	JFM
Sin stocks	sin	Compustat	Annual	Hong & Kacperczyk	2009	JFE
Tax income-to-book income	tb	Compustat	Annual	Lev & Nissim	2004	TAR
Volatility of liquidity (dollar trading vol.)	std_dolvol	CRSP	Monthly	Chordia, Subrahmanyam & Anshuman	2001	JFE
Volatility of liquidity (share turnover)	std_turn	CRSP	Monthly	Chordia, Subrahmanyam, & Anshuman	2001	JFE
Working capital accruals	acc	Compustat	Annual	Sloan	1996	TAR
Zero trading days	zerotrade	CRSP	Monthly	Liu	2006	JFE

Panel B2: Macro-level variables	Code name	Source	Freq.	Author(s)	Year	Jnl.
Book-to-market ratio	b/m	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Default yield spread	dfy	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Dividend-price ratio	dp	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Earnings-price ratio	eq	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Net equity expansion	ntis	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Stock variance	svar	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Term spread	tms	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Treasury bill rate	tbl	Amit Goyal	Monthly	Welch & Goyal	2008	RFS

A.6 Comparison with Gu et al. (2020)

The part of our study that deals with the machine learning approaches draws on Gu et al. (2020). Because we have to ensure the comparability with the theory-based part, we deviate in some aspects, in particular, the selection of stocks, the sample period, and the training and validation strategy. In the following, we explain these differences and the reasons for our choices.

First, as outlined in the main text, the theory-consistent approaches and the hybrid models suggest focusing on S&P 500 constituents. Gu et al. (2020) instead rely on a broader set of NYSE-, AMEX-, and NASDAQ-traded firms and also include penny stocks, yielding an average number of stocks per month of about 6,200. For the purpose of training and testing of non-hybrid models, we could have used such an extended set of stocks, too. However, for the sake of a neat comparison, we focus on S&P 500 constituents both for training and performance evaluation. As argued by Avramov et al. (2023), this restriction can represent a reasonable economic constraint that acknowledges that trading microcaps is costly. They find evidence that the predictive performance of machine learning models deteriorates when excluding microcaps.

Second, the overall sample period used in the present analysis deviates from that of Gu et al. (2020). We use more recent data that became available, but we have also decided to start the training process later. Gu et al. (2020) start training in 1957, the birth year of the S&P 500 index. However, as outlined in the main text, there is a considerable amount of missing values until 1974. In particular, some features, for example *cash flow volatility*, are not available earlier. We follow Gu et al. (2020) who replace a feature's missing value with the cross-sectional median at a given point in time, but in the case of a variable like *cash flow volatility*, this strategy amounts to setting all its values to zero before 1974. More sophisticated imputation methods could be considered, but then results could be biased by overly restrictive assumptions about the structure of the missing data.

Third, we have to consider an out-of-sample testing period that facilitates the comparison with the theory-consistent approaches. Gu et al. (2020) report their out-of-sample results on a testing period that ranges from 1987 to 2016, but this is not tenable here. The option data used to construct the theory-consistent forecasts are only available from 1996 onward. Accordingly, our out-of-sample testing period ranges from 1996 (long training) or 1998 (short training) to 2018.

A.7 Alternative feature transformation

As described in the main text, we apply standard mean-variance or robust median-interquartile range scaling to the firm characteristics z_t^i, pooling across i and t. To prevent future information from leaking into the validation and test sets, the transformation of a feature within those sets is based on the mean, variance, median, and interquartile range in the associated training sets. In contrast, GKX scale firm characteristics to the interval $[-1, 1]$ period-by-period using cross-sectional ranks, as

advocated by Freyberger et al. (2020). More specifically, they transform their set of firm characteristics according to

$$\tilde{c}_t^i = 2 \cdot \frac{\text{rank}(c_t^i)}{N_t + 1} - 1, \tag{A-18}$$

where N_t is the number of sample firms in period t.[34] The macroeconomic features x_t are not scaled, because for the individual time series there is no cross-section on the basis of which a rank transformation could be performed. As a consequence, the set of combined firm-level and macro features originates from

$$\tilde{z}_t^i = (1, x_t')' \otimes \tilde{c}_t^i. \tag{A-19}$$

Which feature scaling strategy is more suitable for the present application? The rank transformation in Equation (A-18) invokes the idea of portfolio sorting, the hallmark of which is that "[one is] typically not interested in the value of a characteristic in isolation, but rather in the rank of the characteristic in the cross section" (Freyberger et al., 2020, pp. 16-17). In the same vein, Kozak et al. (2020) argue that by transforming firm characteristics according to their rank, they can focus on the "purely cross-sectional aspects of return predictability." However, the present study does not exclusively focus on the cross-section, but is also concerned with the *level* of stock risk premia. Using rank-transformed features, one cannot account for structural changes in the level of firm characteristics.[35]

Kelly et al. (2019) and Gu et al. (2021), point out that the rank transformation renders models less susceptible to outliers. However, Kelly et al. (2019) also report that the "results are qualitatively unchanged" compared to those obtained without rank transformation. Da et al. (2022) arrive at a similar conclusion, reporting that the rank transformation "barely changes any follow-up results." As we aim at finding the model that delivers MSE-optimal excess return predictions, the question of how to transform and scale firm characteristics is ultimately a matter of out-of-sample forecast performance (cf. Freyberger et al., 2020). Accordingly, we leave it up to the validation process whether to apply standard or robust scaling, noting that the latter mitigates the issue of outlier susceptibility.

To investigate whether our conclusions from the main analysis are affected by the chosen feature transformation strategy, we perform a supplementary analysis using rank-transformed firm-level features according to Equations (A-18) and (A-19). We thereby acknowledge the code of conduct for research in empirical finance formulated by Arnott et al. (2019).

[34] GKX give no indications as to their treatment of stocks that are tied in the ranking. We assume that they rank tied stocks as in Kozak et al. (2020) by assigning the average rank to each of the stocks.

[35] An obvious thing to note is that without scaling the macro features, the $\tilde{z}_t^i$ are not elements of $[-1, 1]$.

Results using rank-transformed firm-level features

Table A.2 contains the *long training* results for both horizons. It is the counterpart of Panels B of Tables 2.2 and 2.3 from the main analysis.

At the *one-month horizon* (Panel A of Table A.2), RF and GBRT perform worse than the zero forecast, while ANN and ENet benefit from using rank transformed features. Compared to the main analysis, the predictive R^2 increase from 0.2% to 0.4% in case of the ANN and, quite conspicuously, from -0.3% to 0.5% in case of the ENet. Figure A.2 depicts the results for prediction-sorted portfolios. It should be compared with Figure 2.4, the counterpart from the main analysis. The plots confirm the conclusion that the theory-based approach is difficult to beat at the one-month horizon, but also that the ENet is emerging as a new competitor.

Panel B of Table A.2 shows that at the *one-year horizon* ENet, GBRT, and ANN by and large maintain their performance levels from the main analysis (cf. Panel B of Table 2.3). The ENet's R^2_{oos} increases from 5.5% to 6.9%, the predictive R^2 of ANN (from 9.0% to 8.1%) and GBRT (from 10.6% to 9.7%) decrease. In terms of R^2_{oos}, the RF is not as conspicuous as in the main analysis. The R^2_{oos} decreases from 19.5% to 9.6%, but with a Sharpe ratio of 0.67 (increasing from 0.58), the RF is the best approach when prediction-sorted portfolios are used for performance assessment. The ANN is ranked second with a Sharpe ratio of 0.63 (increasing from 0.50) and a favorable alignment of the prediction-sorted portfolios (see Figure A.3, the counterpart of Figure 2.5 from the main analysis).

As can be seen in Table A.3 – which should be compared to Table 2.5 from the main analysis – the *short training* effect is somewhat mitigated at the *one-month horizon*. Although still negative, the predictive R^2 delivered by the machine learning approaches no longer tend to extremes. As in the main analysis, the inclusion of theory features does not improve the one-month horizon results.

At the *one-year horizon* with *short training*, our assessment of the model performances does not differ substantially from that of the main analysis (compare Table A.4 with Table 2.6). In terms of R^2_{oos} and Sharpe ratio, the RF is the preferred model. Its R^2_{oos} increases from 12.4% to 15% and the Sharpe ratio of 0.59 remains unchanged. The ANN ranks second according to both criteria, with an R^2_{oos} of 11.5% (down from 14.1% in the main analysis) and a Sharpe ratio of 0.50 (up from 0.47). As in the main analysis, GBRT (deteriorating) and ENet (though notably improving) are no strong competitors.

Table A.4 further shows that the inclusion of theory features does not improve the performance of the machine learning models, at least when a monthly forecast frequency is considered.The conclusions regarding the *theory assisted by ML* strategy also hold with rank-transformed features, insofar as the predictive R^2 of 9.1% and the Sharpe ratio of 0.37 delivered by MW are notably improved by RF assistance. The MW+RF hybrid delivers an R^2_{oos} of 13.0%, a Sharpe ratio of 0.58, and a favorable alignment of the prediction-sorted portfolios (see Figure A.4). Similar to the main analysis, the ANN assistance proves useful, too (the R^2_{oos} of MW+RF is 11.2%, the

Sharpe ratio is 0.45), while GBRT or ENet assistance does not.

Overall, we find that the conclusions of the main analysis are also supported when using rank-transformed firm-level features.

Table A.2: Performance comparison, monthly forecast frequency: long training, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the forecast horizon is one month and for Panel B, it is one year. In both panels, forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in November 2018. The features are rank-scaled as described in Appendix A.7. The machine learning results are obtained using the long training scheme depicted in Figure 2.2.

Panel A: one-month horizon

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.2	3.2	0.154	0.30
	KT	−1.8	6.9	0.704	0.30
Machine Learning	ENet	0.5	3.5	0.073	0.65
	ANN	0.4	3.4	0.053	0.34
	GBRT	−0.8	4.3	0.300	0.37
	RF	−0.8	4.8	0.294	0.17

Panel B: one-year horizon

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	8.8	16.3	0.051	0.37
	KT	3.1	47.6	0.694	0.37
Machine Learning	ENet	6.9	22.5	0.174	0.49
	ANN	8.1	22.1	0.097	0.63
	GBRT	9.7	23.1	0.086	0.49
	RF	9.6	43.3	0.361	0.67

Figure A.2: Prediction-sorted portfolios, one-month horizon: long training, rank transformation. The stocks are sorted into deciles according to the one-month horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot the predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), ENet (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to November 2018. The features are rank-scaled as described in Appendix A.7. Machine learning results are based on the long training scheme depicted in Figure 2.2.

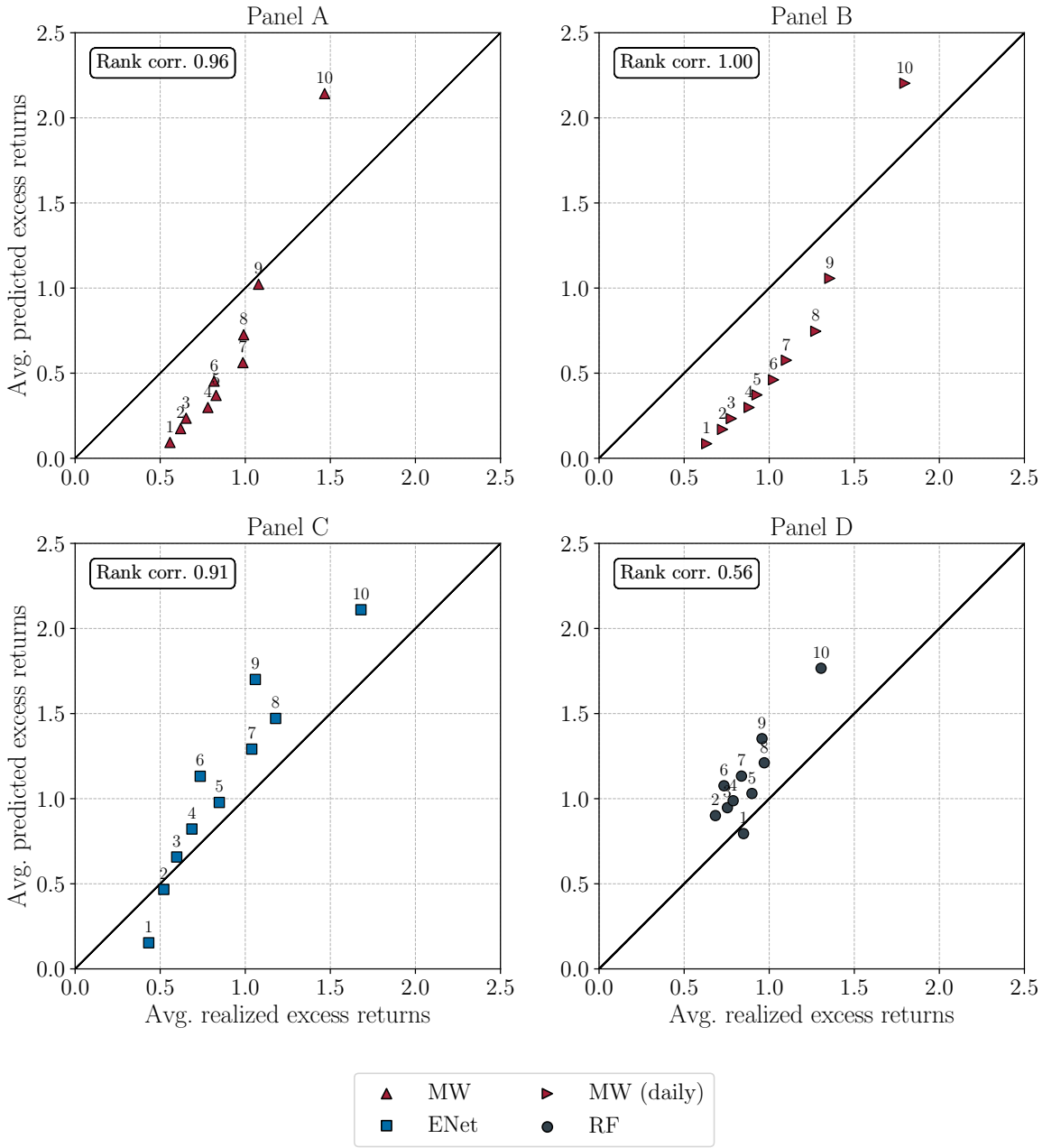


Figure A.3: Prediction-sorted portfolios, one-year horizon: long training, rank transformation. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to December 2017. The features are rank-scaled as described in Appendix A.7. Machine learning results are based on the long training scheme depicted in Figure 2.2.

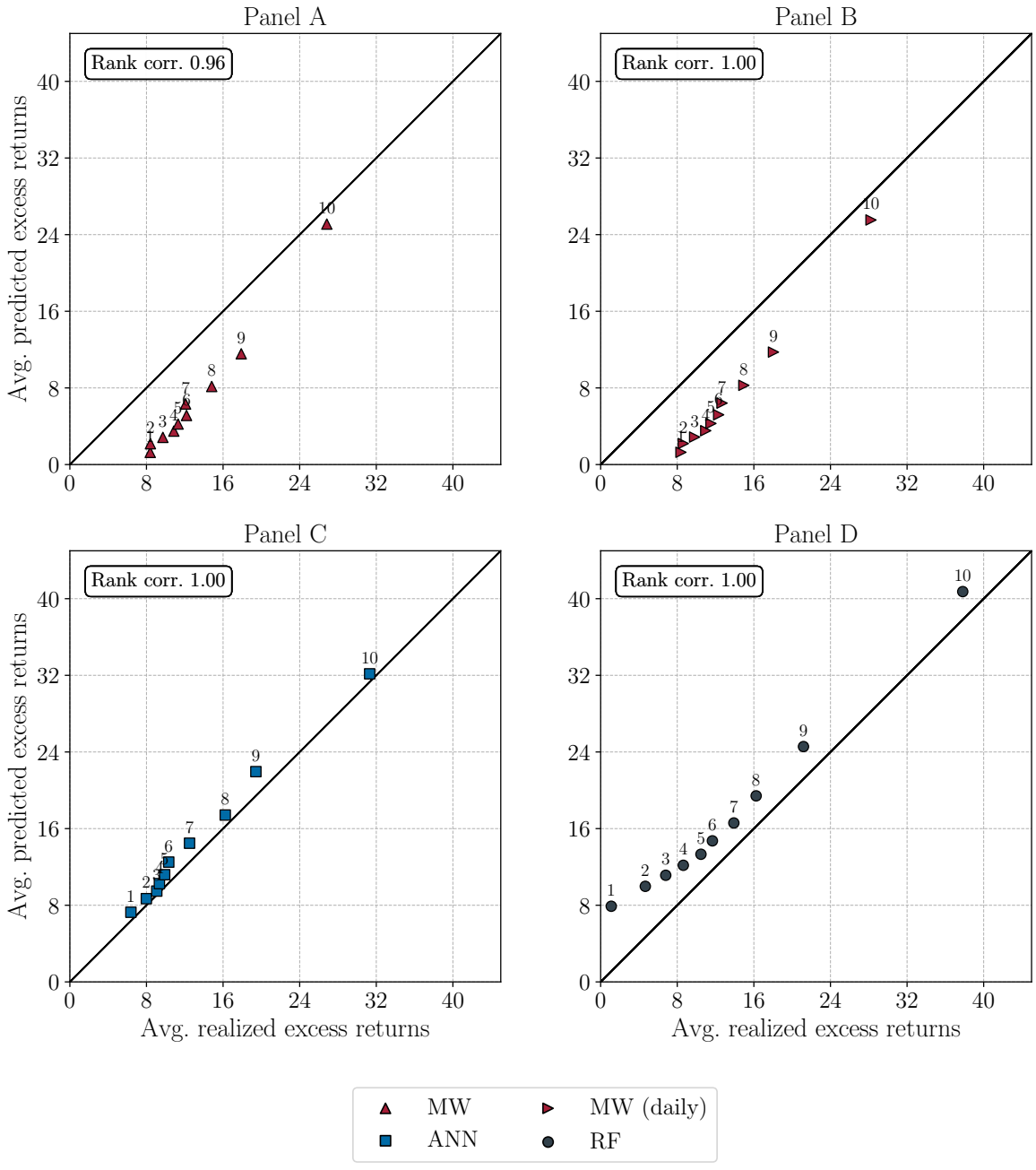


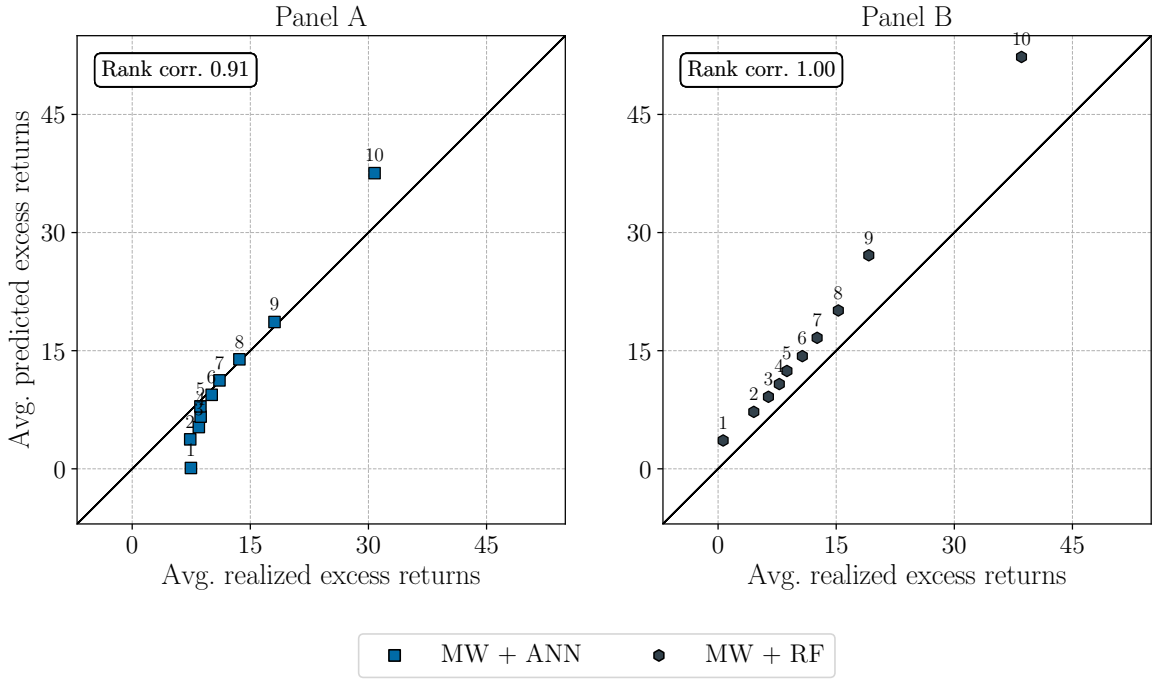
Table A.3: Performance comparison, one-month horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approach, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, the four machine learning models, and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. The one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1998 and ends in November 2018. The features are rank-scaled as described in Appendix A.7. The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	0.1	3.4	0.206	0.32
	KT	−2.0	7.2	0.739	0.32
Machine Learning	ENet	−0.1	2.8	0.277	0.26
	ANN	−0.1	2.9	0.163	0.04
	GBRT	−2.5	5.3	0.914	0.17
	RF	−4.7	8.3	0.898	−0.06
ML with theory features	ENet	−0.1	2.8	0.277	0.26
	ANN	−0.2	3.0	0.214	0.15
	GBRT	−8.5	15.9	0.926	0.19
	RF	−5.7	9.8	0.943	−0.11

Table A.4: Performance comparison, one-year horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which the machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued monthly (end-of-month). The features are rank-scaled as described in Appendix A.7. The machine learning results are obtained using the short training scheme depicted in Figure 2.3.

		$R^2_{oos} \times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.1	17.1	0.072	0.37
	KT	3.1	49.9	0.706	0.37
Machine Learning	ENet	4.3	25.3	0.388	0.49
	ANN	11.5	22.2	0.048	0.50
	GBRT	6.5	30.9	0.521	0.39
	RF	15.0	35.4	0.186	0.59
ML with theory features	ENet	4.3	25.3	0.385	0.49
	ANN	11.1	23.5	0.096	0.45
	GBRT	6.1	32.8	0.596	0.42
	RF	14.0	35.7	0.236	0.57
Theory assisted by ML	ENet	8.6	31.4	0.331	0.47
	ANN	11.2	27.7	0.183	0.45
	GBRT	6.2	38.7	0.548	0.40
	RF	13.0	42.4	0.320	0.58

Figure A.4: Prediction-sorted portfolios, one-year horizon: theory assisted by machine learning approaches (rank transformation). The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed at the end of each month. The two panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW assisted by an ANN (MW + ANN, Panel A) and MW assisted by RF (MW+RF, Panel B). The out-of-sample period ranges from January 1998 to December 2017. The features are rank-scaled as described in Appendix A.7. Results are based on the short training scheme depicted in Figure 2.3.



A.8 Additional results

In this section, we present additional results regarding alternative rank-transformations, goodness-of-fit measures, and disaggregated analyses.

Table A.5: Performance comparison, one-month horizon: theory-based vs. machine learning approaches (long training). The table reports predictive R^2, EV and XS, and the rank correlation (Kendall's τ) between average expected and realized excess returns of prediction-sorted decile portfolios, and p-values of a Diebold-Mariano test implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. For Panel A, the one-month horizon forecasts are issued at a daily frequency. For Panel B, the one-month horizon forecasts are issued at the end of each month. The Diebold-Mariano test is based on the average $R^2_{oos,s}$ across test samples (using the theory-based forecast by Martin and Wagner (2019) at a daily (Panel A) or end-of-month frequency (Panel B) as a base). The out-of-sample testing period starts in January 1996 and ends in November 2018. The machine learning results are obtained using the long training scheme.

Panel A: daily forecast frequency

		$R^2_{oos}\times100$	$EV_{oos}\times100$	$XS_{oos}\times100$	corr	DM
Theory-Based	MW	0.9	0.9	0.4	1.00	
	KT	−0.5	−0.6	−0.6	1.00	0.094
Machine Learning	ENet	0.0	0.0	−0.8	0.33	0.295
	ANN	0.5	0.5	−0.3	0.38	0.441
	GBRT	0.3	0.3	−0.4	0.87	0.443
	RF	−0.5	−0.6	−1.3	0.73	0.342

Panel B: monthly forecast frequency

		$R^2_{oos}\times100$	$EV_{oos}\times100$	$XS_{oos}\times100$	corr	DM
Theory-Based	MW	0.2	0.1	−0.2	0.96	
	KT	−1.8	−1.8	−1.6	0.96	0.089
Machine Learning	ENet	−0.3	−0.3	−0.9	0.24	0.479
	ANN	0.2	0.2	−0.3	0.56	0.883
	GBRT	−0.6	−0.6	−1.1	0.56	0.353
	RF	−1.6	−1.6	−2.2	0.91	0.301

Table A.6: Performance comparison, annual horizon: theory-based vs. machine learning approaches (long training). The table reports predictive R^2, EV and XS, and the rank correlation (Kendall's τ) between average expected and realized excess returns of prediction-sorted decile portfolios, and p-values of a Diebold-Mariano test implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. For Panel A, the annual horizon forecasts are issued at a daily frequency. For Panel B, the annual horizon forecasts are issued at the end of each month. The Diebold-Mariano test is based on the average $R^2_{oos,s}$ across test samples (using the theory-based forecast by Martin and Wagner (2019) at a daily (Panel A) or end-of-month frequency (Panel B) as a base). The out-of-sample testing period starts in January 1996 and ends in November 2018. The machine learning results are obtained using the long training scheme.

Panel A: daily forecast frequency

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	9.1	9.0	4.3	1.00	
	KT	3.5	3.0	−0.9	1.00	0.299
Machine Learning	ENet	4.0	4.0	−2.2	0.60	0.235
	ANN	8.2	8.2	1.6	1.00	0.626
	GBRT	9.9	9.9	3.2	0.91	0.527
	RF	18.2	18.0	11.8	1.00	0.007

Panel B: monthly forecast frequency

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	8.8	8.7	4.1	0.96	
	KT	3.1	2.6	−1.3	0.96	0.295
Machine Learning	ENet	5.5	5.5	−0.4	0.64	0.259
	ANN	9.0	8.9	2.5	1.00	0.919
	GBRT	10.6	10.6	4.2	0.91	0.195
	RF	19.5	19.3	13.3	1.00	0.003

Table A.7: Performance comparison, one-month horizon: theory-based vs. machine learning approaches vs. a hybrid approach (short training). The table reports predictive R^2, EV and XS, and the rank correlation (Kendall's τ) between average expected and realized excess returns of prediction-sorted decile portfolios, and p-values of a Diebold-Mariano test implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, the four machine learning models, and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). For Panel A, the one-month horizon forecasts are issued at a daily frequency, for Panel B, the one-month horizon forecasts are issued at the end of each month. The Diebold-Mariano test is based on the average $R^2_{oos,s}$ across test samples (using the theory-based forecast by Martin and Wagner (2019) at a daily (Panel A) or end-of-month frequency (Panel B) as a base). The out-of-sample testing period starts in January 1998 and ends in November 2018. The machine learning results are obtained using the short training scheme.

Panel A: daily forecast frequency

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	0.8	0.8	0.4	1.00	
	KT	−0.7	−0.7	−0.5	1.00	0.084
Machine Learning	ENet	−4.0	−4.0	−4.5	0.78	0.087
	ANN	−2.7	−2.8	−3.3	0.69	0.109
	GBRT	−22.6	−22.7	−23.2	0.29	0.198
	RF	−5.4	−5.5	−6.0	−0.24	0.106
ML with theory features	ENet	−3.0	−3.0	−3.5	0.87	0.055
	ANN	−30.7	−32.0	−30.4	0.51	0.262
	GBRT	−10.7	−10.8	−11.3	0.42	0.222
	RF	−3.0	−3.0	−3.6	0.69	0.067

Panel B: monthly forecast frequency

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	0.1	0.1	−0.2	0.96	
	KT	−2.0	−2.1	−1.5	0.96	0.086
Machine Learning	ENet	−4.0	−4.0	−4.2	0.64	0.130
	ANN	−3.1	−3.2	−3.6	0.69	0.117
	GBRT	−29.5	−29.6	−29.8	0.38	0.245
	RF	−8.4	−8.4	−8.8	−0.33	0.173
ML with theory features	ENet	−3.2	−3.2	−3.4	0.69	0.130
	ANN	−36.0	−37.4	−35.1	0.24	0.264
	GBRT	−25.6	−25.7	−25.8	0.29	0.253
	RF	−7.6	−7.6	−8.1	−0.20	0.157

Table A.8: Performance comparison, one-year horizon, monthly forecast frequency: Theory-based vs. machine learning approaches vs. hybrid approaches (short training). The table reports predictive R^2, EV and XS, and the rank correlation (Kendall's τ) between average expected and realized excess returns of prediction-sorted decile portfolios, and p-values of a Diebold-Mariano test implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, and the four machine learning models. Results of a hybrid approach, in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*) and a second hybrid approach, in which machine learning models are trained account for the approximation residuals of MW (*Theory assisted by ML*) are also reported. The Diebold-Mariano test is based on the average $R^2_{oos,s}$ across test samples using the theory-based forecast by Martin and Wagner (2019) at a monthly frequency (end-of-month) as a base. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued monthly (end-of-month). The machine learning results are obtained using the short training scheme.

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	9.1	9.0	4.9	1.00	
	KT	3.1	2.5	0.3	1.00	0.315
Machine Learning	ENet	−31.6	−34.0	−40.0	0.96	0.131
	ANN	14.1	13.9	8.1	1.00	0.130
	GBRT	10.3	9.8	4.1	1.00	0.849
	RF	12.4	11.7	6.0	1.00	0.645
ML with theory features	ENet	−32.6	−35.2	−41.0	0.96	0.139
	ANN	14.1	13.8	8.2	1.00	0.265
	GBRT	9.7	9.1	3.4	0.96	0.973
	RF	14.6	14.0	8.4	1.00	0.387
Theory assisted by ML	MW+ENet	−38.2	−41.3	−47.5	0.96	0.168
	MW+ANN	14.2	13.9	8.5	1.00	0.108
	MW+GBRT	9.2	8.6	3.9	1.00	0.955
	MW+RF	16.1	15.3	10.8	1.00	0.367

Table A.9: Performance comparison, one-year horizon, daily forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches (short training). The table reports predictive R^2, EV, and XS, and the rank correlation (Kendall's τ) between average expected and realized excess returns of prediction-sorted decile portfolios, and p-values of a Diebold-Mariano test implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, and the four machine learning models. Results of a hybrid approach, in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*) and a second hybrid approach, in which machine learning models are trained account for the approximation residuals of MW (*Theory assisted by ML*) are also reported. The Diebold-Mariano test is based on the average $R^2_{oos,s}$ across test samples using the theory-based forecast by Martin and Wagner (2019) at a daily frequency as a base. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued at a daily frequency. The machine learning results are obtained using the short training scheme.

		$R^2_{oos}\times 100$	$EV_{oos}\times 100$	$XS_{oos}\times 100$	corr	DM
Theory-Based	MW	9.5	9.3	5.2	1.00	
	KT	3.4	2.9	0.7	1.00	0.318
Machine Learning	ENet	−35.5	−38.0	−44.5	0.96	0.092
	ANN	12.0	11.8	5.7	1.00	0.476
	GBRT	8.8	8.2	2.2	1.00	0.816
	RF	9.0	8.2	2.0	1.00	0.886
ML with theory features	ENet	−27.4	−29.6	−35.8	0.96	0.123
	ANN	16.1	15.7	10.0	1.00	0.210
	GBRT	11.6	11.0	5.3	0.96	0.716
	RF	18.6	17.9	12.3	1.00	0.155
Theory assisted by ML	MW+ENet	−41.2	−44.4	−51.0	1.00	0.125
	MW+ANN	12.8	12.4	6.7	1.00	0.304
	MW+GBRT	8.2	7.6	2.7	1.00	0.774
	MW+RF	14.1	13.3	8.6	1.00	0.567

Figure A.5: Feature importance, one-year horizon: random forest (short training). The figure depicts feature importances (Panel A: firm-level features, Panel B: macro-level features) for the RF. The forecast horizon is one year, the prediction frequency is end-of-month. A feature's importance is measured by the reduction in the Sharpe ratio of a long-short investment strategy into prediction-sorted portfolios that is induced by setting the feature's values in the test samples to zero. In both panels, the features are sorted in descending order of importance. The dashed vertical line is included for reference and represents the Sharpe ratio that is obtained without setting any feature's values to zero. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme.

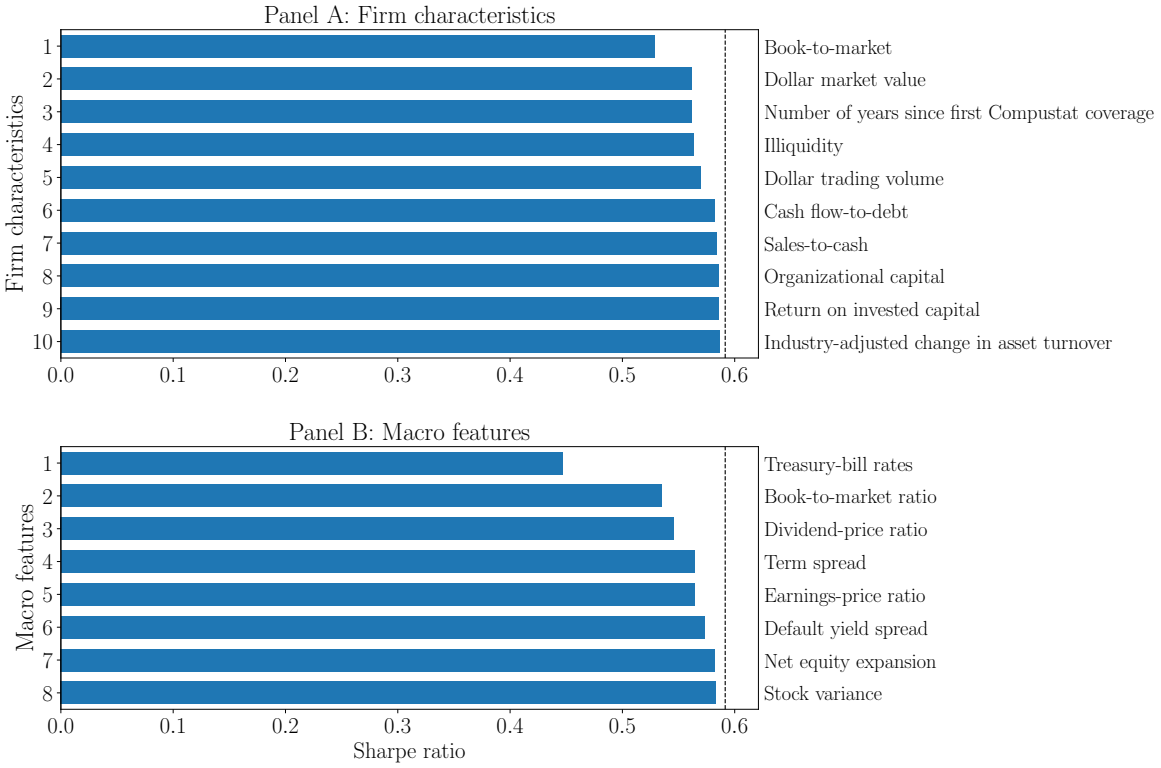


Figure A.6: Feature importance, one-year horizon: MW assisted by random forest (short training). The figure depicts feature importances (Panel A: firm-level features, Panel B: macro-level features) for the MW assisted by RF strategy. The forecast horizon is one year, the prediction frequency is end-of-month. A feature's importance is measured by the reduction in the Sharpe ratio of a long-short investment strategy into prediction-sorted portfolios that is induced by setting the feature's values in the test samples to zero. In both panels, the features are sorted in descending order of importance. The dashed vertical line is included for reference and represents the Sharpe ratio that is obtained without setting any feature's values to zero. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme.

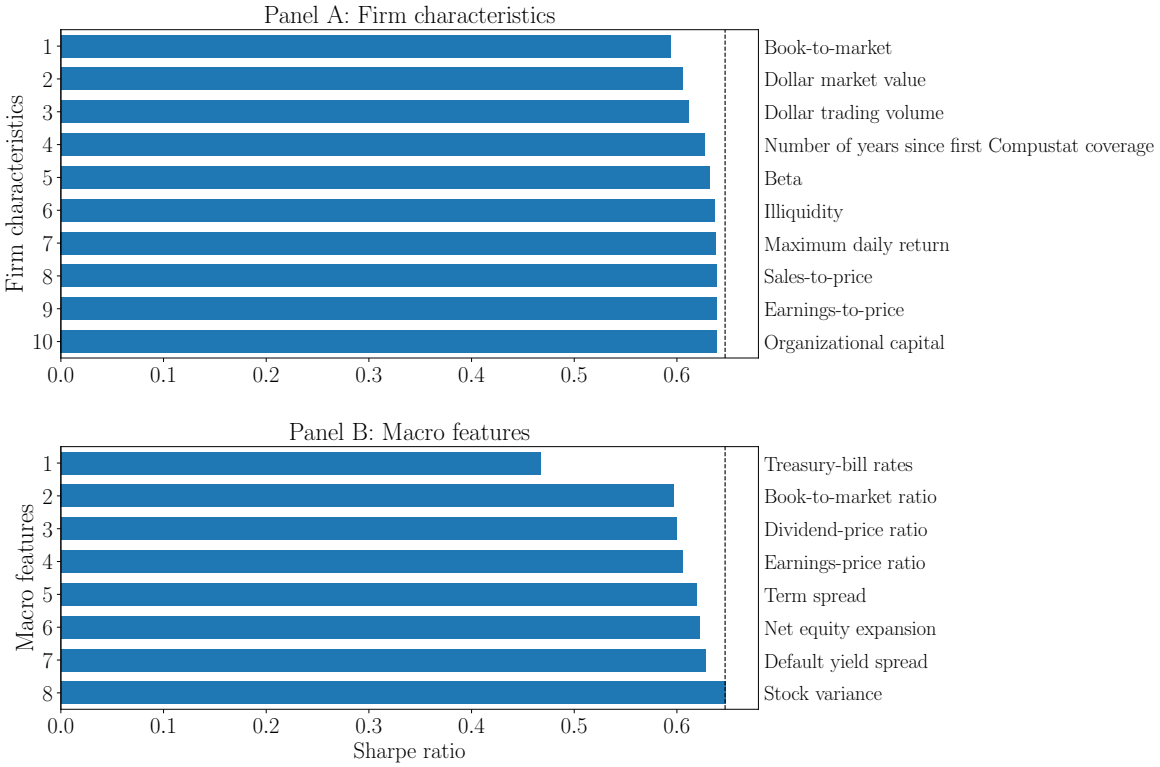
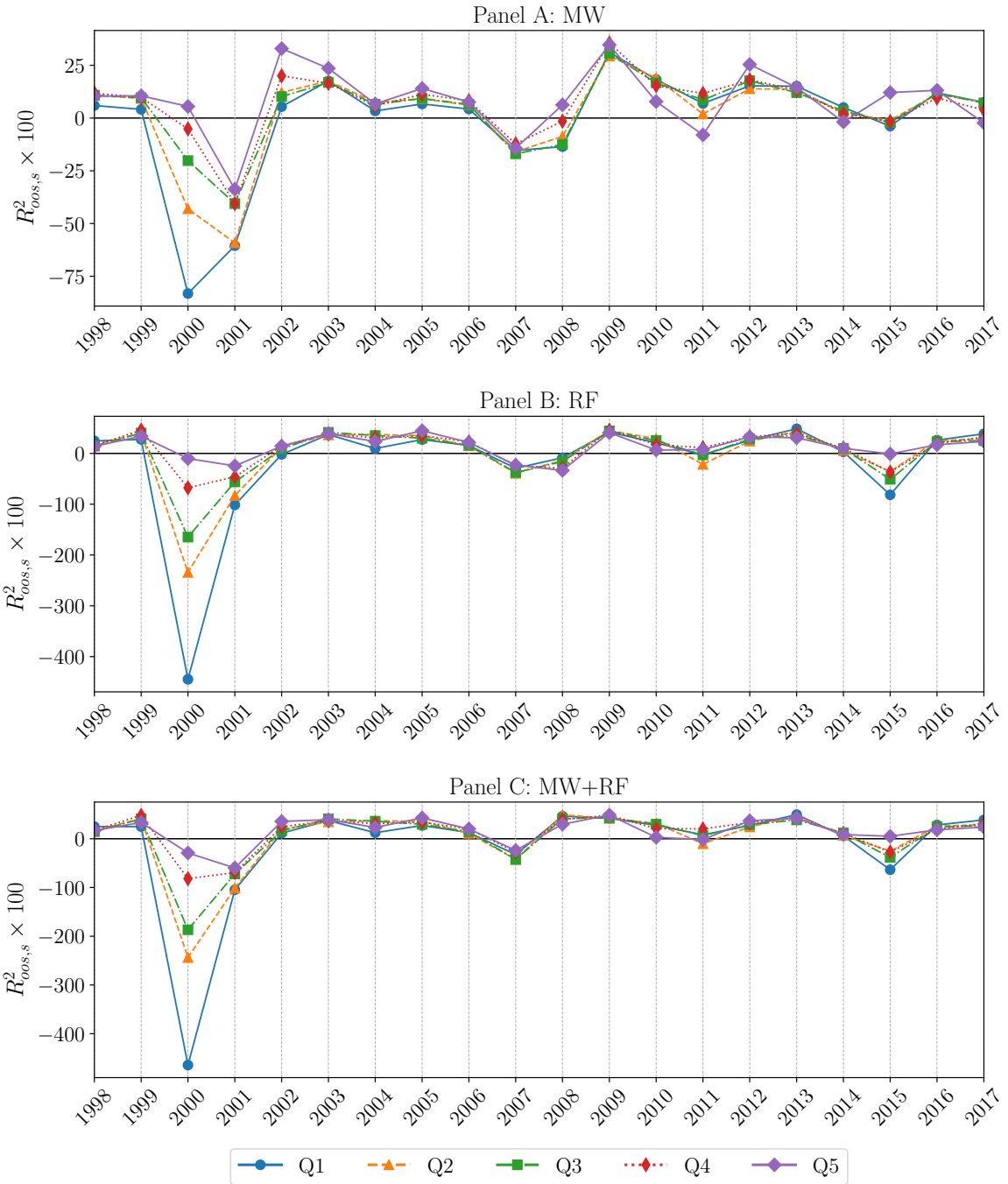


Figure A.7: Time series of predictive R^2, one-year horizon: Quintile portfolios sorted by Amihud illiquidity. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples broken down by quintile portfolios, where the sorting is based on Amihud illiquidity. The forecast horizon is one year, the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1996 to December 2017. Panel A shows the MW results. Panel B shows the pure RF results. Panel C shows the MW+RF results. The machine learning results are obtained using the short training scheme.



Chapter 3

The uncertainty principle in asset pricing

3.1 Motivation

The conditional capital asset pricing model (CAPM) is one of the most controversial models in empirical finance. While some studies conclude that the model performs well in explaining the cross-section of average returns (e.g., Jagannathan and Wang, 1996), there are others suggesting that the conditional CAPM is no more successful than its unconditional counterpart (e.g., Lewellen and Nagel, 2006). A non-negligible source of uncertainty underlying these results is that economic theory provides little guidance on how to obtain measurements of the model's unobserved components – the beta and the equity premium. For this reason, econometricians often resort to estimates of conditional betas from rolling time-series regressions, which are supposed to provide reasonable approximations if the true betas are sufficiently stable over time.[1] Estimates of the conditional equity premium, in turn, are typically obtained from predictive regressions of market excess returns onto a set of predetermined variables, where any variable that outperforms the historical average market excess return is deemed an admissible predictor.[2] However, neither of these approaches is entirely convincing: In the case of the betas, it is unclear over which time period the estimation should be performed (i.e., what the size of the regression window should be), and in the case of the equity premium, the correct state variables are unknown.

In response to these shortcomings, we propose a variant of the conditional CAPM in which the betas and the equity premium are jointly characterized by the information embedded in option prices. An important implication of this approach is that we do not need to assume that the betas are sufficiently stable over time, as they can adjust immediately to changing market conditions through changes in the prices of the underlying options. Similarly, we avoid the multi-dimensional challenge associated with specifying the moments' time variation in terms of firm- and macro characteristics, as we only implicitly refer to the investors' information sets. Because there is no need to estimate *any* structural parameters, we refer to our model as the *fully-implied* CAPM,

[1] Ferson and Harvey (1991), Fama and French (2004), and Frazzini and Pedersen (2014), for example, use betas obtained from rolling-window regressions as instruments for conditional betas.

[2] Frequently used predictors are the dividend-price ratio (e.g., Campbell and Shiller, 1988), interest rate spreads (e.g., Stock and Watson, 1989), the consumption-wealth ratio (e.g., Lettau and Ludvigson, 2001), and the variance-risk premium (e.g., Bollerslev et al., 2009).

or FI-CAPM.

In deriving our model, we build on the findings by Martin and Wagner (2019), who demonstrate that, under certain conditions, the conditional risk premium of a stock can be expressed in terms of risk-neutral variances of returns. Based on this insight, they derive a concise formula for the risk premium that is free of unknown parameters. While we share the same goal, the comparative advantage of our approach lies in its intuitive structure and superior predictive performance. Regarding the latter, we find that the FI-CAPM outperforms the approach by Martin and Wagner (2019) at 7 out of 8 investment horizons. In addition, our identification strategy for the implied beta is inspired by Kempf et al. (2015), whose approach we complement in two respects: 1) we establish a direct link between physical and risk-neutral return distributions, and 2) we provide measurements of both the betas *and* the equity premium.

From an econometric perspective, the novelty of our approach is that it allows us to test the conditional CAPM's unconditional implications in a setting where the betas and the equity premium are defined jointly and in a mutually consistent manner. We test these implications by sorting stocks into portfolios according to the model's predictions and comparing the resulting average implied and realized excess returns. Consistent with previous literature, we find that the relationship between these averages is too flat – a phenomenon that is commonly viewed as evidence that the betas are unable to explain cross-sectional return variation. While this interpretation seems plausible at first glance, it ignores the fact that the betas and the equity premium jointly determine the nature of this relationship. Thus, it may well be that the failure of the conditional CAPM is not due to the betas, but due to the equity premium. To investigate this idea, we propose a modification to the model's testable restrictions that allows us to study their contributions separately.

Comparing the results of the original and modified versions of the moment conditions, we discover that the failure of the conditional CAPM at short investment horizons (1 to 12 months) is due to the inherently unpredictable component of the market excess return, rather than the betas. One might suspect that this finding is driven by the way we measure the model's components, and that using a different approach could alter our results. We address this critique by showing that the above conclusion holds equally well when estimating conditional betas and the equity premium using historical returns, and by arguing that the conditional CAPM's failure in cross-sectional tests at short horizons can be attributed to a property of the equity premium that applies to any admissible specification. To be specific, we find that the positive-sign restriction of the conditional equity premium, which ensures that a risk-averse investor is willing to invest in the market, limits the potential of the betas in cross-sectional tests. Although the betas are successful in describing the stocks' association with the market ex ante, the conditional CAPM fails whenever the realized market excess return is negative, as it is precisely then that high (low) beta portfolios, which are supposed to earn high (low) excess returns, do exactly the opposite.

At longer horizons (beyond 12 months), it is instead the betas that drive the

flat relationship between average predicted and realized excess returns. Although the equity premium accounts for a larger fraction of the variation in market excess returns, and thus constitutes a better forecast in terms of the mean squared error (MSE), the cross-sectional explanatory power of the betas gradually declines. Our reading of this result is that a company's association with the market can change, for example, due to a realignment of the business model, making its prediction increasingly difficult as the investment horizon increases.

In summary, the explanatory power of the betas, when viewed as a function of the investment horizon, is inversely related to the predictive power of the conditional equity premium. The opposing nature of this relationship, in turn, explains the failure of the conditional CAPM across investment horizons. In a way, this finding is reminiscent of Heisenberg's (1927) uncertainty principle in quantum mechanics, which states that two conjugate properties of a particle cannot be measured simultaneously with arbitrary precision. In capital asset pricing, the particle under study is an asset's risk premium, and its properties correspond to the asset-specific and aggregate components of market risk.

The remainder of this article is organized as follows: Section 3.2 gives an overview of the relevant literature. Section 3.3 presents the theoretical foundations of our model, including a comparison with the approach by Martin and Wagner (2019). Section 3.4 provides an evaluation of the model's empirical performance and a discussion of the associated findings. Section 3.5 concludes. Appendix B provides a description of the data and additional analyses.

3.2 Related literature

First, we contribute to a line of research that dates back to the work by French et al. (1983), who examine the usefulness of including option-implied information in measuring betas. Using such information, in general, poses the problem that, while option prices describe moments under the risk-neutral measure, stock risk premia are subject to the physical measure. Any approach that fails to explain the connection between physical and risk-neutral moments hence potentially suffers from a lack in risk adjustment. To draw attention to this issue, Chang et al. (2012) propose a set of assumptions that are needed to estimate beta from implied moments of returns. Buss and Vilkov (2012) define a relationship between objective and risk-neutral correlation that allows them to estimate beta using both option prices and historical returns. Kempf et al. (2015) introduce a family of implied betas based on risk-neutral variance, skewness, and kurtosis, but they make no attempt to risk-adjust implied moments.[3]

Second, we refer to a recent and growing literature that uses information from option prices to approximate conditional risk premia. In a seminal work, Martin (2017) derives an observable lower bound for the conditional expected excess return

[3]Baule et al. (2016) give an overview of the implied-beta literature and compare different identification strategies with regards to their empirical performance.

of the market. Kadan and Tang (2020) investigate the extent to which this bound is applicable to individual stocks. Martin and Wagner (2019) propose a formula for the conditional stock risk premium that is a linear function of risk-neutral stock and market return variances. Schneider and Trojani (2019) provide an extensive family of observable bounds for higher moments of index returns. Bakshi et al. (2020) and Chabi-Yo and Loudis (2020) propose formulas for the expected return of the market which depend on all higher risk-neutral moments of returns. Similarly, Chabi-Yo et al. (2023) consider such bounds for individual stocks.

Third, our conclusions regarding the empirical performance of the conditional CAPM relate to a literature that addresses the question of whether market returns are predictable. Merton (1980) is one of the first to give suggestions on how to measure the conditional equity premium. He states that market returns are predictable if a fair proportion of the variation in realized returns is due to variation in conditional expectations. Fama and French (1988, 1989) argue that conditional expectations vary over business cycles and that, accordingly, variables which forecast business cycles also predict returns. Stambaugh (1999) casts doubt on the statistical significance of these findings, claiming that standard ordinary least squares coefficients are biased. Boudoukh et al. (2006) challenge the conventional wisdom that return predictability is a long-horizon phenomenon, and demonstrate that the results at shorter horizons are no less important if the predictors are persistent. Welch and Goyal (2008) investigate whether market return predictions can be used to engage in market timing strategies, finding that none of the commonly used predictors really outperforms the historical average market excess return. In response, Campbell and Thompson (2008) suggest that economic restrictions, such as requiring a positive equity premium, can help improve the performance of the predictor variables. Cochrane (2008) instead puts forth a theoretical argument, claiming that returns must be predictable because dividend growth is not.

Finally, we tie in with an extensive body of literature that documents various dimensions of CAPM failure. Jensen et al. (1972) test the unconditional CAPM for multiple time periods using portfolios of stocks, finding that some of the slope coefficients obtained from cross-sectional regressions are close to zero or even negative. Blume and Friend (1973) argue, in agreement with Black (1972), that one possible explanation for this is that investors are constrained in the amount of money they can borrow at the risk-free rate. Fama and French (1992) gather empirical evidence that market betas do not explain the cross-section of average returns, finding that stock characteristics such as a firm's size or its book-to-market ratio play an important role in the investors' compensation for risk. Boudoukh et al. (1993) attribute the failure of the conditional CAPM to the equity premium, finding that in some states of the economy its positivity constraint is violated. Lewellen and Nagel (2006), in contrast, reject the conditional model, arguing that the variation in betas and the equity premium would have to be implausibly large to explain asset-pricing anomalies such as momentum or the value premium. Frazzini and Pedersen (2014) extend the argument by

Black (1972) that investors are subject to short-selling constraints and discover that a betting-against-beta strategy can exploit the fact that low-beta portfolios carry higher Sharpe ratios than high-beta portfolios. Savor and Wilson (2014) and Hendershott et al. (2020) argue that the positive relationship between betas and average returns is present only in vicinity of macroeconomic news announcements and during close-to-open intervals, respectively. Ungeheuer and Weber (2021) link the failure of betas to results from behavioral experiments which suggest that investors use a counting heuristic, rather than correlation measures, to assess the dependence between stocks. Hasler and Martineau (2023) examine the contemporaneous relationship between stock and market excess returns implied by the conditional CAPM, finding that betas obtained from rolling-window regressions explain the conditional level of returns.

3.3 Theoretical considerations

3.3.1 A fully-implied capital asset pricing model

At the center of our model economy is a multi-period investor with log utility, whose wealth portfolio W is a claim to all future consumption C_{t+h}. While the subscript t denotes a point in time, h represents a period of time, also referred to as the investment horizon. With log utility, $u(C_t) = \ln(C_t)$, and δ representing the subjective discount factor that accounts for the time preferences of the investor, the price of the wealth portfolio is given by[4]

$$P_t^W = \mathbb{E}_t\left(\sum_{h=1}^{\infty} \delta^h \frac{u'(C_{t+h})}{u'(C_t)}\, C_{t+h}\right) = \frac{\delta}{1-\delta}\, C_t,$$

where the expectation is conditional on the information available in t. Hence, its gross return, $R^W_{t,t+h}$, is proportional to consumption growth

$$R^W_{t,t+h} = \frac{1}{\delta}\frac{C_{t+h}}{C_t},$$

and the reciprocal of $R^W_{t,t+h}$ is a stochastic discount factor

$$M_{t,t+h} = \frac{1}{R^W_{t,t+h}}. \tag{3.1}$$

We use Equation (3.1) to state the price of the payoff $R^i_{t,t+h} \times R^W_{t,t+h}$ in terms of risk-neutral expectations, giving

$$\mathbb{E}_t\big(R^i_{t,t+h}\big) = \frac{1}{R^f_{t,t+h}}\, \mathbb{E}^*_t\big(R^i_{t,t+h} R^W_{t,t+h}\big), \tag{3.2}$$

[4]Detailed expositions of the log utility framework can be found in Kraus and Litzenberger (1975) and Rubinstein (1976). The notation used here is based on Cochrane (2005, p. 160).

where the superscript asterisk indicates that the expected value is subject to the risk-neutral distribution of returns. Because $\mathbb{E}_t^*\big(R_{t,t+h}^i / R_{t,t+h}^f\big) = 1$ is satisfied for any gross return, we can write

$$R_{t,t+h}^f = \frac{1}{R_{t,t+h}^f}\ \mathbb{E}_t^*(R_{t,t+h}^i) \cdot \mathbb{E}_t^*(R_{t,t+h}^W),$$

which we then subtract from either side of Equation (3.2) to obtain

$$\mathbb{E}_t\big(R_{t,t+h}^i - R_{t,t+h}^f\big) = \frac{1}{R_{t,t+h}^f}\ \mathrm{cov}_t^*\big(R_{t,t+h}^i, R_{t,t+h}^W\big). \tag{3.3}$$

The fact that the reciprocal of $R_{t,t+h}^W$ is an SDF thus allows us to establish a direct connection between a stock's conditional expected excess return and its risk-neutral conditional covariance with the return on the wealth portfolio. Because risk-neutral covariances of returns are not directly observable, however, some additional steps are necessary to achieve identification.[5]

In an attempt to recover the risk-neutral conditional covariance in Equation (3.3), we follow Martin and Wagner (2019) and project stock returns onto the return of the wealth portfolio under the risk-neutral measure[6]

$$R_{t,t+h}^i = \alpha_{t,h}^{i,*} + \beta_{t,h}^{i,*} \cdot R_{t,t+h}^W + \varepsilon_{t,t+h}^i, \tag{3.4}$$

such that $\mathbb{E}_t^*(R_{t,t+h}^W \varepsilon_{t,t+h}^i)$ and $\mathbb{E}_t^*(\varepsilon_{t,t+h}^i)$ are equal to zero. Accordingly, the risk-neutral beta in Equation (3.4) is a population regression coefficient, i.e.,

$$\beta_{t,h}^{i,*} = \frac{\mathrm{cov}_t^*\big(R_{t,t+h}^i, R_{t,t+h}^W\big)}{\mathrm{var}_t^*\big(R_{t,t+h}^W\big)}.$$

This allows us to rewrite the expected excess return in Equation (3.3) as

$$\mathbb{E}_t\big(R_{t,t+h}^i - R_{t,t+h}^f\big) = \beta_{t,h}^{i,*} \cdot \frac{1}{R_{t,t+h}^f} \mathrm{var}_t^*\big(R_{t,t+h}^W\big). \tag{3.5}$$

In order to identify $\beta_{t,h}^{i,*}$ in terms of observable quantities, we take risk-neutral variances

[5] A critical assessment of the unobservability of risk-neutral cross-moments of returns is given in Section 3.3.2.

[6] Martin and Wagner (2019) explicitly avoid placing their approach in the context of a consumption-based asset pricing framework. Instead, they start with a portfolio optimization problem in which the weights of the individual stocks are chosen such that the portfolio is growth-optimal. Because they express the objective function in terms of logarithmic returns, they arrive at the same implications for the return on the growth-optimal portfolio as we do for the return on the wealth portfolio – its reciprocal is a stochastic discount factor. While the derivations presented hereafter are in principle consistent with their optimization problem, we believe that the consumption-based framework lends itself more naturally to the derivation of a conditional CAPM.

on both sides of Equation (3.4), so that[7]

$$\text{var}_t^*\left(R_{t,t+h}^i\right) = \left(\beta_{t,h}^{i,*}\right)^2 \cdot \text{var}_t^*\left(R_{t,t+h}^W\right) + \text{var}_t^*(\varepsilon_{t,t+h}^i). \tag{3.6}$$

In addition, we assume that the proportion of systematic risk in Equation (3.6) is constant in the cross-section. That is, we decompose a stock's total risk-neutral variance according to

$$\text{var}_t^*\left(R_{t,t+h}^i\right) = (v_{t,h}^*)^2 \cdot \text{var}_t^*\left(R_{t,t+h}^i\right) + \left(1-(v_{t,h}^*)^2\right) \cdot \text{var}_t^*\left(R_{t,t+h}^i\right), \tag{3.7}$$

where $0 \leq (v_{t,h}^*)^2 < 1$, and define

$$(v_{t,h}^*)^2 \cdot \text{var}_t^*\left(R_{t,t+h}^i\right) \coloneqq \left(\beta_{t,h}^{i,*}\right)^2 \cdot \text{var}_t^*\left(R_{t,t+h}^W\right). \tag{3.8}$$

This strategy is inspired by Kempf et al. (2015), who work with the physical equivalent of Equation (3.6) to estimate between-stock covariances for portfolio optimization. Their approach involves identifying physical conditional betas in terms of physical conditional variances, which they approximate with the help of risk-neutral moments. The change of measure associated with this approximation remains unexplained, however. They note that "although options provide moments under the risk-neutral measure and portfolio selection requires moments under the physical measure, we make no attempt to risk-adjust implied moments in our study." Accordingly, we add to their approach in that we provide a disciplined rationale for why risk-neutral moments should be associated with conditional stock risk premia.[8]

By replacing the first term in Equation (3.7) with the expression in Equation (3.8) and solving for beta we obtain

$$\beta_{t,h}^{i,*} = \alpha \cdot v_{t,h}^* \left(\frac{\text{var}_t^*\left(R_{t,t+h}^i\right)}{\text{var}_t^*\left(R_{t,t+h}^W\right)} \right)^{1/2}, \tag{3.9}$$

where $\alpha \in \{-1, 1\}$.[9] If, in addition, the market return, $R_{t,t+h}^M$, is a sufficient proxy for the return on the wealth portfolio – a standard CAPM assumption – we can solve for

[7] Martin and Wagner (2019) also rely on the orthogonal decomposition shown in Equation (3.4) to derive their version of a fully-implied formula. Unlike them, however, we do not aim at getting rid of beta, which they achieve with a Taylor approximation at $\beta_{t,h}^{i,*} = 1$ and by imposing constraints on the variation of $\varepsilon_{t,t+h}^i$. Rather, we seek to identify beta in terms of risk-neutral variances. In Section 3.3.2 we present details on our motivation for choosing a different path, involving a discussion of the identifying assumptions and associated implications.

[8] In Appendix B.3 we discuss alternative identification strategies involving higher risk-neutral moments of returns. We find that the variance-based identification strategy performs best empirically.

[9] We introduce α because taking the square root of $\beta_{t,h}^{i,*}$ gives a dual solution.

$v^*_{t,h}$ by value-weighting Equation (3.9) and summing over stocks, such that

$$\beta^{M,*}_{t,h} = \sum_j w^j_t \cdot \alpha \cdot v^*_{t,h} \left(\frac{\text{var}^*_t\big(R^j_{t,t+h}\big)}{\text{var}^*_t\big(R^M_{t,t+h}\big)} \right)^{1/2} = 1,$$

and

$$v^*_{t,h} = \frac{\text{var}^*_t\big(R^M_{t,t+h}\big)^{1/2}}{\sum_j w^j_t \cdot \alpha \cdot \text{var}^*_t\big(R^j_{t,t+h}\big)^{1/2}}.$$

Hence, the risk-neutral beta in Equation (3.9) depends on risk-neutral variances only

$$\beta^{i,*}_{t,h} = \frac{\text{var}^*_t\big(R^i_{t,t+h}\big)^{1/2}}{\sum_j w^j_t \cdot \text{var}^*_t\big(R^j_{t,t+h}\big)^{1/2}}, \tag{3.10}$$

and by combining Equations (3.5) and (3.10), we obtain a formula for the expected return of a stock in excess of the risk-free rate that is fully-implied by option prices[10]

$$\mathbb{E}_t\big(R^i_{t,t+h} - R^f_{t,t+h}\big) = \frac{\text{var}^*_t\big(R^i_{t,t+h}\big)^{1/2}}{\sum_j w^j_t \cdot \text{var}^*_t\big(R^j_{t,t+h}\big)^{1/2}} \cdot \frac{1}{R^f_{t,t+h}} \text{var}^*_t\big(R^M_{t,t+h}\big). \tag{3.11}$$

We then value-weight Equation (3.11) and sum over all stocks to recover the conditional equity premium, which is equivalent to Martin's (2017) lower bound

$$\mathbb{E}_t\big(R^M_{t,t+h} - R^f_{t,t+h}\big) = \frac{1}{R^f_{t,t+h}} \text{var}^*_t\big(R^M_{t,t+h}\big). \tag{3.12}$$

By combining Equations (3.10), (3.11) and (3.12), we obtain a variant of the conditional CAPM in which a stock's exposure to the market and the equity premium both are fully-implied by option prices

$$\mathbb{E}_t\big(R^i_{t,t+h} - R^f_{t,t+h}\big) = \underbrace{\beta^i_{t,h}}_{\text{Eq. (3.10)}} \times \underbrace{\mathbb{E}_t\big(R^M_{t,t+h} - R^f_{t,t+h}\big)}_{\text{Eq. (3.12)}}, \tag{3.13}$$

In subsequent sections we refer to Equation (3.13) as the fully-implied capital asset pricing model (or FI-CAPM).[11]

[10] For the case of simple returns, Martin (2017) shows that risk-neutral variances of returns can be computed using a panel of option prices. We provide a detailed description of the necessary data in Appendix B.1 and the corresponding formulas in Appendix B.2.

[11] We are aware of the fact that, with the assumptions we make, we expose ourselves to a long history of CAPM criticism. For example, one might conclude that the log utility assumption is too restrictive, or that a broad-based market portfolio is an inadequate proxy for the wealth portfolio – a concern famously raised by Roll (1977). However, we do not see the contribution of this study in answering such critique. Rather, we intend to complement existing approaches towards motivating the conditional CAPM with a perspective, in which option prices play a pivotal role in specifying the time

To give a first impression, Figure 3.1 presents time series of annualized expected excess returns for Apple Inc. at investment horizons between 30 and 730 calendar days. For reference, we add the corresponding stock risk premia by Martin and Wagner (2019) (henceforth referred to as MW) and those of a conditional CAPM (henceforth referred to as $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$), where the betas are estimated by one-year rolling regressions of daily stock onto market excess returns, and the equity premium is the expanding historical average excess return of CRSP's value weighted index. For the sake of conciseness, we defer a detailed comparison of the approaches underlying the FI-CAPM and MW to Section 3.3.3, and a description of the data used in this study to Appendix B.1.

Figure 3.1 shows that the conditional stock risk premia implied by the FI-CAPM and MW are very similar, except for the period starting in the early 2000s when the dot-com bubble (henceforth DCB) hit the US economy. At first glance, it seems as if the two initially diverged, but then converged again with the peak of the global financial crisis (henceforth GFC) in 2009. We discuss this phenomenon in greater detail in Sections 3.3.3 and 3.4.1. Moreover, compared to $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$, the time-series variation of the option-implied risk premia is considerably higher, which Martin and Wagner (2019) take as an indication that there is substantial variation in returns that cannot be accounted for by traditional estimates of beta.

3.3.2 The assumption of constant correlation

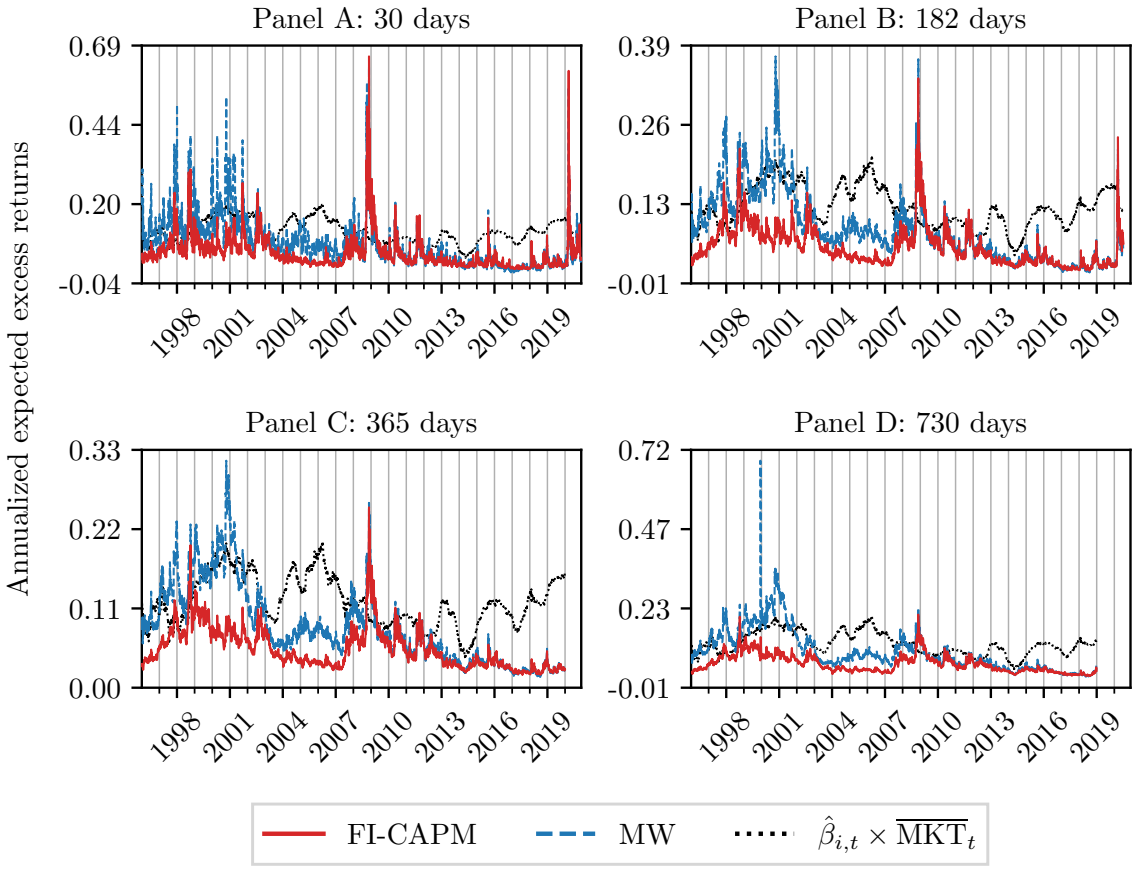
A critical implication of the identifying assumption in Equation (3.8) is that the time-varying risk-neutral conditional correlation of a stock's return with the market is constant in the cross-section. This can be seen by converting the risk-neutral betas in Equation (3.10) into risk-neutral correlations

$$\text{corr}_t^*(R_{t,t+h}^i, R_{t,t+h}^M) = \frac{\text{var}_t^*\left(R_{t,t+h}^M\right)^{1/2}}{\sum_j w_t^j \cdot \text{var}_t^*\left(R_{t,t+h}^j\right)^{1/2}}. \tag{3.14}$$

Because options on the cross-moments of stock returns are neither widely traded nor liquid, estimating conditional risk-neutral correlations (or covariances as in Equation (3.3)) from option prices remains a difficult task. Martin (2018) describes this issue in his survey of Ross (1976) and Breeden and Litzenberger (1978), pointing out the pitfalls that are associated with inferring the joint risk-neutral distribution of two assets from observable option prices. He further admits that it is precisely because of the unobservability of the cross-moments that Martin and Wagner (2019) must resort to a set of harsh approximations in their derivation of a fully-implied formula. As we

variation in conditional moments. Regarding the log utility assumption, we further refer to Rubinstein (1974), who develops sufficient conditions under which individual investors can have heterogeneous preferences and beliefs as long as their aggregate tastes comply with additive generalized log utility. As for the Roll-critique, we concur with Fama and French (2004), who emphasize that such "criticism can be leveled at tests of any economic model when the tests are less than exhaustive or when they use proxies for the variables called for by the model."

Figure 3.1: Annualized stock risk premia. This figure shows time series of annualized expected excess returns for the stock of Apple Inc. with investment horizons of 30, 182, 365, and 730 calendar days between January 1996 and December 2020. In each panel, the solid red line represents the expected excess return calculated according to the FI-CAPM in Equation (3.13). The dashed blue line is the expected excess return proposed by Martin and Wagner (2019) (MW), and the dotted black line is that of a conditional CAPM, where the betas are obtained by rolling-window regression of stock onto market excess returns, and the equity premium corresponds to the historical average market excess return ($\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$). The time series have a daily frequency.



find the implications of their chosen approximations hard to assess, we propose a more direct approach in deriving the FI-CAPM. More specifically, we replace their assumption of a negligible approximation error that is due to the linearization of Equation (3.6) at $\beta_{t,h}^{i,*} = 1$ and setting $\text{var}_t^*(\varepsilon_{t,t+h}^i) - \sum_j w_t^j \cdot \text{var}_t^*(\varepsilon_{t,t+h}^j) = 0$, by the assumption of a constant cross-sectional stock-market correlation, which is not necessarily milder, but more closely related to the problem that the risk-neutral cross-moments of stock returns are unobservable.

To support this idea, we refer to Kempf et al. (2015), who argue that the constant correlation assumption, at least if applied to the physical equivalent of Equation (3.6), allows for a reasonable approximation of the cross-moments of stock returns. Moreover, Chan et al. (1999) note that "factor models yield mean absolute forecast errors that are not notably different from a simple model which assumes that all stocks share the same average pairwise covariance." In a similar vein, Baule et al. (2016) employ Ledoit and Wolf's (2004) shrinkage approach to estimate the covariance matrix of stock returns and conclude that especially shrinkage towards a constant correlation seems promising.

To some extent, Martin and Wagner (2019) also discuss the constant correlation assumption in their Appendix A. Using their SVIX notation, that is,

$$\begin{aligned}
\text{SVIX}_t^2 &= h^{-1} \, \text{var}_t^*\big(R_{t,t+h}^M / R_{t,t+h}^f\big), \\
\text{SVIX}_{i,t}^2 &= h^{-1} \, \text{var}_t^*\big(R_{t,t+h}^i / R_{t,t+h}^f\big), \\
\overline{\text{SVIX}_t^2} &= h^{-1} \sum_j w_t^j \cdot \text{var}_t^*\big(R_{t,t+h}^j / R_{t,t+h}^f\big),
\end{aligned}$$

they argue that $\text{SVIX}_t^2 / \overline{\text{SVIX}_t^2}$ can be conceived of as an approximation of the average risk-neutral correlation between stocks. To illustrate that this claim directly relates to the expression in Equation (3.14), we follow their approach and rewrite the risk-neutral variance of the market return as a weighted sum of covariances

$$\text{var}_t^*\big(R_{t,t+h}^M\big) = \sum_i \sum_j w_t^i w_t^j \cdot \text{cov}_t^*\big(R_{t,t+h}^i, R_{t,t+h}^j\big) = \sum_i w_t^i \cdot \text{cov}_t^*\big(R_{t,t+h}^i, R_{t,t+h}^M\big).$$

With the assumption that the stocks' risk-neutral correlation with the market is constant in the cross-section, and by rearranging terms, we obtain

$$\text{corr}_t^*(R_{t,t+h}^i, R_{t,t+h}^M) = \frac{\text{var}_t^*\big(R_{t,t+h}^M\big)^{1/2}}{\sum_j w_t^j \cdot \text{var}_t^*\big(R_{t,t+h}^j\big)^{1/2}} = \frac{\text{SVIX}_t}{\overline{\text{SVIX}_t}}, \tag{3.15}$$

which corresponds to Equation (3.14).

The key insight here is that, even though Martin and Wagner (2019) sketch a similar idea in their Appendix A, they do not claim a fully-implied CAPM. First, they associate $\text{SVIX}_t^2 / \overline{\text{SVIX}_t^2}$ with the risk-neutral correlation between stocks, rather than the correlation between stocks and the market, which is captured by beta. As shown

above, switching from the one to the other perspective is fairly easy and allows us to establish a more direct connection to the CAPM. Second, for $\text{SVIX}_t^2/\overline{\text{SVIX}_t^2}$ to be a valid approximation of the average risk-neutral stock-stock correlation, they must neglect a Jensen's inequality – something that we do not need to arrive at Equation (3.15), because it is a direct implication of our identification strategy. The MW formula, which we introduce explicitly in Section 3.3.3, instead appears to be unrelated to their finding that $\text{SVIX}_t^2/\overline{\text{SVIX}_t^2}$ can be conceived of as an approximation of the average risk-neutral correlation between stocks.[12]

3.3.3 Martin and Wagner's (2019) formula as a special case

To develop a better understanding of the differences between the FI-CAPM and MW, we find it convenient to resort to their SVIX language, which allows us to rewrite the two competing formulas for the stock risk premium as follows:

$$h^{-1}\left(\mathbb{E}_t\left(\frac{R^i_{t,t+h}}{R^f_{t,t+h}}\right)-1\right)=\begin{cases}\dfrac{\text{SVIX}_{i,t}}{\overline{\text{SVIX}_t}}\cdot\text{SVIX}_t^2 & \text{(FI-CAPM)} \quad (3.16)\\[2ex] \text{SVIX}_t^2+\dfrac{1}{2}\left(\text{SVIX}_{i,t}^2-\overline{\text{SVIX}_t^2}\right) & \text{(MW).} \quad (3.17)\end{cases}$$

Note that in Equation (3.16) the risk premium of a stock is comprised of a multiplication of two components: SVIX_t^2 accounting for the time variation in the equity premium and $\beta^{i,*}_{t,h}$ representing the associated stock-specific risk exposure. In Equation (3.17), the components of the stock risk premium are instead additively separable, meaning that there is no interaction between aggregate and stock-specific sources of risk.

That being said, the differences between the two formulas remain elusive, so we introduce a modified version of the FI-CAPM that allows for a more direct comparison. More specifically, we write Equation (3.16) as a function of $\text{SVIX}_{i,t}^2$

$$g(\text{SVIX}_{i,t}^2)=\frac{\left(\text{SVIX}_{i,t}^2\right)^{1/2}}{\left(w_t^i\cdot\left(\text{SVIX}_{i,t}^2\right)^{1/2}+\sum_{j\neq i}w_t^j\cdot\text{SVIX}_{j,t}\right)^2}\cdot\text{SVIX}_t^2, \quad (3.18)$$

and linearize this function using a first-order Taylor approximation. The first degree polynomial evaluated at some ψ_t is given by

$$G(\text{SVIX}_{i,t}^2,\psi_t)=g(\psi_t)+g'(\psi_t)\cdot(\text{SVIX}_{i,t}^2-\psi_t),$$

[12] Figure B.1 in Appendix B.6 illustrates the differences between Martin and Wagner's (2019) measure of the average stock-stock correlation implied by $\text{SVIX}_t^2/\overline{\text{SVIX}_t^2}$ and ours for the stock-market correlation in Equation (3.14).

and the associated derivative is

$$g'(\text{SVIX}^2_{i,t}) = \frac{1}{2}\frac{\text{SVIX}^2_t \cdot \left(\sum_{j\neq i} w^j_t \cdot \text{SVIX}_{j,t}\right)}{\left(\text{SVIX}^2_{i,t}\right)^{1/2} \cdot \left(\sum_j w^j_t \cdot \text{SVIX}_{j,t}\right)^2}. \tag{3.19}$$

To further simplify matters, we assume that the contribution of asset i to the value-weighted average risk-neutral volatility is negligible, i.e., $\sum_{j\neq i} w^j_t \cdot \text{SVIX}_{j,t} \approx \sum_j w^j_t \cdot \text{SVIX}_{j,t}$. Considering that the average S&P 500 constituent comprises a market capitalization weight of approximately 0.2%, this assumption may not be too far-fetched. Thereby, we obtain an approximate version of the derivative in Equation (3.19), which is

$$g'(\text{SVIX}^2_{i,t}) \approx \frac{1}{2}\frac{\text{SVIX}^2_t}{\left(\text{SVIX}^2_{i,t}\right)^{1/2} \cdot \left(\sum_j w^j_t \cdot \text{SVIX}_{j,t}\right)}. \tag{3.20}$$

Evaluating Equation (3.20) at $\psi_t = \left(\overline{\text{SVIX}_t}\right)^2$ yields

$$g'(\psi_t) \approx \frac{1}{2}\left(\frac{\text{SVIX}_t}{\overline{\text{SVIX}_t}}\right)^2,$$

where the choice of ψ_t is motivated by the fact that Martin and Wagner (2019) rely on a linearization of Equation (3.6) around $\beta^{i,*}_{t,h} = 1$, which in the case of the FI-CAPM is achieved by setting $\psi_t = \left(\overline{\text{SVIX}_t}\right)^2$. Accordingly, a linear approximation of the right-hand side of Equation (3.16) is given by

$$\begin{aligned} G(\text{SVIX}^2_{i,t}) &\approx \text{SVIX}^2_t + \gamma_t \cdot \frac{1}{2}\left(\text{SVIX}^2_{i,t} - \left(\overline{\text{SVIX}_t}\right)^2\right), \\ &\text{where } \gamma_t = \text{corr}^*_t(R^i_{t,t+h}, R^M_{t,t+h})^2. \end{aligned}$$

Note that this expression is already quite similar to the MW formula in Equation (3.17), yet γ_t in their case is equal to 1 and the last term in brackets is $\overline{\text{SVIX}^2_t}$ instead of $\left(\overline{\text{SVIX}_t}\right)^2$. To get even closer, we further neglect Jensen's inequality $\left(\overline{\text{SVIX}_t}\right)^2 \leq \overline{\text{SVIX}^2_t}$, by which we obtain[13]

$$\begin{aligned} \text{FI-CAPM} &\approx \text{SVIX}^2_t + \widetilde{\gamma}_t \cdot \frac{1}{2}\left(\text{SVIX}^2_{i,t} - \overline{\text{SVIX}^2_t}\right), \\ &\text{where } \widetilde{\gamma}_t = \frac{\text{SVIX}^2_t}{\overline{\text{SVIX}^2_t}}. \end{aligned} \tag{3.21}$$

An interesting result of this is that $0 \leq \widetilde{\gamma}_t \leq 1$, because $\text{SVIX}^2_t \leq \overline{\text{SVIX}^2_t}$. $\widetilde{\gamma}_t$ being smaller than 1 implies that the linear approximation of the FI-CAPM in Equation

[13]This is the same inequality that Martin and Wagner (2019) neglect when motivating that $\text{SVIX}^2_t/\overline{\text{SVIX}^2_t}$ can be given the interpretation of an approximate average risk-neutral correlation between stocks.

(3.21) is less sensitive to changes in risk-neutral stock variance and produces less cross-sectional variation in expected excess returns than MW. Moreover, for stocks that carry higher than average risk, i.e., $\left(\text{SVIX}^2_{i,t} - \overline{\text{SVIX}^2_t}\right) > 0$, the expected excess return implied by the FI-CAPM approximation is lower than the one implied by MW. The reverse is true for stocks that carry lower than average risk. As discussed in Section 3.3.2, we can give $\widetilde{\gamma}_t$ the interpretation of an approximate average risk-neutral correlation between stocks. Accordingly, the linear approximation of the FI-CAPM would correspond to MW only if the approximate average risk-neutral correlation between stocks was equal to 1 at all times.[14]

Figure 3.2 corroborates these findings for the sample of S&P 500 constituents, showing the original models' cross-sectional variation in expected stock excess returns as measured by the differences between the 1st and the 10th decile. It can be seen that, especially during the DCB, the GFC, and also the more recent COVID-19 crisis (henceforth COC), the cross-sectional variation generated by MW was significantly higher than the one generated by the FI-CAPM, whereas outside of these periods the differences largely disappeared.

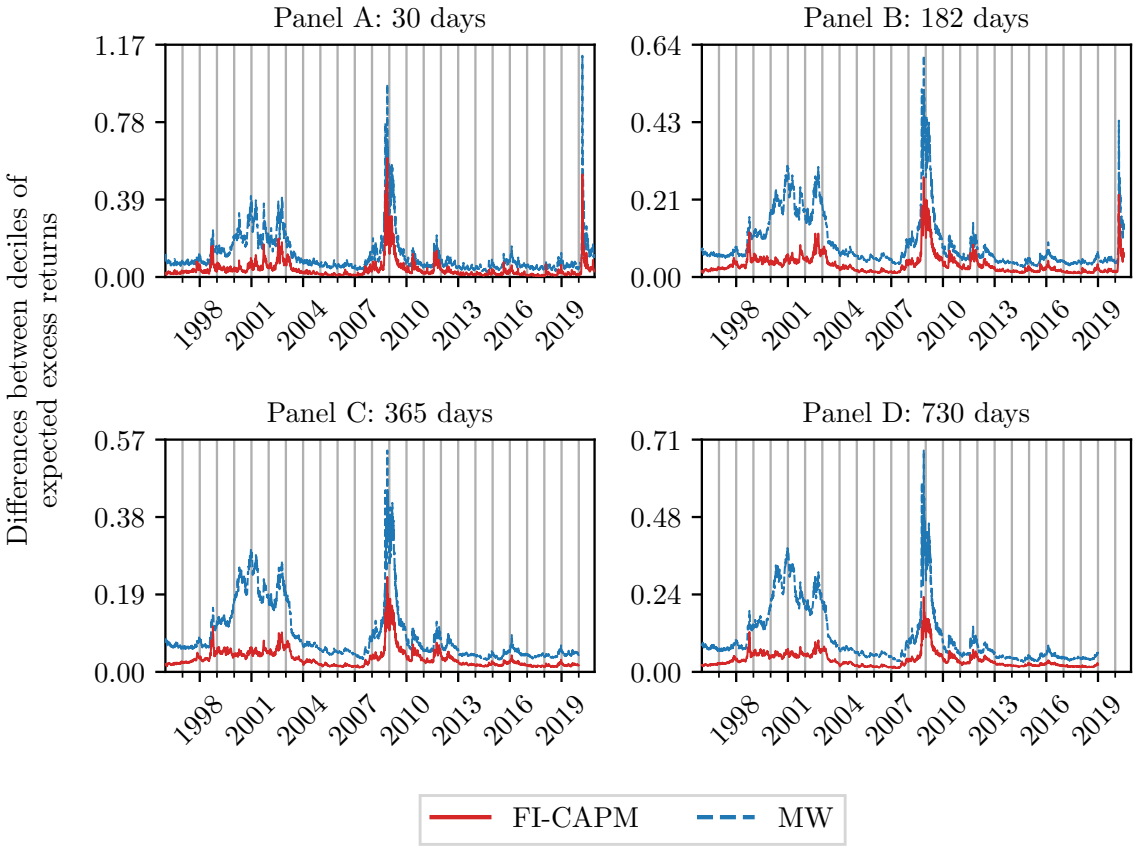
3.4 Model evaluation

In the previous sections, we have argued from a theoretical perspective that the FI-CAPM is to be preferred over MW, not only because the assumptions necessary for its derivation are more transparent, but also because its beta representation provides a more intuitive take on the notion of risk compensation as it is reflected in the cross-section of returns.

A question that inevitably arises therefrom is whether these favorable properties also translate into better empirical performance. In the following, we therefore examine the FI-CAPM both in terms of its average forecast error (Section 3.4.1) and its ability to explain differences in average excess returns across assets (Section 3.4.2). In either case, we compare its performance to established benchmark models, including MW and $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$ as the main competitors. Examining the model's cross-sectional implications, we document a flat relationship between average predicted and realized excess returns of beta-sorted portfolios – a pattern commonly associated with tests of the unconditional CAPM. In Sections 3.4.3 and 3.4.4 we substantiate the conditional CAPM's difficulties in explaining the cross-section of average stock returns by showing that the portfolios' Sharpe ratios decline with beta, regardless of whether we use the FI-CAPM or $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$. In Sections 3.4.5 to 3.4.7 we synthesize the empirical evidence and provide an intertemporal perspective on the CAPM's failure in cross-sectional tests.

[14] In Appendix B.4, we briefly discuss the positivity restriction on beta that we implicitly impose in deriving the FI-CAPM. We compare this restriction to the Negative Correlation Condition (NCC) for individual stocks, as examined by Kadan and Tang (2020).

Figure 3.2: Differences between deciles of stock risk premia. This figure shows time series of differences between the 1st and the 10th decile of expected stock excess returns with investment horizons of 30, 182, 365, and 730 calendar days between January 1996 and December 2020. In each panel, the solid red line represents the expected excess return calculated according to the FI-CAPM in Equation (3.13). The dotted blue line instead represents the approach by Martin and Wagner (2019) (MW). The universe of stocks is confined to securities that are constituents of the S&P 500. The time series have a daily frequency.



3.4.1 Average forecast performance

We evaluate the average forecast performance of the FI-CAPM using the mean square prediction error (MSE), which has the property that it is uniquely minimized by the conditional mean of excess returns, that is,

$$\mathbb{E}_t(R^{e,i}_{t,t+h}) = \underset{\hat{R}^{e,i}_{t,t+h}\in\mathcal{R}}{\arg\min}\ \mathbb{E}_t\big((R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h})^2\big),$$

where $R^{e,i}_{t,t+h}$ denotes the return of asset i in excess of the risk-free rate and $\hat{R}^{e,i}_{t,t+h}$ its time t-conditional forecast from the set of feasible forecasts $\mathcal{R}$. As the conditional expectation is the MSE-optimal predictor of excess returns, we will refer to its measurements simply as *predictions*. Moreover, it is easy to see that by the law of iterated expectations, the conditional expected excess return minimizes not only the conditional but also the unconditional MSE, making it accessible for empirical estimation.

As we are interested in the models' performance over a panel of N stocks rather than a single stock, we estimate their unconditional MSEs by time series averages and then average across stocks. In terms of population moments we thus have

$$\text{MSE}_N = \frac{1}{N}\sum_i \mathbb{E}\big((R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h})^2\big), \tag{3.22}$$

whereas the corresponding sample equivalent is given by[15]

$$\widehat{\text{MSE}}_N = \frac{1}{N}\sum_i \frac{1}{T_i}\sum_t (R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h})^2.$$

In practice, it is often difficult to judge how small the average forecast error of a model should be for it to be considered a reasonable approximation to the true unobserved expectation, so we express the average MSE of the FI-CAPM (ω_1) relative to that of an established competitor (ω_2)

$$R^2_N = 1 - \frac{\frac{1}{N}\sum_i \mathbb{E}\big((R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_1))^2\big)}{\frac{1}{N}\sum_i \mathbb{E}\big((R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_2))^2\big)}, \tag{3.23}$$

such that R^2_N indicates by how much the proposed model performs better (or worse) than the competing model.[16] In subsequent sections, we refer to R^2_N as the predictive or out-of-sample R-squared. Similarly, we use R^2 without the subscript N when

[15] Our sample of security-date observations is unbalanced, so we refer to N as the total number of stocks, to T as the total number of timestamps, to N_t as the number of stocks at a given date and to T_i as the number of timestamps for a given stock.

[16] Note that any benchmark model is yet just another approximation of the true unobserved expectation, so we should never go as far as to reject a potentially interesting model based on a negative R^2_N alone. Even if the level of its forecasts is somewhat off, it may still hold substantial information about the cross-section of stock returns. We defer a discussion of cross-sectional explanatory power to subsequent sections.

forecasting the excess return of the market.

Because stocks frequently enter and leave the market, the time-series averages we use to estimate MSE_N are typically based on a number of observations that is smaller than T. We therefore multiply these averages by their rate of convergence T_i, so that the estimator of R^2_N simplifies to its pooled equivalent

$$\hat{R}^2_N = 1 - \frac{\sum_{i,t}(R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_1))^2}{\sum_{i,t}(R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_2))^2}.$$

To add a formal perspective, we test for superior forecast performance of the FI-CAPM relative to a range of benchmark models using a panel version of the Diebold-Mariano (DM) test.[17] For this purpose, we define $\overline{D}_{t,t+h}$ as the cross-sectional average loss differential between two competing models ω_1 and ω_2

$$\overline{D}_{t,t+h} = \frac{1}{N_t}\sum_i \Big(\big(R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_1)\big)^2 - \big(R^{e,i}_{t,t+h} - \hat{R}^{e,i}_{t,t+h}(\omega_2)\big)^2 \Big).$$

The null hypothesis associated with this test is H_0: $\mathbb{E}(\overline{D}_{t,t+h}) \leq 0$, which makes

$$\text{DM} = \frac{\frac{1}{T}\sum_t \overline{D}_{t,t+h}}{\sqrt{\widehat{\text{var}}\big(\frac{1}{T}\sum_t \overline{D}_{t,t+h}\big)}} \overset{a}{\sim} \mathcal{N}(0,1) \tag{3.24}$$

a natural choice of a test statistic.[18] We obtain estimates of the variance in the denominator of DM using the procedure by Newey and West (1987), thus accounting for serial correlation in the average loss differential.

Table 3.1 provides estimates of R^2_N and p-values associated with a one-sided DM test, both at different forecast horizons and for a range of benchmark models. The latter can be divided into two groups, one that provides cross-sectionally constant forecasts and one that achieves cross-sectional differentiation. The members of the first group are the zero forecast, a constant forecast of 6% p.a., the extending historical average of CRSP's value-weighted index $\overline{\text{MKT}}_t$, and Martin's (2017) lower bound on the equity premium SVIX^2_t. The second group consists of three variants of the conditional CAPM in which the betas are estimated by one-year rolling regressions and the equity premium is chosen as 6% p.a., $\overline{\text{MKT}}_t$, or SVIX^2_t. The second group also includes the Fama-French three factor model (FF3), the option-based approach by Kadan and Tang (2020) (KT), and MW.

What is striking is that, regardless of the choice of the benchmark model, the out-of-sample R^2s all carry positive signs, with the exception of MW at the monthly

[17] For a detailed discussion of comparing forecasts in a panel data setting, see Timmermann and Zhu (2019).

[18] Given that a CLT can be applied and $\mathbb{E}(\overline{D}_{t,t+h}) = 0$, the DM test statistic converges in distribution to $\mathcal{N}(0,1)$. For this to hold, we need to make high-level assumptions about the statistical properties of $\overline{D}_{t,t+h}$, i.e., we assume that the cross-sectional average loss differential is covariance stationary. Moreover, we do not account for the fact that $\overline{D}_{t,t+h}$ itself is subject to estimation.

investment horizon, indicating that the FI-CAPM produces lower average forecast errors than its competitors. In particular, the FI-CAPM appears to outperform not only the models providing a constant cross-sectional forecast, an example of which is given by the comparison with the zero forecast yielding an R^2_N of about 6.1% at the one-year horizon, but also the more traditional estimates of the conditional CAPM where the betas are obtained from one-year rolling regressions. Compared to MW, FI-CAPM performs better at 7 out of 8 forecast horizons, extending its lead as the forecast horizon increases. The R^2_N for this comparison ranges from negative 0.02% at the one-month horizon to 4.4% at the two-year horizon.

As we have seen in Figures 3.1 and 3.2, the differences between the forecasts of the FI-CAPM and MW are largest during the DCB. Therefore, we investigate the effects of clustering R^2_N by time in Tables 3.2 and 3.3.[19] In particular, we present estimates of R^2_N and associated p-values of the DM test for the years 2000 to 2002, including the GFC from 2007 to 2009 for reference.[20] During the DCB, the average forecast error of the FI-CAPM relative to MW was particularly low, precisely when the cross-sectional variation in expected excess returns generated by MW was much higher than that of FI-CAPM, as we discussed previously with reference to Figure 3.2. The estimates of R^2_N in this period range from approximately 1.3% to 32.4% depending on the forecast horizon, and the DM test statistic shows significant differences in the cross-sectional average forecast error for all forecast horizons. Hence, it seems as if the more balanced adjustment of the FI-CAPM to this period of low average risk-neutral stock-market correlation was beneficial in terms of predictive R-squared. For the GFC, the situation is somewhat more complicated: For shorter forecast horizons the FI-CAPM produces lower average forecast errors (the highest R^2_N is approximately 2.5%), whereas for longer forecast horizons MW takes the lead (the lowest R^2_N is approximately negative 2.8%).

Table 3.4 presents results for the predictability of market excess returns when comparing the equity premium of the FI-CAPM from Equation (3.12) with a zero forecast, a constant prediction of 6% p.a., and the historical average excess return, $\overline{\text{MKT}}_t$. Again, the R^2's are positive for each of the benchmark models, except for the 6% p.a. forecast with an investment horizon of 547 calendar days. Compared to the zero forecast, the FI-CAPM appears to perform particularly well, with R^2's ranging from 2.0% to 10.8%, although the DM test indicates that this superior performance is borderline significant. Still, the amount of variation explained increases in the investment horizon and so the results previously reported by Martin (2017) seem to be robust to an expansion of the sample period.

[19] We do not include the more recent COC, as the data coverage for the year 2020 is not the same for all investment horizons. Since the last date available in CRSP is December 31, 2020, we cannot calculate forward-looking returns with an investment horizon of, say, 730 calendar days.

[20] More precisely, we choose March 24, 2000 as the starting date and October 9, 2002 as the ending date for the DCB, as these dates correspond to the high-low values of the S&P 500 during the 2000-2002 period. For the GFC we choose October 9, 2007 and March 9, 2009 as starting and ending dates, respectively.

Table 3.1: Out-of-sample forecast evaluation. This table presents results from an out-of-sample performance comparison between the FI-CAPM in Equation (3.13) and a collection of benchmark models from January 1996 to December 2020. For each investment horizon, ranging from 30 to 730 calendar days, we present estimates of R^2_N (Equation (3.23)) in the rows, thus comparing the MSE of the FI-CAPM with that of its competitors in the columns. A positive sign of R^2_N indicates that the FI-CAPM is superior to the benchmark model in terms of MSE. From left to right, we use the following benchmark models for comparison: a constant zero forecast, a forecast of 6% p.a., the extending historical average of CRSP's value-weighted index $\overline{\text{MKT}}_t$, the lower bound on the equity premium SVIX^2_t, three variants of the conditional CAPM where betas are obtained by one-year rolling regressions and multiplied by either 6% p.a., $\overline{\text{MKT}}_t$, or SVIX^2_t, the Fama-French three factor model (FF3), and the option-implied formulas of expected excess stock returns proposed by Kadan and Tang (2020) (KT) and Martin and Wagner (2019) (MW). The number in brackets below R^2_N represents the p-value that is associated with a one-sided Diebold-Mariano test for superior predictive accuracy of the FI-CAPM, i.e., we reject the null for large values of the DM test statistic (Equation (3.24)). Asterisks denote statistical significance at conventional levels of 10% ($*$), 5% ($**$) and 1% ($***$). The frequency of the forecast errors used in this analysis is daily, and the universe of stocks consists of S&P 500 constituents.

Investment horizon	0% p.a.	6% p.a.	$\overline{\text{MKT}}_t$	SVIX^2_t	$\hat{\beta}_{i,t}\times$6% p.a.	$\hat{\beta}_{i,t}\times\overline{\text{MKT}}_t$	$\hat{\beta}_{i,t}\times\text{SVIX}^2_t$	FF3	KT	MW
30	0.939**	0.414	0.341	0.236*	0.420	0.416	0.153**	0.664*	0.873	−0.021
	(0.050)	(0.158)	(0.186)	(0.092)	(0.137)	(0.137)	(0.025)	(0.054)	(0.221)	(0.531)
60	1.751**	0.778	0.703	0.430*	0.805*	0.882*	0.244**	1.398**	2.101	0.131
	(0.027)	(0.109)	(0.133)	(0.080)	(0.082)	(0.089)	(0.027)	(0.029)	(0.154)	(0.398)
91	2.371**	0.928	0.857	0.554*	0.972*	1.138	0.304**	2.016**	3.467	0.346
	(0.030)	(0.128)	(0.154)	(0.088)	(0.095)	(0.109)	(0.034)	(0.031)	(0.112)	(0.309)
182	4.325**	1.739	1.710	1.060*	1.806*	2.212	0.550**	4.018**	4.964	0.227
	(0.021)	(0.103)	(0.132)	(0.074)	(0.072)	(0.102)	(0.032)	(0.025)	(0.145)	(0.424)
273	5.417**	1.848	1.875	1.343*	1.885	2.481	0.743**	5.180**	7.159	0.294
	(0.025)	(0.163)	(0.205)	(0.080)	(0.139)	(0.170)	(0.037)	(0.042)	(0.142)	(0.435)
365	6.118**	1.657	1.704	1.490*	1.645	2.360	0.847*	5.865*	9.619	0.467
	(0.028)	(0.230)	(0.280)	(0.091)	(0.220)	(0.240)	(0.056)	(0.059)	(0.136)	(0.419)
547	7.397**	1.340	1.387	1.537*	1.345	2.316	0.931**	7.182*	18.743*	2.453
	(0.012)	(0.278)	(0.355)	(0.071)	(0.282)	(0.296)	(0.047)	(0.066)	(0.077)	(0.218)
730	8.554**	1.700	1.978	1.648*	1.831	3.276	1.066*	9.310**	26.527*	4.381
	(0.013)	(0.270)	(0.327)	(0.081)	(0.262)	(0.253)	(0.083)	(0.041)	(0.063)	(0.140)

Table 3.2: Out-of-sample forecast evaluation by time periods (Short investment horizons). This table presents results from an out-of-sample performance comparison between the FI-CAPM in Equation (3.13) and a collection of benchmark models during the years 2000-2002 when the Dotcom Bubble (DCB) hit the US economy and during the Global Financial Crisis (GFC) from 2007-2009. For each investment horizon, ranging from 30 to 182 calendar days, we present estimates of R^2_N (Equation (3.23)) in the rows, thus comparing the MSE of the FI-CAPM with that of its competitors in the columns. A positive sign of R^2_N indicates that the FI-CAPM is superior to the benchmark model in terms of MSE. From left to right, we use the following benchmark models for comparison: a constant zero forecast, a forecast of 6% p.a., the extending historical average of CRSP's value-weighted index $\overline{\text{MKT}}_t$, the lower bound on the equity premium SVIX^2_t, three variants of the conditional CAPM where betas are obtained by one-year rolling regressions and multiplied by either 6% p.a., $\overline{\text{MKT}}_t$ or SVIX^2_t, the Fama-French three factor model (FF3), and the option-implied formulas of expected excess stock returns proposed Kadan and Tang (2020) (KT) and Martin and Wagner (2019) (MW). The number in brackets below R^2_N represents the p-value that is associated with a one-sided Diebold-Mariano (DM) test for superior predictive accuracy of the FI-CAPM, i.e., we reject the null for large values of the DM test statistic (Equation (3.24)). Asterisks denote statistical significance at conventional levels of 10% ($*$), 5% ($**$) and 1% ($***$). The frequency of the forecast errors used in this analysis is daily, and the universe of stocks consists of S&P 500 constituents.

Investment horizon	Period	0% p.a.	6% p.a.	$\overline{\text{MKT}}_t$	SVIX^2_t	$\hat{\beta}_{i,t}\times$6% p.a.	$\hat{\beta}_{i,t}\times\overline{\text{MKT}}_t$	$\hat{\beta}_{i,t}\times\text{SVIX}^2_t$	FF3	KT	MW
30	DCB	−0.084	0.359	0.844***	−0.135	0.924***	1.860***	0.172*	1.412***	5.484***	1.326**
		(0.541)	(0.139)	(0.003)	(0.731)	(<0.001)	(<0.001)	(0.057)	(0.003)	(0.007)	(0.019)
	GFC	−2.842	−1.341	−0.283	−0.670	−0.969	0.353	−0.086	1.137	8.579**	1.514**
		(0.844)	(0.728)	(0.557)	(0.790)	(0.679)	(0.415)	(0.594)	(0.239)	(0.019)	(0.045)
60	DCB	−0.440	0.509	1.549***	−0.326	1.623***	3.530***	0.263*	2.638**	11.368***	2.851***
		(0.644)	(0.198)	(0.010)	(0.787)	(<0.001)	(<0.001)	(0.088)	(0.012)	(<0.001)	(0.002)
	GFC	−3.133	−0.749	0.986	−0.649	−0.204	1.936	−0.031	3.149**	12.388**	2.151*
		(0.834)	(0.635)	(0.304)	(0.730)	(0.546)	(0.124)	(0.536)	(0.016)	(0.030)	(0.081)
91	DCB	−0.973	0.616	2.331***	−0.555	2.361***	5.376***	0.453**	4.108**	16.837***	4.348***
		(0.745)	(0.139)	(0.009)	(0.867)	(<0.001)	(<0.001)	(0.041)	(0.014)	(<0.001)	(<0.001)
	GFC	−3.715	−0.441	1.975	−0.598	0.204	3.121*	−0.025	4.704**	15.126*	2.468
		(0.783)	(0.564)	(0.234)	(0.650)	(0.482)	(0.098)	(0.527)	(0.015)	(0.050)	(0.150)
182	DCB	−3.746	0.237	4.180***	−1.564	3.646***	9.826***	0.745**	7.958***	30.739***	8.657***
		(0.996)	(0.361)	(0.006)	(0.999)	(<0.001)	(<0.001)	(0.013)	(0.001)	(<0.001)	(<0.001)
	GFC	0.894	4.999	8.274***	1.185	5.702**	9.599***	0.869	11.936***	14.242*	0.675
		(0.453)	(0.108)	(<0.001)	(0.324)	(0.036)	(<0.001)	(0.222)	(<0.001)	(0.076)	(0.423)

Table 3.3: Out-of-sample forecast evaluation by time periods (Long investment horizons). This table presents results from an out-of-sample performance comparison between the FI-CAPM in Equation (3.13) and a collection of benchmark models during the years 2000-2002 when the Dotcom Bubble (DCB) hit the US economy and during the Global Financial Crisis (GFC) from 2007-2009. For each investment horizon, ranging from 273 to 730 calendar days, we present estimates of R^2_N (Equation (3.23)) in the rows, thus comparing the MSE of the FI-CAPM with that of its competitors in the columns. A positive sign of R^2_N indicates that the FI-CAPM is superior to the benchmark model in terms of MSE. From left to right, we use the following benchmark models for comparison: a constant zero forecast, a forecast of 6% p.a., the extending historical average of CRSP's value-weighted index $\overline{\text{MKT}}_t$, the lower bound on the equity premium SVIX^2_t, three variants of the conditional CAPM where betas are obtained by one-year rolling regressions and multiplied by either 6% p.a., $\overline{\text{MKT}}_t$ or SVIX^2_t, the Fama-French three factor model (FF3), and the option-implied formulas of expected excess stock returns proposed by Kadan and Tang (2020) (KT) and Martin and Wagner (2019) (MW). The number in brackets below R^2_N represents the p-value that is associated with a one-sided Diebold-Mariano (DM) test for superior predictive accuracy of the FI-CAPM, i.e., we reject the null for large values of the DM test statistic (Equation (3.24)). Asterisks denote statistical significance at conventional levels of 10% ($*$), 5% ($**$) and 1% ($***$). The frequency of the forecast errors used in this analysis is daily, and the universe of stocks consists of S&P 500 constituents.

Investment horizon	Period	0% p.a.	6% p.a.	$\overline{\text{MKT}}_t$	SVIX^2_t	$\hat{\beta}_{i,t}\times$6% p.a.	$\hat{\beta}_{i,t}\times\overline{\text{MKT}}_t$	$\hat{\beta}_{i,t}\times\text{SVIX}^2_t$	FF3	KT	MW
273	DCB	−4.734	0.724	6.321***	−1.902	5.048***	13.286***	0.996***	11.414***	40.547***	11.898***
		(0.970)	(0.327)	(0.002)	(0.950)	(<0.001)	(<0.001)	(0.010)	(<0.001)	(<0.001)	(<0.001)
	GFC	3.672	8.166*	12.009***	2.351	8.812**	13.325***	1.598	16.111***	13.342	−1.011
		(0.363)	(0.072)	(<0.001)	(0.251)	(0.024)	(<0.001)	(0.166)	(<0.001)	(0.141)	(0.574)
365	DCB	−5.738	1.074	8.236***	−2.252	5.628***	15.441***	0.825*	14.779***	49.045***	15.698***
		(0.938)	(0.291)	(<0.001)	(0.903)	(<0.001)	(<0.001)	(0.065)	(<0.001)	(<0.001)	(0.002)
	GFC	5.826	10.437**	14.728***	3.371	11.063***	16.095***	2.495	19.319***	14.060	−2.756
		(0.310)	(0.036)	(<0.001)	(0.194)	(0.008)	(<0.001)	(0.114)	(<0.001)	(0.149)	(0.673)
547	DCB	−5.633	1.205	9.807***	−2.229	4.833***	15.760***	0.205	18.038***	61.072***	21.637***
		(0.807)	(0.302)	(<0.001)	(0.784)	(<0.001)	(<0.001)	(0.418)	(<0.001)	(<0.001)	(0.010)
	GFC	11.847	14.451**	18.892***	5.693	15.163***	20.753***	4.557**	25.327***	31.577***	−2.116
		(0.182)	(0.011)	(<0.001)	(0.109)	(<0.001)	(<0.001)	(0.045)	(<0.001)	(<0.001)	(0.634)
730	DCB	−8.046	0.051	11.298***	−3.885	3.999***	17.588***	−1.178	21.841***	72.977***	32.388***
		(0.860)	(0.493)	(<0.001)	(0.910)	(<0.001)	(<0.001)	(0.756)	(<0.001)	(<0.001)	(<0.001)
	GFC	18.457*	16.610***	18.654***	7.735**	16.689***	19.756***	5.627**	24.282***	33.304***	−2.507
		(0.067)	(0.009)	(<0.001)	(0.054)	(0.002)	(<0.001)	(0.022)	(<0.001)	(<0.001)	(0.644)

Table 3.4: Out-of-sample forecast evaluation (S&P 500 index). This table presents results from an out-of-sample performance comparison between the equity premium in Equation (3.12) (FI-CAPM) and a collection of benchmark models from January 1996 to December 2020. For each investment horizon, ranging from 30 to 730 calendar days, we present estimates of R^2 (Equation (3.23)) in the rows, thus comparing the MSE of the FI-CAPM with that of its competitors in the columns. A positive sign of R^2 indicates that the FI-CAPM is superior to the benchmark model in terms of MSE. From left to right, we use the following benchmark models for comparison: a constant zero forecast, a forecast of 6% p.a., and the extending historical average of CRSP's value-weighted index $\overline{\text{MKT}}_t$. The number in brackets below R^2 represents the p-value that is associated with a one-sided Diebold-Mariano test for superior predictive accuracy of the FI-CAPM, i.e., we reject the null for large values of the DM test statistic (Equation (3.24)). Asterisks denote statistical significance at conventional levels of 10% ($*$), 5% ($**$) and 1% ($***$). The frequency of the forecast errors used in this analysis is daily.

Investment horizon	0% p.a.	6% p.a.	$\overline{\text{MKT}}_t$
30	2.008*	0.879	1.425
	(0.091)	(0.199)	(0.123)
60	3.405*	1.365	2.776
	(0.097)	(0.211)	(0.110)
91	4.650	1.563	3.866
	(0.117)	(0.251)	(0.120)
182	8.442*	3.098	7.645
	(0.068)	(0.131)	(0.108)
273	9.682*	2.402	9.132
	(0.094)	(0.286)	(0.177)
365	9.775	1.051	9.808
	(0.133)	(0.425)	(0.228)
547	10.446	−0.062	11.732
	(0.180)	(0.503)	(0.253)
730	10.803	0.577	14.907
	(0.217)	(0.474)	(0.229)

3.4.2 Average excess returns of prediction-sorted portfolios

Although the results from Section 3.4.1 indicate that the average forecast error associated with the FI-CAPM is relatively small compared to a number of benchmark models, this does not necessarily mean that the FI-CAPM is able to explain differences in returns *across* assets. The reason is that, while R^2_N penalizes the error in predicting the level of excess returns, it does not explicitly account for the error in predicting the cross-sectional ranking of stocks. Hence, a model that is completely uninformative for the cross-section can have a positive R^2_N as long as the level component of its prediction is better than that of the benchmark model.

Therefore, we turn to the model's unconditional cross-sectional implications, which we obtain from Equation (3.13) by the law of total expectations

$$\mathbb{E}\big(R^{e,i}_{t,t+h} - \beta^i_{t,h} \cdot \mathbb{E}_t(R^{e,M}_{t,t+h})\big) = 0 \qquad i = 1, 2, \ldots, N. \tag{3.25}$$

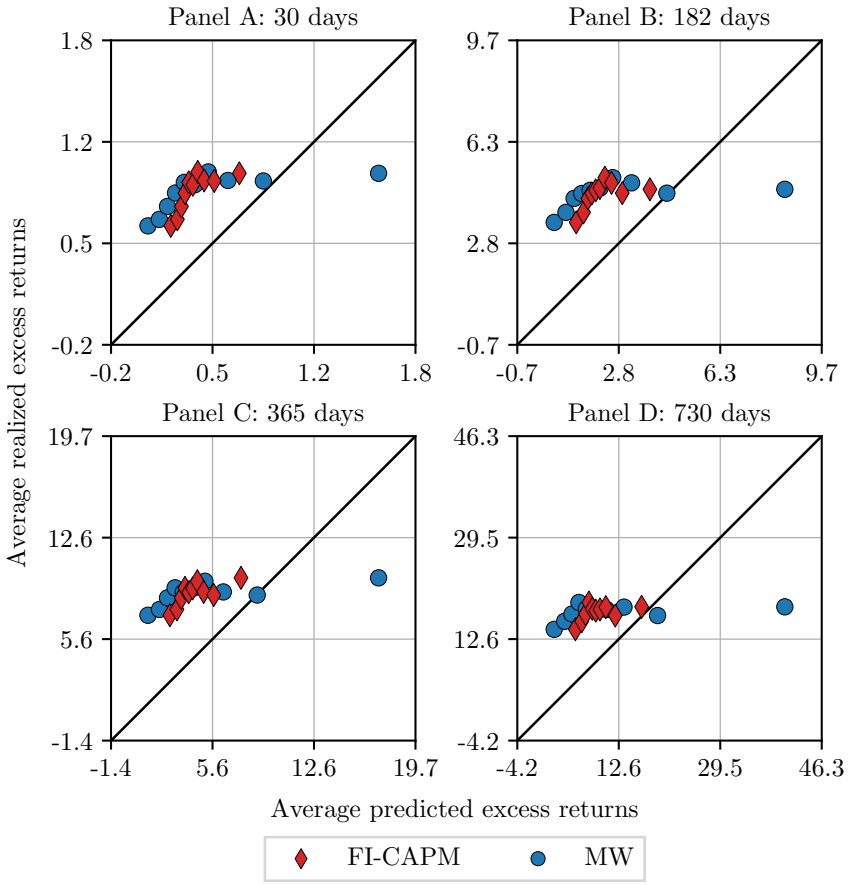
Equation (3.25) states that the realized excess return of an asset i must on average be equal to the asset's exposure to market risk, $\beta^i_{t,h}$, times the MSE-optimal prediction of the market excess return, $\mathbb{E}_t(R^{e,M}_{t,t+h})$.

In Figure 3.3, we examine the moment conditions from Equation (3.25) by contrasting the average predicted and realized excess returns of 10 prediction-sorted portfolios. The portfolios are formed daily on the basis of excess return predictions of the FI-CAPM and MW, and for investment horizons of 30, 182, 365, and 730 calendar days. In the case of the FI-CAPM, we use the terms *prediction-sorted* and *beta-sorted* interchangeably because the equity premium, as part of the stock risk premium, has no effect on the composition of the portfolios. As is common in the literature, we resort to portfolios rather than individual stocks to account for potential measurement error in the betas. An advantage of using prediction- rather than characteristic-sorted portfolios is that a model's cross-sectional performance is then determined not only by the alignment of the portfolios along the 45 degree reference line, but also by the variation produced in average realized returns across portfolios.

Regardless of the investment horizon, we observe that the average realized excess returns of the two models broadly increase across portfolios, but that the corresponding slopes are far too flat. This is emphasized by the fact that both the FI-CAPM and MW have trouble generating return differentiation in the higher deciles, as is reflected in the associated portfolios showing only little variation along the ordinate.

It is worth noting that neither of the two models can claim an advantage in terms of cross-sectional explanatory power, as both achieve their cross-sectional differentiation through risk-neutral variances of returns only, and so the portfolios' compositions are exactly the same at each point in time. Nevertheless, the FI-CAPM-based portfolios are closer to the 45-degree reference line, which confirms our finding from Section 3.4.1 that the aggregate level of its stock risk premia is more appropriate than that of MW.

Figure 3.3: Predicted and realized excess returns of prediction-sorted portfolios. This figure contrasts average predicted and realized excess returns of 10 prediction-sorted portfolios for investment horizons of 30, 182, 365, and 730 calendar days between January 1996 and December 2020. The models compared are the FI-CAPM from Equation (3.13) and the expected excess return by Martin and Wagner (2019) (MW). The portfolios are formed at a daily frequency by sorting stocks into portfolios according to the respective models' predictions. The universe of stocks is confined to securities that are constituents of the S&P 500. The 45-degree line is included for reference.



3.4.3 Sharpe ratios of prediction-sorted portfolios

To get a better understanding of the flat relationship between average predicted and realized excess returns, we estimate the portfolios' unconditional Sharpe ratios (SR). In this way, we focus on the variation in average *realized* rather than *expected* excess returns, while taking into account the estimation uncertainty associated with calculating time-series averages of returns. If $R^{e,q}_{t,t+h}$ denotes the excess return on the q'th prediction-sorted portfolio, the corresponding Sharpe ratio is given by

$$\text{SR} = \frac{\mathbb{E}(R^{e,q}_{t,t+h})}{\sqrt{\text{var}(R^{e,q}_{t,t+h})}} \qquad q = 1, 2, \dots K, \tag{3.26}$$

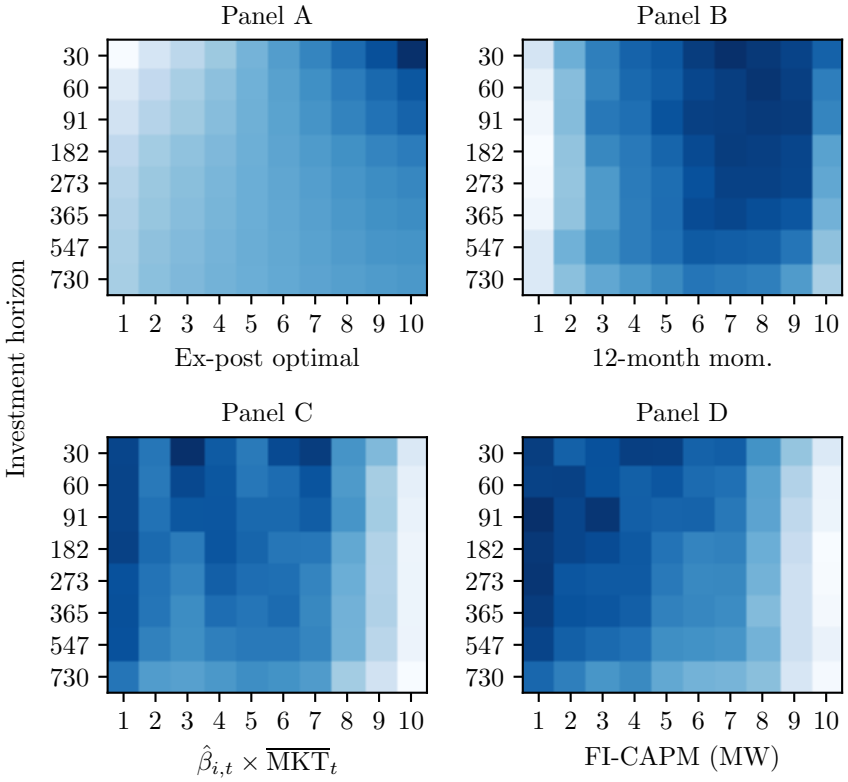
where K is the number of portfolios.

Figure 3.4 highlights the differences between the SRs across portfolios and investment horizons using a heatmap plot in which, for each model, the lowest annualized SR is associated with a light blue color and the highest SR with a dark blue color. For reference, we present in Panel A the SRs of an ex post optimal model that assigns stocks to portfolios according to their ex post realized returns, thus generating the maximum achievable SRs for the given sample. In Panel B we display the SRs of portfolios formed on 12-month momentum, Panel C shows the ones of $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$, and Panel D those of the FI-CAPM and MW, respectively.

In Panel A, the SRs increase steadily across portfolios, indicating that, in the case of the ex post optimal model, higher mean returns are not offset by higher time-series variation. The SRs across portfolios range from -5.68 to 6.13 at the 30 days investment horizon, whereas at the 365 days investment horizon they range from -1.92 to 1.91. For portfolios formed by 12-month momentum, the overall SR pattern closely resembles that of the ex post optimal model, yet the spread between the portfolios is less pronounced, ranging from 0.19 (182 days) to 0.61 (30 days), only. Also, the highest SR is obtained by the 7'th instead of the 10'th portfolio. In Panels C and D, the two versions of the conditional CAPM, the FI-CAPM and $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$, produce very similar SRs across portfolios and investment horizons. However, compared to the ex post optimal model, their SR patterns are inverted, meaning that the 10th portfolio produces a lower SR than the 1st at all investment horizons considered. The highest SR for both the FI-CAPM and MW ($\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$) is 0.62 at the 91 days investment horizon (0.61 at the 30 days investment horizon) and obtained with the 1st (3rd) portfolio, whereas the lowest SR is 0.25 (0.23) at the 182 days (730 days) investment horizon obtained with the 10th portfolio. The moderate increase in average realized returns across portfolios seen in Figure 3.3 must therefore be taken with a grain of salt – the underlying time-series variation in the higher decile portfolios is considerable.[21]

[21] Frazzini and Pedersen (2014) confirm this pattern for various international markets using betas obtained from rolling-window regressions.

Figure 3.4: Sharpe ratios of prediction-sorted portfolios. This figure displays Sharpe ratios of portfolios (Equation (3.26)) formed on the basis of a model's excess return predictions. In Panel A, we present the SRs of an ex post optimal model that assigns stocks to portfolios according to their ex post realized returns, thus generating the maximum achievable SRs for the given sample. In Panel B, we display the SRs of portfolios formed on 12-month momentum. Panel C presents the SRs of a conditional CAPM, where the betas are obtained by one-year rolling regressions and the equity premium corresponds to CRSP's value-weighted index. Panel D shows the SRs obtained by the FI-CAPM and MW, respectively, as both achieve their cross-sectional differentiation through risk-neutral variances of returns. The universe of stocks is confined to securities that are constituents of the S&P 500. The 45-degree line is included for reference. The investment horizons correspond to 30, 60, 91, 182, 273, 365, 547, and 730 calendar days, and the sample period ranges from January 1996 to December 2020.



3.4.4 Pairwise tests of relative portfolio performance

We follow up on the observation of an inverted SR pattern for both the FI-CAPM and MW (but also $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$) by proposing a formal test to examine whether the differences in the mean returns of the portfolios in Figure 3.3 are statistically significant. This should be the case if the relationship between average predicted and realized excess returns were indeed strictly monotonically increasing, as implied by the moment conditions in Equation (3.25). More specifically, we test the null hypothesis

$$H_0 : \mathbb{E}(R^{e,q}_{t,t+h}) \geq \mathbb{E}(R^{e,p}_{t,t+h})$$

for $q < p$ and $p, q \in \{1, 2, \ldots, K\}$, where $K = 10$ is the number of prediction-sorted portfolios. This involves testing multiple hypotheses simultaneously, so we need to account for multiplicity, that is, the fact that the probability of falsely rejecting at least one of the hypothesis increases in the number of portfolios being compared. To this end, we extend the testing framework by Hothorn et al. (2008) to a panel context, which allows us to examine the above hypotheses both jointly and individually, while controlling the familywise error rate.[22] The results are presented in Table 3.5 for an investment horizon of 365 calendar days.[23]

Even though the average realized returns (3rd column) generated by the FI-CAPM and MW are broadly increasing from the 1st to the 10th portfolio, we have a hard time rejecting the individual null hypotheses for the higher deciles. Significant differences compared to the 1st portfolio, for example, are obtained only up to the 8th (with the exception of the 2nd portfolio), which is in line with the Sharpe ratios (4th column) steadily decreasing from the 1st to the 10th portfolio. Hence, we find no statistical evidence that the FI-CAPM or MW produce a strictly monotonic relationship between average predicted and realized excess returns.

3.4.5 Puzzling evidence of failure in cross-sectional tests

The results presented in Sections 3.4.2 to 3.4.4 are as puzzling as they are devastating: Neither the FI-CAPM, MW, nor $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$ seem to be able to provide a sufficient explanation for why average returns vary across assets. Because previous attempts to explain this phenomenon offer at best partial relief, however, we do not simply reject those models, but instead seek to contextualize the evidence.[24]

[22] We present details on the methodology in Appendix B.5.

[23] The results remain essentially unchanged when these tests are performed at different investment horizons or when $\hat{\beta}_{i,t} \times \overline{\text{MKT}}_t$ is used instead of the FI-CAPM.

[24] An explanation put forth by Black (1972) and Frazzini and Pedersen (2014) is that investors are constrained in the amount of money they can borrow at the risk-free rate. However, as financial markets have become more complete, one would expect the relationship to have steepened over time. To the best of our knowledge, no such change has ever been reported in the literature. Similar considerations apply to the rationale that additional factors might affect the investor's compensation for risk. Few of the factors proposed in the past have been stable enough to accomodate the flat relationship's persistence. Harvey and Liu (2021), for example, find that the "market factor is by far the most important factor in explaining the cross-section of expected returns."

Table 3.5: Pairwise tests of relative portfolio performance. This table presents the results from testing $H_0 : \mathbb{E}(R^{e,q}_{t,t+h}) \geq \mathbb{E}(R^{e,p}_{t,t+h})$ for all $q < p$ and $p, q \in \{1, 2, \ldots, 10\}$. We consider portfolios formed on the excess return predictions of the FI-CAPM, yet the results equally apply to MW, because both achieve their cross-sectional differentiation through risk-neutral variances of returns. The 1st column (Portf.) indicates which of the decile portfolios is considered in the rows, the 2nd and 3rd columns present average predicted (Pred.) and realized (Real.) excess returns, the 4th column shows the associated Sharpe ratios (SR), and from the 5th to the 13th column we present one-sided p-values (adjusted for multiplicity) associated with the aforementioned $K(K-1)/2$ hypotheses. Asterisks denote statistical significance at conventional levels of 10% ($*$), 5% ($**$) and 1% ($***$). The sample period ranges from January 1996 to December 2020, and the universe of stocks is confined to S&P 500 constituents. The frequency of the data used in this analysis is daily and the investment horizon is 365 calendar days.

					Portf.								
	Portf.	Pred.	Real.	SR	2	3	4	5	6	7	8	9	10
FI-CAPM (MW)	1	2.68	7.66	0.64	0.625	0.004^{***}	$<0.001^{***}$	0.005^{***}	0.002^{***}	0.001^{***}	0.065^{*}	0.340	0.182
	2	3.18	7.89	0.59		$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.075^{*}	0.401	0.205
	3	3.48	8.63	0.57			$<0.001^{***}$	0.169	0.040^{**}	0.011^{**}	0.472	0.820	0.406
	4	3.74	9.34	0.56				1.000	0.911	0.363	0.984	0.995	0.666
	5	4.01	9.15	0.52					0.486	0.072^{*}	0.913	0.979	0.561
	6	4.30	9.45	0.51						0.378	0.997	0.999	0.651
	7	4.63	9.83	0.50							1.000	1.000	0.796
	8	5.08	9.34	0.43								0.998	0.483
	9	5.80	9.20	0.34									0.167
	10	7.75	10.69	0.28									

To deliver an explanation for the CAPM's failure in cross-sectional tests, we zoom in on the changing compositions of the beta-sorted portfolios over time. In order for Equation (3.25) to hold on average, the allocation of stocks to portfolios should be such that, at each point in time, the realized deciles correspond to the ones predicted. To test this assertion, we sort stocks into deciles according to the FI-CAPM's predictions and record the relative frequency with which each possible combination of predicted and realized deciles occurs. Table 3.6 reports the results of this exercise, showing a very unique pattern across investment horizons: If one imagines a line between the 5th and the 6th decile, whether vertically or horizontally, the relative frequencies appear to mirror each other along this line. At an investment horizon of 30 days, this phenomenon is particularly striking for the 1st and the 10th predicted decile, as the stocks with the supposed lowest (highest) betas end up in the correct realized decile in 27% (32%) of the cases, but in the exact opposite highest (lowest) decile in 22% (34%) of the cases.

In Figure 3.5, we investigate the periodicity of this switching behavior by contrasting the predicted and realized deciles of excess returns over time. The color gradient that we plot at each date represents the realized order of the deciles relative to the order predicted by the FI-CAPM. For example, if the darkest blue color appears at the top of the graph, and the lighest white color at its bottom, this indicates that the respective 1st and 10th deciles coincide. Conversely, if the darkest blue color appears at the bottom, and the lightest white color at the top, the predicted order is reversed compared to the realized order.

As can be seen from Panels A to D of Figure 3.5, the state of the portfolio alignment (either correct or inverted) changes quite frequently, but gets more persistent as the investment horizon increases. At the annual horizon, the alignment appears to be almost block-shaped across dates, suggesting that this pattern is driven by some cyclical, macroeconomic component. Adding the annualized realized market excess return to the graph, reveals that the order of the deciles tends to change whenever the S&P 500 transitions from a bear to a bull market, and vice versa.

As a matter of fact, this observation is not entirely new. Pettengill et al. (1995), for example, find "a consistent and highly significant relationship between unconditional beta and cross-sectional portfolio returns", when conditioning on the sign of realized market excess returns. Their reading of this result is that the failure of conventional tests is due to the fact that the positive relationship between returns and beta asserted by the unconditional CAPM is based on expected rather than realized returns. Under rational expectations (and rather mild regularity conditions), however, this should not matter much as time-series averages of returns constitute consistent estimators of expected returns. Placing their observation in the context of a conditional CAPM, instead, yields a different and, in our view, more coherent interpretation of this phenomenon.

Table 3.6: Predicted and realized deciles of excess returns. This table shows the relative frequencies (in %) of all possible combinations of predicted and realized deciles of excess returns. The predicted deciles (columns) are formed on a daily basis by sorting stocks into portfolios according to the expected excess returns of the FI-CAPM in Equation (3.13). The realized deciles (rows) instead represent the order that is achieved by investing in these portfolios. Note that the numbers presented also apply to MW, because both achieve their cross-sectional differentiation through risk-neutral variances of returns. Panels A to D show the results for investment horizons of 30, 182, 365, and 730 calendar days. The sample period ranges from January 1996 to December 2020. The universe of stocks corresponds to the extended sample of all common shares available in CRSP/OptionMetrics. The frequency of the data used in this analysis is daily.

Panel A: 30 days

Realized deciles \ Predicted deciles	1	2	3	4	5	6	7	8	9	10
1	27	7	4	3	3	3	4	6	9	34
2	10	19	9	5	5	6	7	10	21	6
3	7	10	17	10	8	9	10	16	9	4
4	6	8	12	15	13	14	14	9	6	3
5	5	8	10	15	18	16	12	9	5	3
6	5	8	11	14	18	17	12	8	5	2
7	5	9	12	15	13	12	15	10	5	4
8	6	10	14	11	10	10	12	16	9	4
9	8	16	8	7	8	7	9	11	19	8
10	22	5	5	5	4	5	5	6	11	32

Panel B: 182 days

Realized deciles \ Predicted deciles	1	2	3	4	5	6	7	8	9	10
1	29	6	2	3	3	3	3	5	9	38
2	11	22	6	4	5	5	5	10	24	7
3	7	11	17	8	7	8	11	17	9	3
4	5	7	11	17	12	13	13	10	7	3
5	6	7	10	13	19	17	11	9	5	3
6	5	7	10	15	18	19	12	7	5	3
7	5	7	12	16	12	13	14	11	5	4
8	5	9	16	12	10	10	14	14	6	5
9	7	17	9	8	8	7	9	9	19	6
10	21	6	6	5	6	5	7	7	10	29

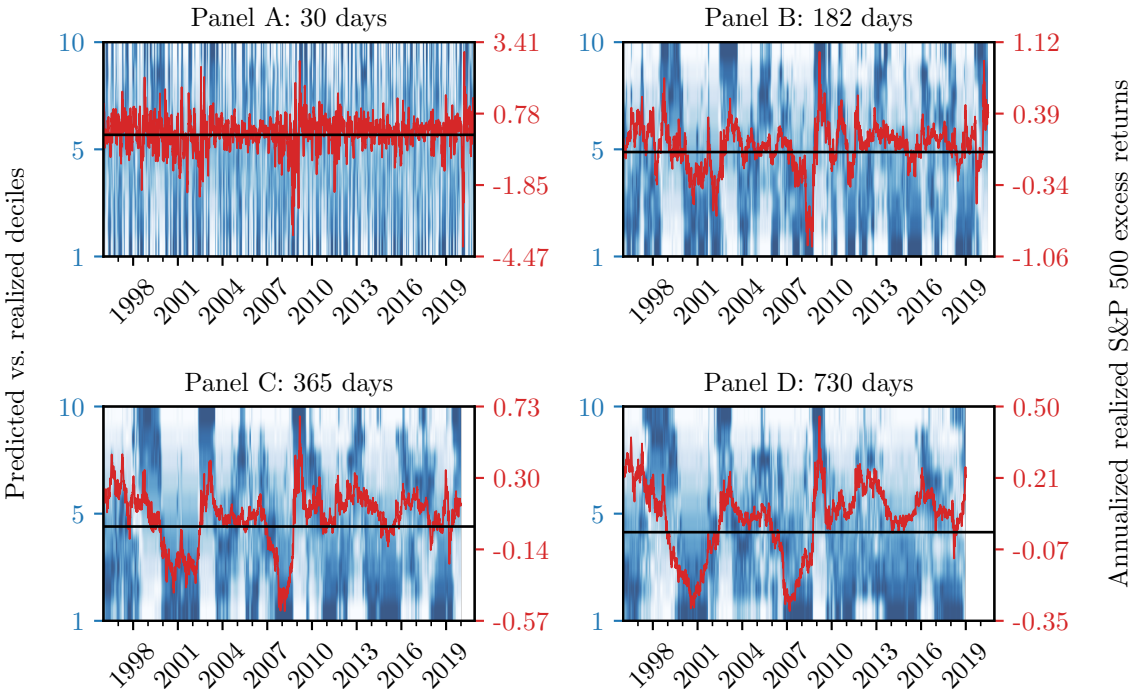
Panel C: 365 days

Realized deciles \ Predicted deciles	1	2	3	4	5	6	7	8	9	10
1	26	7	4	2	4	2	2	5	9	39
2	12	20	6	3	5	5	5	9	25	9
3	8	10	17	7	8	8	9	19	9	5
4	6	8	11	13	12	14	14	11	6	4
5	6	7	8	14	17	18	12	9	6	3
6	5	7	8	15	18	17	13	9	5	2
7	5	8	12	17	14	13	14	9	6	2
8	5	9	16	12	10	11	14	13	8	3
9	7	16	11	9	8	7	11	10	18	4
10	21	7	7	6	5	4	6	7	9	29

Panel D: 730 days

Realized deciles \ Predicted deciles	1	2	3	4	5	6	7	8	9	10
1	25	7	4	1	2	2	4	6	11	38
2	10	16	8	2	5	6	9	10	21	12
3	7	10	12	6	9	10	12	17	11	7
4	5	9	9	13	13	16	13	10	8	4
5	4	8	9	13	16	17	13	7	7	3
6	5	8	11	15	17	14	13	8	6	3
7	6	9	13	15	15	13	12	9	7	3
8	7	11	15	12	10	10	11	12	8	3
9	10	14	12	11	7	7	8	10	13	7
10	20	9	7	11	6	4	5	10	8	20

Figure 3.5: Predicted and realized excess return deciles over time. This figure contrasts the predicted and realized excess returns of 10 prediction-sorted portfolios on the primary vertical axis (blue) with the annualized excess return of the S&P 500 index (red) on the secondary vertical axis. The portfolios are formed on a daily basis using the expected excess returns of the FI-CAPM in Equation (3.13). The labels from 1 to 10 along the primary vertical axis indicate the predicted deciles. The plotted color gradient shows the corresponding realized deciles at each date, with the lightest white color representing the 1st decile and the darkest blue color representing the 10th decile. Panels A to D show the results for investment horizons of 30, 182, 365, and 730 calendar days. The sample period ranges from January 1996 to December 2020. The universe of stocks corresponds to the extended sample of all common shares available in CRSP/OptionMetrics. The frequency of the data used in this analysis is daily.



3.4.6 The uncertainty principle in asset pricing

To provide a formal explanation for the stylized facts presented in Section 3.4.5, we return to the conditional CAPM's unconditional implications from Equation (3.25), and interpret the term $\beta^i_{t,h} \times \mathbb{E}_t(R^{e,M}_{t,t+h})$ as the MSE-optimal prediction of future stock excess returns. It is easy to see that, because of its composition, any variation in the resulting forecasts stems from either the betas *or* the conditional equity premium. The failure of the CAPM, however, is commonly associated with a failure of the betas and/or the absence of additional risk factors.[25]

We challenge this view arguing that the flat relationship between average predicted and realized excess returns at short horizons is rather due to the inherent uncertainty associated with forecasting market excess returns. Moreover, it is the mismatch between the predicted and realized signs of market excess returns that explains the cyclical behavior of the portfolios shown in Figure 3.5.

To formalize this claim, we decompose the return of the market portfolio according to

$$R^{e,M}_{t,t+h} = \mathbb{E}_t\big(R^{e,M}_{t,t+h}\big) + \varepsilon^M_{t,t+h},$$

where $\mathbb{E}_t\big(\varepsilon^M_{t,t+h}\big) = \mathbb{E}(\varepsilon^M_{t,t+h}|\mathcal{F}_t) = 0$, and $\mathcal{F}_t$ represents the information set available to investors when forming expectations. Solving for $\mathbb{E}_t(R^{e,M}_{t,t+h})$ allows us to express Equation (3.25) in terms of the realized market excess return and its inherently unpredictable component $\varepsilon^M_{t,t+h}$, so that

$$\mathbb{E}\big(R^{e,i}_{t,t+h} - \beta^i_{t,h} \cdot R^{e,M}_{t,t+h}\big) + \mathbb{E}\big(\beta^i_{t,h} \cdot \varepsilon^M_{t,t+h}\big) = 0.$$

Note that, as a result of $\mathbb{E}(\varepsilon^M_{t,t+h}|\mathcal{F}_t) = 0$, the residual component $\varepsilon^M_{t,t+h}$ is uncorrelated with any $\mathcal{F}_t$-measurable random variable. Provided that $\beta^i_{t,h}$ is a function of such random variables, and can thus be calculated from option prices, it follows that $\mathbb{E}(\beta^i_{t,h} \cdot \varepsilon^M_{t,t+h}) = 0$.[26] As a result, alternative moment conditions to test the conditional CAPM's cross-sectional implications are given by

$$\mathbb{E}\big(R^{e,i}_{t,t+h} - \beta^i_{t,h} \cdot R^{e,M}_{t,t+h}\big) = 0 \qquad i = 1, 2, \ldots, N. \tag{3.27}$$

The crucial difference between Equations (3.25) and (3.27) is that for the latter, the conditional market premium has been replaced by the market excess return and its inherently unpredictable component $\varepsilon^M_{t,t+h}$. When testing the conditional CAPM with

[25] Cochrane (2011), for example, illustrates the CAPM's shortcomings by contrasting the average returns and betas of 10 book-to-market sorted portfolios. In doing so, he neglects the role of the conditional CAPM and its unconditional implications, according to which average returns should actually be set against the model's average excess return predictions, not just the betas. While this distinction seems subtle at first, its omission disguises the fact that the failure of the CAPM may be due to the equity premium rather than to the betas.

[26] Any procedure that estimates beta based on information available at t relies on $\beta^i_{t,h}$ being a function of $\mathcal{F}_t$-measurable random variables.

respect to Equation (3.27), we therefore abstract from the question of whether market returns are predictable and instead concentrate on the betas' cross-sectional explanatory power.

The insight that we gain from this is presented in Figure 3.6. Comparing the quality of the results from matching the moments in Equations (3.25) and (3.27), we find that the alignment of the sample averages improves significantly when leaving aside predicting market excess returns – the displayed averages are closer to the 45 degree reference line and their dispersion is much larger. The alignment of the excess return deciles, however, deteriorates as the investment horizon increases, indicating that the betas' cross-sectional explanatory power gradually declines.

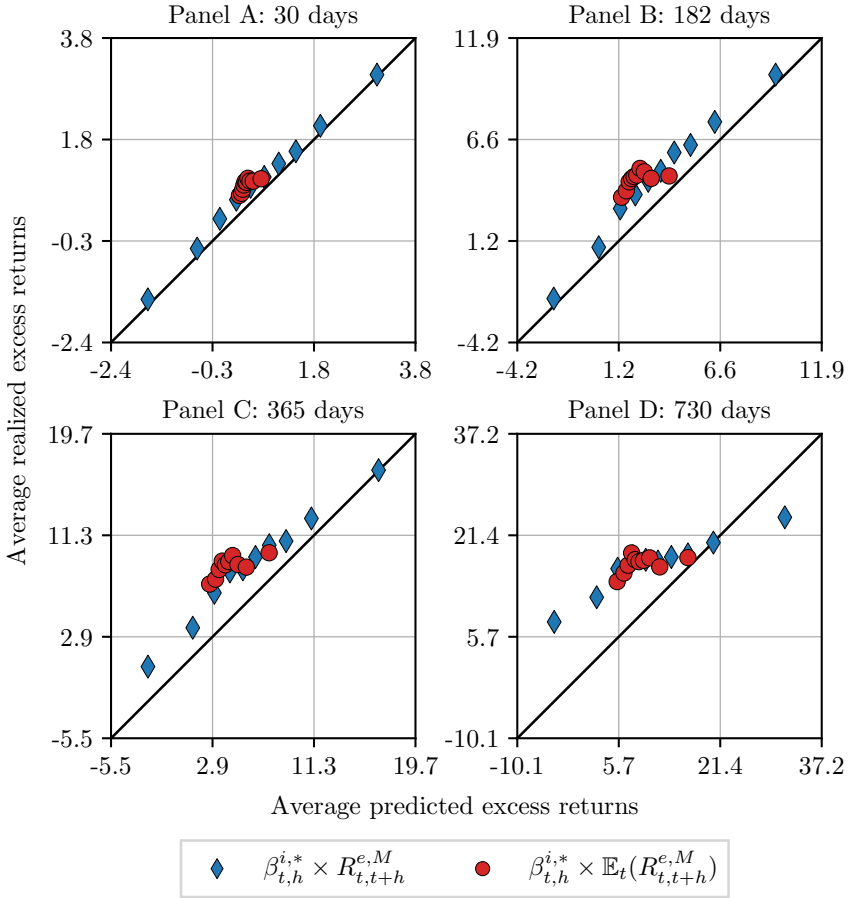
Together with the results from Figure 3.5 and Table 3.6, these findings suggest that at shorter horizons (1 to 12 months), the failure of the conditional CAPM is less due to the betas than to the existence of the unexpected news component $\varepsilon^M_{t,t+h}$, which accounts for most of the variation in $R^{e,M}_{t,t+h}$. Moreover, it seems as if the flat relationship could only be overcome if the conditional market premium were informative about the sign of future market excess returns. Any equilibrium asset pricing model that assumes risk-averse investors, however, requires the equity premium to be positive at all times, and so the conditional CAPM is doomed to fail in cross-sectional tests of Equation (3.25).[27] Although the conditional CAPM is thus rather limited in its ability to predict future stock returns, its betas still contribute to explaining the variation in returns across assets. Consequently, there is little to be said against using betas as ex ante measures of exposure to market risk, as long as the investment horizon is sufficiently short.

At longer horizons (beyond 12 months), the equity premium implied by the FI-CAPM accounts for a larger fraction of the variation in realized market excess returns, meaning that there is less uncertainty associated with their prediction (see Table 3.4). At the same time, however, the cross-sectional signal contained in the betas deteriorates, which is why, again, the FI-CAPM is unable to satisfy the moment conditions in Equation (3.25). Our reading of this result is that a company's association with the market can change, for example, due to a realignment of the business model, making its prediction increasingly difficult as the investment horizon increases.

Considering the results at shorter and longer horizons together, we find that the stock return predictions of the FI-CAPM are subject to two different types of uncertainties – one associated with forecasting realized betas, and the other associated with forecasting market excess returns. As functions of the investment horizon, these two types of uncertainties are inversely related, which explains the flat relationship's persistence. In analogy to the uncertainty principle in quantum physics, we refer to

[27] The equity premium is the compensation a risk-averse investor demands ex ante for holding risk. It is subject to a positivity constraint, which ensures that investing in the market is a sensible strategy (cf. Merton, 1982). A negative equity premium is thus difficult to reconcile with equilibrium asset pricing models. We therefore believe that allowing for a negative market premium is not a viable solution to the situation described above.

Figure 3.6: Portfolio alignment when using alternative moment conditions. This figure illustrates the results from matching the moment conditions in Equations (3.25) and (3.27) for 10 prediction-sorted portfolios. The portfolios are formed daily by sorting stocks into portfolios according to the excess return predictions of the FI-CAPM (red circle), and the product of the implied betas and realized market excess returns (blue diamond). Panels A to D represent investment horizons of 30, 182, 365, and 730 calendar days, and the sample period ranges from January 1996 to December 2020. The universe of stocks is confined to securities that are constituents of the S&P 500. The 45-degree line is included for reference.



this observation as the uncertainty principle in asset pricing.[28]

3.4.7 The cross-sectional explanatory power of the betas

We conclude our discussion of the conditional CAPM's failure in cross-sectional tests with a positive note on the betas. By contrasting Figures 3.3 and 3.6, we have seen that, while measurements of the market premium are not very informative for the variation in future market excess returns at shorter horizons, betas still provide substantial cross-sectional explanatory power. We have attributed this power to the observation that the betas generate substantial variation between portfolios once one abstracts from the question of market return predictability, as shown in Figure 3.6. To this point, however, we have refrained from asking how market betas compare, in this respect, to other stock characteristics.

To investigate this matter, we propose a novel procedure to quantify the degree of cross-sectional differentation that is achieved in the creation of prediction- or characteristic-sorted portfolios. Taking into account the evidence from Section 3.4.6, we abstract from the predictability of market excess returns and concentrate on the period-by-period return variation across portfolios. Using this procedure, we hope to gain insight into whether established characteristics such as 12-month momentum or a company's book-to-market ratio provide additional information for the cross-section of returns. We consider 12-month momentum, in particular, an interesting competitor, because Kelly et al. (2021) have recently shown that, when used as an instrument for conditional factor loadings, it outperforms many other stock characteristics.

To construct our measure of cross-sectional explanatory power, we start by defining $R^{e,p}_{t,t+h}$ as the return of an equally-weighted portfolio that comprises the entire cross-section of stocks in t, such that $\sum_i w^i_t = 1$ and $w^i_t = 1/N_t$. This portfolio, which we refer to as the *cross-sectional* portfolio, can be conceived of as a portfolio of prediction-sorted portfolios formed by some model ω_m

$$R^{e,p}_{t,t+h} = \sum_i w^i_t R^{e,i}_{t,t+h} = \sum_q w^q_t \sum_i w^{i,q}_t R^{e,i}_{t,t+h} = \sum_q w^q_t R^{e,q}_{t,t+h}(\omega_m), \tag{3.28}$$

where w^q_t represents the weight of portfolio q in the cross-sectional portfolio, and $w^{i,q}_t$ is the weight of asset i in portfolio q.[29] We then use Equation (3.28), together with

[28] A study that in its conclusions appears to be similar to ours it that by Hasler and Martineau (2023), who seek to explain the failure of the unconditional CAPM by the success of the conditional CAPM. However, in testing the conditional model they focus on the *contemporaneous* relationship between stock and market excess returns shown in Equation (3.27), thus ignoring the fact that the conditional CAPM should also be able to predict stock returns, as required by the moment conditions in Equation (3.25). Accordingly, Hasler and Martineau (2023) find a flat relationship between betas and average excess returns only for the case of the unconditional model. We, instead, provide evidence that it also applies to the conditional model. Hence, while their approach reconciles the success of the conditional CAPM with the failure of the unconditional CAPM, it tells us little about the origins of the flat relationship shown in Figure 3.3.

[29] Note that $w^q_t = N^q_t/N_t$ and $w^{i,q}_t = \mathbb{1}(\{i,t\} \in \mathbf{\Omega}_q)/N^q_t$, where N^q_t is the number of assets in portfolio q, $\mathbf{\Omega}_q$ is the set of security-date indices belonging to the q'th portfolio and $\mathbb{1}(\cdot)$ is the

the law of total variance, to decompose the cross-sectional variance of returns in t according to

$$\text{cvar}_t(R^{e,i}_{t,t+h}) = \sum_i w^i_t \big(R^{e,i}_{t,t+h} - \sum_i w^i_t R^{e,i}_{t,t+h}\big)^2 = \text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_m)\big) + \eta_t, \tag{3.29}$$

where the term

$$\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_m)\big) = \sum_q w^q_t \big(R^{e,q}_{t,t+h}(\omega_m) - \sum_q w^q_t R^{e,q}_{t,t+h}(\omega_m)\big)^2 \tag{3.30}$$

represents the between-portfolio variation produced by some model ω_m, and

$$\eta_t = \sum_q w^q_t \Big(\sum_i w^{i,q}_t \big(R^{e,i}_{t,t+h} - R^{e,q}_{t,t+h}(\omega_m)\big)^2 \Big)$$

is the residual average within-portfolio variation.

The appeal of this decomposition is that the expression in Equation (3.30) can be conceived of as the proportion of cross-sectional variation that is explained by the predictions of ω_m. To illustrate this idea, we present in Figure 3.7 an example of a cross-sectional return distribution that is partitioned by $K = 5$ prediction-sorted portfolios, once for the case of a non-informative model where the between-portfolio variation is low (Panel A), and once for an informative model where the between-portfolio variation is high (Panel B).

As follows from Equation (3.29) and can be seen in Figure 3.7 when moving from Panel A to Panel B, an increase in between-portfolio variation must coincide with a decrease in average within-portfolio variation. The stocks associated with the portfolios in Panel B are thus more similar in terms of their future realized returns than the stocks in Panel A. As a consequence, the model among competing models that generates the highest between-portfolio variation is best at discriminating between stocks and achieves the highest level of cross-sectional explanatory power. Under the null hypothesis that the FI-CAPM betas provide a sufficient explanation for the variation in returns across assets, we would expect no other characteristic to be able to generate greater between-portfolio variation.

Considering that the amount of cross-sectional variation can change over time, it remains unclear how large the between-portfolio variation of a model should be for it to be considered economically relevant. With reference to the expression in Equation (3.29), we emphasize that, if the number of portfolios is such that $1 < K < N_t$ and also $\text{cvar}_t(R^{e,i}_{t,t+h}) > 0$, we have $\eta_t > 0$, regardless of how well the underlying model captures differences in stock risk premia. Accordingly, even with an ex post optimal model ω_0 satisfying

$$\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_m)\big) \leq \text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_0)\big), \tag{3.31}$$

indicator function giving 1 if a certain $\{i,t\}$-combination belongs to the q'th portfolio and 0 else.

there is an irreducible, greater-than-zero expected within-portfolio variation.[30]

We take advantage of this mathematical fact and express the between-portfolio variation of any sub-optimal model relative to that of the ex post optimal model, which results in a normalization of the between-portfolio variation, so that

$$\text{BPV}_t = \frac{\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_m)\big)}{\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_0)\big)} \in [0,1]. \tag{3.32}$$

Accordingly, one may think of BPV_t as the ratio of two cross-sectional R^2s

$$\text{BPV}_t = \frac{\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_m)\big)}{\text{cvar}_t(R^{e,i}_{t,t+h})} \cdot \frac{\text{cvar}_t(R^{e,i}_{t,t+h})}{\text{cvar}_t\big(R^{e,q}_{t,t+h}(\omega_0)\big)} = \frac{R^2_t(\omega_m)}{R^2_t(\omega_0)}$$

where $R^2_t(\omega_m)$ is the sub-optimal model's R-squared and $R^2_t(\omega_0)$ is that of the ex post optimal model.[31]

Figure 3.8 presents time series of BPV_t for both the FI-CAPM and 12-month momentum at investment horizons of 30, 182, 365, and 730 calendar days. In either case, the amount of cross-sectional variation explained is largest during periods of market turmoil (almost 70% during the DCB), and it gains persistence as the investment horizon increases. The highest average BPV_t associated with the FI-CAPM is 9.5% at an investment horizon of 182 calendar days, whereas that of 12-month momentum is 8.8% (30 days). Moreover, the two time series of BPV_t are very similar across panels, meaning that they indicate comparable levels of cross-sectional explanatory power over time.

Based on these findings, we conclude that 12-month momentum provides little, if any, additional information to explaining the cross-section of stock returns once one abstracts from the uncertainty associated with forecasting market excess returns.[32]

[30] We think of the ex post optimal model as a stylized model, that assigns stocks to portfolios according to their ex post realized returns.

[31] If the number of portfolios is equal to the number of stocks in the cross-section ($K = N_t$), each of the N_t portfolios consists of a single stock only, so $\text{BPV}_t = 1$. Similarly, if there is only one portfolio ($K = 1$), BPV_t does not exist. Therefore, BPV_t is only meaningful if the number of portfolios is chosen such that $1 < K < N_t$.

[32] Similar conclusions can be drawn when using the book-to-market ratio in lieu of 12-month momentum. The associated results are presented in Figure B.2 of Appendix B.6.

Figure 3.7: Between- and within-portfolio variation of returns (Illustration). This figure shows an example of a cross-sectional return distribution partitioned by $K = 5$ prediction-sorted portfolios, once for the case of a non-informative model with low between-portfolio variation (Panel A), and once for an informative model with high between-portfolio variation (Panel B). In either Panel, the solid line represents the distribution of the assets' excess returns associated with the cross-sectional portfolio, the dashed lines correspond to those of the 5 prediction-sorted portfolios, and the crosses along the abscissa represent the portfolios' excess returns.

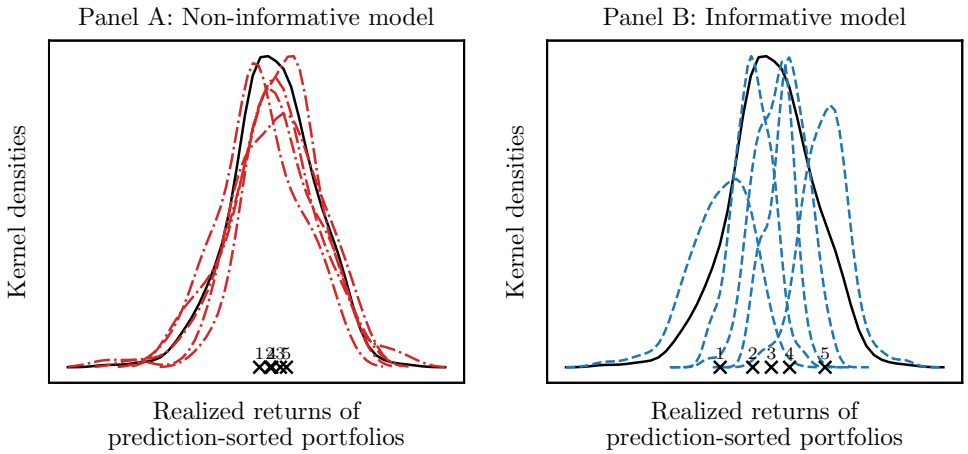
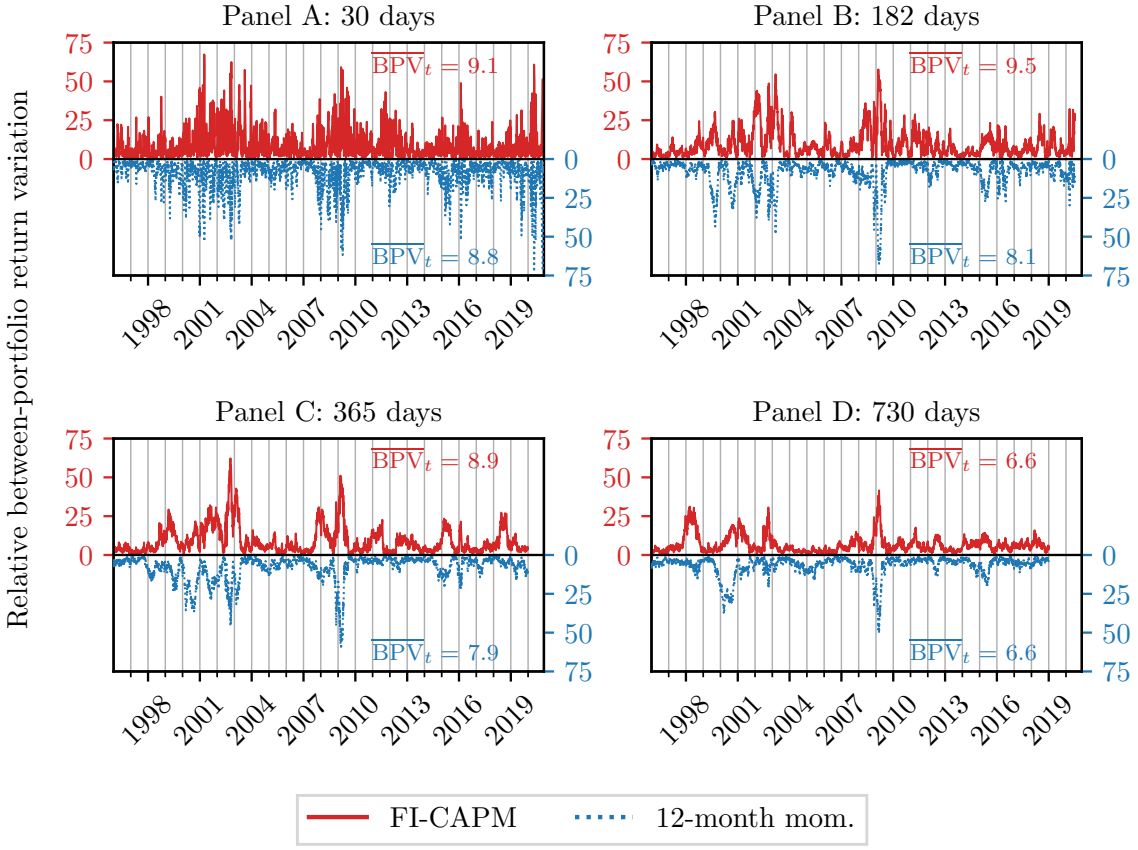


Figure 3.8: Relative between-portfolio return variation (12-month momentum). This figure shows time series of relative between-portfolio return variation (BPV_t from Equation (3.32)) for 10 prediction-sorted portfolios at investment horizons of 30, 182, 365, and 730 calendar days between January 1996 and December 2020. The underlying portfolios are formed on the excess return predictions of FI-CAPM in Equation (3.13) (solid red line) and the stocks' 12-month momentum (dotted blue line). For the sake of clarity, the time series are plotted on two different vertical axes, mirrored by the horizontal zero line. The numbers associated with $\overline{\text{BPV}}_t$ represent time-series averages of relative between-portfolio variation. The universe of stocks is confined to securities that are constituents of the S&P 500. The time series have a daily frequency.



3.5 Conclusions

In this chapter, we develop a conditional CAPM that is fully-implied by option prices, meaning that neither the betas nor the equity premium need to be estimated econometrically. We refer to this specification as the *fully-implied* CAPM, or FI-CAPM. While the approach presented is more favorable in its assumptions and yields lower out-of-sample forecast errors than a number of competing models, it appears to struggle with explaining the variation in average returns across assets. Similar to the unconditional CAPM, this manifests itself in a flat relationship between average predicted and realized excess returns of beta-sorted portfolios.

Rather than simply rejecting the model, we provide an explanation for this phenomenon, arguing that, at shorter investment horizons, it is the uncertainty associated with forecasting market excess returns that renders the conditional CAPM unsuccessful in cross-sectional tests. That is, once we abstract from the question of market return predictability, the betas of the FI-CAPM exhibit substantial cross-sectional explanatory power, to the extent that both momentum and value characteristics fail to provide additional information.

At longer horizons, the equity premium accounts for a larger fraction of the variation in realized market excess returns, indicating that there is less uncertainty associated with their prediction. Despite this reduction in uncertainty, however, the conditional CAPM remains unsuccessful as the betas' cross-sectional explanatory power simultaneously declines. We conclude that, depending on the investment horizons, one of the two components of the conditional CAPM is always subject to a degree of uncertainty that prevents the other from showing its potential. In analogy to the uncertainty principle in quantum mechanics, we refer to this observation as the uncertainty principle in asset pricing.

B Appendix

B.1 Database

For the period between January 1996 and December 2020 we obtain daily price and return data for the US stocks available via CRSP (The Center for Reseach in Security Prices) and the S&P 500 index. For every stock we add the company's name, CRSP's permanent security identifier, the trading volume, the number of shares outstanding, the share code, the Standard Industrial Classification (SIC) code and, if available, delisting information. From Compustat, we retrieve a list of historical S&P 500 constituents which we use to determine whether a particular date-security combination in CRSP was part of the S&P 500. Since CRSP and Compustat rely on different permanent security identifiers, we link securities across databases using the linking suite provided by WRDS (Wharton Research Data Services).

Using the series of daily returns, we compute multi-period returns for investment horizons of 30, 182, 365, and 730 calendar days. The investment horizons are chosen such that they match the standardized maturities in the volatility surface maintained by OptionMetrics. Because daily returns are recorded at a business day frequency, we require the multi-period returns to cover at least the amount of calendar days associated with a given investment horizon. Thus, whenever the liquidation date falls on a weekend or holiday, the holding period will exceed the number of calendar days that comprise the investment horizon by the number of days to the next business day. For stocks being delisted during January 1996 and December 2020, we adjust the last available daily return by the associated delisting return provided by CRSP. As the delisting event is unexpected from the perspective of the investor, the number of calendar days over which the multi-period returns are calculated shrink toward zero as we approach the delisting date. Returns with a holding period exceeding December 31, 2020 are removed from the sample.[33]

For calculating realized and expected excess returns, we also require a term structure of risk-free interest rates, which we obtain from OptionMetrics. As the grid of the provided term structure does not always match the investment horizons (or maturities) of interest, we employ linear interpolation between neighbouring zero-coupon rates as well as constant extrapolation. After that, we retrieve an implied volatility surface from OptionMetrics for the security-dates and investment horizons in our sample. The volatility surface is derived from American put and call option contracts and covers deltas ranging from -0.9 to 0.9 in steps of 0.05. We link securities between CRSP and OptionMetrics using the linking suite provided by WRDS, and remove observations with missing or invalid values in implied volatilities, strike prices, closing prices, returns and/or risk-free rates.

With the panel of implied volatilities, we compute prices of equivalent European

[33]For more information on the bias that is due to missing delisting information from CRSP please refer to Shumway (1997).

call and put options at given deltas and investment horizons using the Black-Scholes-Merton (BSM) formula. These prices form the basis for calculating the risk-neutral moments of simple stock returns, which we use to construct the different variants of option-based excess return forecasts. We provide details on the approximation of risk-neutral moments in Appendix B.2. Following Martin and Wagner (2019), we remove security-date observations where the risk-neutral variance is not monotonically increasing in the time-to-maturity of the underlying options.

We further obtain stock characteristics that we use as benchmarks in our empirical analysis. More specifically, we compute 12-month momentum as the rolling return over the past 12 months preceeding the dates of interest, log-size, which we derive from a stock's market capitalization, conditional CAPM betas, which we estimate by one-year rolling regressions, and the betas of the Fama-French three factor model, where we obtain the market, SMB, and HML factors directly from CRSP. We also construct book-to-market ratios following the procedure by Daniel et al. (1997). Each of the stock characteristics is made available at a daily frequency.

Tables B.1 and B.2 provide descriptive statistics for the samples of S&P 500 constituents and common shares available in CRSP/OptionMetrics for an investment horizon of 365 calendar days.

B.2 Approximating risk-neutral moments of returns

In order to calculate the individual components of the FI-CAPM, i.e, the betas and the equity premium, we need to express the risk-neutral variances in Equation (3.11) as functions of option prices. To this end, we refer to Bakshi and Madan (2000), Martin (2018), and Chabi-Yo et al. (2023), who state the m'th uncentered risk-neutral moment of a simple return as

$$A_m = \mathbb{E}_t^*\big((R_{t,t+h}^i)^m\big) = \underbrace{\frac{(F_{t,t+h}^i)^m}{(S_t^i)^m}}_{=(R_{t,t+h}^f)^m} + \frac{m(m-1)R_{t,t+h}^f}{(S_t^i)^m}$$
$$\times \left(\int_0^{F_{t,t+h}^i} K^{(m-2)}\ \mathrm{put}_{t,t+h}(K)\ dK + \int_{F_{t,t+h}^i}^{\infty} K^{(m-2)}\ \mathrm{call}_{t,t+h}(K)\ dK \right),$$

where m denotes the order of the moment, S_t^i is the price of the underlying asset, $F_{t,t+h}^i$ is the associated forward contract with maturity in $t+h$, $\mathrm{call}_{t,t+h}$ and $\mathrm{put}_{t,t+h}$ are the prices of European call and put options, respectively, and K is the strike price.

The risk-neutral variance of an asset's gross return can thus be expressed as

$$\begin{aligned}\mathrm{var}_t^*(R_{t,t+h}^i) &= \mathbb{E}_t^*\big((R_{t,t+h}^i)^2\big) - \mathbb{E}_t^*(R_{t,t+h}^i)^2 \\ &= A_2 - A_1^2,\end{aligned}$$

Table B.1: Sample descriptives (S&P 500 constituents). This table presents sample descriptives for S&P 500 constituents with returns over 365 calendar days. The returns in our sample are aligned in a forward-looking manner, so that long positions with a timestamp from 2020 are closed out in 2021. As a consequence, our sample excludes observations from 2020 for which the holding periods would exceed the last date available in CRSP, which is December 31, 2020. For annual sub-samples, the table presents the number of security-date observations (Observations), the average number of securities per day (Securities), the average-per-day market capitalization in millions of US Dollars (Mcap.), the average-per-day trading volume in units of a thousand shares (Volume), and the average-per-day return on investments over 365 calendar days in percent (Return).

Year	Observations	Securities	Mcap.	Volume	Return
1996	115,095	457	10,490	737	27.4
1997	118,636	473	13,684	986	19.7
1998	120,079	478	17,423	1,298	14.3
1999	122,215	485	21,978	1,734	6.2
2000	119,145	473	25,423	2,780	8.1
2001	120,053	484	21,806	3,559	−7.2
2002	121,032	484	18,367	3,945	6.6
2003	122,441	490	17,892	3,609	29.1
2004	121,753	487	21,039	3,509	14.0
2005	120,736	483	22,793	3,845	14.2
2006	120,975	484	24,717	4,462	15.0
2007	119,833	481	27,477	5,519	−16.8
2008	121,579	484	22,179	7,922	−13.7
2009	124,083	494	17,273	9,114	37.7
2010	124,501	494	21,274	7,829	16.2
2011	123,105	489	24,011	6,749	8.5
2012	121,332	485	26,030	5,461	27.0
2013	122,247	485	30,866	4,619	21.7
2014	120,358	478	36,556	4,297	8.4
2015	120,178	477	38,685	4,581	2.3
2016	120,736	479	38,432	4,700	18.4
2017	120,165	479	44,292	4,037	12.2
2018	118,287	471	49,970	4,579	6.2
2019	119,738	475	51,887	4,002	2.5

Table B.2: Sample descriptives (Common shares). This table presents sample descriptives for all common shares available in CRSP/OptionMetrics with share codes equal to 10, 11, 12, or 18 and returns over 365 calendar days. The returns in our sample are aligned in a forward-looking manner, so that long positions with a timestamp from 2020 are closed out in 2021. As a consequence, our sample excludes observations from 2020 for which the holding periods would exceed the last date available in CRSP, which is December 31, 2020. For annual sub-samples, the table presents the number of security-date observations (Observations), the average number of securities per day (Securities), the average-per-day market capitalization in millions of US Dollars (Mcap.), the average-per-day trading volume in units of a thousand shares (Volume), and the average-per-day return on investments over 365 calendar days in percent (Return).

Year	Observations	Securities	Mcap.	Volume	Return
1996	427,816	1,698	3,601	408	19.3
1997	515,193	2,053	4,007	452	10.9
1998	580,497	2,313	4,519	522	14.7
1999	602,792	2,392	5,498	657	30.2
2000	540,177	2,144	7,167	1,115	−6.6
2001	487,044	1,964	6,506	1,431	−12.7
2002	489,514	1,958	5,554	1,450	10.3
2003	467,893	1,872	5,839	1,470	33.6
2004	498,677	1,995	6,552	1,470	11.9
2005	525,499	2,102	6,831	1,460	15.9
2006	557,092	2,228	7,112	1,635	13.4
2007	589,682	2,368	7,516	1,856	−19.4
2008	591,426	2,356	6,182	2,497	−12.6
2009	589,563	2,349	4,939	2,790	41.5
2010	616,925	2,448	5,898	2,391	15.9
2011	628,953	2,496	6,450	2,153	4.3
2012	621,359	2,485	6,770	1,802	26.0
2013	663,414	2,633	7,607	1,599	18.8
2014	648,499	2,573	8,928	1,615	2.4
2015	645,182	2,560	9,318	1,625	−3.7
2016	663,594	2,633	8,860	1,634	21.8
2017	655,327	2,611	10,335	1,501	14.0
2018	626,042	2,494	11,904	1,666	−0.5
2019	616,539	2,447	12,606	1,609	3.7

where

$$A_2 = \underbrace{\frac{(F^i_{t,t+h})^2}{(S^i_t)^2}}_{=(R^f_{t,t+h})^2} + \frac{2R^f_{t,t+h}}{(S^i_t)^2}\left(\int_0^{F^i_{t,t+h}} \text{put}_{t,t+h}(K)\ dK + \int_{F^i_{t,t+h}}^{\infty} \text{call}_{t,t+h}(K)\ dK\right)$$

and

$$A_1 = \frac{F^i_{t,t+h}}{S^i_t} = R^f_{t,t+h},$$

such that

$$\text{var}^*_t(R^i_{t,t+h}) = \frac{2R^f_{t,t+h}}{(S^i_t)^2}\left(\int_0^{F^i_{t,t+h}} \text{put}_{t,t+h}(K)\ dK + \int_{F^i_{t,t+h}}^{\infty} \text{call}_{t,t+h}(K)\ dK\right). \quad \text{(B-1)}$$

To approximate the integrals in Equation (B-1), we follow the conservative approach by Martin (2017). For convenience, we denote the price of an out-of-the-money option with strike price K_j as[34]

$$Z_j(K_j) = \begin{cases} \text{put}_j(K_j) & \text{if} \quad K_j < F_j \\ \text{call}_j(K_j) & \text{if} \quad K_j \geq F_j. \end{cases}$$

The approximation of the sum of the two integrals in Equation (B-1) is then obtained by

$$\int_0^{\infty} K^{m-2} Z(K)\ dK \approx \sum_j K_j^{m-2} Z_j(K_j) \Delta K_j,$$

where $m = 2$ for the case of risk-neutral variances, and

$$\begin{aligned} \Delta K_j &= \frac{K_{j+1} - K_{j-1}}{2} \quad j = 2, \ldots, n-1, \\ \Delta K_1 &= K_2 - K_1, \\ \Delta K_n &= K_n - K_{n-1}. \end{aligned}$$

B.3 Alternative identification strategies for beta

In addition to the risk-neutral variance-based strategy, Kempf et al. (2015) propose two alternative ways to identify beta using risk-neutral skewness and kurtosis. In doing so, they preempt a recent literature that emphasizes the importance of higher risk-neutral moments in describing the cross-section of stock returns. An example of this literature is given by Chabi-Yo et al. (2023), who account for all higher risk-neutral moments when deriving a generalized lower bound on the expected excess return of individual

[34] For notational convenience, we drop the security, time, and maturity indices.

stocks. A data-driven approach guided by a similar idea is that by Wang (2018), who connects higher risk-neutral cumulants with expected stock risk premia through latent risk-factors obtained by partial least squares.

The two alternative identification strategies proposed by Kempf et al. (2015) rely on taking the third and fourth centered moments of Equation (3.4), as a result of which they obtain a skewness-based beta

$$\beta_{t,h}^{i,*,\text{skew}} = \frac{\text{skew}_t^*(R_{t,t+h}^i/R_{t,t+h}^f)^{1/3}}{\sum_j w_t^j \cdot \text{skew}_t^*(R_{t,t+h}^j/R_{t,t+h}^f)^{1/3}} \tag{B-2}$$

and a kurtosis-based beta, respectively[35]

$$\beta_{t,h}^{i,*,\text{kurt}} = \frac{\text{kurt}_t^*(R_{t,t+h}^i/R_{t,t+h}^f)^{1/4}}{\sum_j w_t^j \cdot \text{kurt}_t^*(R_{t,t+h}^j/R_{t,t+h}^f)^{1/4}}.$$

The two alternative specifications can be viewed as drop-in replacements for the risk-neutral variance-based beta that we use to calculate the FI-CAPM. Which of the betas is to be preferred, is ultimately a matter of empirical performance. In accordance with Baule et al. (2016), we find that betas based on risk-neutral variance perform best. Skewness-based betas exhibit substantial volatility, making the fully-implied formula's predictions take on values that are beyond reasonable.[36] Kurtosis-based betas, instead, provide expected excess returns that are similar to the ones obtained when using risk-neutral variances. As a result, they contribute only little, if any, additional information. Similarly, Chabi-Yo et al. (2023) find that the contributions of skewness and kurtosis to changes in their generalized lower bound on expected stock excess returns are much weaker than that of risk-neutral variance.[37]

B.4 The positive-sign restriction

Another implication of Equation (3.10) is that it rules out scenarios in which risk-neutral betas are negative. While this assumption is arguably harsh, it is still milder than claiming that the negative correlation condition (NCC) applies to individual stocks. Kadan and Tang (2020) allow for a comparison by showing that the stock-level NCC holds up to a linear approximation if both $\text{cov}_t(R_{t,t+h}^i, 1/R_{t,t+h}^M) < 0$ and $\gamma \geq \text{var}_t(R_{t,t+h}^i)/\text{cov}_t(R_{t,t+h}^i, R_{t,t+h}^M)$, where γ is the level of relative risk aversion of the investor. Hence, for the NCC to hold, not only must the beta be non-negative, but γ must be high enough to offset the conditional variance of stock returns.

[35]Information on how to approximate risk-neutral skewness and kurtosis of simple returns using option prices can be found in Appendix B.2.

[36]The reason for this volatility is that the term $\sum_j w_t^j \cdot \text{skew}_t^*(R_{t,t+h}^j/R_{t,t+h}^f)^{1/3}$ in the denominator of Equation (B-2) is frequently close to zero, causing the betas to be highly unstable.

[37]According to Chabi-Yo et al. (2023), a one-standard-deviation shock in implied variance at a monthly investment horizon moves their bound by 70% of its standard deviation. Skewness and kurtosis, instead, contribute only with -14.4% and 1.3%, respectively.

B.5 Pairwise tests of relative portfolio performance: Methodology

To test the null hypotheses $H_0 : \mathbb{E}(R^{e,q}_{t,t+h}) \geq \mathbb{E}(R^{e,p}_{t,t+h})$ for $q < p$ and $p, q \in \{1, 2, \ldots, K\}$, with $K = 10$ being the number of prediction-sorted portfolios, we state the excess return of the q'th prediction-sorted portfolio as follows:

$$R^{e,q}_{t,t+h} = \frac{1}{\sum_i \mathbb{1}\big(\{i,t\} \in \mathbf{\Omega}_q\big)} \sum_i R^{e,i}_{t,t+h} \cdot \mathbb{1}\big(\{i,t\} \in \mathbf{\Omega}_q\big).$$

$\mathbf{\Omega}_q$ represents the set of security-date indices belonging to the q'th portfolio and $\mathbb{1}(\cdot)$ is the indicator function giving 1 if a certain $\{i,t\}$-combination belongs to the q'th portfolio and 0 else. Note that, by summing over all stocks and portfolios at a given point in time, we obtain the number of stocks in the cross-section $N_t = \sum_{i,q} \mathbb{1}(\{i,t\} \in \mathbf{\Omega}_q)$, whereas by iterating over all three dimensions we obtain the total number of observations $\sum_t N_t = \sum_{i,t,q} \mathbb{1}(\{i,t\} \in \mathbf{\Omega}_q)$.

In a next step, we collect each of the mean portfolio returns $\mathbb{E}(R^{e,q}_{t,t+h}) = \mu_q$ in a $K \times 1$ column vector $\boldsymbol{\theta} = [\mu_1 \ \mu_2 \ \ldots \ \mu_K]'$. As a result, the above hypotheses can be expressed as a system of linear equations, $H_0 : \mathbf{R}\boldsymbol{\theta} \leq \mathbf{r}$, where

$$\mathbf{R} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & -1 & 0 & 1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

is a matrix of dimensions $(K(K-1)/2) \times K$, and $\mathbf{r}$ is an appropriate vector of zeros. Accordingly, the first row of $\mathbf{R}$ represents the hypothesis that the mean excess return of the 2nd decile portfolio is lower than or equal to that of the 1st, whereas the last row represents the hypothesis that the mean excess return of the 10th decile portfolio is lower than or equal to that of the 9th.

Estimation of $\boldsymbol{\theta}$ involves computing time-series averages of the form

$$\hat{\mu}_q = \frac{1}{T} \sum_t R^{eq}_{t,t+h},$$

which in the case of a balanced panel can be achieved by regressing individual stock excess returns $R^{e,i}_{t,t+h}$ on a $K \times 1$ vector of dummy variables $\mathbf{x}_{it} = [x_{it1} \ x_{it2} \ \ldots \ x_{itK}]'$. Each element $x_{itq} = \mathbb{1}(\{i,t\} \in \mathbf{\Omega}_q)$ of this vector indicates whether a given security-date combination belongs to the q'th portfolio, as determined by a model's forecast.[38]

Given that our panel of stocks and dates is unbalanced, however, we additionally need to account for the fact that the size of the portfolios can change over time.

[38] More formally, the value of x_{itq} follows from an $\mathcal{F}_t$-measurable function Q_t that assigns stocks according to a model's predictions to the K-dimensional set of prediction-sorted portfolios, that is $Q_t : \{\hat{R}^1_{t,t+h}, \hat{R}^2_{t,t+h}, \ldots, \hat{R}^{N_t}_{t,t+h}\} \to \{1, 2, \ldots, K\}$.

Therefore, we stack both the dependent and the independent variables across stocks and time, yielding a $\sum_t N_t \times 1$ vector of individual stock excess returns $\mathbf{y}$ and a $\sum_t N_t \times K$ matrix of dummy variables $\mathbf{X}$. We then define a diagonal matrix $\mathbf{V}$ of dimension $\sum_t N_t \times \sum_t N_t$ that contains the elements $v_{itq} = (\sum_i \mathbb{1}(\{i,t\} \in \mathbf{\Omega}_q))^{-1}$ on its diagonal and 0 else. Decomposing $\mathbf{V}$ according to $\mathbf{V} = \mathbf{C}'\mathbf{C}$ such that $\mathbf{C} = \mathbf{V}^{1/2}$, we state the previous regression problem in terms of transformed variables $\tilde{\mathbf{y}} = \mathbf{C}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{C}\mathbf{X}$, and $\tilde{\boldsymbol{\varepsilon}} = \mathbf{C}\boldsymbol{\varepsilon}$, such that $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{\varepsilon}}$ represents the transformed regression residuals. With pre-determined regressors, the vector of mean excess return estimates is given by[39]

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y} = \Big(\sum_{i,t} \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\Big)^{-1} \sum_{i,t} \tilde{\mathbf{x}}_{it}\tilde{y}_{it}.$$

To derive the test statistic and its limiting distribution, we require a heteroskedasticity and autocorrelation-consistent estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$ that is robust to arbitrary forms of spatial dependence in the transformed regression residuals. The reason is that in financial predictive regressions, macroeconomic shocks typically generate dependencies between firms at given points in time, but also between different firms at different points in time, the latter due to the overlapping nature of the multi-period returns being explained. Hence, we resort to an estimator that was originally proposed by Driscoll and Kraay (1998), discussed by Petersen (2009) in a general asset pricing context, and by Thompson (2011) in the special case of predictive regressions.[40] Following the notation by Thompson (2011), the estimator and its individual components are given by

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \widehat{\text{var}}(\hat{\boldsymbol{\theta}})_{\text{time},0} + \sum_l b_l \big(\widehat{\text{var}}(\hat{\boldsymbol{\theta}})_{\text{time},l} + \widehat{\text{var}}(\hat{\boldsymbol{\theta}})'_{\text{time},l}\big)$$

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}})_{\text{time},l} = \Big(\sum_{i,t} \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\Big)^{-1} \sum_{i,j,t} \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{jt-l}' \hat{\tilde{\varepsilon}}_{it}\hat{\tilde{\varepsilon}}_{jt-l} \Big(\sum_{i,t} \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\Big)^{-1},$$

where l is the maximum lag length and b_l is the distance-decreasing Bartlett kernel.[41]

Provided that a CLT holds and $\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \widehat{\text{var}}(\hat{\boldsymbol{\theta}}))$, we obtain the test statistic under H_0 as

$$\mathbf{J} = \hat{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{R}\boldsymbol{\theta}) \overset{a}{\sim} \mathcal{N}(0, \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{S}}\hat{\mathbf{D}}^{-\frac{1}{2}}), \tag{B-3}$$

where $\hat{\mathbf{S}} = \mathbf{R}\widehat{\text{var}}(\hat{\boldsymbol{\theta}})\mathbf{R}'$ and $\hat{\mathbf{D}}$ is the diagonal matrix obtained by the diagonal elements

[39] To economize on notation, we define $\sum_i \sum_t \coloneqq \sum_{i,t}$.

[40] Alternatively, we could use the block-bootstrap procedure that is used by Martin and Wagner (2019) to estimate the variance-covariance matrix of $\boldsymbol{\theta}$. In an unreported simulation study we compare the two and find that the estimates do not differ much.

[41] For details concerning the implementation, we refer to Hoechle (2007) and Millo (2017).

of $\hat{\mathbf{S}}$. One can test the above null hypotheses globally using

$$\mathbf{J}'(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{S}}\hat{\mathbf{D}}^{-\frac{1}{2}})^{+}\mathbf{J} \xrightarrow{d} \chi^2\big(\text{rank}(\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}})\big),$$

where the superscript + indicates the Moore-Penrose inverse, or individually using a max-t type test. For the latter, we obtain p-values adjusted for multiplicity for the j'th two-sided hypothesis by $p_j = 1 - P(\max(\mathbf{J}) \leq t_j)$, where the t_j are the observed elements of the test statistic and the probability is obtained by multiple integration of the limiting normal in Equation (B-3) over the interval $(-\infty, t_j]$. The results for the max-t type test are reported in Table 3.5 for the FI-CAPM (and MW) at an investment horizon of 365 calendar days.[42]

[42] Patton and Timmermann (2010) propose a similar procedure to test the monotonicity relationship implied by Equation (3.25). The two approaches have in common that they both assume under H_0 that expected returns are identical or weakly declining. Unlike them, however, we focus on the differences between individual portfolios, for the reason that a joint test does not give any indication as to the origin of a (non-)rejection. As pointed out by Romano and Wolf (2005) and Hothorn et al. (2008), we could increase the power of our test by using a stepwise rather than a single-step procedure. We refrain from doing so, however, as this goes beyond the scope of this study.

B.6 Additional figures

Figures B.1 and B.2 present time series of average risk-neutral correlations, and time series of the BPV_t for portfolios formed on the firms' book-to-market ratios, respectively.

Figure B.1: Average risk-neutral correlations. This figure displays cross-sectionally constant risk-neutral stock-market ($\text{SVIX}_t/\overline{\text{SVIX}_t}$) and stock-stock correlations ($\text{SVIX}_t^2/\overline{\text{SVIX}_t^2}$) between January 1996 and December 2020. The panels present daily time series with investment horizons of 30, 182, 365, and 730 calendar days. In each panel, the solid red line denotes average stock-market and the dotted blue line average stock-stock correlations. The universe of stocks is confined to securities that are constituents of the S&P 500.

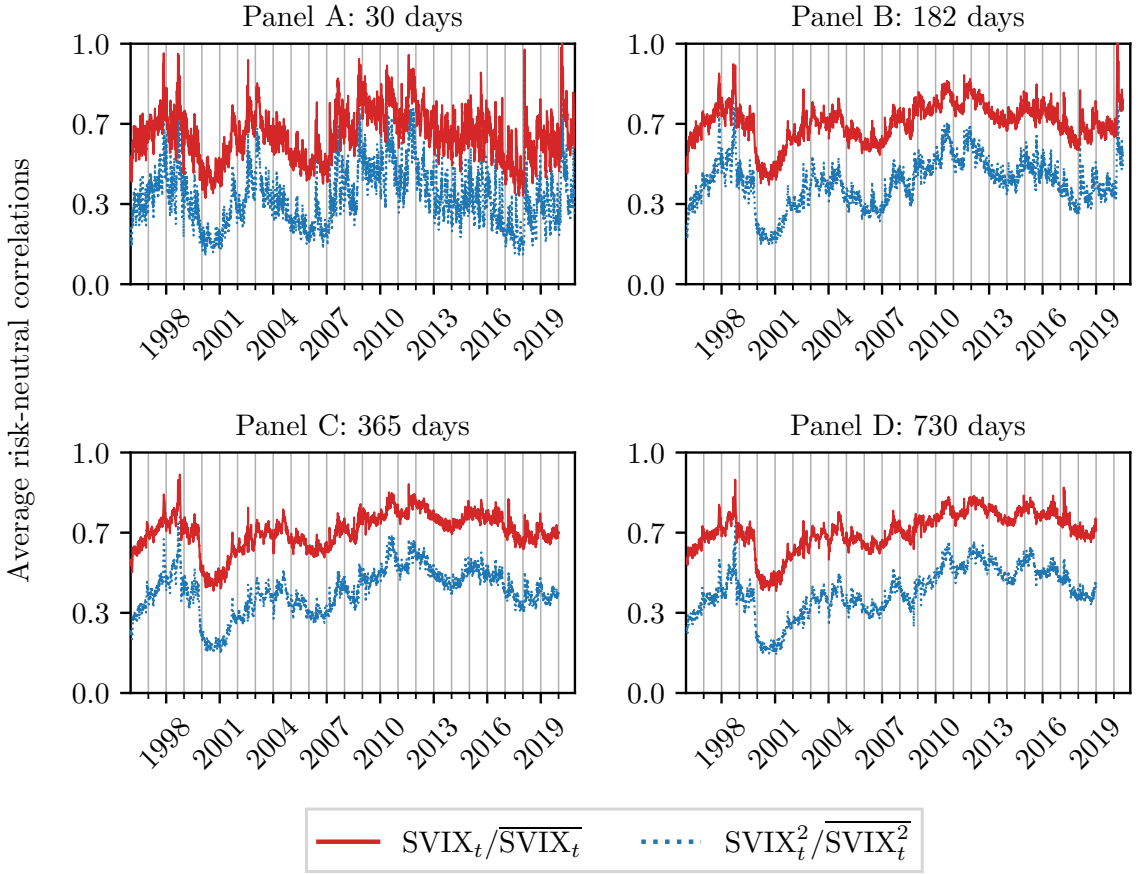
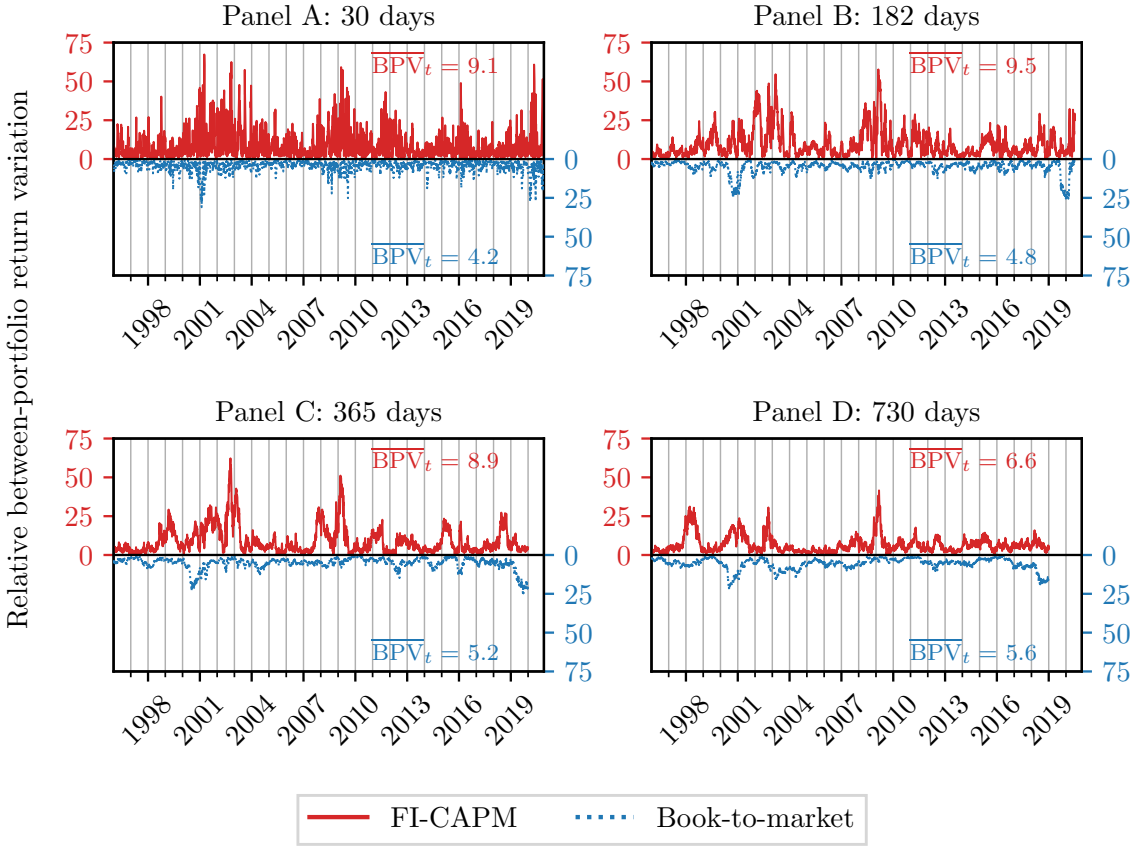


Figure B.2: Relative between-portfolio return variation (Book-to-market). This figure shows time series of relative between-portfolio return variation (BPV_t from Equation (3.32)) for 10 prediction-sorted portfolios at investment horizons of 30, 182, 365, and 730 calendar days between January 1996 and December 2020. The underlying portfolios are formed on the excess return predictions of the FI-CAPM in Equation (3.13) (solid red line) and the companies' book-to-market ratios (dotted blue line). For the sake of clarity, the time series are plotted on two different vertical axes, mirrored by the horizontal zero line. The numbers associated with $\overline{\text{BPV}}_t$ represent time-series averages of relative between-portfolio variation. The universe of stocks is confined to securities that are constituents of the S&P 500. The time series have a daily frequency.



Chapter 4

Multi-task learning in cross-sectional regressions

> "... Only if asset returns depend on *how you behave*, not *who you are* – on betas rather than characteristics – can a market equilibrium survive..." – Cochrane (2005, p. 79)

4.1 Motivation

Cross-sectional regressions have long been a popular tool in empirical finance. Originally introduced by Fama and MacBeth (1973) as an efficient means to account for cross-sectional correlation in the residuals when testing the unconditional CAPM, they are now widely used to construct factor models with time-varying loadings. Fama and French (2020), for example, show in a recent application that cross-sectional regressions of returns onto a small set of stock characteristics yield cross-sectional factors that compare favorably to the classical Fama and French (2015) time-series factors. Similarly, Kelly et al. (2019) perform cross-sectional regressions on linear combinations of characteristics to reduce the dimensionality of the characteristic space.

Although empirically successful, factor models with (functions of) characteristics as loadings raise serious theoretical concerns. Cochrane (2005), for instance, argues that a market equilibrium could hardly survive if differences in stock risk premia were truly driven by stock characteristics. Using the *Size* characteristic as an example, he illustrates that managers could earn arbitrage profits from the difference between the high average returns of small firms and the low average returns of large firms by consolidating the former into a large holding company. He concludes that the right betas (or loadings), although likely correlated with many observable characteristics, should drive out any such characteristics in cross-sectional regressions.

Inspired by this controversy, we revisit the idea of using cross-sectional regressions for model *evaluation* rather than model *construction*. Following the seminal paper by Fama and MacBeth (1973), we focus on testing the CAPM, but refer to its conditional representation. While the unconditional CAPM has been criticized for its empirical shortcomings since the early 1970s, its conditional version has recently received new impetus from studies that emphasize the importance of using time-varying specifications of beta and the market premium.[1]

[1]Examples of such studies are those by Hollstein et al. (2020) and Hasler and Martineau (2023), who arrive at a more positive assessment of the model's performance by allowing its components to

From an econometric point of view, however, testing the conditional model poses significant methodological challenges. One such challenge is to provide time-varying measurements of the model's components in a way that is consistent with financial economic theory. Common approaches, such as the use of time-series regressions to estimate beta or predictive regressions to estimate the market premium, are only loosely grounded in theory and ambiguous with respect to certain specifications, such as setting the size of the regression window or choosing the set of exogenous predictors.

Against this background, we consolidate earlier work by Kempf et al. (2015) and Martin and Wagner (2019) to derive a representation of the conditional CAPM that is fully-implied by option prices. The term *fully-implied* refers to the fact that both the beta and the market premium are measurable functions of option prices that are defined jointly and in a consistent manner across investment horizons. As such, they are directly computable from observable quantities without requiring any econometric estimation or explicit reference to the investor's information set. An interesting feature of our model is that, despite the assumed risk aversion of the representative investor, there is no need to adjust the risk-neutral moments that define beta and the market premium. That is, we establish a direct link between physical and risk-neutral return distributions.[2]

To test this specification empirically, we derive moment conditions from period-by-period cross-sectional regressions that include the market beta and other stock characteristics as regressors. As shown by Fama and French (2020), the coefficients of such regressions represent returns of zero-investment portfolios, which can be interpreted as factors with pre-specified loadings. To derive our null hypothesis, we exploit the fact that the only relevant factor in the conditional CAPM is the excess return of the market portfolio. Therefore, including additional factor-generating characteristics should not improve the market beta's description of cross-sectional return variation. As we will demonstrate, this insight allows us to derive testable restrictions on the unconditional means of the factors, which can be easily tested using a generalized method of moments approach.

One of the more challenging questions associated with this test is how to identify a small set of meaningful characteristics (or moment conditions) from a potentially large set of predictors. Characteristics are considered meaningful if they provide incremental information for the cross-section of returns over the entire sample period. Using hundreds of characteristics simultaneously is not feasible because testing the significance of the associated factors requires estimating a large covariance matrix. Fama

vary over time.

[2]Option prices describe moments under the risk-neutral measure, whereas stock risk premia are subject to the physical measure. Any approach that fails to explain the connection between physical and risk-neutral distributions thus potentially suffers from a lack in risk adjustment. Chang et al. (2012) propose a set of assumptions that are needed to estimate beta from implied moments of returns. Buss and Vilkov (2012) define a relationship between objective and risk-neutral correlation that allows them to estimate beta using both option prices and historical returns. Kempf et al. (2015) introduce a family of implied betas based on risk-neutral variance, skewness, and kurtosis, but they make no attempt to risk-adjust implied moments.

and French (2020) remain silent on this issue and simply use the five usual suspects for factor construction: Size, Value, Operating profitability, Investment, and Momentum. However, given the plethora of candidate predictors, it may be beneficial to approach this matter more systematically. In this study, we do so by leveraging insights from high-dimensional statistics for both the selection of characteristics and the ensuing problem of post-selection inference.

Starting with the selection problem, a natural approach may seem to be to apply standard ℓ_1-regularization to the pooled sample of returns and characteristics, also known as the *pooled Lasso* (PL).[3] However, as we demonstrate in simulations, this is not a viable option in a setting in which the regression coefficients, i.e., the factor realizations, vary over time. As it turns out, the pooled Lasso is unable to recover the true set of predictors even if the true data-generating process is linear and the usual irrepresentable condition for the covariates is satisfied.[4] The same is true for ℓ_1-regularization at the level of each individual cross-sectional regression (referred to as the *individual Lasso*, IL), as this can lead to an unstable selection of characteristics, which is fundamentally at odds with the goal of identifying stable factors of returns.

Our solution to the selection problem is to employ a combination of ℓ_1- and ℓ_2-regularization (or equivalently ℓ_{12}), which dates back to the work by Obozinski et al. (2010) and is known as the *multi-task Lasso* (MTL). A useful feature of the MTL is that it enforces a joint sparsity pattern for the covariates across multiple tasks, while allowing the coefficients to be task-specific. In our application, each cross-sectional regression qualifies as a task, and the idea is to leverage information from all tasks to simultaneously select a meaningful set of factor-generating characteristics. To further distinguish between *stable* and *anomalous* return predictive signals, we combine the MTL objective function with standard ℓ_1-regularization, as proposed by Jalali et al. (2010), which allows us to shrink the coefficients of the two groups separately. In this way, we contribute to a literature that examines the robustness of stock return predictive signals (e.g., McLean and Pontiff, 2016).

As the irrepresentable condition is typically hard to defend in empirical applications, the MTL may not be able to reliably identify a stable set of characteristics, despite its advantages over the PL and IL. This manifests itself in the fact that slight variations in the data can lead to considerable changes in the set of selected characteristics. Consequently, the importance of a characteristic for describing the cross-section of returns cannot simply be inferred from the fact that its MTL coefficients are nonzero. For this reason, we complement our selection strategy with the *repeated subsampling* approach that was introduced by Meinshausen and Bühlmann (2010). Their observation is that individual runs of the Lasso usually do not yield a stable set of predictors because noise variables tend to overshadow the regularization paths of the truly relevant covariates. As a solution, they propose to run the Lasso repeatedly on random

[3] Variants of the pooled Lasso are employed, for example, by Gu et al. (2020) and Freyberger et al. (2020) to approximate conditional stock risk premia.

[4] We refer to the definition of the irrepresentable condition as it is used in the paper by Zhao and Yu (2006).

subsamples of the data to obtain different sets of selected covariates. Based on these sets, they estimate selection probabilities for each of the candidate predictors, which, by theoretical arguments, constitute more reliable indicators of variable importance than standard regularization paths. However, a shortcoming of their approach is that it can suffer from multicollinearity, as the selection probabilities of correlated predictors are shared among the members of the associated clusters. As a consequence, entire groups of variables may appear unimportant simply because they reflect similar information. To address this issue, we randomly draw subsets of characteristics from correlation clusters prior to the MTL selection, which results in a fair representation of their relative importance.

As opposed to Kozak et al. (2020), we do not augment our objective function with an additional ℓ_2-penalty term, as would be considered standard in the machine learning literature for dealing with predictor redundancy. The reason is that we do not aim to identify the stochastic discount factor (SDF) under reasonably chosen economic priors, but to select a small set of factor-generating characteristics that act as competitors to the market beta in cross-sectional regressions. Using the ℓ_2-penalty for the latter purpose can be unfavorable in terms of factor stability because highly correlated predictors tend to be selected jointly, leading to nearly rank deficient design matrices in post-selection regressions.

For the present application, the *repeated subsampling* approach is interesting not only because it allows us to assess which characteristics have explanatory power beyond the market beta, but also because it provides the basis for valid post-selection inference. The term *post-selection inference* typically refers to the problem that standard econometric methods for quantifying estimation uncertainty fail when researchers pre-select variables using procedures that rely on in-sample information. This is relevant in empirical finance because data are abundant and a pre-screening of variables is often necessary to separate the signal of interest from noise. Solutions to this problem have been proposed by Feng et al. (2020) and Harvey and Liu (2021), who examine a large number of observable factors to explain differences in expected returns. Conceptually, the former rely on the *double-selection* paradigm by Belloni et al. (2014), which has proven useful for estimating treatment effects in the presence of high-dimensional confounders. The latter, instead, propose a forward stepwise procedure that accounts for the *multiple-testing* problem associated with the selection of factors. While either of these approaches achieves uniform validity under certain conditions, we argue that a third approach, which has become popular for its simplicity and general applicability, is more favorable in the present application – the use of *sample splitting* for post-selection inference.[5]

The simple but powerful idea underlying this approach is that one can achieve uniformly valid inference by dividing the data into two halves, using the first part (the auxiliary set) for model selection and the second part (the main set) for statistical

[5] A useful comparison of the different perspectives on post-selection inference is given by Kuchibhotla et al. (2022).

inference. According to Rinaldo et al. (2019), this applies to situations in which the targets of inference are the parameters of a statistical model that approximates certain aspects of the true data-generating process. In the present application, these target parameters correspond to the means of the cross-sectional factors that are optimized with respect to the selected characteristics. As these parameters exist independently of the ground truth and are defined conditional on the set of selected covariates, this approach is valid even in the presence of model misspecification, including variable selection mistakes.

In contrast, the double-selection framework used by Feng et al. (2020), is designed for situations where inference is drawn on a small subset of the true parameter space, which is typically the treatment effect in the causal inference literature (cf. Chernozhukov et al., 2018). The remainder of the true model is considered irrelevant for answering the research question and therefore treated as a nuisance function that can be approximated using machine learning techniques. Although in principle this allows for approximation errors, the estimation may still be sensitive to misspecifications of the assumed functional form. This can be problematic if the true data-generating process is globally – rather than partially – non-linear, not all relevant information is observable, or the components of the true model are entangled by multicollinearity, so that parameter identification is virtually impossible.[6]

Which perspective is the right one for the present application? According to Berk et al. (2013), the approximation perspective is preferable to the true-model perspective in situations where model uncertainty is high and predictor redundancy is an issue. In asset pricing, this is typically the case because the exact functional form of the return generating process (or the SDF) is unknown and the data available for its estimation are highly correlated. For this reason, many researchers resort to arbitrage pricing theory (APT), which derives its appeal from the fact that assumptions regarding functional forms can be replaced by a purely statistically motivated decomposition of returns. The resulting factor models, however, should not be taken as realistic depictions of the return-generating process, but rather as useful approximations to the actual macroeconomic forces driving asset prices – a distinction that Harvey and Liu (2021) make explicit by referring to the significant factors in their study as *useful* rather than *true*. Given these caveats, we consider the true-model perspective inappropriate for the construction and evaluation of factor models, even if some of the characteristics used in this process are grounded in rational pricing theories. This view is shared by Jagannathan and Wang (1998), who note that if cross-sectional regressions are used for factor identification, misspecification of the assumed beta-model can have tremendous effects on statistical inference. At the same time, they advocate adding characteristics to cross-sectional regressions to detect model misspecification, which we adopt as a guiding principle for our study.

An interesting alternative to sample splitting is the approach by Harvey and Liu (2021), which is often referred to as *simultaneous inference* because it allows multiple

[6]For a detailed exposition of the above arguments, please refer to the study by Berk et al. (2013).

model specifications to be examined jointly. Just like the sample-splitting approach, simultaneous inference achieves uniform validity regardless of the procedure that is used for model selection, and it also allows for model misspecification. A disadvantage, however, is that the derived inference can be conservative because all possible model specifications need to be taken into account, which is criticized by Jensen et al. (2021). Moreover, the set of selections must be specified in advance, which prohibits the subsequent inclusion of characteristics.

Finally, we reconcile the idea of using multiple subsamples of the data with the single-split inference approach by Rinaldo et al. (2019) by resorting to what Chernozhukov et al. (2023) refer to as *quantile-aggregated inference* (QAI). The central idea of QAI is to split the data not just once, but several times, to obtain multiple auxiliary and main sets over which inferences can be aggregated. In this way, they aim to account not only for the usual estimation uncertainty, but also for the uncertainty induced by sample splitting, which in our case is important because different partitions of the data can lead to different selections of characteristics and thus to different cross-sectional factors. To arrive at a uniform statement about the null hypothesis of interest, Chernozhukov et al. (2023) suggest to aggregate the individual p-values across subsamples by their median and to adjust the nominal significance level accordingly.

For the empirical implementation of our testing procedure, we use a collection of 7,665 stocks traded on major US exchanges during the period from January 1996 to December 2021. For each of these stocks, we compute market betas from the prices of European call and put options using the volatility surface by OptionMetrics, extract a selection of 78 characteristics from Chen and Zimmermann's (2022) *Open Source Asset Pricing* repository, and obtain monthly gross returns from CRSP with a holding period of 30 calendar days, resulting in a panel of 667,113 security-date observations. After compiling the data, we proceed as follows: First, we randomly divide the set of stocks into two halves based on their permanent security identifier and then assign their entire data histories to either the auxiliary or the main set. In this way, we ensure that the same stock does not appear in both sets, which would violate the notion of independent sampling. Second, as many stock characteristics contain missing values, we apply the imputation procedure by Bryzgalova et al. (2022), but account for the induced uncertainty by performing it separately for each partition of the data. Third, we use the auxiliary set to cluster characteristics by their pairwise correlation coefficients and randomly draw a single constituent from each cluster. Based on this subset, we then select meaningful characteristics using the MTL in combination with 3-fold cross-validation. Fourth, we resort to the main set to construct factors using period-by-period cross-sectional regressions of gross returns onto both the selected characteristics and the implied market beta. Fifth, we test whether the means of the resulting characteristic-based factors are jointly zero, which would be the case if the market portfolio were the only driving factor of return variation. Finally, we account for the uncertainty induced by sample splitting by creating multiple auxiliary and main partitions of the data, repeating the above procedure, and aggregating the individual

p-values according to the QAI approach.

We employ this procedure to address the following research questions:

1.) How does the inclusion of implied beta in cross-sectional regressions affect the other 78 stock characteristics? Do they retain their importance or are they driven out as the conditional CAPM suggests? To answer this question, we run the above selection procedure twice, once using the original 78 stock characteristics without taking market beta into account, and once using the same characteristics but orthogonalized with respect to beta. If the conditional CAPM holds, the latter should result in a considerable reduction in the characteristics' selection probabilities.

2.) Are the means of the cross-sectional factors jointly significantly different from zero, so that the conditional CAPM is rejected? Previous results by Lewellen and Nagel (2006) suggest that the conditional CAPM is unable to explain well-known asset-pricing anomalies, such as momentum. Hasler and Martineau (2023), however, challenge this conclusion, arguing that the conditional CAPM successfully explains the returns of various characteristic-sorted portfolios as well as those of individual stocks.

3.) What is the effect of using individual stocks as test assets instead of characteristic-sorted portfolios? Freyberger et al. (2020) illustrate that using individual stocks is in principle sufficient because regressions of returns onto rank-transformed characteristics are, up to some approximation error, equivalent to constructing conditional portfolio sorts. However, as using portfolios remains popular, we additionally assign stocks at each point in time to 500 beta-sorted portfolios as well as to $25 \times 78 = 1{,}950$ univariate characteristic-sorted portfolios, and repeat the above procedure to see if our conclusions change with respect to the chosen set of test assets.

4.) Harvey and Liu (2021) show in their study that the market factor is by far the strongest predictor of return variation among a large set of observable factors. Does this result carry over to the implied market beta if used alongside other stock characteristics in cross-sectional regressions? And if so, how does the market beta's performance compare to that of the other characteristics? To investigate this, we decompose the amount of variation explained using the concept of Shapley values (cf. Shapley, 1951). This decomposition allows us to assess the relative contribution of each characteristic individually, while taking into account that many of them are highly correlated.

5.) Are the selected characteristics stable predictors of returns or are they subject to temporal instability? As mentioned above, we generate insights into the prevalence of anomalous return predictability by adding a regression-wise ℓ_1-penalty term to the MTL objective function, which allows us to separately shrink the coefficients of stable and anomalous factor-generating characteristics.

Our findings are as follows: If we do not account for the market beta, many characteristics are useful to explain the cross-section of returns. This observation is at odds with the results by Freyberger et al. (2020), who use a variant of ℓ_1-regularization known as the *group Lasso* to select non-linear expansions of characteristics, finding that only few of them are informative. This discrepancy can be explained by the fact that, in addition to imposing approximate sparsity in the selection step, we generate variation at each iteration of the selection procedure by randomly subsampling characteristics from disjoint correlation clusters, which accounts for the instability of ℓ_1-regularization in the presence of highly correlated predictors. Empirically, this leads to both a minimization of the information overlap among candidate predictors and a much larger set of characteristics with high selection probabilities. Based on these findings, it appears that returns are not so much sparse in the set of factor-generating characteristics, but rather in the latent information driving asset prices, as reflected in the small number of correlation clusters in the data. This conclusion is in line with the results by Kozak et al. (2020), who compare a characteristics-sparse specification of the SDF with an ℓ_2-penalized alternative, finding that the latter is better at summarizing the cross-section of returns.

Once we account for the market beta in the MTL-based selection step, we observe a substantial reduction in the selection probabilities of several characteristics that were previously identified as important. That is, the conditional CAPM succeeds in driving out many of the stock characteristics considered. Nevertheless, a small fraction of the predictors, mainly related to the momentum effect, remain important even after accounting for beta, which translates into a rejection of the model when using individual stocks as test assets. The results are less conclusive, however, in the case of portfolios: While the rejection is borderline in the case of characteristic-sorted portfolios, we cannot reject the conditional CAPM at any conventional level for beta-sorted portfolios. We conclude from these results that the returns of individual stocks pose a greater challenge to our model than the returns of portfolios.

Should we thus cease to consider market risk in cross-sectional regressions and instead focus exclusively on other stock characteristics, as is done by Fama and French (2020) and Kozak and Nagel (2022)? – The answer is no. In fact, the implied beta accounts for most of the cross-sectional variation in returns, leaving all other characteristics, including variants of the momentum effect, far behind. This is evident when comparing the Shapley contributions of each characteristic to the explained variation with that of the market beta. In light of this evidence, and given the fact that the proposed implied beta is directly observable at any point in time, we strongly encourage its use for testing the incremental contribution of newly proposed predictors in cross-sectional regressions.

Finally, we find that most of the informative characteristics are stable predictors of return variation. This is reflected in the fact that the proportion of characteristics associated with the part of the objective function that imposes a joint sparsity pattern across tasks is much higher than the proportion of characteristics associated with the

time-specific ℓ_1-penalty term. In agreement with the study by Kelly et al. (2019), we conclude that there is not much anomalous return predictability left after accounting for stable factor exposure.

This study is related to a recent and growing literature that uses information from option prices to approximate conditional risk premia. In a seminal work, Martin (2017) derives a lower bound for the conditional expected excess return of the market that is fully-implied by option prices. Kadan and Tang (2020) investigate the extent to which this bound is applicable to individual stocks. Martin and Wagner (2019) propose a formula for the conditional stock risk premium that is a linear function of risk-neutral stock and market return variances. Schneider and Trojani (2019) provide an extensive family of observable bounds for higher moments of index returns. Bakshi et al. (2020) and Chabi-Yo and Loudis (2020) propose formulas for the expected return of the market that depend on all higher risk-neutral moments of returns. Similarly, Chabi-Yo et al. (2023) consider such bounds for individual stocks.

Moreover, we contribute to a strand of research that uses machine learning and high-dimensional statistics for the construction and evaluation of asset pricing models. Gu et al. (2020) examine the performance of a suite of machine learning models for approximating conditional stock risk premia. Kozak et al. (2020) use penalized regressions and economically motivated priors to identify the SDF. Freyberger et al. (2020) select groups of nonlinear basis functions of characteristics using a form of block-norm regularization. Kelly et al. (2019) and Gu et al. (2021) use instrumented principal components analysis and autoencoder neural networks to determine functions of stock characteristic that serve as time-varying loadings of latent factors. Giglio and Xiu (2021) consider estimating risk premia of observable factors in the presence of omitted variables using a three-pass method that exploits principal components of returns. Bryzgalova et al. (2021) and Chen et al. (2023) use decision trees and adversarial learning, respectively, to construct optimal sets of test assets. Feng et al. (2020) and Harvey and Liu (2021) use double machine learning and simultaneous inference to examine the extent to which pre-specified factors provide independent information for the cross-section of returns.

The remainder of this chapter is structured as follows: Section 4.2 introduces our model. Section 4.3 derives testable restrictions from cross-sectional regressions. Section 4.4 introduces the multi-task learning paradigm. Section 4.5 discusses methods to assess the characteristics' relative importance. Section 4.6 presents the QAI approach for post-selection inference. Section 4.7 presents the data and our empirical results, and Section 4.8 concludes. Appendix C provides further analyses.

4.2 A fully-implied representation of the conditional CAPM

The conditional CAPM consists of two components that jointly determine stock risk premia: 1) the beta, which represents the quantity of risk associated with an individual

asset, and 2) the market premium, which is commonly referred to as the market price of risk. Formally, this corresponds to the following relationship

$$\mathbb{E}_t(R^i_{t+1} - R^f_{t+1}) = \beta^i_t \cdot \mathbb{E}_t(R^M_{t+1} - R^f_{t+1}), \tag{4.1}$$

where R^i_{t+1} is the gross return of an asset i, β^i_t is the asset's exposure to market risk, R^f_{t+1} is the risk-free rate, and R^M_{t+1} is the gross return of the market portfolio.

An important property of Equation (4.1) is that each of the terms involved carries a time index – that is, they are specified in terms of conditional return distributions. From an econometric point of view, this poses significant problems because standard econometric methods are not particularly well suited to capture time-varying moments. For this reason, estimating the conditional CAPM is often simplistically viewed as a two-step problem: In a first step, one obtains an estimate of the conditional beta using rolling time-series regressions,[7] and in a second step, one approximates the market premium by predictive regressions of market excess returns onto predetermined variables, which can be asset-specific characteristics or macroeconomic variables.[8] Neither of these steps, however, draws substantially on financial economic theory, which is why tests of the conditional CAPM can always be challenged based on the chosen empirical implementation.

We take a different route in this study. In Chapter 3, we have shown that under certain conditions both the market beta and its risk premium arise jointly as functions of risk-neutral return variances, which, following the logic of Breeden and Litzenberger (1978), can be replicated using a panel of option prices.[9] More precisely, the conditional beta is given by

$$\beta^i_t = \frac{\text{var}^*_t\big(R^i_{t+1}\big)^{1/2}}{\sum_j w^j_t \cdot \text{var}^*_t\big(R^j_{t+1}\big)^{1/2}}, \tag{4.2}$$

where w^j_t denotes the market capitalization weight of asset j, and the market premium is given by

$$\mathbb{E}_t(R^M_{t+1} - R^f_{t+1}) = \frac{1}{R^f_{t+1}} \text{var}^*_t\big(R^M_{t+1}\big). \tag{4.3}$$

The asterisks in Equations (4.2) and (4.3) indicate that the variances refer to the risk-neutral distribution of returns. Note, however, that both the beta and the market

[7] Ferson and Harvey (1991), Fama and French (2004), and Frazzini and Pedersen (2014), for example, use betas obtained from rolling-window regressions as instruments for conditional betas.

[8] Frequently used macroeconomic variables are the dividend-price ratio (e.g., Campbell and Shiller, 1988), interest rate spreads (e.g., Stock and Watson, 1989), the consumption-wealth ratio (e.g., Lettau and Ludvigson, 2001), and the variance-risk premium (e.g., Bollerslev et al., 2009). Kelly and Pruitt (2013), instead, use disaggregated valuation ratios.

[9] We refrain from a detailed exposition of the underlying assumptions and mathematical derivations at this point, as they have been discussed previously in Chapter 3. For information on how to replicate risk-neutral moments using a panel of option prices, see Chapter 3, Appendix B.2.

premium on the left are given in terms of physical distributions, meaning that the corresponding implied quantities represent direct measurements of the physical moments of interest – no further risk adjustment is needed. This is in contrast to the related study by Kempf et al. (2015), who arrive at numerically equivalent measurements of beta, but do not explain the associated change of measure.

4.3 Testable restrictions from cross-sectional regressions

The following section describes how cross-sectional regressions can be used to establish testable restrictions for the conditional CAPM. We begin with providing an overview of the work by Fama and French (2020), who consider cross-sectional regressions as a means to construct factor models with time-varying loadings. The factors in their approach correspond to the projection coefficients obtained when regressing stock returns onto a set of predetermined characteristics. The characteristics, in turn, represent the loadings with respect to which the factors are optimized. Building on this interpretation of cross-sectional regressions, we show that if the conditional CAPM holds and we include the market beta as an explanatory variable, the factor means associated with any other characteristics must be zero.

For each cross-sectional regression, we consider N_t left-hand-side test assets that can be individual stocks or portfolios, and K explanatory characteristics, which, for the time being, are treated as given. Among the explanatory characteristics, there is also a constant regressor. The regression equation that applies to each of the T cross-sections is composed of a vector of gross returns, $\boldsymbol{R}_{t+1} = [R^i_{t+1}]_{N_t \times 1}$, a matrix of stock characteristics, $\boldsymbol{C}_t = [c^i_{k,t}]_{N_t \times K}$, a vector of coefficients, $\boldsymbol{f}_{t+1} = [f^k_{t+1}]_{K \times 1}$, and a vector of regression residuals $\boldsymbol{\varepsilon}_{t+1} = [\varepsilon^i_{t+1}]_{N_t \times 1}$

$$\boldsymbol{R}_{t+1} = \boldsymbol{C}_t \boldsymbol{f}_{t+1} + \boldsymbol{\varepsilon}_{t+1} \qquad t = 1, \dots, T. \tag{4.4}$$

To give meaning to this equation, we follow Fama and French (2020) and impose the usual least squares orthogonality conditions, providing us with a closed-form solution for the vector of projection coefficients

$$\begin{aligned} \boldsymbol{f}_{t+1} &= \boldsymbol{W}_t' \boldsymbol{R}_{t+1}, \\ \text{where} \quad \boldsymbol{W}_t' &= (\boldsymbol{C}_t' \boldsymbol{C}_t)^{-1} \boldsymbol{C}_t'. \end{aligned} \tag{4.5}$$

The notation indicates that each of the coefficients in Equation (4.5) represents the return of a zero-investment portfolio comprising the left-hand side test assets from Equation (4.4) whose weights are contained in the columns of $\boldsymbol{W}_t$. Thus, when stacking the coefficients over t, they can be viewed as factor time series that are optimized with respect to the pre-specified loadings $\boldsymbol{C}_t$.

An interesting property of the factors in Equation (4.5) is that each reflects the

contribution of the associated characteristic in isolation (cf. Fama, 1976). That is, the k'th column of the matrix $\boldsymbol{W}_t$, denoted by $\boldsymbol{w}_{k,t}$, is chosen such that the portfolio's value of the characteristic is equal to one for the k'th characteristic and zero for all other characteristics. Formally, this corresponds to

$$\boldsymbol{w}_{k,t}'\boldsymbol{c}_{l,t} = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise,} \end{cases} \tag{4.6}$$

where the $\boldsymbol{c}_{l,t}$ denotes the l'th column of $\boldsymbol{C}_t$. This property is useful because it has certain implications for the factors that are not accounted for by our theory, i.e., the factors that are generated by characteristics other than the market beta. To illustrate this, we include the market beta as an explanatory variable in Equation (4.4) and isolate both the constant regressor and the market beta from the other characteristics

$$\boldsymbol{R}_{t+1} = \boldsymbol{\iota} f_{t+1}^1 + \boldsymbol{\beta}_t f_{t+1}^2 + \boldsymbol{C}_t \boldsymbol{f}_{t+1} + \boldsymbol{\varepsilon}_{t+1} \qquad t = 1, \dots, T. \tag{4.7}$$

The $\boldsymbol{\iota} = [1]_{N_t \times 1}$ refers to the vector of ones that was previously included in the characteristics matrix, $\boldsymbol{\beta}_t = [\beta_t^i]_{N_t \times 1}$ denotes the vector of market betas, and $\boldsymbol{C}_t$ is the matrix of stock characteristics, now with reduced dimensions $N_t \times (K-2)$, so that in total we still have K regressors. Moreover, for the definition of the factors from Equation (4.5) to remain valid, we need to express the matrix of portfolio weights in terms of the extended collection of regressors $\boldsymbol{X}_t = [\boldsymbol{\iota}, \boldsymbol{\beta}_t, \boldsymbol{C}_t]$, so that $\boldsymbol{W}_t' = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'$.

Recall that, according to the conditional CAPM in Equation (4.1), an asset's risk premium is driven exclusively by its exposure to market risk. Using R_{t+1}^{eM} as a shorthand for the excess return of the market portfolio, we can express this implication equivalently using vector notation

$$\mathbb{E}_t(\boldsymbol{R}_{t+1} - \boldsymbol{\iota} R_{t+1}^f) = \boldsymbol{\beta}_t \mathbb{E}_t(R_{t+1}^{eM}). \tag{4.8}$$

Because the factors in Equation (4.7) are themselves portfolio returns, the condition in Equation (4.8) must also apply to their risk premia. Hence, the factors' conditional expectations can be written as functions of the associated portfolio betas

$$\mathbb{E}_t(f_{t+1}^k) = \begin{cases} \boldsymbol{w}_{k,t}'\boldsymbol{\iota} R_{t+1}^f + \boldsymbol{w}_{k,t}'\boldsymbol{\beta}_t \mathbb{E}_t(R_{t+1}^{eM}) & \text{if } k = 1 \\ \boldsymbol{w}_{k,t}'\boldsymbol{\beta}_t \mathbb{E}_t(R_{t+1}^{eM}) & \text{otherwise,} \end{cases}$$

where the distinction between the first and the remaining factors stems from the fact that the projection coefficient associated with the constant regressor is a return, while the others are excess returns. Utilizing the property of the coefficients in Equation

(4.6), we obtain the following sets of restrictions on the factor risk premia

$$R1 \begin{cases} \mathbb{E}_t(f^1_{t+1}) = \underbrace{\boldsymbol{w}'_{1,t}\boldsymbol{\iota}}_{=1} R^f_{t+1} + \underbrace{\boldsymbol{w}'_{1,t}\boldsymbol{\beta}_t}_{=0} \mathbb{E}_t(R^{eM}_{t+1}) = R^f_{t+1} \\ \mathbb{E}_t(f^2_{t+1}) = \underbrace{\boldsymbol{w}'_{2,t}\boldsymbol{\beta}_t}_{=1} \mathbb{E}_t(R^{eM}_{t+1}) = \mathbb{E}_t(R^{eM}_{t+1}) \end{cases} \tag{4.9}$$

$$R2 \begin{cases} \mathbb{E}_t(f^3_{t+1}) = \underbrace{\boldsymbol{w}'_{3,t}\boldsymbol{\beta}_t}_{=0} \mathbb{E}_t(R^{eM}_{t+1}) = 0 \\ \qquad\vdots \qquad\qquad\qquad \vdots \\ \mathbb{E}_t(f^K_{t+1}) = \underbrace{\boldsymbol{w}'_{K,t}\boldsymbol{\beta}_t}_{=0} \mathbb{E}_t(R^{eM}_{t+1}) = 0. \end{cases} \tag{4.10}$$

These restrictions embody the essence of the conditional CAPM: The only risk factor for which investors demand compensation is the excess return of the market portfolio (*R1*), hence the risk premia of the other characteristic-based factors must be zero (*R2*).

Note, however, that we can neither directly observe nor readily estimate the involved factor risk premia, as they are defined in terms of conditional return distributions. Therefore, we apply the law of total expectations to either side of the constraints in (4.9) and (4.10), allowing us to express the above restrictions in terms of unconditional expectations

$$\underbrace{\mathbb{E}(f^1_{t+1} - R^f_{t+1}) = \mathbb{E}(f^2_{t+1} - R^{eM}_{t+1})}_{R1} = \underbrace{\mathbb{E}(f^3_{t+1}) = \cdots = \mathbb{E}(f^K_{t+1})}_{R2} = 0.$$

Moreover, we focus exclusively on testing the conditions in *R2*, because those in *R1* are only of interest if indeed the restrictions in *R2* are satisfied, as pointed out by Fama (1976, p.329 ff.).

To formally test the conditions in *R2*, we write the associated null hypothesis in terms of the target parameters $\theta_k = \mathbb{E}(f^k_{t+1})$

$$H_0 : \theta_3 = \cdots = \theta_K = 0, \tag{4.11}$$

and set up an appropriate Wald statistic that is χ^2-distributed under the conditions of the H_0, i.e.,

$$W = \hat{\boldsymbol{\theta}}' \widehat{\text{var}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\theta}} \underset{d}{\to} \chi^2(K-2). \tag{4.12}$$

The $\hat{\boldsymbol{\theta}} = [\hat{\theta}_k]_{(K-2)\times 1}$ are obtained by time-series regressions of the factors on a constant and the estimator of the covariance matrix is given by

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \frac{1}{T^2} \sum_t \hat{\boldsymbol{u}}_{t+1} \hat{\boldsymbol{u}}'_{t+1},$$

where the $\hat{\boldsymbol{u}}_{t+1} = [f^k_{t+1} - \hat{\theta}_k]_{(K-2)\times 1}$ denote the corresponding regression residuals.

A few remarks are in order at this point: First, we do not consider the regressions in Equation (4.7) as a method for estimating population parameters, as is common in econometrics. Rather, we employ them as a tool to construct factor portfolios with desirable properties (cf. Fama, 1976, p.326 ff.). This distinction is crucial, because the targets of inference in our application are not the projection coefficients $\boldsymbol{f}_{t+1}$ but their unconditional means $\boldsymbol{\theta}$. The reason we rely on OLS algebra for factor construction, is that in this way the risk premia of the characteristic-based factors are zero if the conditional CAPM holds. In other words, the purpose of the above cross-sectional regressions is to provide a statistical decomposition of returns that encompasses the hypothesized model, giving us the opportunity to gauge the limitations of our theory.

Second, we do not need to adjust the covariance matrix estimator $\widehat{\text{var}}(\hat{\boldsymbol{\theta}})$ for first-stage estimation errors, as suggested by Shanken (1992). This is because we refer to the definition of the market beta from Equation (4.2), which allows us to obtain direct measurements of beta using a panel of option prices. If the beta were instead estimated via time-series regressions prior to being employed as an explanatory variable, such corrections would be necessary to account for the additional estimation uncertainty.

Third, the above representation gives no indication as to whether we should use individual stocks or portfolios as test assets. Using portfolios is a common way to mitigate the errors-in-variables problem that arises if the betas are measured with error. The beta in Equation (4.2) is subject to measurement error because we are constrained to a discrete grid of strike prices to approximate the underlying risk-neutral variances. Consequently, the associated regression coefficients (or factors) are drawn toward zero, leading to the usual attenuation bias (cf. Kim, 1995). However, one drawback of working with portfolios is that our conclusions regarding the CAPM's performance may heavily depend on the chosen set of portfolios (cf. Lewellen et al., 2010). In fact, this concern has recently sparked increased efforts to return to testing beta models with individual stocks, as evidenced by the contributions of Gagliardini et al. (2016) and Chordia et al. (2019). A persuasive argument that is put forth by Freyberger et al. (2020), is that running cross-sectional regressions with individual stocks on rank-transformed characteristics is, in principle, equivalent to examining a continuum of univariate characteristic-sorted portfolios. The idea is that the factors represent the marginal effects by which a one percent increase in the quantile of the characteristic's cross-sectional distribution affects the expected return. The crucial difference from classical portfolio sorting is that the regression approach can additionally account for the effects of other characteristics. That is, if both the *Size* and *Value* characteristics are included as explanatory variables, the *Size* effect is adjusted for the contribution of *Value* according to the conditions in Equation (4.6). Double- or triple-sorted portfolios can only mimic this behavior to a limited extent, as the number of portfolios drastically increases with the number of characteristics being considered. Given these pros and cons, we focus on individual stocks as test assets, but also report the results for a selection of portfolios to assess the robustness of our findings.

Finally, Fama and French (2020) treat the set of characteristics in their cross-sectional regressions as given. That is, they consider only a small set of handpicked characteristics that align with their preconceived notions about investor preferences, without referring to any principled selection mechanism. Given the abundance of candidate characteristics, a more efficient strategy is to employ statistical methods that are specifically designed to deal with the variable selection problem in high-dimensional settings. To meet the requirements posed by the regressions in Equation (4.4), however, we need to answer the following two questions: 1) how can we achieve an optimal selection of characteristics in a setting in which the set of explanatory variables remains constant over time, but the coefficients are time-dependent? And 2) how can we account for the fact that the classical approach to quantifying estimation uncertainty breaks down if we pre-select characteristics systematically? The solutions we propose in Sections 4.4 and 4.6 are quite universal and can be used not only to test the conditional CAPM, but also to construct and evaluate characteristic-based factor portfolios along the lines of Fama and French (2020).

4.4 Selecting characteristics using the multi-task Lasso

The variable selection problem posed by the regressions in Equation (4.4) and (4.7) is peculiar and different from the one encountered with individual cross-sections. The challenge is to identify a set of characteristics that are useful to explain the observed differences in returns over the entire sample period, while taking into account that the relationship between the returns and characteristics can change over time, as indicated by the time-indexed factors $\boldsymbol{f}_{t+1}$. We tackle this challenge by treating the selection of characteristics as a multi-task learning problem, where each cross-sectional regression represents a task and the goal is to leverage information from all tasks to establish a common representation. Conceptually, we draw on the approach by Obozinski et al. (2010), who introduce a block-norm penalty on the regression coefficients that enforces a joint sparsity pattern across tasks, but allows the coefficients to be task-specific. This approach is commonly known as the multi-task Lasso (or MTL) as it extends the idea of using an ℓ_1-penalty for the selection of covariates to the context of multiple related regression tasks. Formally, the selection across tasks is achieved by setting a constraint on the magnitude of the coefficients, which is expressed in the following optimization problem:

$$\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{C}_t \boldsymbol{f}_{t+1}||_2^2 + \underbrace{\lambda ||\boldsymbol{F}||_{1,2}}_{\ell_{12}\text{-penalty}} \tag{4.13}$$

$$||\boldsymbol{F}||_{1,2} = \sum_{k=1}^{K} ||\boldsymbol{f}^k||_2,$$

where the matrix $\boldsymbol{F} = [f_{t+1}^k]_{T\times K}$ is obtained by stacking the transposed vector of coefficients $\boldsymbol{f}_{t+1}$ over t, the vector $\boldsymbol{f}^k = [f_{t+1}^k]_{T\times 1}$ denotes the k'th column of $\boldsymbol{F}$, $||\cdot||_p$ is the ℓ_p-vector norm, and λ accounts for tightness of the constraint.[10] The ℓ_{12}-penalty added to the sum of the individual least-squares objective functions can be conceived of as an ℓ_1-penalty on the ℓ_2-norms of the covariate-specific coefficient vectors. Intuitively speaking, this penalty ensures that a characteristic is either selected across all tasks or disregarded altogether, where in the latter case the entire coefficient time series is set to zero. For a given value of λ, selecting characteristics by the MTL thus corresponds to creating a set of indices $\hat{\mathrm{S}} = \{k \in \{1,\dots,K\} : \hat{\boldsymbol{f}}^k \neq \boldsymbol{0}\}$, where the associated characteristics can be used as explanatory variables in the regressions of Equations (4.4) or (4.7).[11]

It is worth noting that Freyberger et al. (2020) use a similar block-norm regularization approach to approximate conditional risk premia within a pooled regression framework. The focus of their work is on introducing nonlinearities in the approximation of conditional risk premia by means of nonparametric splines. That is, they associate each characteristic with a collection of basis functions for which they jointly shrink the coefficients towards zero via the *group Lasso* procedure by Yuan and Lin (2006). The crucial difference between the group Lasso and the multi-task Lasso is that the former penalizes the collection of coefficients associated with the basis functions, whereas the latter acts on the characteristics' coefficient time series. For the present application, the group Lasso is not particularly well suited as it does not simultaneously account for both time-varying coefficients and a robust selection of characteristics. Either it is used period-by-period, which introduces time-varying coefficients at the expense of an unstable set of selected characteristics, or it is used in a pooled fashion, so that the resulting conditional mean approximation can only vary through changes in the characteristics' values, but not through changes in the coefficients. The latter is important though, as the coefficients in our framework represent (excess) returns of factor portfolios that are supposed to capture time-varying risks.

In addition to the above formulation of the MTL, we consider adding a regression-wise ℓ_1-penalty term to the objective function in Equation (4.13), which allows us to distinguish between *stable* and *anomalous* return predictive signals. The motivation for this originates from the observation by McLean and Pontiff (2016) that some of the characteristics proposed in the past are only important within certain periods of time, so that considering them in subsets of the regression tasks would lead to an increase in predictive accuracy. In empirical finance, such characteristics are typically referred to as *anomalies* as they do not generate persistent risk premia and are thus hard to sell within rational pricing theories. To account for the existence of anomalous return predictability, we modify the objective function in Equation (4.13) in the spirit of

[10] To keep the notation simple, we do not include a constant regressor in the MTL objective function, but assume that the characteristics in $\boldsymbol{C}_t$ are scaled appropriately. Thus, in contrast to previous sections, both the number of characteristics and the number of regressors are given by K.

[11] The optimization problem in Equation (4.13) must be solved numerically. For this purpose, we rely on the proximal gradient descent procedure by Liu et al. (2009).

Jalali et al. (2010). That is, we decompose the task-specific vector of coefficients into a component $\boldsymbol{f}_{t+1}$ that, as before, captures stable but time-varying factors of return variation, and an anomalous component $\boldsymbol{a}_{t+1}$ that exhibits a task-specific sparsity pattern

$$\hat{\boldsymbol{F}}, \hat{\boldsymbol{A}} = \underset{\boldsymbol{F},\boldsymbol{A}}{\arg\min} \sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{C}_t(\boldsymbol{f}_{t+1} + \boldsymbol{a}_{t+1})||_2^2 + \lambda_1 ||\boldsymbol{F}||_{1,2} + \underbrace{\lambda_2 ||\boldsymbol{A}||_{1,1}}_{\ell_{11}\text{-penalty}} \tag{4.14}$$

$$||\boldsymbol{A}||_{1,1} = \sum_{t=1}^{T} ||\boldsymbol{a}_{t+1}||_1.$$

Consistent with this objective function, we refer to characteristics as anomalous predictive signals if their columns in $\boldsymbol{F}$ are zero, but the columns in $\boldsymbol{A}$ are scattered with nonzero entries. To distinguish between the two variants of the MTL in Equations (4.13) and (4.14), we follow Jalali et al. (2010) and denote the latter as the *dirty multi-task Lasso* (DMTL).

In the following, we highlight the advantages of the MTL-based selection approach by conducting a simulation study within which we compare the MTL's performance to that of two seemingly compelling alternatives. The first alternative is referred to as the *individual Lasso* (or IL) as it employs a standard ℓ_1-penalty term that operates at the level of each individual cross-sectional regression. In this way, it accounts for the time-varying nature of the coefficients, but entirely disregards the stable-selection aspect. The optimization problem for the IL is given by

$$\hat{\boldsymbol{F}} = \underset{\boldsymbol{F}}{\arg\min} \sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{C}_t \boldsymbol{f}_{t+1}||_2^2 + \lambda ||\boldsymbol{F}||_{1,1}.$$

The crucial difference compared to Equation (4.13) is that the IL shrinks the coefficients for each cross-sectional regression separately, meaning that the sparsity pattern in the rows of $\boldsymbol{F}$ can vary over time. Consequently, the set of selected characteristics, $\hat{\mathrm{S}}_t$, carries a time index.

The second alternative is referred to as the *pooled Lasso* (or PL) as it employs an ℓ_1-penalty to the pooled data of returns and characteristics under the simplifying assumption of a constant coefficient vector $\boldsymbol{f} = [f^k]_{K\times 1}$, thereby enforcing the same sparsity pattern across tasks. For a given value of λ, the coefficients for the PL are determined according to

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\arg\min} \sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{C}_t \boldsymbol{f}||_2^2 + \underbrace{\lambda ||\boldsymbol{f}||_1}_{\ell_1\text{-penalty}}$$

$$||\boldsymbol{f}||_1 = \sum_{k=1}^{K} |f^k|.$$

Variants of the PL are popular in empirical finance and used, among others, by Freyberger et al. (2020) in form of the group Lasso, by Kozak et al. (2020) to characterize the SDF subject to economically motivated priors, and by Feng et al. (2020) to select observable factors in a two-step procedure that accounts for model selection mistakes.

To determine the optimal amount of regularization, we perform 5-fold cross-validation for each of the above selection methods. That is, we repeatedly divide the data into a training and a validation set, use the training data to create sets of selected characteristics $\hat{\mathrm{S}}_t(\boldsymbol{\lambda}, p)$ conditional on the partition p, and evaluate these sets based on the validation data.[12] For the latter, we run T cross-sectional regressions for each $\hat{\mathrm{S}}_t(\boldsymbol{\lambda}, p)$ and estimate the average mean-squared error across partitions by

$$\widehat{\mathrm{MSE}} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}||\boldsymbol{R}_{t+1}(p) - \boldsymbol{C}_t(p)\big(\hat{\boldsymbol{f}}_{t+1}(p) + \hat{\boldsymbol{a}}_{t+1}(p)\big)||_2^2.$$

As the vector $\hat{\boldsymbol{a}}_{t+1}$ exists only for the DMTL, we set it to zero for the other approaches. Finally, we determine the optimal $\hat{\boldsymbol{\lambda}}$ that minimizes the validation $\widehat{\mathrm{MSE}}$, and solve the individual optimization problems for the combined training and validation data, thus obtaining a unified set $\hat{\mathrm{S}}_t(\hat{\boldsymbol{\lambda}})$ for each of the methods considered.[13]

For the simulation, we consider $T = 500$ tasks or individual cross-sections, $N =$ 5,000 test assets per task, and $K = 100$ candidate characteristics. The characteristics' values, $c^i_{k,t}$, are obtained as independent draws from a standard normal distribution. To account for the findings by McLean and Pontiff (2016), we further distinguish between three different types of characteristics: The first 10 belong to the group of *stable* predictors that are useful throughout the sample period, another 10 represent the group of *anomalous* predictors that are important only within certain periods of time, and the remaining 80 are completely *irrelevant*. For the stable predictors, the coefficients (or factors) evolve according to an autoregressive process

$$f^k_{t+1} = \phi_k f^k_t + \eta^k_{t+1},$$

where the innovations are independent normally distributed random variables, $\eta^k_{t+1} \sim \mathcal{N}(0, 0.05)$, and the parameter ϕ_k is uniformly distributed, $\phi_k \sim \mathcal{U}(0, 0.9)$. The coefficients of the anomalous predictors, a^k_{t+1}, follow the same distribution as the innovations η^k_{t+1}, but are allowed to be nonzero for at most 50 consecutive periods, where the initial periods are chosen at random. Finally, we collect the simulated coefficients in the vectors $\boldsymbol{f}_{t+1}$ and $\boldsymbol{a}_{t+1}$ and generate 5,000 return observations according to

$$\boldsymbol{R}_{t+1} = \boldsymbol{C}_t(\boldsymbol{f}_{t+1} + \boldsymbol{a}_{t+1}) + \boldsymbol{\varepsilon}_{t+1} \qquad t = 1, \ldots, 500,$$

where $\varepsilon^i_{t+1} \sim \mathcal{N}(0, 1)$. Note that, for the irrelevant predictors, the entries in both $\boldsymbol{f}_{t+1}$

[12] The regularization parameter $\boldsymbol{\lambda}$ is scalar-valued for the IL, PL, and MTL and a vector containing λ_1 and λ_2 for the DMTL.

[13] Note that the set $\hat{\mathrm{S}}(\hat{\boldsymbol{\lambda}})$ is time-constant for the PL, MTL, and DMTL.

and $\boldsymbol{a}_{t+1}$ are zero, whereas for the stable (anomalous) predictors only the vector $\boldsymbol{f}_{t+1}$ ($\boldsymbol{a}_{t+1}$) contains nonzero entries.

For the purpose of visualization, we collect the simulated coefficients in a $T \times K$-dimensional matrix that holds the characteristics' coefficient time series in the columns and the coefficients for the individual cross-sections in the rows. The ground truth is displayed in Panel A of Figure 4.1, where each pixel represents an element of this matrix and grey (white) pixels indicate that the respective coefficient values are nonzero (zero).

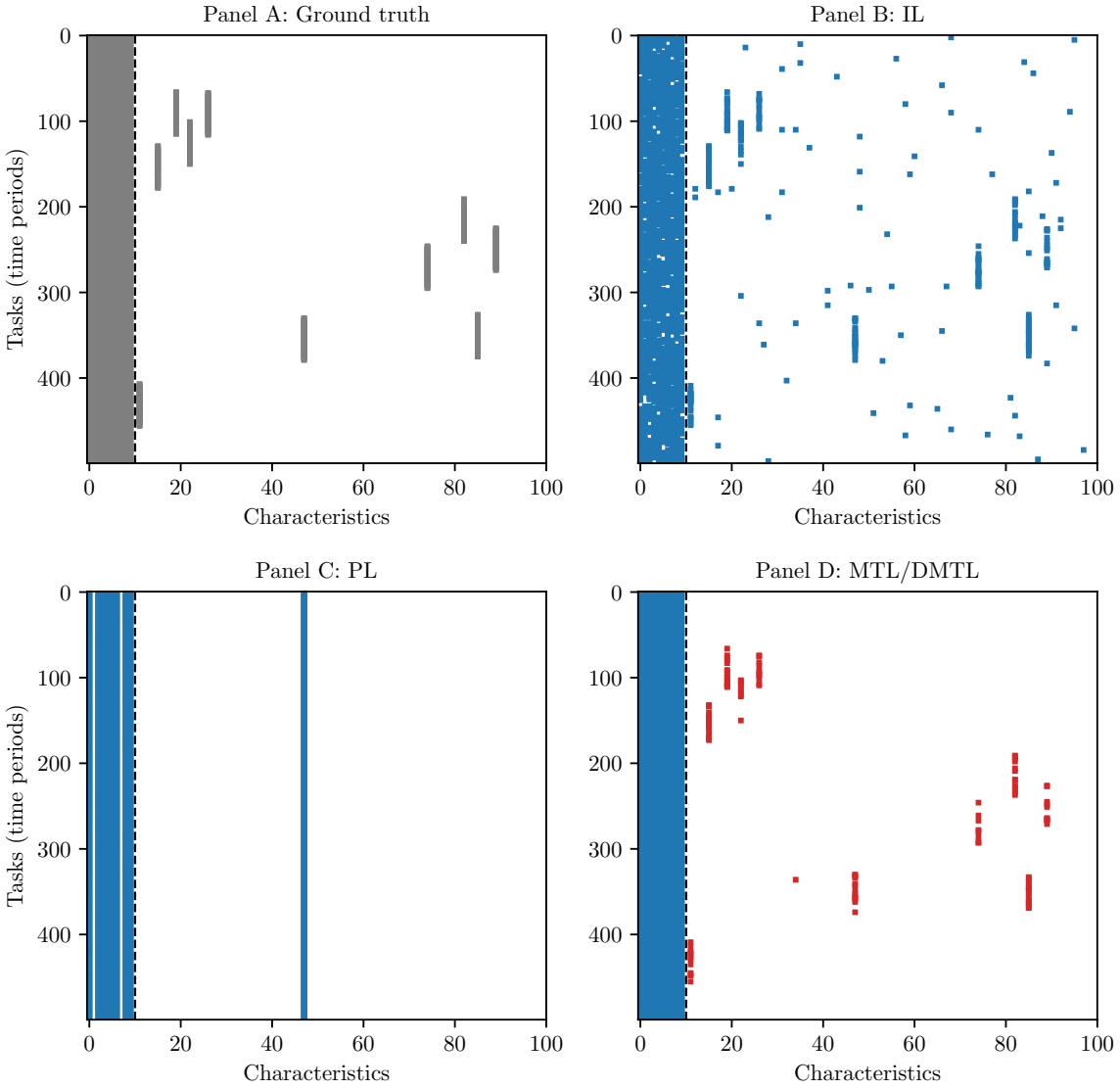
The sparsity patterns obtained by the different selection methods are shown in Panel B to D. The MTL and DMTL solutions are displayed jointly in Panel D, where the nonzero entries in $\hat{\boldsymbol{f}}_{t+1}$ are shown in blue, and the DMTL's nonzero entries in $\hat{\boldsymbol{a}}_{t+1}$ are shown in red.

When examining the sets of selected characteristics for the IL, it is evident that the flexibility of treating each cross-section separately comes at the cost of an unstable selection of characteristics. Although many of the stable predictors are identified as such, there remain zero entries within the first 10 columns. Moreover, the IL cannot distinguish between stable and anomalous return predictive signals because both are subsumed in $\hat{\mathrm{S}}_t(\hat{\boldsymbol{\lambda}})$. The IL also occasionally selects characteristics that are completely irrelevant, which is due to the fact that the number of observations used to determine the sparsity pattern is limited by the number of test assets available for each cross section, which is $N = 5{,}000$. In contrast, the PL employs the full $5{,}000 \times 500$ observations to simultaneously determine the set $\hat{\mathrm{S}}(\hat{\boldsymbol{\lambda}})$ for all time periods. Although the PL makes more efficient use of the available information, its selection is still inconsistent as it disregards the time-varying nature of the data-generating process. That is, the PL misses some of the stable predictors and is unable to account for anomalous return predictability. In contrast, both the MTL and the DMTL correctly classify the first 10 covariates as stable predictors, and the DMTL additionally succeeds in highlighting the presence of anomalous return predictability via $\hat{\boldsymbol{a}}_{t+1}$, which exemplifies the effectiveness of our multi-task approach.

4.5 Assessing the importance of characteristics

The following two subsections center on how to evaluate the importance of characteristics for describing the cross-section of returns. Section 4.5.1 introduces the notion of *conditional selection probabilities* that are estimated over subsamples of the data and serve as a basis for investigating whether the implied market beta from Equation (4.2) drives out any stock characteristics in cross-sectional regressions. Section 4.5.2 presents a more refined approach that is based on the concept of Shapley values and allows us to distinguish between characteristics with similar selection probabilities. We employ these two measures of variable importance to explore whether there are characteristics that provide incremental information beyond what is captured by the market beta.

Figure 4.1: The variable selection problem in cross-sectional regressions (Simulation). This figure illustrates the results of the simulation study discussed in Section 4.4. Panel A shows the entries of the coefficient matrix that is used to generate the panel of returns with $T = 500$ time periods, $N = 5{,}000$ test assets, and $K = 100$ characteristics. Each pixel represents an entry in the coefficient matrix and grey (white) pixels indicate that the respective coefficient values are nonzero (zero). Panels B to D show the sparsity patterns obtained by the IL, PL, and MTL/DMTL in conjunction with 5-fold cross-validation. The MTL and DMTL solutions are displayed jointly in Panel D: The blue pixels represent the nonzero values in the coefficient matrix $\hat{\boldsymbol{F}}$ from Equations (4.13) and (4.14), whereas the red pixels represent the task-specific entries of $\hat{\boldsymbol{A}}$.



4.5.1 Conditional selection probabilities

The MTL-based selection approach described in Section 4.4 aims to identify characteristics that are stable predictors of return variation across all the regressions shown in Equation (4.7). Unfortunately, real-world data-generating processes are rarely as stylized as in the simulation study outlined above. In practice, ℓ_1-based procedures often fail to produce a stable selection due to the non-trivial correlation structure governing the predictors, as pointed out by Zhao and Yu (2006) and Meinshausen and Bühlmann (2006). This is reflected in the fact that even small variations in the data can lead to substantial changes in the set of selected characteristics, $\hat{\mathrm{S}}$.

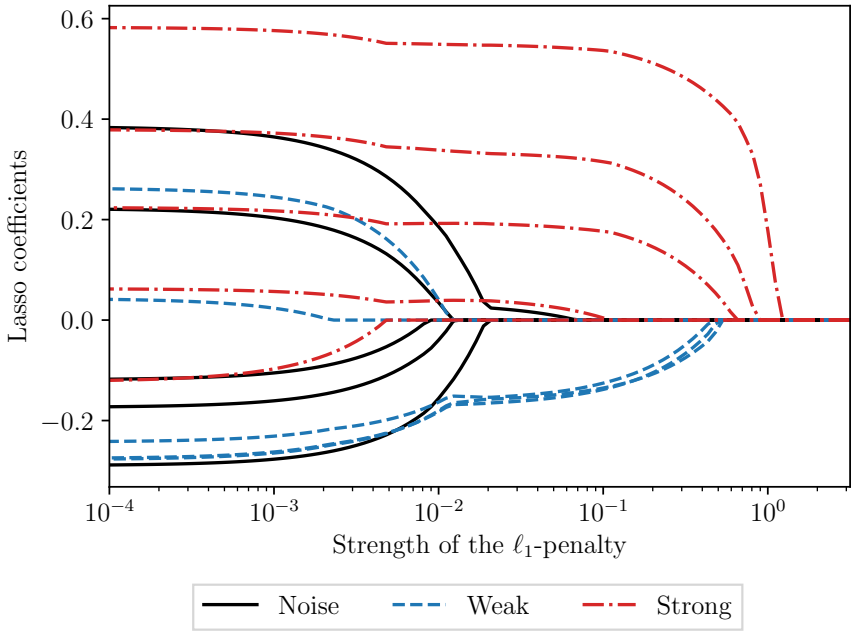
To demonstrate the instability of the MTL in the presence of highly correlated predictors, we conduct another simulation study that is more realistic in terms of the assumed dependence between characteristics. For simplicity, we generate a single cross section of $N = 5{,}000$ gross returns that is governed by two latent signals, which can be conceived of as unobserved loadings (or characteristics).[14] The return-generating process is such that the first signal strongly affects the level of gross returns, whereas the latter generates only little variation. In addition, we create 15 observable characteristics that constitute imperfect proxies for the two latent signals, five of which are correlated with the strong signal, another five are correlated with the weak signal, and the remaining five represent pure noise. Based on these observable characteristics, we compute Lasso coefficient paths for a sequence of ℓ_1-penalty parameters and display them in Figure 4.2.

Strikingly, there is no level of regularization for which the Lasso (or the MTL) is able to identify all relevant characteristics. This is reflected in the fact that some of the coefficient paths of the relevant predictors are drawn towards zero even faster than those of the noise variables. While eliminating as much redundant information as possible is consistent with our goal of arriving at a parsimonious representation of returns, caution must be exercised when using standard regularization paths as a means of assessing variable importance. For any value of λ, the set $\hat{\mathrm{S}}$ obtained by the Lasso contains only few of the relevant characteristics, suggesting that the ones omitted are completely unimportant. However, if we were to repeat the selection process based on a newly generated sample, the composition of $\hat{\mathrm{S}}$ would change, as different combinations of informative characteristics achieve similar levels of predictive accuracy. For a conclusive assessment of the characteristics' relative importance, it is therefore crucial that we account for the instability of the MTL.

Before we continue with a discussion of possible solutions, we demonstrate in Figure 4.3 that the 78 stock characteristics used in this study are indeed highly correlated. A detailed description of the data and the meaning of the characteristics' labels can be found in Section 2.3. To illustrate the underlying dependence structure, we group the characteristics according to their similarity by forming correlation clusters based on the K-means algorithm, using $\boldsymbol{d} = 1 - |\boldsymbol{\rho}|$ as a measure of dissimilarity between

[14] When there is only a single cross-section, the MTL collapses to the standard Lasso.

Figure 4.2: The instability of Lasso regularization paths (Simulation). This figure presents Lasso regularization paths for simulated data that features strongly correlated predictors. The prediction target is subject to a latent representation that consists of two unobserved signals: one that strongly affects the outcome variable and one that is only weakly influential. For the ℓ_1-regularized prediction model, we create 15 predictors, of which five load on the strongly predictive signal (Strong), another five load on the weakly predictive signal (Weak), and the remaining five are completely irrelevant (Noise). Using these variables, we run the Lasso optimization algorithm and generate coefficient paths for a reasonably chosen sequence of ℓ_1-regularization parameters. The resulting coefficient paths are shown below, where the dotted red lines represent the group of strong predictors, the dashed blue lines represent the group of weak predictors, and the noise variables are associated with solid black lines.



characteristics, and $\boldsymbol{\rho} = [\rho_{k,l}]_{K\times K}$ as the matrix of pairwise correlation coefficients.[15] Specifically, the set of clusters $\mathrm{G} = \{\mathrm{G}_1, \ldots, \mathrm{G}_Z\}$ is found by minimizing the following optimization problem:

$$\hat{\mathrm{G}} = \arg\min_{\mathrm{G}} \sum_{z=1}^{Z} \sum_{k\in \mathrm{G}_z} ||\boldsymbol{d}_k - \bar{\boldsymbol{d}}_z||_2^2, \tag{4.15}$$

where $\boldsymbol{d}_k$ is the k-th column (or row) of $\boldsymbol{d}$, and Z denotes the number of clusters, which we determine via Silhouette scores (Rousseeuw, 1987). An attractive feature of the K-means algorithm is that it considers all correlation coefficients jointly, so that two characteristics within the same cluster are not only similar to each other, but also exhibit similar correlation patterns with all other characteristics.

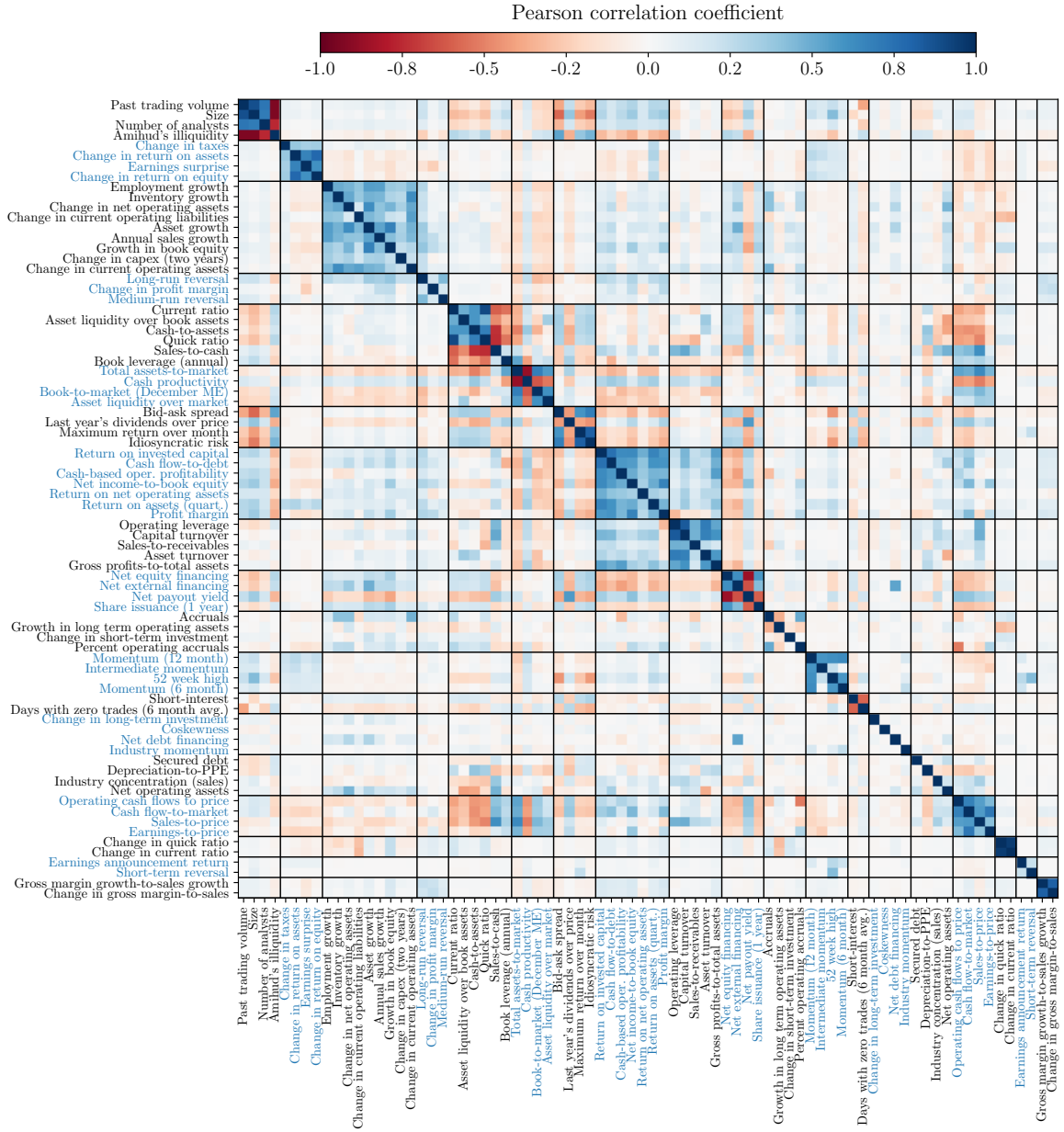
As can be seen in Figure 4.3, the number of correlation clusters found ($Z = 19$) is considerably lower than the total number of characteristics ($K = 78$), indicating that they are governed by a lower-dimensional latent structure. Intuitively, one can think of this as many characteristics representing alternative quantifications of the same unobserved signal. For example, consider the cluster comprising *Past trading volume*, *Size*, *Number of analysts*, and *Amihud's illiquidity*: All of these characteristics can be conceived of as alternative quantifications of an asset's liquidity. However, even if they were individually informative for the cross-section of returns, the MTL would not select the entire group, but rather shrink some of the coefficient time series to zero. The resulting $\hat{\mathrm{S}}$ would indicate that only a fraction of these liquidity proxies are important, when in fact they all are to a certain extent.

To overcome the instability of ℓ_1-based selection methods, various solutions have been proposed. One such solution is the *adaptive Lasso* by Zou (2006), which has been used by Freyberger et al. (2020) in an asset pricing context. One advantage of the adaptive Lasso is that it enjoys the oracle property, i.e., it achieves consistent variable selection if the true underlying model is sparse. Compared to the standard Lasso, the objective function of the adaptive Lasso includes additional parameters (or weights) that control the tightness of the ℓ_1-penalty individually for each regressor. These parameters must be determined prior to solving the modified optimization problem, and are usually defined in terms of the reciprocal of the associated OLS coefficients. However, as noted by Meinshausen and Bühlmann (2010), the disadvantage of this two-step procedure is that it involves an additional tuning parameter, for which it is unclear how it should be chosen in practice. Moreover, the adaptive Lasso assumes that the data-generating process is linear and sparse in the available characteristics – an assumption that is difficult to defend, considering the complexity of financial markets.

Alternatively, one could introduce an additional ℓ_2-penalty term that ties together the coefficient paths of the correlated characteristics, thus increasing the probability

[15] Note that applying the K-means algorithm to $\boldsymbol{d}$ instead of $\boldsymbol{\rho}$ ensures that highly correlated characteristics are assigned to the same cluster, irrespective of whether said correlation is positive or negative.

Figure 4.3: Correlation clusters of stock characteristics. This figure illustrates the pairwise correlation coefficients of $K = 78$ stock characteristics that are cross-sectionally rank-transformed to the unit interval. Each pixel of the below heatmap corresponds to an entry of the matrix $\boldsymbol{\rho} = [\rho_{k,l}]_{K\times K}$, which collects the characteristics' pairwise correlation coefficients. To capture the similarity of the characteristics, we assign them to disjoint correlation clusters that are determined by the K-means algorithm using $\boldsymbol{d} = 1-|\boldsymbol{\rho}|$ as a measure of dissimilarity. The optimal number of clusters, $Z = 19$, is determined via Silhouette scores (Rousseeuw, 1987).



that they are selected jointly (cf. Hastie et al., 2015, p. 56). In this way, the variables' common variation is employed more efficiently, which can have a positive impact on predictive accuracy. In a related study, Kozak et al. (2020) use such a combination of ℓ_1- and ℓ_2-regularization to model the SDF subject to economically motivated priors, finding that adding the ℓ_2-penalty improves the SDF's out-of-sample performance. However, in the present application, we intend to stick to the definition of cross-sectional regressions by Fama and French (2020), where the characteristics serve as pre-specified loadings. When highly correlated characteristics are selected jointly and used alongside each other as covariates in the regressions of Equation (4.4) and (4.7), the resulting factors may be unstable because the design matrix is nearly rank-deficient, calling into question the appropriateness of employing an ℓ_2-penalty in this case.

Last but not least, one could resort to classical methods of dimensionality reduction, such as principal components regression, to avoid the selection problem altogether. Kelly et al. (2019) pursue this idea and demonstrate that the loadings in cross-sectional regressions can be represented as linear combinations of multiple characteristics using an approach they call *instrumented* PCA (or IPCA). While this approach is statistically efficient, interpreting the resulting factors is difficult, as the associated loadings represent artificial combinations of the original variables that often lack economic meaning. The financial industry, however, appears to have a strong interest in attributing importance to individual characteristics, as evidenced by the fact that many institutional investors form *style portfolios* based on individual characteristics. In this regard, Fama and French's (2020)'s OLS factors have the advantage that they accommodate the financial markets' desire for interpretability.

In this study, we take a different route and employ the methodology proposed by Meinshausen and Bühlmann (2010), which allows us to preserve the integrity of the MTL-based selection procedure. Their approach is based on creating multiple random subsamples of the data, which we refer to as *auxiliary* samples. We generate these subsamples cross-sectionally, that is, we randomly pick stock indices from $\{1, \dots, N\}$ and assign them to the subset A, so that $|\mathrm{A}| = N/2$.[16] For each of these auxiliary samples, we generate different sets of selected characteristics, $\hat{\mathrm{S}}_{\mathrm{A}}$, and estimate their selection probabilities, $\mathbb{P}(k \in \hat{\mathrm{S}}_{\mathrm{A}})$, across subsamples, the latter serving as quantifications of the characteristics' relative importance. The appeal of this approach is that, even though the MTL is not able to produce a stable selection for any given A, it is still possible to identify the relevant characteristics by aggregating the results across subsamples.

For the present application, however, the original approach by Meinshausen and Bühlmann (2010) needs to be modified, as it tends to underestimate the importance of correlated characteristics. This is because the MTL selects different collections of correlated predictors across subsamples, resulting in the selection probability of the entire group being distributed among its constituents. To overcome this problem, we propose to first cluster the characteristics by their pairwise correlation coefficients us-

[16] N denotes the total number of stocks across all time periods.

ing the K-means procedure described in Equation (4.15), and then randomly draw representatives ω from each cluster $\hat{\mathrm{G}}_z$, so that $\hat{\Omega}_{\mathrm{A}} = \{\omega \in_R \hat{\mathrm{G}}_z : z \in \{1,\dots,Z\}\}$ represents the resulting collection of characteristics with cardinality $|\hat{\Omega}_{\mathrm{A}}| = Z$.[17] In this way, we provide each group of correlated characteristics with a guarantee of being represented by one of its constituents, and additionally minimize the information overlap between the predictors within each subsample. As opposed to Meinshausen and Bühlmann (2010), we eventually assess a characteristic's importance by considering its *conditional* selection probability, $\mathbb{P}(\omega \in \hat{\mathrm{S}}_{\mathrm{A}} | \omega \in \hat{\Omega}_{\mathrm{A}})$, thus accounting for the fact that characteristics are only available for selection if they are contained in $\hat{\Omega}_{\mathrm{A}}$.

With respect to testing the conditional CAPM, this repeated-subsampling approach is particularly interesting as it allows us to assess whether the implied market beta drives out any characteristics in cross-sectional regressions. This is achieved by performing the MTL selection procedure twice, once based on the original characteristics without considering the implied market beta, and once using characteristics that have been orthogonalized with respect to beta, i.e.,

$$\begin{aligned} \tilde{\boldsymbol{C}}_t &= \boldsymbol{C}_t - \boldsymbol{\beta}_t \hat{\boldsymbol{\gamma}}_t, \\ \text{where } \hat{\boldsymbol{\gamma}}_t &= (\boldsymbol{\beta}_t' \boldsymbol{\beta}_t)^{-1} \boldsymbol{\beta}_t' \boldsymbol{C}_t. \end{aligned} \tag{4.16}$$

In this way, we ensure that a characteristic is selected only if it provides an incremental improvement over the market beta's description of cross-sectional return variation. A sizable reduction in a characteristic's conditional selection probability would thus be considered evidence against the hypothesis that this characteristic carries additional information.

4.5.2 Shapley decompositions of cross-sectional R-squared

The conditional selection probabilities from Section 4.5.1 have their limitations when its comes to discriminating between characteristics with similar selection probabilities. For example, it may be the case that two characteristics are selected equally often, but one of them still contributes more to the explained variation than the other. Hence, conditional selection probabilities are rather crude measurements of variable importance that merely give an indication as to whether characteristics are relevant *at all*.

To overcome these limitations, we complement our assessment of the importance of individual characteristics by decomposing the cross-sectional R^2_{A}, which is defined as follows:

$$R^2_{\mathrm{A}} = 1 - \frac{\sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{\iota} f^1_{t+1} - \boldsymbol{\beta}_t f^2_{t+1} - \tilde{\boldsymbol{C}}_t \boldsymbol{f}_{t+1}||_2^2}{\sum_{t=1}^{T} \frac{1}{N_t} ||\boldsymbol{R}_{t+1} - \boldsymbol{\iota} \bar{R}_{t+1}||_2^2},$$

[17] Although $\hat{\Omega}_{\mathrm{A}}$ is not subject to an estimation in the classical sense, the notation aims to highlight that its composition depends on the clusters found by the K-means algorithm.

where the $\bar{R}_{t+1}$ denotes the average return across stocks in $t+1$. In practice, we compute the R^2_{A} *out-of-sample*, i.e., we generate factor realizations using the auxiliary set A and hold them fixed to compute the R^2_{A} based on the leftover stocks. In what follows, we will refer to this $N/2$-dimensional collection of leftover stocks as the *main* set, or M.

To quantify the characteristics' contributions to the cross-sectional R^2_{A}, we utilize a decomposition approach that is based on the concept of *Shapley values*.[18] In cooperative game theory, Shapley values are used to determine a player's contribution to the total surplus generated by a coalition of players (cf. Shapley, 1951). In the present context, each characteristic represents a player, and the value of the coalition is given by the cross-sectional R^2_{A}. The Shapley value of the k'th characteristic is then defined as its average contribution to the value of R^2_{A} over all possible regression specifications that can be generated using subsets of the characteristics in $\hat{\text{S}}_{\text{A}}$ without k. Formally, this corresponds to

$$R^2_{\text{A},k} = \sum_{\text{V}\subseteq\text{W}\setminus\{k\}} \frac{|\text{V}|!(|\text{W}|-|\text{V}|-1)!}{|\text{W}|!}\big(R^2_{\text{A}}(\text{V}\cap\{k\})-R^2_{\text{A}}(\text{V})\big), \tag{4.17}$$

where W denotes the set of players (or characteristics), which in this case is $\hat{\text{S}}_{\text{A}}$, and V represents one possible subset of W excluding the k'th characteristic.

An attractive feature of this approach is that it naturally accounts for the dependence between regressors, as the contribution of a characteristic depends on the contribution provided by others. To give an intuitive example, suppose that the *Size* characteristic is correlated with the market beta. In this case, we would expect its contribution to the cross-sectional R^2_{A} to be high when the market beta is not accounted for, but to decrease once the beta is included as a regressor. This decrease is accounted for in the above definition of Shapley values, as we average over the contributions of *Size* over all permutations of additional regressors.[19]

As we perform this decomposition separately for each random partition of the data, represented by A and M, we need to aggregate the characteristics' contributions across subsamples. We do so by reporting the median Shapley value for each characteristic.

4.6 Post-selection inference via repeated sample splitting

In Section 4.3, we explained how cross-sectional regressions of returns onto a small set of characteristics can be used to establish testable restrictions for the conditional

[18] For an in-depth discussion of the theoretical properties of this approach, please refer to Grömping (2007).

[19] Note that performing this decomposition using all 78 stock characteristics is computationally infeasible, as we would have to run $T\times 2^K$ cross-sectional regressions. In this respect, the MTL-based pre-selection step may be viewed as a means of reducing the complexity of this decomposition approach.

CAPM, initially assuming that the set of characteristics is known. We then introduced an MTL-based selection procedure in Section 4.4 to identify characteristics that are stable predictors of return variation and thus suitable candidates for the above cross-sectional regressions.

Unfortunately though, if these two steps are performed on the same data, the Wald statistic in Equation (4.12) is no longer approximately χ^2-distributed, but distorted in such a way that a correct null hypothesis would be rejected too often. This size distortion is due to the fact that classical inference procedures do not account for the underlying selection bias, i.e., they ignore the fact that only those statistical relationships are tested that were found to be sufficiently strong in the given sample. This is problematic because even in large samples, strong in-sample relationships can arise by chance.

Moreover, we have seen in Section 4.5 that, even in the very stylized case where the true DGP is linear and all relevant information is observable, determining the *true* set of predictors can be a challenging task. In real-world applications, the situation may be even worse: For example, if the true DGP is highly nonlinear and subject to interaction effects, or if some of the relevant predictors are simply unobserved, consistent model selection is off the table. In these cases, all we can hope for is that the regressions we employ capture important aspects of the true DPG.

It seems to us that many recent contributions to empirical finance do not give due consideration to the problem of model misspecification and its implications for statistical inference. Instead, researchers simply assume that their statistical models are correctly specified, which allows them to use inferential procedures that target the parameters of the true DGP. The studies by Feng et al. (2020) and Freyberger et al. (2020) exemplify this approach: Both operate under the assumption that the true return-generating process is linear, either in factors or basis functions, and approximately low-dimensional. However, if these conditions are not met, the reported standard errors and p-values are simply wrong.

Although the *true-model* perspective has been dominating econometric philosophy for years, it is by no means without an alternative. White (1980, 1981), for example, argues in a series of papers that, in a non-experimental science such as economics, the functional form of a statistical model is typically chosen not out of certainty about the true causal (or equilibrium) relationship, but rather on the basis of mathematical convenience. Thus, rather than focusing on the true DGP, it is often more useful to aim for a consistent estimation of the parameters of an *approximate* model. Berk et al. (2013) complement this view by pointing out that, under certain circumstances, it is not possible to identify the parameters of the true model, especially when there are many redundant predictors. In such cases, one solution may be to focus on an approximate model whose parameters exist independently of the true DGP.

In what follows, we depart from the true-model perspective and present an approach towards post-selection inference that 1) accounts for the distributional distortions induced by variable selection and 2) accommodates arbitrary forms of model

misspecification. The basic idea is to obtain valid inference on the parameters of an approximate model by performing the selection and estimation steps separately on different parts of the data. The theoretical foundations for this approach have been established by Rinaldo et al. (2019), who demonstrate that the resulting standard errors and p-values are valid regardless of the underlying data-generating distribution or selection procedure chosen. For the present application, we implement the sample-splitting approach as follows: Initially, we partition the universe of stocks cross-sectionally into an auxiliary set A and a main set M, just as described in Sections 4.5.1 and 4.5.2. We then employ the MTL to select meaningful characteristics using A, and construct the corresponding factor time series using M. To emphasize that the means of the cross-sectional factors are defined conditional on the set of selected characteristics, $\hat{\mathrm{S}}_{\mathrm{A}}$, we will henceforth denote them by $\boldsymbol{\theta}_{\mathrm{A}}$. Finally, we compute the Wald statistic and the associated p-value for the joint hypothesis in Equation (4.11) using M.[20]

The sample-splitting approach is compelling as it allows us to compute the Wald statistic and the associated p-values using standard formulas, without having to make restrictive assumptions about the true DGP. However, as we have seen in Section 4.5.1, different partitions of the data can lead to different sets of selected characteristics, and thus to different conclusions about the null hypothesis under investigation. To additionally account for the uncertainty that is due to randomly splitting the data, we follow Chernozhukov et al. (2023) and partition the data not just once but multiple times. Consequently, the target parameters $\boldsymbol{\theta}_{\mathrm{A}}$ are random variables, where the randomness stems from the fact that we draw multiple pairs of auxiliary and main sets. To obtain a uniform statement about the null hypothesis, Chernozhukov et al. (2023) propose to perform *quantile aggregation*. Following their approach, we obtain sample-splitting adjusted p-values, p_{med}, by taking the median of the individual p-values that we obtain for the different partitions of the data, i.e.,

$$p_{\mathrm{med}} = 2 \cdot \mathrm{med}(p_{\mathrm{A}}), \tag{4.18}$$

where the multiplication by two can be conceived of as the price of splitting the data.

4.7 Empirical strategy

Before delving into the specifics of how we compile our data, let us briefly review the research questions that we posed at the beginning and provide an overview of our testing strategy. The main objective of this study is to use cross-sectional regressions to test the fully-implied specification of the conditional CAPM that we proposed in Section 4.2. Although the underlying testing principle is not new and has been studied quite extensively in the past, we believe that the challenges associated with

[20] To illustrate the logic behind the sample-splitting approach, we present a simulation study in Appendix C.1, where we plot the approximate distributions of the t- and Wald statistics, once for the case where both selection and inference are performed on the full sample, and once for the case where these steps are distributed across subsamples.

time-varying coefficients and high-dimensional sets of competing stock characteristics have not been adequately addressed. Table 4.1 provides a concise summary of our testing strategy, which interprets the selection of characteristics as a multi-task learning problem (Section 4.4) and employs quantile-aggregation to obtain valid post-selection inference (Section 4.6).

In the following, we will use this procedure to answer the following research questions: 1.) Does the implied market beta drive out any of the competing stock characteristics in cross-sectional regressions? 2.) Can we reject the null hypothesis that there are no additional nonzero factor means? 3.) Do the results change if we use characteristic-sorted portfolios instead of individual stocks as test assets? 4.) Where does the implied market beta rank in terms of its cross-sectional explanatory power relative to other characteristics? 5.) What can be said about the nature of the selected characteristics – are they stable predictors of return variation, or are they informative only within certain periods of time?

4.7.1 Database

To address these questions, we require an extensive data set that includes monthly stock returns, prices of European call and put options with matching maturities, and a comprehensive collection of stock characteristics. For the return data, we draw on the daily security file provided by CRSP (The Center for Research in Security Prices) and compute monthly gross returns for all common shares (CRSP codes 10 and 11) traded on major US exchanges (CRSP codes 1, 2, and 3) during the period from January 1996 to December 2021. For stocks that are delisted during this period, we adjust the last available daily return by the associated delisting return that is provided by CRSP. The delisting event is assumed to be unexpected, so the number of calendar days over which the monthly returns are calculated converges to zero as we approach the delisting date. Returns for which the holding period extends beyond December 31, 2021 are excluded from the sample.

The implied market betas from Equation (4.2) are functions of risk-neutral return variances, which can be approximated using a collection of European call and put option prices. However, options on US stocks are typically traded American-style, so we need to convert between the prices of American and European option contracts. This is accomplished by using the implied volatility surface provided by OptionMetrics as an input to the Black-Scholes-Merton formula to compute the prices of equivalent European options with maturities of 30 calendar days and deltas ranging from -0.9 to 0.9 in steps of 0.05. In addition, we obtain monthly gross risk-free rates using the zero-coupon yield curve provided by OptionMetrics. For more details on the approximation of risk-neutral moments and the construction of the risk-free rate, please refer to Chapter 3, Appendix B.1 and B.2.

An extensive collection of stock characteristics that have been used as return predictive signals in previous literature is available from Chen and Zimmermann's (2022) *Open Source Asset Pricing* repository. From this repository, we download a selection

Table 4.1: Testing strategy. This table provides an overview of our testing strategy, which employs multi-task learning for the selection of characteristics and sample splitting for post-selection inference. The first step of our procedure is to randomly partition the cross-section of stocks into 1,000 pairs of auxiliary and main sets, denoted by A and M. Steps 2 to 9 refer to the auxiliary sets and describe how the MTL is used to select stable predictors of return variation. Steps 10 to 17 refer to the main sets and cover the construction of the cross-sectional factors and the computation of quantile-aggregated p-values. Steps 18 to 21 present the various methods that we use to aggregate our results across partitions.

1. Create $1,000$ random partitions (A,M) of $\{1,\dots,N\}$
2. For each auxiliary set A:
3. Find a set of correlation clusters $\mathrm{G} = \{\mathrm{G}_1,\dots,\mathrm{G}_Z\}$ according to
$$\hat{\mathrm{G}} = \arg\min_{\mathrm{G}} \sum_{z=1}^{Z} \sum_{k\in\mathrm{G}_z} ||\boldsymbol{d}_k - \bar{\boldsymbol{d}}_z||_2^2$$
where $\boldsymbol{d}_k = 1-|\boldsymbol{\rho}_k|$ and Z is chosen via Silhouette scores
4. Randomly draw a single characteristic from each $\hat{\mathrm{G}}_z$, such that $\hat{\Omega}_{\mathrm{A}} = \{\omega \in_R \hat{\mathrm{G}}_z : z \in \{1,\dots,Z\}\}$
5. Collect the drawn characteristics in a sub-matrix $\boldsymbol{C}_t = [c^i_{\omega,t}]_{|\mathrm{A}_t|\times|\hat{\Omega}_{\mathrm{A}}|}$, with A_t being the set of stocks in A available in t
6. Orthogonalize each characteristic with respect to the implied market beta using $\tilde{\boldsymbol{C}}_t = \boldsymbol{C}_t - \boldsymbol{\beta}_t\hat{\boldsymbol{\gamma}}_t$, where $\hat{\boldsymbol{\gamma}}_t = (\boldsymbol{\beta}_t'\boldsymbol{\beta}_t)^{-1}\boldsymbol{\beta}_t'\boldsymbol{C}_t$
7. Select characteristics by the CMTL (or DMTL) using 3-fold CV:
8. $\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=1}^{T} \frac{1}{|\mathrm{A}_t|}||\boldsymbol{R}_{t+1} - \tilde{\boldsymbol{C}}_t\boldsymbol{f}_{t+1}||_2^2 + \lambda||\boldsymbol{F}||_{1,2}$
9. $\hat{\mathrm{S}}_{\mathrm{A}} = \{\omega \in \hat{\Omega}_{\mathrm{A}} : \hat{\boldsymbol{f}}^{\omega} \neq \boldsymbol{0}\}$
10. For each main set M:
11. Construct factor time series via OLS:
12. Define $\boldsymbol{C}_t = [c^i_{s,t}]_{|\mathrm{M}_t|\times|\hat{\mathrm{S}}_{\mathrm{A}}|}$ and $\boldsymbol{X}_t = [\boldsymbol{\iota}, \boldsymbol{\beta}_t, \boldsymbol{C}_t]$
13. Create factors according to $\boldsymbol{f}_{t+1} = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{R}_{t+1}$
14. Employ quantile-aggregation for inference:
15. Estimate the factor means: $\hat{\boldsymbol{\theta}}_{\mathrm{A}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_{t+1}$
16. Compute the value of the test statistic: $w_{\mathrm{A}} = \hat{\boldsymbol{\theta}}_{\mathrm{A}}'\widehat{\mathrm{var}}(\hat{\boldsymbol{\theta}}_{\mathrm{A}})^{-1}\hat{\boldsymbol{\theta}}_{\mathrm{A}}$
17. Get the p-value: $p_{\mathrm{A}} = \mathbb{P}(W \geq w_{\mathrm{A}})$, where $W \overset{H_0}{\sim} \chi^2(|\hat{\mathrm{S}}_{\mathrm{A}}|)$
18. Across partitions A and M:
19. Estimate conditional selection probabilities: $\mathbb{P}(\omega \in \hat{\mathrm{S}}_{\mathrm{A}} | \omega \in \hat{\Omega}_{\mathrm{A}})$
20. Decompose the cross-sectional R^2_{A} using Shapley values
21. Compute sample splitting-adjusted p-values: $p_{\mathrm{med}} = 2\cdot\mathrm{med}(p_{\mathrm{A}})$

of 78 characteristics, the composition of which is determined by the following two criteria: Popularity and importance in previous studies, and availability and quality of the data.[21] For the latter, we examine the structure of missing values and keep only those characteristics whose proportion of missing values does not exceed 50%. The distributions of missing values for the remaining 78 characteristics are shown in Figure 4.4, once for annual subsamples (Panel A) and once for the entire sample period (Panel B). The characteristics' names and information about the publications in which they were introduced as predictive signals can be found in Table 4.2.

To deal with the remaining missing values, we employ the imputation procedure by Bryzgalova et al. (2022), which captures the characteristics' cross-sectional dependencies by means of a latent factor representation. Their approach involves estimating the following cross-sectional factor models, each of which describes the statistical relationship between characteristics at a given point in time t:

$$\boldsymbol{C}_t = \boldsymbol{\Gamma}_t \boldsymbol{\Lambda}_t' + \boldsymbol{u}_t \qquad t = 1, \ldots, T, \tag{4.19}$$

where $\boldsymbol{\Gamma}_t = [\gamma_{l,t}^i]_{N_t \times L}$ is a matrix comprising L stock-level factors, $\Lambda_t = [\lambda_{l,t}^k]_{K \times L}$ is the corresponding loadings matrix, and $\boldsymbol{u}_t = [u_{k,t}^i]_{N_t \times K}$ is a matrix of residuals.[22] The estimation of the models' components is performed in two steps: First, one obtains preliminary loadings $\tilde{\boldsymbol{\Lambda}}_t$ as the eigenvectors of the L largest eigenvalues of the characteristics' covariance matrix. The individual entries of this covariance matrix are calculated based on the stocks for which the associated pair of characteristics is observed. The stock-specific factor estimates are then obtained by regressing the characteristics vector $\boldsymbol{C}_t^i = [c_{k,t}^i]_{K \times 1}$ onto the preliminary loadings

$$\hat{\boldsymbol{\Gamma}}_t^i = (\tilde{\boldsymbol{\Lambda}}_t' \boldsymbol{W}_t^i \tilde{\boldsymbol{\Lambda}}_t)^{-1} (\tilde{\boldsymbol{\Lambda}}_t' \boldsymbol{W}_t^i \boldsymbol{C}_t^i) \qquad i = 1, \ldots, N_t, \tag{4.20}$$

where $\hat{\boldsymbol{\Gamma}}_t^i = [\hat{\gamma}_{l,t}^i]_{L \times 1}$ is the resulting vector of factor estimates, and $\boldsymbol{W}_t^i$ is a symmetric matrix containing ones on its diagonal if the corresponding characteristic is observed and zeros otherwise. Second, the final loadings are estimated using

$$\hat{\boldsymbol{\Lambda}}_t^k = (\hat{\boldsymbol{\Gamma}}_t' \boldsymbol{W}_t^k \hat{\boldsymbol{\Gamma}}_t)^{-1} (\hat{\boldsymbol{\Gamma}}_t' \boldsymbol{W}_t^k \boldsymbol{C}_t^k) \qquad k = 1, \ldots, K,$$

where the transpose of $\hat{\boldsymbol{\Lambda}}_t^k = [\hat{\lambda}_{l,t}^k]_{L \times 1}$ is the k'th row of $\hat{\boldsymbol{\Lambda}}_t$, and $\boldsymbol{C}_t^k = [c_{k,t}^i]_{N_t \times 1}$ is a vector comprising the cross-section of the k'th characteristic. The missing values in $\boldsymbol{C}_t$ are eventually replaced by their corresponding entries in $\hat{\boldsymbol{C}}_t = \hat{\boldsymbol{\Gamma}}_t \hat{\boldsymbol{\Lambda}}_t'$. To prevent information from the auxiliary set from leaking into the main set, we perform the imputation procedure separately for each subsample of the data.

Note, however, that computing the inverse in Equation (4.20) can be unstable if the number of missing characteristics is large. For this reason, we additionally exclude

[21] The data are part of the March 2022 update and retrieved according to the instructions that are given on the repository's website: https://www.openassetpricing.com/.

[22] We follow Bryzgalova et al. (2022) and set the number of factors to $L = 6$.

Figure 4.4: Structure of missing values. This figure illustrates the quality and availability of the characteristics data that we obtain from Chen and Zimmermann's (2022) *Open Source Asset Pricing* repository. Panel A presents the share of missing values per year and characteristic, and Panel B gives an overview of the overall share of missing values. The characteristics are sorted in ascending order according to the percentages displayed in Panel B. Characteristics for which the share of missing values is greater than 50% are excluded from the sample.

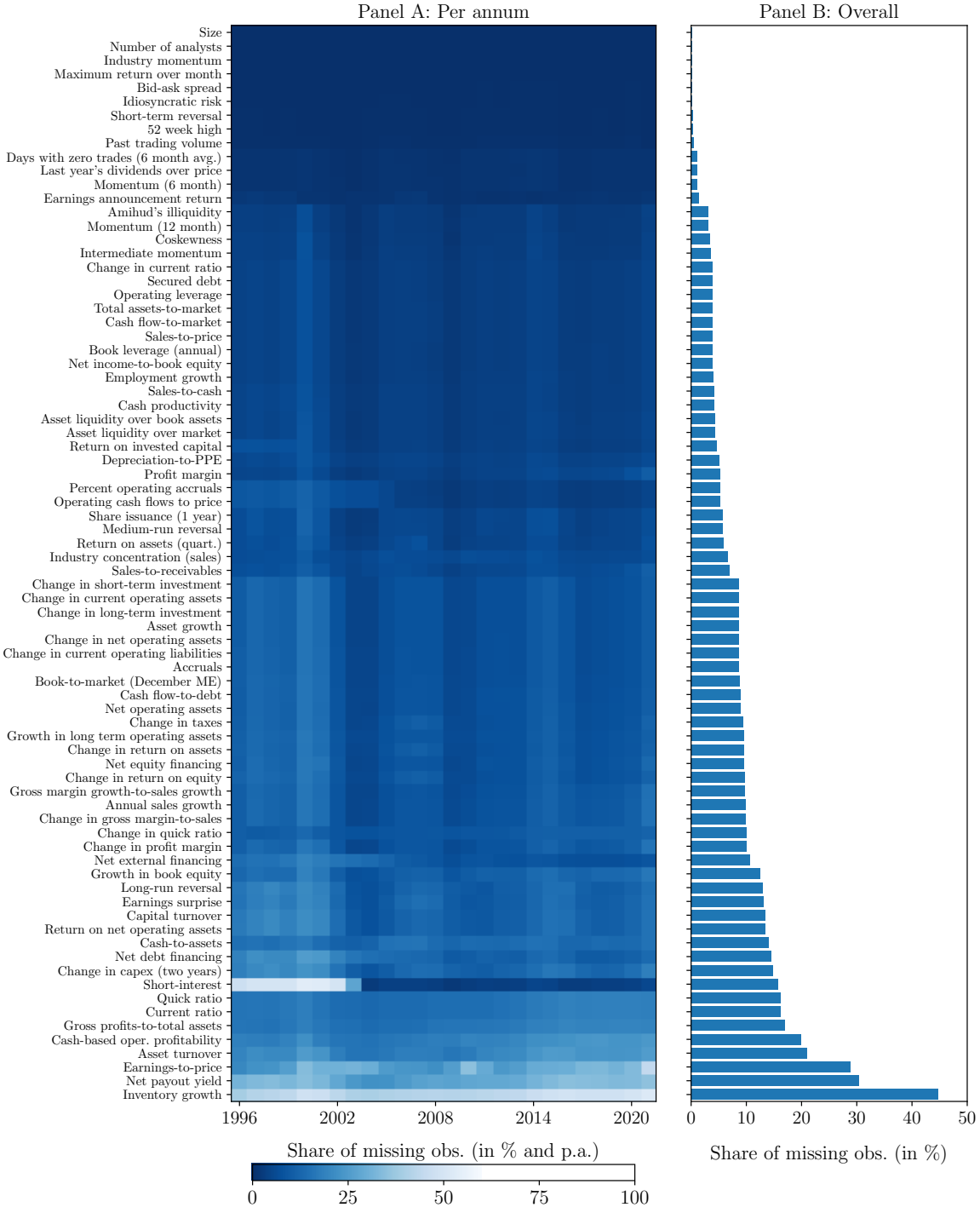


Table 4.2: Stock characteristics. This tables provides detailed information on the 78 stock characteristics that we obtain from Chen and Zimmermann's (2022) *Open Source Asset Pricing* repository. The first column indicates the economic category of the characteristic. The second column contains the characteristic's name. The third column presents the authors by whom the characteristic was first introduced, and the fourth and fifth columns the journal and year of the associated publication. Note that the economic categories have been slightly changed compared to the original documentation file.

Economic category	Characteristic	Authors	Journal	Year
Accruals	Accruals	Sloan	AR	1996
	Percent operating accruals	Hafzalla, Lundholm, Van Winkle	AR	2011
Asset Composition	Cash productivity	Chandrashekar and Rao	WP	2009
	Cash-to-assets	Palazzo	JFE	2012
	Depreciation-to-PPE	Holthausen and Larcker	JAE	1992
	Inventory growth	Belo and Lin	RFS	2012
	Net operating assets	Hirshleifer et al.	JAE	2004
Cash Flow	Cash flow-to-debt	Ou and Penman	JAR	1989
	Cash flow-to-market	Lakonishok, Shleifer, Vishny	JF	1994
	Operating cash flows to price	Desai, Rajgopal, Venkatachalam	AR	2004
Corporate Information	Change in taxes	Thomas and Zhang	JAR	2011
	Employment growth	Bazdresch, Belo and Lin	JPE	2014
	Industry concentration (sales)	Hou and Robinson	JF	2006
	Short-interest	Dechow et al.	JFE	2001
Corporate Liquidity	Asset liquidity over book assets	Ortiz-Molina and Phillips	JFQA	2014
	Asset liquidity over market	Ortiz-Molina and Phillips	JFQA	2014
	Change in current ratio	Ou and Penman	JAR	1989
	Change in quick ratio	Ou and Penman	JAR	1989
	Current ratio	Ou and Penman	JAR	1989
	Quick ratio	Ou and Penman	JAR	1989
Dividends	Last year's dividends over price	Naranjo, Nimalendran, Ryngaert	JF	1998
	Net payout yield	Boudoukh et al.	JF	2007
Earnings	Earnings announcement return	Chan, Jegadeesh and Lakonishok	JF	1996
	Earnings surprise	Foster, Olsen and Shevlin	AR	1984
	Earnings-to-price	Basu	JF	1977
External Financing	Change in current operating liabilities	Richardson et al.	JAE	2005
	Net debt financing	Bradshaw, Richardson, Sloan	JAE	2006
	Net equity financing	Bradshaw, Richardson, Sloan	JAE	2006
	Net external financing	Bradshaw, Richardson, Sloan	JAE	2006
	Secured debt	Valta	JFQA	2016
	Share issuance (1 year)	Pontiff and Woodgate	JF	2008
Investment	Asset growth	Cooper, Gulen and Schill	JF	2008
	Change in capex (two years)	Anderson and Garcia-Feijoo	JF	2006
	Change in current operating assets	Richardson et al.	JAE	2005
	Change in long-term investment	Richardson et al.	JAE	2005
	Change in net operating assets	Hirshleifer, Hou, Teoh, Zhang	JAE	2004
	Change in short-term investment	Richardson et al.	JAE	2005
	Growth in book equity	Lockwood and Prombutr	JFR	2010

Table 4.2 continued...

Economic category	Characteristic	Authors	Journal	Year
Leverage	Book leverage (annual)	Fama and French	JF	1992
	Operating leverage	Novy-Marx	ROF	2010
Momentum	52 week high	George and Hwang	JF	2004
	Industry momentum	Grinblatt and Moskowitz	JFE	1999
	Intermediate momentum	Novy-Marx	JFE	2012
	Long-run reversal	De Bondt and Thaler	JF	1985
	Medium-run reversal	De Bondt and Thaler	JF	1985
	Momentum (12 month)	Jegadeesh and Titman	JF	1993
	Momentum (6 month)	Jegadeesh and Titman	JF	1993
	Short-term reversal	Jegadeesh	JF	1989
Profitability	Cash-based oper. profitability	Ball et al.	JFE	2016
	Change in profit margin	Soliman	AR	2008
	Change in return on assets	Balakrishnan, Bartov and Faurel	JAE	2010
	Change in return on equity	Balakrishnan, Bartov and Faurel	JAE	2010
	Gross margin growth-to-sales growth	Abarbanell and Bushee	AR	1998
	Gross profits-to-total assets	Novy-Marx	JFE	2013
	Net income-to-book equity	Haugen and Baker	JFE	1996
	Profit margin	Soliman	AR	2008
	Return on assets (quart.)	Balakrishnan, Bartov and Faurel	JAE	2010
	Return on invested capital	Brown and Rowe	WP	2007
	Return on net operating assets	Soliman	AR	2008
Risk Measures	Coskewness	Harvey and Siddique	JF	2000
	Idiosyncratic risk	Ang et al.	JF	2006
	Maximum return over month	Bali, Cakici, and Whitelaw	JF	2010
Sales	Annual sales growth	Lakonishok, Shleifer, Vishny	JF	1994
	Change in gross margin-to-sales	Abarbanell and Bushee	AR	1998
	Sales-to-cash	Ou and Penman	JAR	1989
	Sales-to-price	Barbee, Mukherji and Raines	FAJ	1996
	Sales-to-receivables	Ou and Penman	JAR	1989
Trading	Amihud's illiquidity	Amihud	JFM	2002
	Bid-ask spread	Amihud and Mendelsohn	JFE	1986
	Days with zero trades (6 month avg.)	Liu	JFE	2006
	Number of analysts	Elgers, Lo and Pfeiffer	AR	2001
	Past trading volume	Brennan, Chordia, Subra	JFE	1998
Turnover	Asset turnover	Soliman	AR	2008
	Capital turnover	Haugen and Baker	JFE	1996
Valuation	Book-to-market (December ME)	Fama and French	JPM	1992
	Size	Banz	JFE	1981
	Total assets-to-market	Fama and French	JF	1992

security-date observations with a share of missing values greater than 50%. This threshold is chosen such that the number of missing values is significantly reduced, but overall only few observations are lost. The associated trade-off is illustrated in Figure 4.5: By setting the maximum share of missing characteristics to 50%, the number of security-date observations is reduced by only 4.13%.

Furthermore, Freyberger et al. (2020) have shown that there is an equivalence between cross-sectional regressions and classical portfolio sorts if the characteristics' values are scaled to the unit interval. For this reason, and to avoid problems associated with non-stationarities, it has become common practice in the literature to rank-transform characteristics cross-sectionally before using them as explanatory variables.[23] The factor models in Equation (4.19), however, do not necessarily map to the unit interval. Therefore, we introduce an additional step in which we apply the inverse normal cumulative distribution function to the rank-transformed characteristics. We then estimate the above factor models based on these normalized characteristics, impute their missing values, and reverse the normal transformation afterwards. In this way, we ensure that both the observed and imputed values fall into the range between zero and one.[24]

Table 4.3 provides detailed information on the final panel of returns and characteristics. Panel A gives an overview of the structure of our data, Panel B lists the number of securities per month for various time periods, and Panel C presents descriptive statistics for the implied market beta from Equation (4.2).

4.7.2 Results

One of the central implications of the conditional CAPM is that the market beta is the only stock-level quantity needed to explain the cross-section of returns. We examine this claim by running the selection procedure (Table 4.1, Step 1 to 9) twice: In the first run, we determine the set of characteristics, $\hat{\mathrm{S}}_{\mathrm{A}}$, without considering the implied beta (i.e., excluding step 6), while in the second run, we orthogonalize each characteristic with respect to beta. The latter ensures that a characteristic is selected only if it provides incremental information for the regression tasks from Equation (4.7). In both cases, we use 1,000 random partitions of the data to estimate the characteristics' conditional selection probabilities, $\mathbb{P}(\omega \in \hat{\mathrm{S}}_{\mathrm{A}} | \omega \in \hat{\Omega}_{\mathrm{A}})$, and use these estimates to evaluate the effects of orthogonalization. If the conditional CAPM holds, we would expect the characteristics to be selected much less frequently once we control for the test assets' exposure to market risk.

Figure 4.6 presents the characteristics' conditional selection probabilities for various sets of test assets, including individual stocks (Stocks), $25 \times 78 = 1{,}950$ univariate characteristic-sorted portfolios (CPF), and 500 beta-sorted portfolios (BPF). The lat-

[23] For example, Freyberger et al. (2020), Kozak et al. (2020), and Gu et al. (2020) employ such rank transformations.

[24] Figure C.2 in the appendix illustrates the cross-sections of observed and imputed values for the normalized characteristic *Asset growth* in January 31, 1996.

Figure 4.5: Distribution of missing characteristics. This figure shows the empirical distribution of the share of missing characteristics for the panel of security-date observations. It thus illustrates the trade-off associated with excluding observations whose share of missing characteristics exceeds a certain threshold. We set this threshold to 50%, which reduces the number of security-date observations from 695,884 to 667,113, i.e., by 4.13%.

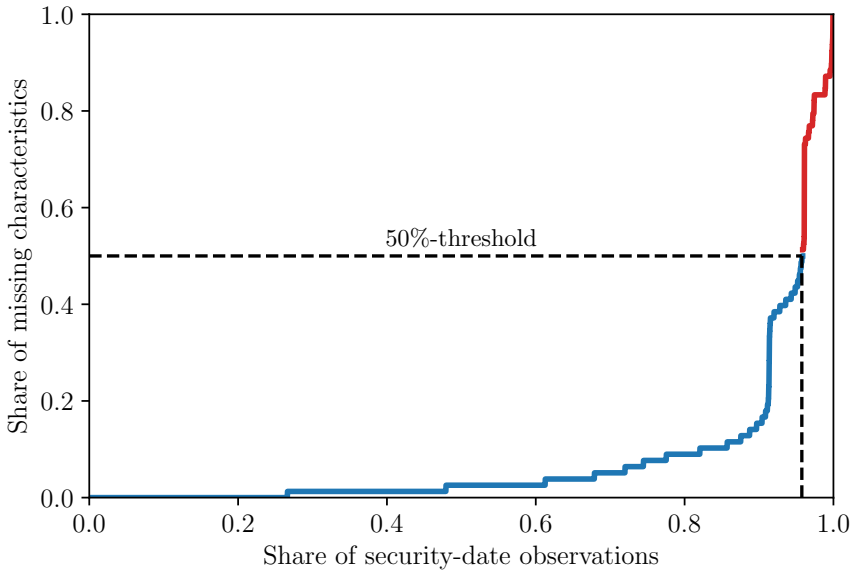


Table 4.3: Descriptive statistics. This table presents detailed information on the panel of returns and characteristics (Panel A), distributional measures for the number of securities per month, broken down by five time periods (Panel B), and averages, standard deviations, as well as a range of quantiles for the implied beta from Equation (4.2) (Panel C).

Panel A: General information

Sample period: January 1996 - December 2021
Frequency: Monthly
Investment horizon: 30 calendar days
Number of months: 311
Number of securities: 7,655
Number of security-date observations: 695,884
$\leq$ 50% missing characteristics: 667,113 (-4.13%)

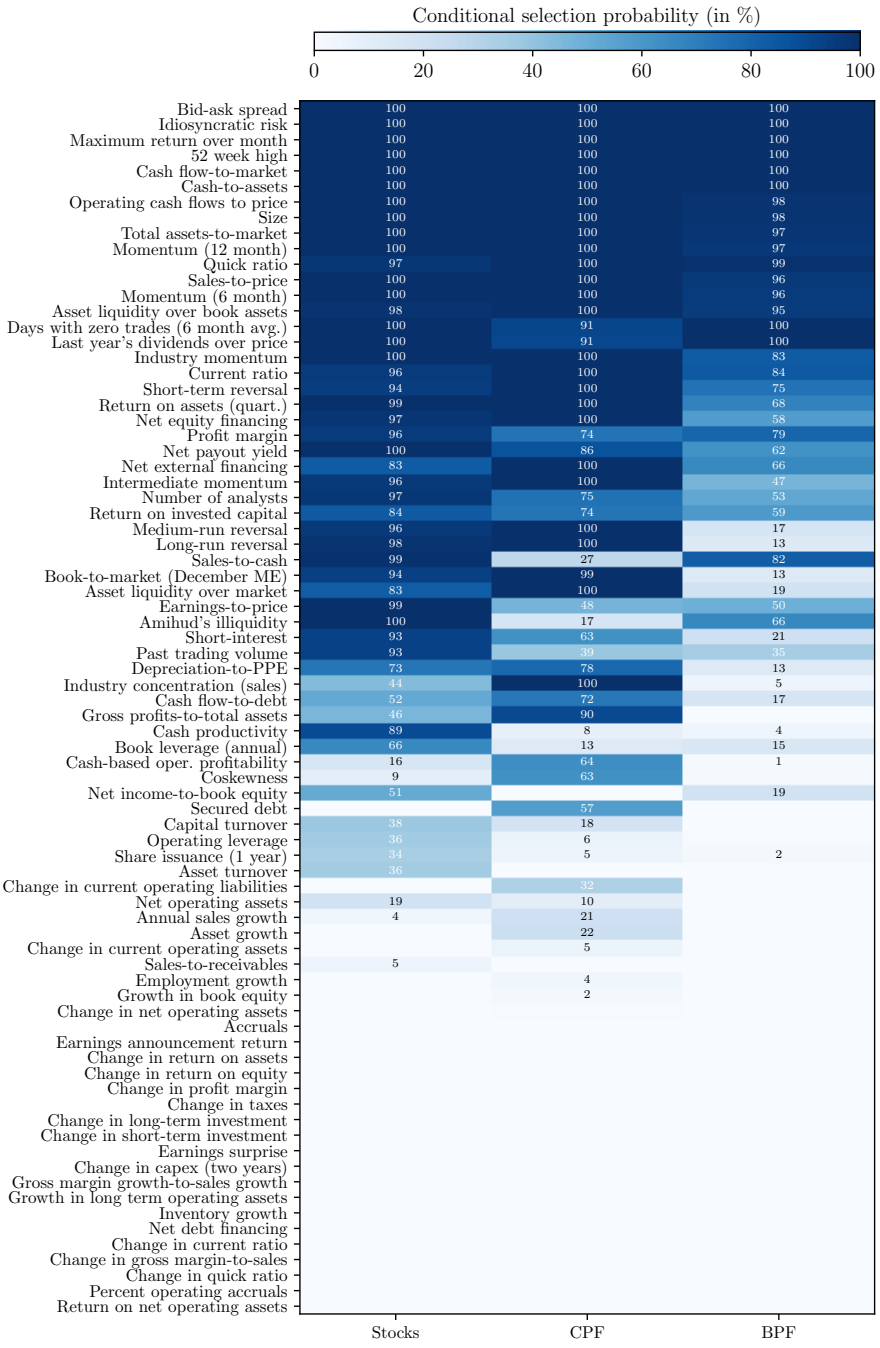
Panel B: Number of securities per month

Period	Avg.	Min.	Max.
1996 - 2000	2,060	1,537	2,316
2001 - 2005	1,889	1,760	2,041
2006 - 2010	2,170	2,000	2,251
2011 - 2015	2,455	2,265	2,655
2016 - 2021	2,555	2,343	2,799

Panel C: Descriptive statistics for the implied beta

Period	Avg.	Std.	10%	25%	50%	75%	90%
1996 - 2000	1.45	0.62	0.78	0.98	1.33	1.83	2.29
2001 - 2005	1.52	0.66	0.82	1.03	1.38	1.88	2.40
2006 - 2010	1.50	0.67	0.85	1.07	1.36	1.75	2.26
2011 - 2015	1.79	0.96	0.89	1.15	1.56	2.15	2.93
2016 - 2021	1.89	1.03	0.90	1.17	1.64	2.31	3.19

Figure 4.6: Conditional selection probabilities. This figure shows estimates of conditional selection probabilities (in %) for the 78 stock characteristics from Table 4.2, without orthogonalization. The estimates are obtained by counting the number of times that a characteristic is contained in the set of selected characteristics, $\hat{S}_A$, divided by the number of times it is drawn from one of the correlation clusters. The estimation is based on 1,000 random auxiliary sets, and the results are displayed for three different sets of test assets, including individual stocks (Stocks), $25 \times 78 = 1{,}950$ univariate characteristic-sorted portfolios (CPF), and 500 beta-sorted portfolios (BPF). The characteristics are arranged in descending order according to their average selection probabilities across columns.



ter two are formed on a monthly basis by sorting stocks into portfolios according to the characteristics' cross-sectional distributions. The characteristics are displayed in descending order by their average selection probabilities across columns.

In the case of individual stocks, we observe that 29 characteristics have conditional selection probabilities greater than 95% – a number that is relatively large compared to the 13 characteristics identified as important in the study by Freyberger et al. (2020). This discrepancy can be attributed to the fact that Freyberger et al. (2020) apply the group Lasso only once to their entire data set, whereas we perform the selection step repeatedly using different subsamples of our data. As discussed in Section 4.5.1, the latter serves to account for the instability of ℓ_1-regularization in the presence of highly correlated predictors. This instability is due to the fact that the solution to the selection problem is inherently ambiguous, in the sense that many different combinations of selected characteristics produce very similar levels of predictive accuracy. Therefore, it is crucial to rely on multiple sets of selected characteristics to obtain a reliable estimate of the number of informative characteristics.

Considering the results for the CPF and the BPF, we find that broadly the same characteristics are important as in the case of individual stocks, although the number of characteristics with conditional selection probabilities greater than 95% decreases from 29 to 26 for the CPF and to 16 for the BPF. This result suggests that certain aspects of return variation are lost when forming portfolios. Another interesting observation is that *Idiosyncratic risk* is always among the important characteristics, no matter which set of test assets is considered. Given that one of the central tenets of the CAPM is that idiosyncratic risk is not priced by investors, we would expect to see a sharp decline in the selection probability for this characteristic in particular.

The changes in conditional selection probabilities resulting from the use of orthogonalized characteristics are shown in Figure 4.7. What is striking is that many of the characteristics that were previously identified as important are now selected with a much lower frequency. In the case of individual stocks, the number of characteristics with conditional selection probabilities greater than 95% drops from 29 to 20 once we introduce orthogonalization. However, there are some characteristics that retain their high selection probabilities even after adjusting for beta, most of which are associated with the momentum effect (e.g., *Industry Momentum*). Interestingly, when comparing the results for individual stocks and portfolios, it appears that far more predictors are redundant in the latter case: The number of characteristics with conditional selection probabilities greater than 95% decreases from 26 to 7 for the CPF, and from 16 to 4 for the BPF. To give an example: *Idiosyncratic risk* is (almost) completely driven out in the case of the BPF and the CPF, but affected to a much lesser extent when using individual stocks. The heterogeneity of these results suggests that explaining the returns of individual stocks is more challenging for the conditional CAPM than explaining the returns of characteristic-sorted portfolios.

This conclusion is confirmed in Figure 4.8, which shows the distribution of the number of selected characteristics across partitions (Panel A), as well as the distribu-

Figure 4.7: Changes in conditional selection probabilities. This figure shows the changes in conditional selection probability estimates (in %-points) that are due to orthogonalizing the characteristics with respect to the implied beta from Equation (4.2). The estimation is based on 1,000 random auxiliary sets, and the results are displayed for three different sets of test assets, including individual stocks (Stocks), $25 \times 78 = 1{,}950$ univariate characteristic-sorted portfolios (CPF), and 500 beta-sorted portfolios (BPF). For ease of comparison, the characteristics are arranged in the same way as in Figure 4.6.

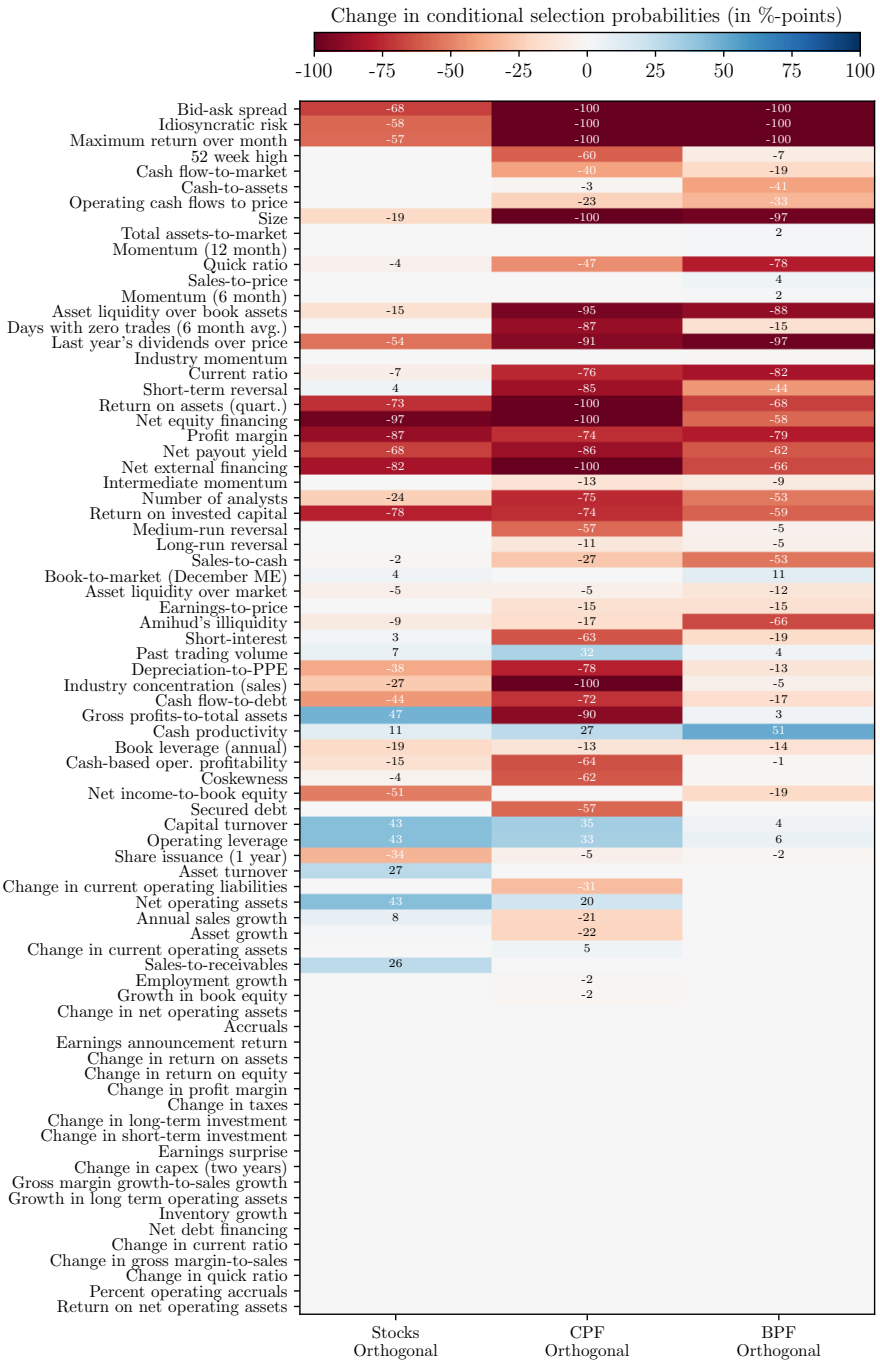
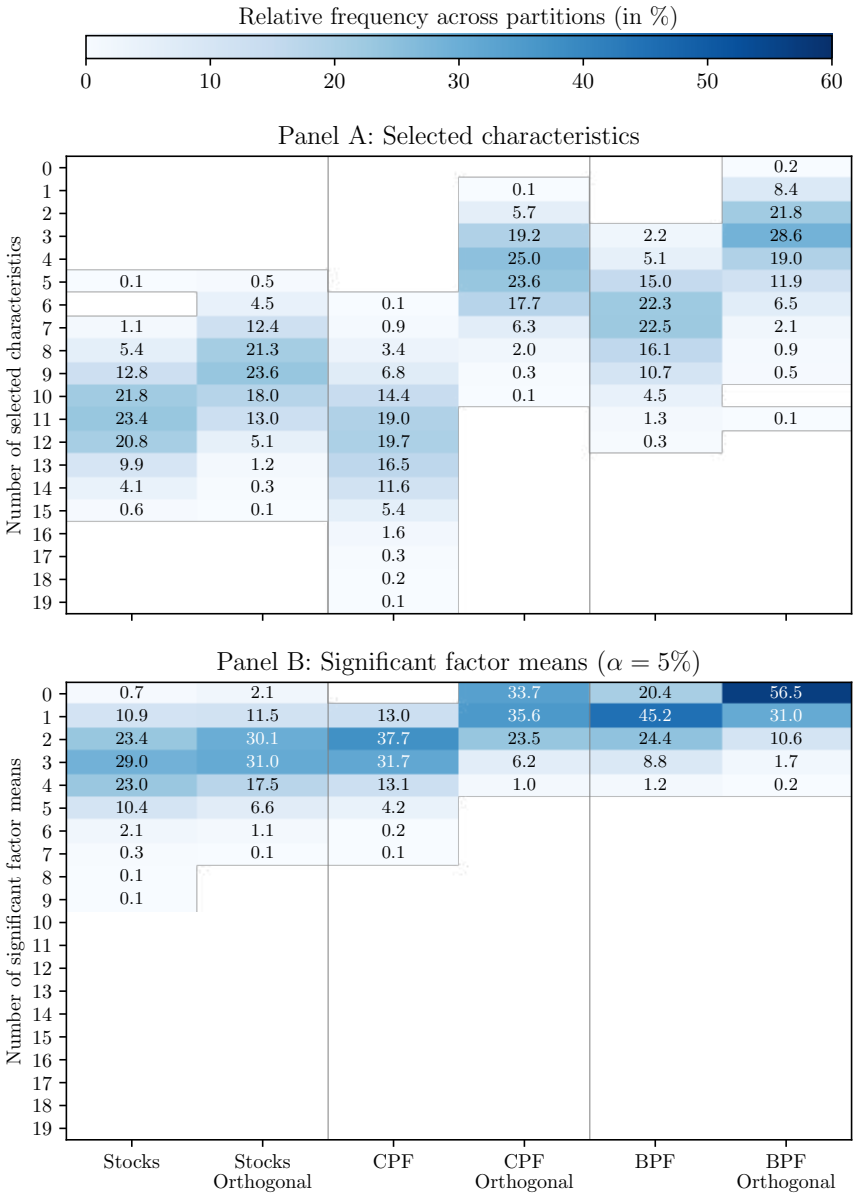


Figure 4.8: Numbers of selected characteristics and significant factor means. This figure shows the distribution of the number of selected characteristics (Panel A), as well as the distribution of the number of factor means that are significant at a 5% level for different sets of test assets. The first two columns in either panel present the results for individual stocks, once for the case where the characteristics are orthogonalized and once without orthogonalization. Columns three to six display the results for characteristic-sorted portfolios (CPF) and beta-sorted portfolios. The significance of the factor means is determined for each of the 1,000 main sets individually using a standard t-test.



tion of the number of factor means that are significant at a 5% level (Panel B), both with and without orthogonalization.[25] As can be seen from the first two columns of Panel A, introducing beta only slightly shifts the distribution of the number of selected characteristics in the case of individual stocks. In contrast, the shift is much more pronounced for the CPF (BPF), where the mode of the distribution declines from 12 characteristics to 4 (from 7 to 3). Similarly, the number of significant factor means is hardly affected for individual stocks (the mode of the distribution is 3 in both cases), but drawn towards zero for both the CPF and the BPF. We conclude that, while the CAPM is quite successful in driving out characteristics in the case of portfolios, the results for individual stocks suggest that additional stock-level features may be needed to capture all aspects of return variation.

We continue with a discussion of the results obtained from testing the joint hypothesis in Equation (4.11). Table 4.4 presents quantile-aggregated p-values for the Wald statistic, once for the case where the regression specifications do not contain the market beta (Equation 4.4), and once for the case where the market beta is included (Equation 4.7). In the case of individual stocks, the p-value associated with the hypothesis that the additional factors derived from cross-sectional regressions have zero means is below any conventional significance level, leading us to reject the conditional CAPM. This is consistent with the previous observation that the implied beta is unable to drive out all the competing stock characteristics in cross-sectional regressions. The results are less conclusive, however, in the case of portfolios: While the rejection for the CPF is borderline ($p_{\text{med}} = 0.02$), we cannot reject the conditional CAPM for the BPF at any conventional level ($p_{\text{med}} = 0.18$). Again, these results are in line with the observation that, especially in the case of the BPF, many predictors do not provide incremental information.

For comparison, we additionally conduct this test for the model specifications in Equation (4.4), which do not account for the market beta. As can be seen from the results in Table 4.4, the level of the p-values strongly depends on whether the market beta is included: For the BPF, the p-value is 0.01, which is much lower than the corresponding p-value of 0.18. Similarly, the p-value for the CPF drops from 0.02 to a value that is below any conventional level. These findings substantiate our impression that, although the conditional CAPM appears to be quite successful in describing the returns of portfolios, the returns of individual stocks are better explained by a multi-factor model.

Next, we address the question of where the implied beta ranks in terms of its cross-sectional explanatory power relative to the other characteristics. To this end, we employ the characteristics' median Shapley contributions to cross-sectional R^2_{A}, which are displayed in Figures 4.9 to 4.11 for individual stocks, the CPF, and the BPF. Strikingly, the implied beta is by far the most important predictor of return variation, regardless of whether we use individual stocks or portfolios as test assets. In

[25] We test for significant factor means using a standard t-test that is performed for each of the 1,000 main sets.

Figure 4.9: Shapley contributions to R^2_{A} (Stocks). This figure shows the individual characteristics' Shapley contributions to R^2_{A} as defined in Equation (4.17) and obtained when using individual stocks as test assets. The underlying decomposition is performed for each of the 1,000 random partitions of the data. The red circles correspond the median contributions across partitions, while the blue bars indicate the range between the highest and lowest deciles. The characteristics are sorted in descending order according to their relative importance.

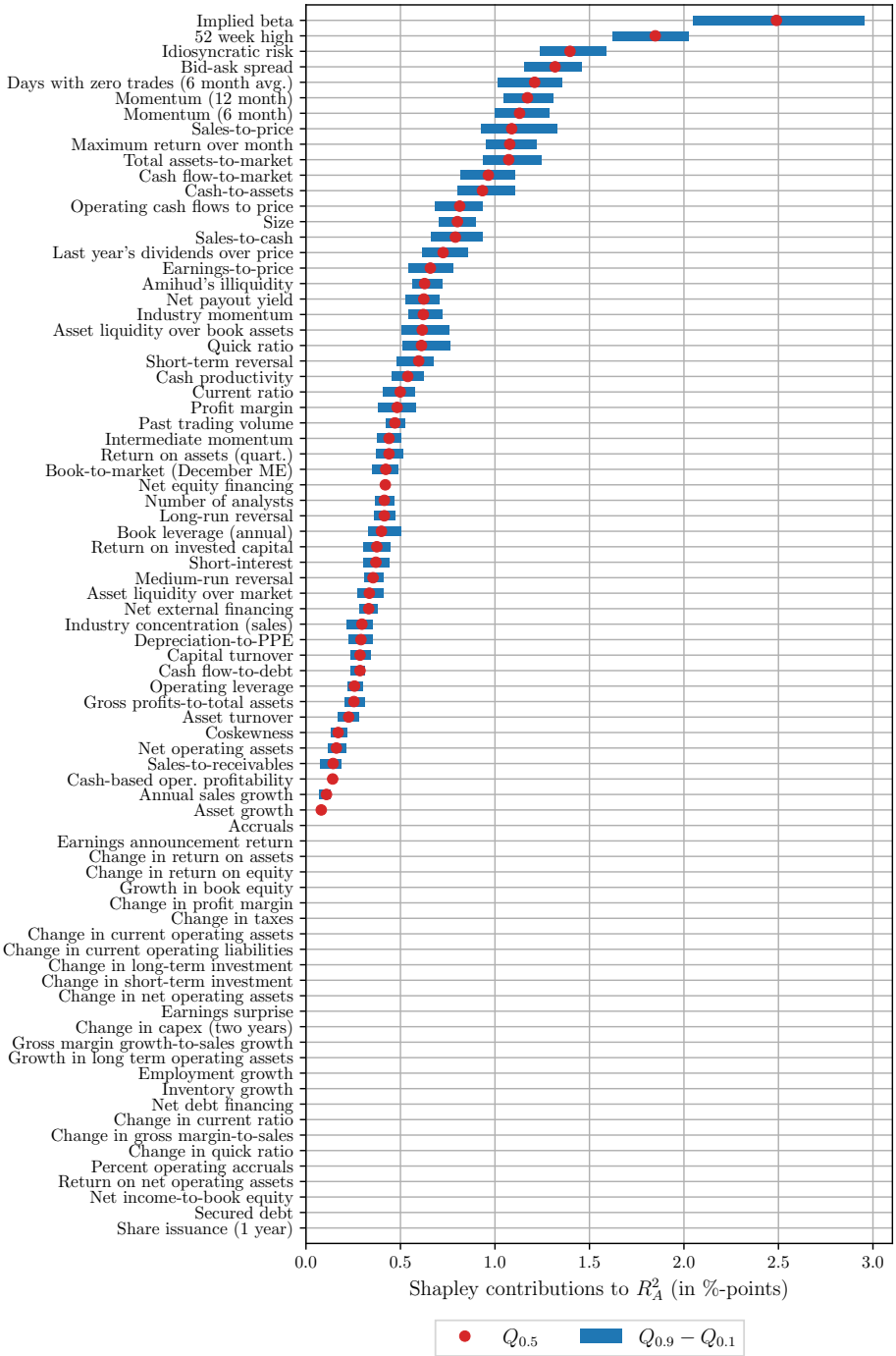


Figure 4.10: Shapley contributions to R^2_A (CPF). This figure shows the individual characteristics' Shapley contributions to R^2_A as defined in Equation (4.17) and obtained when using characteristic-sorted portfolios (CPF) as test assets. The underlying decomposition is performed for each of the 1,000 random partitions of the data. The red circles correspond the median contributions across partitions, while the blue bars indicate the range between the highest and lowest deciles. The characteristics are sorted in descending order according to their relative importance.

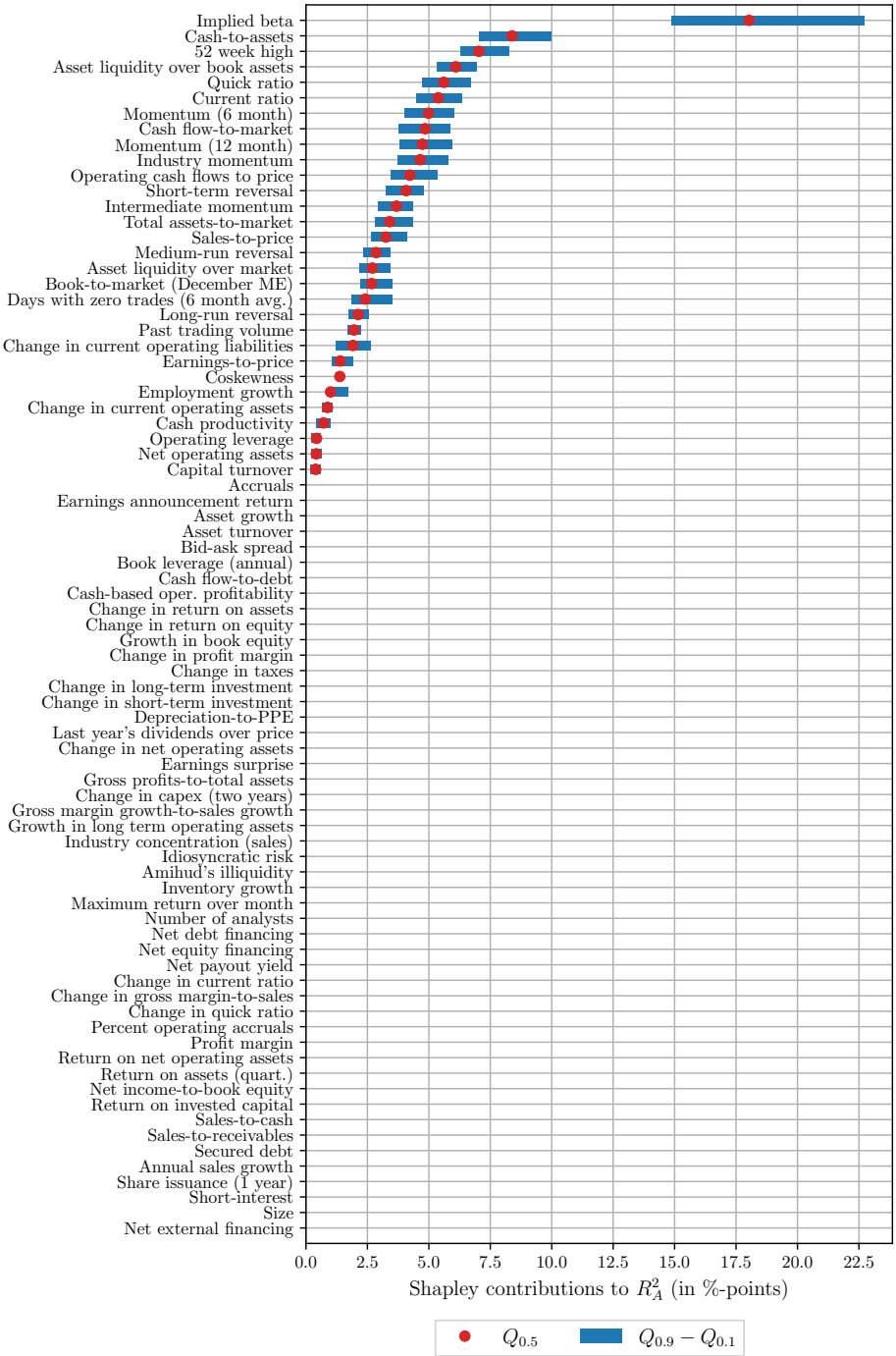
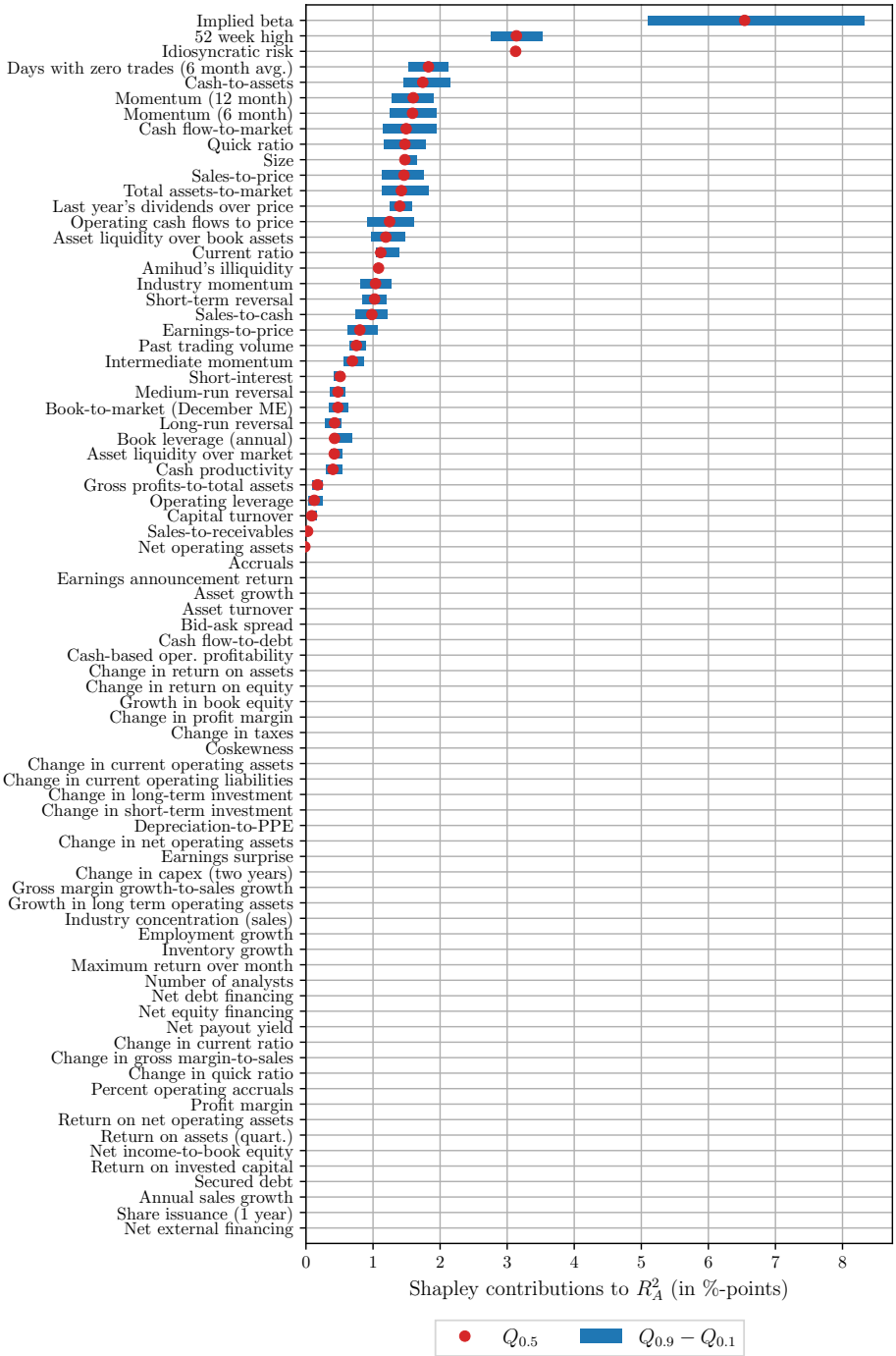


Figure 4.11: Shapley contributions to R^2_{A} (BPF). This figure shows the individual characteristics' Shapley contributions to R^2_{A} as defined in Equation (4.17) and obtained when using beta-sorted portfolios (BPF) as test assets. The underlying decomposition is performed for each of the 1,000 random partitions of the data. The red circles correspond the median contributions across partitions, while the blue bars indicate the range between the highest and lowest deciles. The characteristics are sorted in descending order according to their relative importance.



the case of individual stocks, its median contribution to the R^2_{A} is about 2.5 percentage points, which is half a percentage point higher than that of the second most important characteristic (*52 week high*). This outperformance is even more pronounced for the CPF and the BPF, where in both cases the median contribution of the implied beta is more than twice that of the second most important characteristic. Given these results, it seems reasonable to assume that the cross-sectional factor models presented by Fama and French (2020), which do not account for the assets' exposure to market risk, could have been further improved by adding a time-varying estimate of beta.

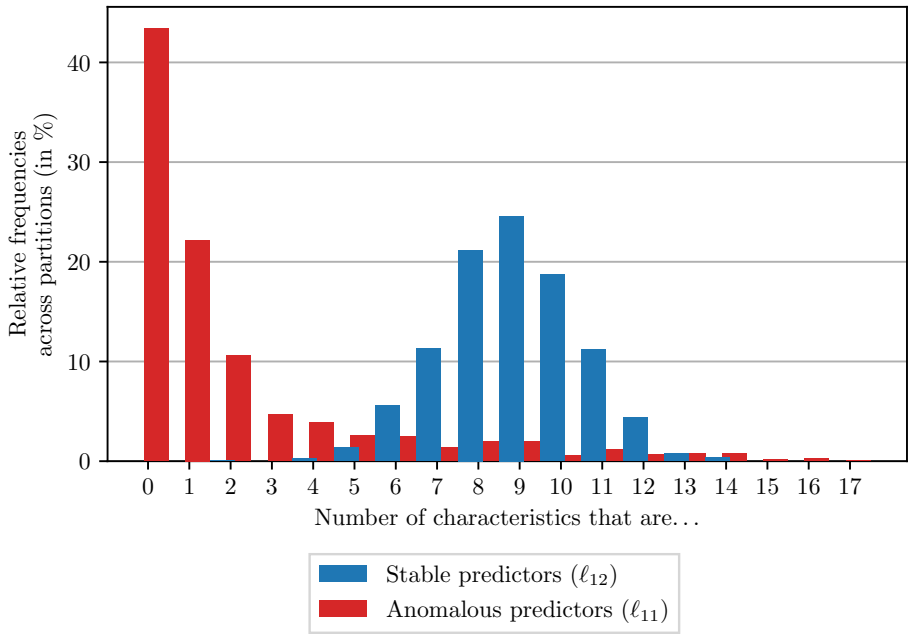
Finally, we investigate whether there are stock characteristics that generate anomalous return predictability. We address this question using the DMTL from Equation (4.14), which comprises two different regularization terms: One of these terms, the ℓ_{12}-penalty, enforces a joint sparsity pattern along the time series dimension of the coefficient matrix $\boldsymbol{F}$, while the other, the ℓ_{11}-penalty, shrinks the values of the coefficient matrix $\boldsymbol{A}$ for each cross-section individually. Hence, we consider characteristics anomalous predictive signals if their coefficient time series in $\boldsymbol{F}$ are zero (i.e., if they do not qualify as stable predictors), but the corresponding coefficients in $\boldsymbol{A}$ are nonzero for at least some of the regression tasks (i.e., if they are informative only within certain periods of time). As before, we account for the instability of the DMTL by solving the corresponding optimization problem for each of the 1,000 auxiliary sets and then counting how many of the characteristics are classified as either stable or anomalous return predictive signals. The resulting distributions are displayed in Figure 4.12.

As evident from the results, the number of characteristics exhibiting anomalous return predictability is rather small compared to the number of stable factor-generating characteristics. In more than 40% of the solutions, the DMTL shrinks all the coefficients in $\boldsymbol{A}$ to zero, indicating that none of the characteristics are classified as anomalous predictive signals. In contrast, in over 60% of the cases, the DMTL identifies between 8 and 10 characteristics as stable predictors. These findings align closely with the conclusions drawn by Kelly et al. (2019), who assert that characteristics "contain little (if any) anomalous return predictability once their explanatory power for factor exposures has been accounted for."

Table 4.4: Results from testing the conditional CAPM. This table presents quantile-aggregated p-values, p_{med}, for testing the joint hypothesis in Equation (4.11). The rows contain results for different sets of test assets, including individual stocks (Stocks), characteristic-sorted portfolios (CPF), and beta-sorted portfolios (BPF). The columns are used to distinguish between the regression specifications in Equations (4.4) and (4.7). The two are different in that the former does not include the market beta (*Without beta*), while the latter does (*With beta*). The quantile-aggregation step is performed as described in Section 4.6 using 1,000 random partitions of the data.

	Quantile-aggregated p-values (p_{med})	
	Without beta	With beta
Stocks	<0.001	<0.001
CPF	<0.001	0.023
BPF	0.013	0.184

Figure 4.12: Stable vs. anomalous return predictability. This figure shows the distributions of the number of characteristics that are classified as either stable or anomalous return predictive signals. The distributions are obtained by solving the DMTL optimization problem from Equation (4.14) for each of the 1,000 auxiliary sets. The objective function that is minimized is subject to two different types of regularization terms, denoted by ℓ_{12} and ℓ_{11}, where the former enforces a joint sparsity pattern across regression tasks, while the latter operates at the level of each individual cross-section. The distinction between stable and anomalous return predictors is derived from the coefficient matrices that are associated with the two penalty terms. Accordingly, characteristics are considered stable predictors if they are assigned nonzero coefficient time series in the columns of $\boldsymbol{F}$, and anomalous return predictors if their column entries in $\boldsymbol{F}$ are zero, but nonzero in $\boldsymbol{A}$ for at least some of the regression tasks.



4.8 Conclusions

The conditional CAPM is one of the most widely studied models in empirical finance. Despite this extensive research, however, there is still untapped potential for improvement, both in terms of economic theory and testing methodology. This study aims to fill some of the remaining gaps related to the measurement of the model's time-varying components and their evaluation in cross-sectional regressions.

In a first step, we propose a novel representation of the conditional CAPM that is fully-implied by option prices, allowing us to compute both the beta and the market premium without any econometric estimation. An interesting feature of this representation is that it directly relates physical and risk-neutral return distributions without the need for further risk-adjustment.

In a second step, we test our model using cross-sectional regressions that include the implied beta and other stock characteristics as regressors. Although the underlying testing principle is not new, we contribute to the existing literature in at least two respects: First, we select competing characteristics using the multi-task Lasso, which allows us to determine a joint representation of the regressions over the entire sample period, while taking into account that each of them is subject to a time-specific parameterization. Second, we obtain misspecification-robust post-selection inference by performing the selection and estimation steps separately on multiple random partitions of the data. In addition, we employ conditional selection probabilities and Shapley decompositions to assess the incremental contribution of characteristics beyond beta, as well as an extension of the multi-task Lasso that serves to distinguish between stable and anomalous return predictive signals.

Our findings are as follows: Although many characteristics appear to be useful for explaining the cross-section of returns, only few provide incremental information beyond beta. This is reflected in a sizable reduction in the characteristics' conditional selection probabilities once we include the beta in cross-sectional regressions. Depending on the chosen set of test assets, the few remaining characteristics are either too important to be ignored, as in the case of individual stocks, or not important enough for the conditional CAPM to be rejected, as in the case of characteristic- and beta-sorted portfolios. One observation that is robust to the choice of test assets is that no other characteristic contributes as much to explaining the cross-section of returns as the implied market beta, as evidenced by the fact that its contribution to the explained variation exceeds that of all the other characteristics by orders of magnitude. Finally, we find that most of the characteristics that have been found to be informative in the previous literature are stable rather than anomalous predictors of returns, which can be taken as evidence against the often claimed replicability crisis in empirical finance.

C Appendix

C.1 Post-selection inference via sample splitting: Simulation study

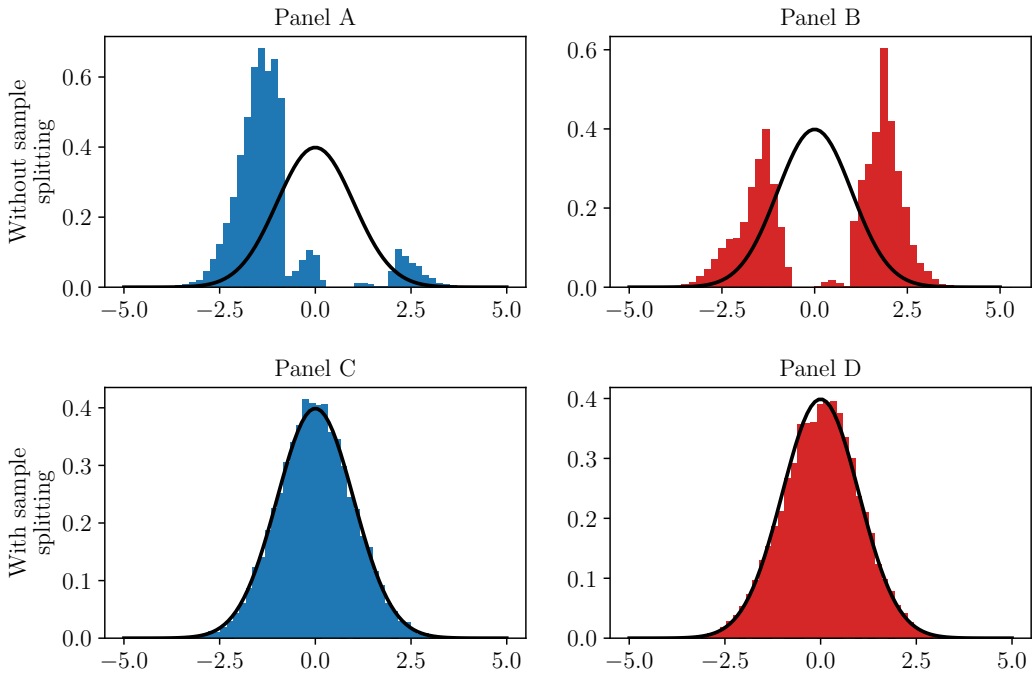
To provide some intuition for the sample-splitting approach, we briefly consider the following simulation: Suppose there is a single cross-section of returns that is governed by a linear model. This linear model consists of two correlated characteristics ($\rho = 0.8$), the parameter values of which are given by $\boldsymbol{\theta}_0 = [-0.1, 0.05]$. From this model, we generate 100,000 samples of returns and characteristics, each of size $N = 10{,}000$. For each of these samples, we employ the Lasso to determine the best linear specification (including either a constant regressor, one of the two characteristics, or both characteristics) and estimate the associated model parameters. However, we distinguish between two scenarios: In the first scenario, the selection and estimation are both performed on the full sample of N observations. In the second scenario, we split the data in half, use the first half for selection, and the second half for estimation as well as inference. We then run post-selection regressions for each scenario and collect the resulting parameter estimates. Note, however, that these estimates do not target the true model parameters, $\boldsymbol{\theta}_0$, but rather the OLS projection coefficients that are defined conditional on the sets of selected characteristics, $\boldsymbol{\theta}_{\hat{\mathrm{S}}}$ and $\boldsymbol{\theta}_{\hat{\mathrm{S}}_{\mathrm{A}}}$, where the former refers to the full-sample scenario and the latter to the scenario in which we use sample-splitting.

Figure C.1 shows the approximate null distributions of the t-statistic for each characteristic, once with and once without sample splitting. In addition, we construct a Wald statistic for the joint hypothesis that the OLS projection coefficients are equal to their population values and estimate the probabilities with which the associated p-values fall below the significance level of $\alpha = 5\%$. If this probability exceeds α, the test statistic is subject to size distortions, meaning that the probability of rejecting a correct null hypothesis is too high.

Panels A and B show that, without sample splitting, the null distributions of the t-statistics are not approximately Gaussian as expected, but distorted in such a way that we too often commit a type I error. Similar results are obtained for the joint hypothesis test, where the probability for rejecting a correct H_0 is substantially higher (≈ 0.26) than the targeted significance level of 5%. Note, however, that these deviations are not small-sample phenomena: The sample comprises $N = 10{,}000$ observations.

Panels C and D tell a different story: If we separate the selection and estimation steps across different subsamples, the approximate distributions are no longer distorted. Likewise, the probability for rejecting a correct joint hypothesis (≈ 0.05) does not exceed the targeted significance level. These results indicate that the sample-splitting approach is effective in counteracting the bias that is induced by selection.

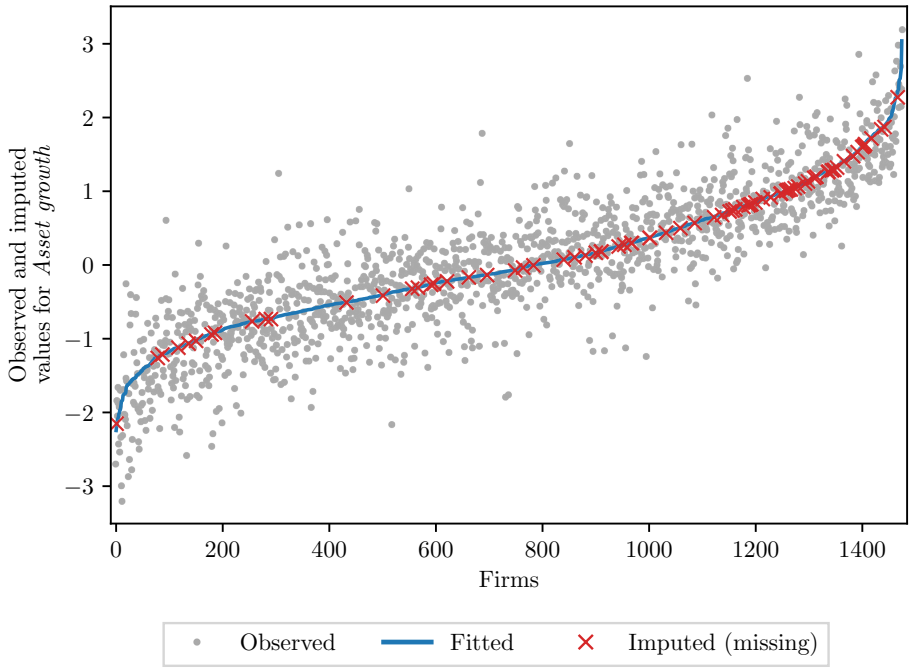
Figure C.1: Post-selection inference with(out) sample splitting (Simulation). This figure shows approximate null distributions of the t-statistic, once with (Panels A and B) and once without sample splitting (Panels C and D). The results are obtained by generating a single cross-section of $N = 10{,}000$ returns based on a linear model comprising two positively correlated characteristics with $\rho = 0.8$. The true model parameters are given by $\boldsymbol{\theta}_0 = [-0.1, 0.05]$. We then employ the Lasso to select characteristics, once using the full sample (Panels A and B) and once using half of the data (Panels C and D). After that, we use the full sample (the other half of the sample) to estimate the OLS projection parameters $\boldsymbol{\theta}_{\hat{\mathrm{S}}}$ ($\boldsymbol{\theta}_{\hat{\mathrm{S}}_\mathrm{A}}$), which are defined conditional on the set of selected characteristics. Finally, we repeat the sampling, selection, and estimation steps 100,000 times to approximate the null distributions of the associated t-statistics. Panels A and C (blue color) present the distributions for the first characteristic, whereas Panels B and D (red color) show the distributions for the second characteristic. The solid black lines represent the probability density of a standard normal random variable.



C.2 Additional figures

Figure C.2 shows observed and imputed values for the characteristic *Asset growth* for the cross-section of firms available in January 31, 1996. The imputed values are obtained using the imputation procedure by Bryzgalova et al. (2022).

Figure C.2: Examples of observed and imputed values. This figure shows the observed and imputed values of *Asset growth* for the cross-section of firms available in January 31, 1996. For the imputation, we employ the procedure by Bryzgalova et al. (2022), which operates on the panel of rank-transformed characteristics. In addition, we normalize the characteristics' values to ensure that the factor model in Equation (4.19) is unconstrained in its mapping. After the imputation, we undo this normalization so that the characteristics values again lie in the unit interval. The individual firms are displayed on the horizontal axis and arranged in ascending order according to their fitted values. The grey dots indicate the observed values, the blue line represents the fitted values, and the red crosses are the imputed values for which the characteristic's value is missing.



Chapter 5
Conclusions

This dissertation examines various aspects of two (currently) separate strands of financial economic research that are concerned with the quantification of conditional stock risk premia. On the one hand, there is the statistical approach pursued by Gu et al. (2020), which adopts a predictive perspective and attempts to approximate conditional risk premia using flexible, highly parameterized statistical models that are evaluated on the basis of their ability to predict future realized excess returns. On the other hand, there is a theory-based approach that aims to recover conditional risk premia from financial economic paradigms by exploiting the information contained in option prices, as is done, for example, by Martin and Wagner (2019). The central theme of this dissertation is to compare these two approaches, to thoroughly assess their respective strengths and weaknesses, and to identify situations in which combining the two can be useful. In pursuit of these goals, particular emphasis is placed on the integration of financial economic paradigms into statistical modeling in order to return to the principles of econometrics at a time when machine learning methods are gaining in popularity.

In Chaper 1, we act on this maxim by proposing a hybrid approach that aims to capture the approximation errors induced by the structural assumptions employed by Martin and Wagner (2019) through various machine learning methods. In this way, we explore the limitations of their theory and assess the added value of incorporating economic considerations into statistical measurements of conditional risk premia. Our main finding is that the relative performance of the two strategies strongly depends on the investment horizon: The theory-based approach seems to benefit from the informational efficiency of options markets at shorter investment horizons, while the machine-learning approach can take advantage of the signals embedded in a variety of stock- and macro-level features at longer investment horizons.

In Chapter 2, we draw on the findings by Martin and Wagner (2019) to derive a representation of the conditional CAPM in which both the beta and the equity premium are fully-implied by option prices. The comparative advantage of our approach lies in its intuitive exposition of the risk-return relationship, as well as its superior predictive performance. We employ this novel representation to shed light on a well-known phenomenon that has preoccupied economic research for decades: the flat relationship between average returns and betas. Our contribution is a novel interpretation of this phenomenon, according to which the failure of the conditional CAPM can be attributed to an inverse relationship between two different types of uncertainties – one related to the inherently unpredictable component of market excess returns, and the

other related to forecasting the assets' exposure to market risk.

In Chapter 3, we aim to improve the way cross-sectional regressions are used to test the validity of dynamic factor models. Our methodological contribution consists of two parts: First, we present a systematic approach to the selection of stock characteristics based on a variant of block-norm regularization that takes into account that the individual regression problems are interrelated. Second, we provide valid post-selection inference by performing the selection and estimation steps separately on multiple random subsamples of the data, thus avoiding restrictive assumptions about the return-generating process. We apply this testing strategy to evaluate the conditional CAPM presented in Chapter 2, and find that while the model is rejected when using individual stocks as test assets, the implied beta is by far the most important predictor of cross-sectional return variation.

One lesson we believe is important for future work is that these two strands of research should no longer be perceived as antithetical. While many studies currently refer to either the one or the other approach, this barrier needs to be overcome to make room for a productive convergence of the two. We suspect that the lack of progress in this regard is partly due to the misnomer *machine learning*, which fails to recognize that the methods subsumed under this category are nothing more than a logical evolution of classical statistical approaches that have always played an important role in empirical finance research. Rather than viewing machine learning as an innovation that renders theoretical considerations obsolete, we should explore how its principles can be integrated into financial economic theory, especially when large amounts of data are involved or considerable flexibility is demanded.

Equally important as incorporating high-dimensional statistical methods into the econometric repertoire is recognizing how crucial theoretical contributions such as that by Martin and Wagner (2019) are in shaping our understanding of the world. The statistical methods currently available are rather limited in this respect, as their strength lies in exploiting statistical relationships rather than discovering fundamental economic principles. Thus, for the foreseeable future, the most fruitful application of statistical methods will continue to be in making theoretical propositions amenable to empirical analysis. Whether statistical methods will ever be able to discover equilibrium or causal relationships on their own, without recourse to expert knowledge, is a question that we must leave to future research.

Bibliography

Amihud, Y. (2002): "Illiquidity and Stock Returns: Cross-section and Time-series Effects," *Journal of Financial Markets*, 5(1), 31–56.

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

Avramov, D., S. Cheng, and L. Metzker (2023): "Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability," *Management Science*, 69(5), 2587–2619.

Bakshi, G., J. Crosby, X. Gao, and W. Zhou (2020): "A New Formula for the Expected Excess Return of the Market," Working Paper.

Bakshi, G., and D. Madan (2000): "Spanning and Derivative-Security Valuation," *Journal of Financial Economics*, 55(2), 205–238.

Baule, R., O. Korn, and S. Sassning (2016): "Which Beta Is Best? On the Information Content of Option-implied Betas," *European Financial Management*, 22(3), 450–483.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81(2), 608–650.

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

Black, F. (1972): "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business*, 45(3), 444–455.

Blume, M. E., and I. Friend (1973): "A New Look at the Captial Asset Pricing Model," *Journal of Finance*, 28(1), 19–34.

Bollerslev, T., G. Tauchen, and H. Zhou (2009): "Expected Stock Returns and Variance Risk Premia," *Review of Financial Studies*, 22(11), 4463–4492.

Boudoukh, J., M. Richardson, and T. Smith (1993): "Is the Ex Ante Risk Premium Always Positive?: A New Approach to Testing Conditional Asset Pricing Models," *Journal of Financial Economics*, 34(3), 387–408.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (2006): "The Myth of Long-Horizon Predictability," *Review of Financial Studies*, 21(4), 1577–1605.

Breeden, D. T., and R. H. Litzenberger (1978): "Prices of State-contingent Claims Implicit in Option Prices," *Journal of Business*, 51(4), 621–651.

Brennan, M. J., A. W. Wang, and Y. Xia (2004): "Estimation and Test of a Simple Model of Intertemporal Capital Asset Pricing," *Journal of Finance*, 59(4), 1743–1775.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2022): "Missing Financial Data," Working Paper.

Bryzgalova, S., M. Pelger, and J. Zhu (2021): "Forest Through the Trees: Building Cross-sections of Stock Returns," Working Paper.

Buss, A., and G. Vilkov (2012): "Measuring Equity Risk with Option-implied Correlations," *Review of Financial Studies*, 25(10), 3113–3140.

Campbell, J. Y., and R. J. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1(3), 195–228.

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

Chabi-Yo, F., C. Dim, and G. Vilkov (2023): "Generalized Bounds on the Conditional Expected Excess Return on Individual Stocks," *Management Science*, 69(2), 922–939.

Chabi-Yo, F., and J. Loudis (2020): "The Conditional Expected Market Return," *Journal of Financial Economics*, 137(3), 752–786.

Chan, L. K., J. Karceski, and J. Lakonishok (1999): "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model," *Review of Financial Studies*, 12(5), 937–974.

Chang, B.-Y., P. Christoffersen, K. Jacobs, and G. Vainberg (2012): "Option-Implied Measures of Equity Risk," *Review of Finance*, 16(2), 385–428.

Chen, A. Y., and T. Zimmermann (2022): "Open Source Cross-sectional Asset Pricing," *Critical Finance Review*, 11(2), 207–264.

Chen, L., M. Pelger, and J. Zhu (2023): "Deep Learning in Asset Pricing," *Management Science*, 0(0).

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21(1), C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2023): "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India," Working Paper 24678, National Bureau of Economic Research.

Chordia, T., A. Goyal, and J. A. Shanken (2019): "Cross-sectional Asset Pricing with Individual Stocks: Betas versus Characteristics," Working Paper.

Cochrane, J. H. (2005): *Asset Pricing: Revised Edition*. Princeton University Press.

——— (2008): "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21(4), 1533–1575.

——— (2011): "Presidential Address: Discount Rates," *Journal of Finance*, 66(4), 1047–1108.

Da, R., S. Nagel, and D. Xiu (2022): "The Statistical Limit of Arbitrage," Working Paper.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers (1997): "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *Journal of Finance*, 52(3), 1035–1058.

Driscoll, J. C., and A. C. Kraay (1998): "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics*, 80(4), 549–560.

Fama, E. F. (1976): *Foundations of Finance*. Basic Books: New York.

Fama, E. F., and K. R. French (1988): "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 22(1), 3–25.

——— (1989): "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, 25(1), 23–49.

——— (1992): "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2), 427–465.

——— (2004): "The Capital Asset Pricing Model: Theory and Evidence," *Journal of Economic Perspectives*, 18(3), 25–46.

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

——— (2020): "Comparing Cross-Section and Time-Series Factor Models," *Review of Financial Studies*, 33(5), 1891–1926.

Fama, E. F., and J. D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81(3), 607–636.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, 75(3), 1327–1370.

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

Frazzini, A., and L. H. Pedersen (2014): "Betting Against Beta," *Journal of Financial Economics*, 111(1), 1–25.

French, D. W., J. C. Groth, and J. W. Kolari (1983): "Current Investor Expectations and Better Betas," *Journal of Portfolio Management*, 10(1), 12–17.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2022): "Missing Data in Asset Pricing Panels," Working Paper 30761, National Bureau of Economic Research.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, 33(5), 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet (2016): "Time-Varying Risk Premium in Large Cross-Sectional Equity Data Sets," *Econometrica*, 84(3), 985–1046.

Giglio, S., B. Kelly, and D. Xiu (2022): "Factor Models, Machine Learning, and Asset Pricing," *Annual Review of Financial Economics*, 14(1), 337–368.

Giglio, S., and D. Xiu (2021): "Asset Pricing with Omitted Factors," *Journal of Political Economy*, 129(7), 1947–1990.

Grammig, J., C. Hanenberg, C. Schlag, and J. Sönksen (2022): "Diverging Roads: Theory-based vs. Machine Learning-implied Stock Risk Premia," Working Paper.

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

Grömping, U. (2007): "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition," *The American Statistician*, 61(2), 139–147.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), 2223–2273.

——— (2021): "Autoencoder Asset Pricing Models," *Journal of Econometrics*, 222(1, Part B), 429–450, Annals Issue: Financial Econometrics in the Age of the Digital Economy.

Harvey, C. R., and Y. Liu (2021): "Lucky Factors," *Journal of Financial Economics*, 141(2), 413–435.

Hasler, M., and C. Martineau (2023): "Explaining the Failure of the Unconditional CAPM with the Conditional CAPM," *Management Science*, 69(3), 1835–1855.

Hastie, T., R. Tibshirani, and J. Friedman (2017): *The Elements of Statistical Learning*, Springer Series in Statistics.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

Heisenberg, W. (1927): "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik," *Zeitschrift für Physik*, 43, 172–198.

Hendershott, T., D. Livdan, and D. Rösch (2020): "Asset pricing: A Tale of Night and Day," *Journal of Financial Economics*, 138(3), 635–662.

Hoechle, D. (2007): "Robust Standard Errors for Panel Regressions with Cross-sectional Dependence," *The Stata Journal*, 7(3), 281–312.

Hollstein, F., M. Prokopczuk, and C. Wese Simen (2020): "The Conditional Capital Asset Pricing Model Revisited: Evidence from High-Frequency Betas," *Management Science*, 66(6), 2474–2494.

Hothorn, T., F. Bretz, and P. Westfall (2008): "Simultaneous Inference in General Parametric Models," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3), 346–363.

Jagannathan, R., and Z. Wang (1996): "The Conditional CAPM and the Cross-Section of Expected Returns," *Journal of Finance*, 51(1), 3–53.

——— (1998): "An Asymptotic Theory for Estimating Beta-Pricing Models Using Cross-Sectional Regression," *Journal of Finance*, 53(4), 1285–1309.

Jalali, A., S. Sanghavi, C. Ruan, and P. Ravikumar (2010): "A Dirty Model for Multi-task Learning," in *Advances in Neural Information Processing Systems*, ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, vol. 23. Curran Associates, Inc.

Jensen, M. C., F. Black, and M. S. Scholes (1972): "The Capital Asset Pricing Model: Some Empirical Tests," *Studies in the Theory of Capital Markets*.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2021): "Is There A Replication Crisis In Finance?," Working Paper 28432, National Bureau of Economic Research.

Kadan, O., and X. Tang (2020): "A Bound on Expected Stock Returns," *Review of Financial Studies*, 33(4), 1565–1617.

Kelly, B., and S. Pruitt (2013): "Market Expectations in the Cross-Section of Present Values," *Journal of Finance*, 68(5), 1721–1756.

Kelly, B. T., T. J. Moskowitz, and S. Pruitt (2021): "Understanding Momentum and Reversal," *Journal of Financial Economics*, 140(3), 726–743.

Kelly, B. T., S. Pruitt, and Y. Su (2019): "Characteristics are Covariances: A Unified Model of Risk and Return," *Journal of Financial Economics*, 134(3), 501–524.

Kempf, A., O. Korn, and S. Sassning (2015): "Portfolio Optimization Using Forward-Looking Information," *Review of Finance*, 19(1), 467–490.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang (2016): "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," *arXiv*, 1609.04836.

Kim, D. (1995): "The Errors in the Variables Problem in the Cross-Section of Expected Stock Returns," *Journal of Finance*, 50(5), 1605–1634.

Kozak, S., and S. Nagel (2022): "When do Cross-sectional Asset Pricing Factors Span the Stochastic Discount Factor?," Working Paper.

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the Cross-section," *Journal of Financial Economics*, 135(2), 271–292.

Kraus, A., and R. H. Litzenberger (1975): "Market Equilibrium in a Multiperiod State Preference Model with Logarithmic Utility," *Journal of Finance*, 30(5), 1213–1227.

Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner (2022): "Post-Selection Inference," *Annual Review of Statistics and Its Application*, 9(1), 505–527.

Ledoit, O., and M. Wolf (2004): "Honey, I Shrunk the Sample Covariance Matrix," *Journal of Portfolio Management*, 30(4), 110–119.

Lettau, M., and S. Ludvigson (2001): "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," *Journal of Political Economy*, 109(6), 1238–1287.

Lewellen, J., and S. Nagel (2006): "The Conditional CAPM Does Not Explain Asset-pricing Anomalies," *Journal of Financial Economics*, 82(2), 289–314.

Lewellen, J., S. Nagel, and J. Shanken (2010): "A Skeptical Appraisal of Asset Pricing Tests," *Journal of Financial Economics*, 96(2), 175–194.

Light, N., D. Maslov, and O. Rytchkov (2017): "Aggregation of Information About the Cross Section of Stock Returns: A Latent Variable Approach," *Review of Financial Studies*, 30(4), 1339–1381.

Lioui, A., and P. Maio (2014): "Interest Rate Risk and the Cross Section of Stock Returns," *Journal of Financial and Quantitative Analysis*, 49(2), 483–511.

Liu, J., S. Ji, and J. Ye (2009): "Multi-Task Feature Learning Via Efficient l2,1-Norm Minimization," *The Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*.

Lundberg, S. M., and S.-I. Lee (2017): "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30. Curran Associates, Inc.

Maio, P., and P. Santa-Clara (2012): "Multifactor Models and their Consistency with the ICAPM," *Journal of Financial Economics*, 106(3), 586–613.

——— (2017): "Short-Term Interest Rates and Stock Market Anomalies," *Journal of Financial and Quantitative Analysis*, 52(3), 927–961.

Martin, I. (2018): "Options and the Gamma Knife," *Journal of Derivatives*, 25(4), 71–79.

Martin, I. W. R. (2011): "Simple Variance Swaps," Working Paper.

——— (2017): "What is the Expected Return on the Market?," *Quarterly Journal of Economics*, 132(1), 367–433.

Martin, I. W. R., and S. Nagel (2022): "Market Efficiency in the Age of Big Data," *Journal of Financial Economics*, 145(1), 154–177.

Martin, I. W. R., and C. Wagner (2019): "What Is the Expected Return on a Stock?," *Journal of Finance*, 74(4), 1887–1929.

McLean, R. D., and J. Pontiff (2016): "Does Academic Research Destroy Stock Return Predictability?," *Journal of Finance*, 71(1), 5–32.

Meinshausen, N., and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34(3), 1436 – 1462.

——— (2010): "Stability Selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41(5), 867–887.

——— (1980): "On Estimating the Expected Return on the Market: An Exploratory Investigation," *Journal of Financial Economics*, 8(4), 323–361.

——— (1982): "On the Microeconomic Theory of Investment under Uncertainty," vol. 2 of *Handbook of Mathematical Economics*, pp. 601–669. Elsevier.

Millo, G. (2017): "Robust Standard Error Estimators for Panel Models: A Unifying Approach," *Journal of Statistical Software*, 82(3), 1–27.

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

Newey, W. K., and K. D. West (1987): "Hypothesis Testing with Efficient Method of Moments Estimation," *International Economic Review*, pp. 777–787.

Obozinski, G., B. Taskar, and M. I. Jordan (2010): "Joint Covariate Selection and Joint Subspace Selection for Multiple Classification Problems," *Statistics and Computing*, 20, 231–252.

Patton, A. J., and A. Timmermann (2010): "Monotonicity in Asset Returns: New Tests with Applications to the Term Structure, the CAPM, and Portfolio Sorts," *Journal of Financial Economics*, 98(3), 605–625.

Petersen, M. A. (2009): "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies*, 22(1), 435–480.

Petkova, R. (2006): "Do the Fama-French Factors Proxy for Innovations in Predictive Variables?," *Journal of Finance*, 61(2), 581–612.

Pettengill, G. N., S. Sundaram, and I. Mathur (1995): "The Conditional Relation between Beta and Returns," *Journal of Financial and Quantitative Analysis*, 30(1), 101–116.

Rinaldo, A., L. Wasserman, and M. G'Sell (2019): "Bootstrapping and Sample Splitting for High-dimensional, Assumption-lean Inference," *Annals of Statistics*, 47(6), 3438 – 3469.

Roll, R. (1977): "A Critique of the Asset Pricing Theory's Tests Part I: On Past and Potential Testability of the Theory," *Journal of Financial Economics*, 4(2), 129–176.

Romano, J. P., and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73(4), 1237–1282.

Ross, S. A. (1976): "Options and Efficiency," *Quarterly Journal of Economics*, 90(1), 75–89.

Rousseeuw, P. J. (1987): "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, 20, 53–65.

Rubinstein, M. (1974): "An Aggregation Theorem for Securities Markets," *Journal of Financial Economics*, 1(3), 225–244.

——— (1976): "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics*, pp. 407–425.

Savor, P., and M. Wilson (2014): "Asset Pricing: A Tale of Two Days," *Journal of Financial Economics*, 113(2), 171–201.

Schneider, P., and F. Trojani (2019): "(Almost) Model-Free Recovery," *Journal of Finance*, 74(1), 323–370.

Shanken, J. (1992): "On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, 5(1), 1–33.

Shapley, L. S. (1951): *Notes on the N-person Game*. Rand Corporation.

Shumway, T. (1997): "The Delisting Bias in CRSP Data," *Journal of Finance*, 52(1), 327–340.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15(56), 1929–1958.

Stambaugh, R. F. (1999): "Predictive Regressions," *Journal of Financial Economics*, 54(3), 375–421.

Stock, J. H., and M. W. Watson (1989): "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 4, 351–394.

Thompson, S. B. (2011): "Simple Formulas for Standard Errors that Cluster by both Firm and Time," *Journal of Financial Economics*, 99(1), 1–10.

Timmermann, A., and Y. Zhu (2019): "Comparing Forecasting Performance with Panel Data," Working Paper.

Ungeheuer, M., and M. Weber (2021): "The Perception of Dependence, Investment Decisions, and Stock Prices," *Journal of Finance*, 76(2), 797–844.

Wang, K. (2018): "Risk-Neutral Cumulants, Expected Risk Premia, and Future Stock Returns," Working Paper.

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

White, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21(1), 149–170.

——— (1981): "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76(374), 419–433.

Yuan, M., and Y. Lin (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Zhao, P., and B. Yu (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7(90), 2541–2563.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429.