Missing Financial Data[☆]

Svetlana Bryzgalova

London Business School, CEPR, sbryzgalova@london.edu

Sven Lerner

Stanford University, Institute for Computational and Mathematical Engineering, svenl@stanford.edu

Martin Lettau

University of California at Berkeley, Haas School of Business, NBER, CEPR, lettau@berkeley.edu

Markus Pelger

Stanford University, Department of Management Science & Engineering, mpelger@stanford.edu

Abstract

We document the widespread nature and structure of missing observations of firm fundamentals and show how to systematically handle them. Missing financial data affects more than 70% of firms that represent about half of the total market cap. Firm fundamentals have complex systematic missing patterns, invalidating traditional ad-hoc approaches to imputation. We propose a novel imputation method to obtain a fully observed panel of firm fundamentals that exploits both time-series and cross-sectional dependency of data to impute missing values and allows for general systematic patterns of missingness. We document important implications for risk premia estimates, cross-sectional anomalies, and portfolio construction.

Keywords: Missing data, firm characteristics, cross-sectional asset pricing, PCA, factor model, big data, asset pricing

JEL classification: C14, C38, C55, G12

This draft: November 27, 2023

First draft: March 22, 2022

[☆]We thank Nina Boyarchenko (discussant), Andrew Chen (discussant), Serhiy Kozak (discussant), Allan Timmermann, Michael Weber (discussant), Dacheng Xiu (discussant), Lingxiao Zhao (discussant), Guofu Zhou (discussant), and seminar and conference participants at the University of California San Diego, Swiss Federal Institute of Technology Lausanne, University of Lausanne, University of Maryland, Stanford University, the London Business School, Technical University of Munich, Shanghai Advanced Institute of Finance, NBER Big Data and Securities Markets Conference, NBER Asset Pricing Meeting, Annual Meeting of the American Finance Association, AI & Big Data in Finance Research Forum, NBER Forecasting and Empirical Methods Summer Institute, NBER-NSF Time-Series Conference, Annual Meeting of the American Economic Association, Annual Meeting of the European Finance Association, SFS Cavalcade North America, Machine Learning and Quantitative Finance Workshop at Oxford, Society for Financial Econometrics Annual Conference, RCEA Big Data and Machine Learning Conference, California Econometric Conference, Hong Kong Conference for Fintech, AI and Big Data in Business, German Economists Abroad Conference, International Conference on Computational and Financial Econometrics, Brandes Center, NVIDIA, PanAgora Asset Management and Schroders Asset Management.

1. Introduction

This paper studies a widespread yet little-researched phenomenon in finance: missing data in firm characteristics, which are the cornerstone of academic research in finance. The standard source of fundamental firm-level data is the Compustat database, which includes more than 1,000 individual variables. Many Compustat variables are sparsely populated; for example, Koh and Reeb (2015) report that R&D information of 42% of all firms is missing during the period of 1980 to 2006.[1] The coverage of other important variables, such as current assets and liabilities, physical assets, investment, profits, and taxes, among others, is also limited, whereas other variables are present for almost all firms.[2] As a result, the patterns of "missingness" vary substantially across characteristics. The problem of missing data is of course not unique to finance, and it is more broadly relevant for economics: Abrevaya and Donald (2017)'s data from top empirics economic journals indicates that close to 40% of all the papers published in those journals had to deal with missing data and about 70% of those simply drop missing observations. Similarly, most papers in finance ignore the issue of missing data and either exclude all the firms with missing observations or use some form of ad-hoc filled-in values.

Missing characteristic data have several potential implications on the outcome of empirical tests in asset pricing and corporate finance. First, if firms with missing observations are excluded, it can substantially reduce the size of the data, and hence the precision of estimates. For example, it reduces the number of stocks in portfolios that are constructed by sorting on characteristics. More importantly, it can lead to a selection bias, when stocks with missing observations are systematically different from those with observed values, and if the sources for systematic missingness are not accounted for. Hence, asset pricing factor models constructed with stocks that have only observed data might neglect systematic pricing information. Second, using all the observations, yet with wrongly imputed values, can lead to a significant imputation bias. This can affect parameter estimates of both structural and reduced-form models.

Our paper provides three main contributions: *(i)* it provides a comprehensive analysis of missing data in firm characteristics, *(ii)* it builds an effective statistical model for imputing missing values, and, *(iii)* it investigates the impact of missingness on asset returns in a variety of applications. In doing so, our paper also provides guidelines for handling missing observations in many other potential

[1] We confirm their finding and find similar results in our updated sample.

[2] Compustat codes ACT, LCT, PPEGT, CAPX, GP, and variables starting with TX.

settings, such as international data.

Our first contribution is a comprehensive analysis of the issue of missing firm fundamentals that establishes the following stylized facts:

Fact #1: Missing financial data is extremely prevalent for most characteristics. The number of missing fundamentals is large, both statistically and economically. Our dataset includes 45 of the the most popular and widely used characteristics in asset pricing. From the beginning of our sample period, 1967 to 1981, more than 25% of observations across all stocks are missing, whereas 10% of observations are missing between 1990 and 2020. Before 1975, all stocks have at least one missing characteristic in any given year, and only 25% of all stocks have no missing characteristics in any year since 2000. There is, of course, substantial heterogeneity in the cross-section and over time, with particular characteristics and time periods, for which more than 90% of the data is missing. Missingness is a feature of firms, whether small or large, young or mature, and profitable or in financial distress.

Fact #2: The problem of missing data becomes substantially severe when one requires observations of multiple characteristics at the same time. A study of return predictability relying on a fully observed panel of 45 firm characteristics would have to omit more than 70% of firms, representing about one-half of the total market capitalization. The issue remains in subsets of characteristics. Consider five of the most widely-studied characteristics: book-to-market (B2M), earnings-to-price (EP), momentum (MOM), operating profitability (OP), and investment (INV). During the period of 1967 to 1980, only 50% of all stocks have a complete record of all five characteristics. The number increases to 80% toward the end of our sample, so that one-fifth of all stocks miss at least one of the five characteristics in a given year. Hence, considering only firms with a fully observed set of fundamentals neglects a substantial amount of data and, as we show, leads to severe sample selection.

Fact #3: Data has systematic complex missing patterns. There is strong heterogeneity and dependency in the distribution of missing observations, which creates clusters both cross-sectionally and over time. Naturally, some of the missingness patterns arise mechanically, for example, different fundamentals might require similar accounting variables, or young firms lack a prior history for constructing certain characteristics (e.g., momentum or long-term reversal). At the same time, there is a substantial number of characteristics missing during any stage of the firm's life cycle. Other clusters arise because firms with missing data have a similar underlying latent structure. In particu-

lar, we note that small-cap companies generally have a higher propensity for missing data, and more extreme realization of characteristics are often more likely to be unobserved.

Fact #4: Returns on their own depend on whether a firm has missing fundamentals. We show that returns of stocks with observed and unobserved characteristics are different, which drives a substantial selection bias, if one focuses only on the data with observed characteristic values. On average, we find that stocks with a missing characteristic value have lower overall returns than their counterparts when the same variable is observed.

Our second contribution is a novel approach for imputing missing firm fundamentals and a comprehensive empirical comparison study of imputation methods. Any imputation method has to solve two problems. First, it requires a good model for characteristics. As characteristics have complex dependencies in both the cross-section and over time, omitting any of this information leads to an omitted variable bias. Second, the model needs to be estimated from partially observed data, but has to be valid on the missing data. This means that ignoring the systematic patterns in missing firm characteristics can lead to a selection bias in the imputation. An ad-hoc imputation like a simple cross-sectional median incurs both biases. We provide a solution to both problems.

First, we use a robust latent factor model to capture contemporaneous cross-sectional dependency in characteristics, while taking advantage of all observed characteristics. Our procedure has two key benefits: It remains valid even in the presence of the complex missing patterns that we have documented. We can reliably recover the latent characteristic factor model when the probability of missing data varies over time, for different characteristics and different stocks, and can even depend on the factor model itself. The second key feature is its robust estimation, which can also leverage local cross-sectional correlation patterns and leads to small variances in the imputed values.

Second, we use a time-series model to capture the persistence in characteristics. Our model combines the cross-sectional factors and time-series observation, and, hence, can extract slow persistent movements from the time series while capturing fast changes from contemporaneous factor realizations.

Our comprehensive empirical study shows that our imputation method strongly dominates leading conventional approaches. The most widely used imputation approach for firm characteristics is a simple cross-sectional median (of the whole market or the industry the firm belongs to). We show that our model allows us to achieve a 50% reduction in the out-of-sample imputation error compared to using both types of medians. Another popular approach, especially for persistent characteristics,

is to use their last observed stale values. This also leads to a subpar empirical performance, particularly when there are blocks of consecutively missing observations. Overall, we conclude that even though our imputation method is transparent and easy-to-use, it uniformly dominates leading empirical approaches.

Our third contribution shows that missing financial data can have a profound impact on asset pricing. Missing firm characteristics can have two fundamental effects on asset pricing. First, using only the subsample of stocks with fully observed characteristics can lead to a selection bias in asset pricing metrics. This is reflected in the substantially higher out-of-sample Sharpe ratio of the stochastic discount factor based on conditional latent factors that are extracted from all stocks instead of the non-representative subsample with fully observed data. This effect is also present in univariate portfolio sorts. In many cases, requiring more characteristic observations lowers expected returns, while at the same time the lower diversification with fewer stocks increases the volatility, resulting in overall lower Sharpe ratios. The second effect is the imputation bias in the estimation of parameters of an asset pricing model when using a biased imputation method. We study the fundamental problem of estimating the risk premium of characteristics from cross-sectional characteristic regressions. Biased imputation methods like the median imputation lead to uniformly and substantially larger risk premia errors compared to our more precise imputation approach. The imputation bias of the median is also manifested in the distorted time-series of factor-mimicking pure-play portfolios, whereas, in contrast, our imputation method provides precise estimates of their full time-series.

The method and insights of our paper have wider implications for other strands of literature in economics and finance. For example, it is widely known that the issue of missing observations is even more pronounced in international accounting data. Furthermore, we expect the problem to become even more relevant due to the growing importance of Big Data and availability of new data sources that also often provide only partial coverage. We hope that our paper lays out foundations for imputing missing data that could be applied in many different contexts.[3]

Closely Related Literature

There is a vast body of literature on the topic of missing data in statistics and data science. We refer to Little and Rubin (2020) for an excellent introduction and focus our discussion only

[3] In principle, one could use our method also to compare observed characteristic values with model implied values to identify errors. This direction can be further improved with an appropriate robustness objective for our model.

on the most closely related literature in economics and finance.[4] Naturally, our work is related to the actively growing field of econometrics literature on missing data in panels, which directly focuses on the different assumptions regarding missing data. The most widespread and successful solutions, developed recently, rely on the estimation of a low-rank model, which is then used to impute missing values. The cross-sectional factor model, proposed in this paper, builds on the work of Xiong and Pelger (2023), who provide an all-purpose estimator for latent factors that allows for general missing patterns. Importantly, their approach allows the missing pattern to depend on the latent factor model, which is crucial for our application. Our novel estimator can be viewed as a combination of the factor model estimator of Xiong and Pelger (2023) for missing data, with the robust regularized factor model estimator of Bai and Ng (2019). Bai and Ng (2021), Cahan et al. (2023), and Jin et al. (2021) develop alternative latent factor estimators with different assumptions on the missing pattern. The imputation of missing values in a panel is closely related to conducting causal inference in a panel (see Athey et al. (2022)), because unobserved counterfactual outcomes can be modeled as missing values. Hence, the common challenge consists of uncovering a low-rank model that could be used to impute missing data, when the missingness or treatment depends on unobserved confounders.[5]

The most widely and recently used approaches to deal with missing data in the finance literature, with an application to firm fundamentals, use either cross-sectional median imputation (e.g., Green et al. (2017), Light et al. (2017), Kozak et al. (2020), and Gu et al. (2020)), or use only the subset of fully observed data (e.g., Lewellen (2015), Freyberger et al. (2020), and Kelly et al. (2019)). The

[4]Seminal contributions in statistics include Yates (1933), Dagenais (1973), Rubin (1976), Little (1992), and Rao and Toutenburg (1999). Popular Bayesian solutions to data imputation typically rely on strong parametric assumptions and very specific types of missingness in the data. A separate strand of literature focuses on down-weighting missing observations and/or on a specific pattern of missingness, that is, assuming that either all the data is observed or only a specific subset of it is missing, see Dagenais (1973), Abrevaya and Donald (2017), Robins et al. (1994), and Wooldridge (2007). With the widespread introduction of machine-learning techniques, alternative data-driven methods using K-nearest neighbor, ensemble-based imputation, and Support Vector Machine (SVM) have also gained popularity in empirical work. For recent examples and an overview, see Emmanuel et al. (2021), Devlin et al. (2018), and Raja and Thangavel (2020).

[5]Our paper is also related to latent factor models in financial data. Usually, factor models are directly applied to a panel of returns to estimate either unconditional factor models of principal component analyses (Connor and Korajczyk (1988), Pelger (2020), Lettau and Pelger (2020a), Lettau and Pelger (2020b), Giglio and Xiu (2021) and Dello Preite et al. (2022)) or conditional ones (Fan et al. (2016), Kelly et al. (2019), Gagliardini et al. (2016), Zaffaroni (2022)). Our paper also relies on the existence of a factor structure, but not in returns. Instead, it leverages strong dependence in firm fundamentals. Importantly, the factors are extracted from only partially observed data. Another distinguishing element is that we use a three-dimensional dataset (that is, including both cross-sectional and temporal dependence in characteristics), rather than the conventional two-dimensional panel. It is, therefore, also related to an emerging body of literature in finance that directly examines three-dimensional data, for example Lettau (2022), who recently proposed a tensor factor model for mutual fund characteristics.

issue of missing firm fundamentals is exacerbated in the fast growing body of literature on asset pricing with a large number of predictors that addresses the multidimensional challenge brought to attention by Cochrane (2011); see, among others, Gu et al. (2020), Freyberger et al. (2020), Kelly et al. (2019), Chen et al. (2022), and Bryzgalova et al. (2019). The methods used in those papers require the presence of multiple characteristics, and, as such, lead to either some form of data selection or data imputation. Our systematic study of missing firm fundamentals and the better imputation tools that we provide help to further improve the work in this research direction. In particular, Kaniel et al. (2023) use a version of our model to impute missing fundamentals in the holdings of mutual funds.

Unfortunately, there is little work that systematically addresses the problem of missing financial data. There are a few contemporaneous papers, that pursue different goals and are therefore complementary. We discuss them in detail in the main text and provide a short overview here. An important independent contribution is the paper of Freyberger et al. (2022), who consider missing firm characteristics in asset pricing, and show how to adjust conventional estimation within the General Method of Moment (GMM) framework. They focus on how to weight conditional moment conditions, that use imputed values, in order to increase efficiency. Whereas their imputation approach uses only the subset of always observed cross-sectional covariates, our method is more general and takes advantage of all the partially observed data. Furthermore, our imputation method could be combined with their weighting scheme and used within a GMM context. Their empirical findings on return prediction are complementary and confirm the importance of taking into account missing data. Chen and McCoy (2022) rely on the expectations maximization (EM) algorithm to impute missing characteristic data, which requires firm characteristics to be jointly normally distributed. Similar to our pure cross-sectional model, it leverages cross-sectional dependence in characteristics but does not include time-series information in the imputation, which, as we show, is important in many cases. Empirically, they study return prediction based portfolios. Beckmeyer and Wiedemann (2022) apply attention-based algorithms from machine learning to impute characteristic values from available cross-sectional and time-series data on firm fundamentals. Although this method is capable to capture non-linearities in a complex and flexible empirical model, like with many modern machine learning techniques, it is not clear under what assumptions it works and what theoretical properties it possesses. Blanchet et al. (2022) analyze the trade-off between look-ahead-bias and variance in an imputation used for out-of-sample investment. Fundamentally, we provide a systematic study of missing data in finance and asset pricing, establish the magnitude and stylized features of

this phenomenon, and provide a "general purpose" solution to it, with a complete data set of firm fundamentals, which can then be used in any of the follow-up applications.

2. Missing values

2.1. Data

We obtain the data from the CRSP/Compustat universe with the standard filters for outliers and exchanges.[6] Our sample consists of 648 months from January 1967 through December 2020 and includes 22,630 individual stocks. We consider 45 characteristics related to value, investment, profitability, intangibles, past returns, and trading frictions, as summarized in Table C.1 in the Appendix. The raw characteristics are converted into centered rank quantiles and scaled to be in the $[-0.5, 0.5]$ interval.

We construct characteristics if the required variables are available in CRSP and Compustat. Otherwise, we consider a characteristic missing. Characteristics are updated either monthly or at a lower frequency, which is typically quarterly. For quarterly updated characteristics, we do not observe the monthly observations in between the quarters, which are, therefore, mechanically missing. To avoid these mechanical effects, we report evaluation metrics for characteristics that are updated quarterly based on quarterly data points. We do not fill the months between the quarters with stale values, nor do we count those as missing values in our summary statistics about missingness.[7] We emphasize that our imputation method provides imputed values in between the quarters and, hence, is also a solution to mixed-frequency observations. As a result our imputed data set that we use for our asset pricing studies has all characteristics available at a monthly frequency.

We use the most up-to-date last-observed values as current characteristics. For characteristics based on the ratio of variables with different updating frequencies, we use the most up-to-date information of each variable, and the variable with the slowest updating frequency determines the updating frequency of the characteristic. For example, the quarterly updated book-to-market ratio

[6] The sample only includes stocks listed on the NYSE, NASDAQ, and AMEX exchanges (exchange codes 11, 12 and 14 respectively) with share codes 1, 2, or 3 (common stock, foreign incorporated, ADR, respectively) and at least one entry in the Compustat accounting tables. We do not filter out stocks based on share price, nor do we filter out financial firms. However, we show in an extensive robustness study that our results are not affected by these choices, that is, the results are robust to including or excluding either of those subsets.

[7] Using stale values in between observations of characteristics with low updating frequency is a form of data imputation itself. Using stale values as the actual monthly characteristics values would also lead to mechanical trivial predictability.

divides the book value from the most recent quarter by the last observed monthly market capitalization. Asset pricing applications, which condition on characteristics, usually lag characteristics by several months to ensure that the information is available to investors. Our data imputation uses the most recent information; however, we lag characteristics in asset pricing applications.

2.2. How much data is missing?

Missing financial data is prevalent, and almost all characteristics have missing observations. The number of missing fundamentals is large, both statistically and economically. Figures 1 and D.1 summarize some patterns in missing values over time. The black line in Panel (a) of Figure 1 shows the number of firms in our sample over time. It is well-known that the number of listed stocks has declined over the last 25 years. At its peak in November 1997, our sample includes 7,784 stocks but only 4,241 in December 2020. The spike in January 1973 is due to the inclusion of the NASDAQ. The plot also shows the number of firms with observed values of five important characteristics: book-to-market (B2M), operating profitability (OP), investment (INV), and leverage (LEV). We also include the ratio of real investment to book value of assets (DPI2A, Lyandres et al. (2008)) because it has the most missing values among all 45 characteristics. Panel (b) of Figure 1 shows the percentage of stocks with missing values for each of the five characteristics.

Panels (a) and (b) of Figure 1 show substantial cross-sectional and time variation in missingness. The proportion of missing values has, on average, decreased over time, which is not surprising since both the coverage of Compustat has improved throughout the sample and changes in regulations led to more comprehensive and more frequent disclosures of accounting information. Consider first the four accounting variables B2M, OP, INV, and LEV. Missing data is particularly prevalent throughout the early 1980s for all four characteristics. Between 30% and 95% of observations are missing between 1967 and 1981.[8] Approximately 15% to 20% of observations are missing between 1982 and 1992, followed by a further decline throughout the 2000s. At the end of the sample in 2020, 14%, 10%, 8%, and 3% of OP, INV, LEV, and BEME data, respectively, are missing. Fewer book-to-market observations are missing than those of the other variables because the definition of BEME includes several alternatives and fall-back options, if individual component variables are not in Compustat.[9]

[8] During this period, most stocks report many accounting variables only once per year, which accounts for the spikes in the plots. As the reporting frequency increases over the sample, this pattern largely vanishes.

[9] Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholders' equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock

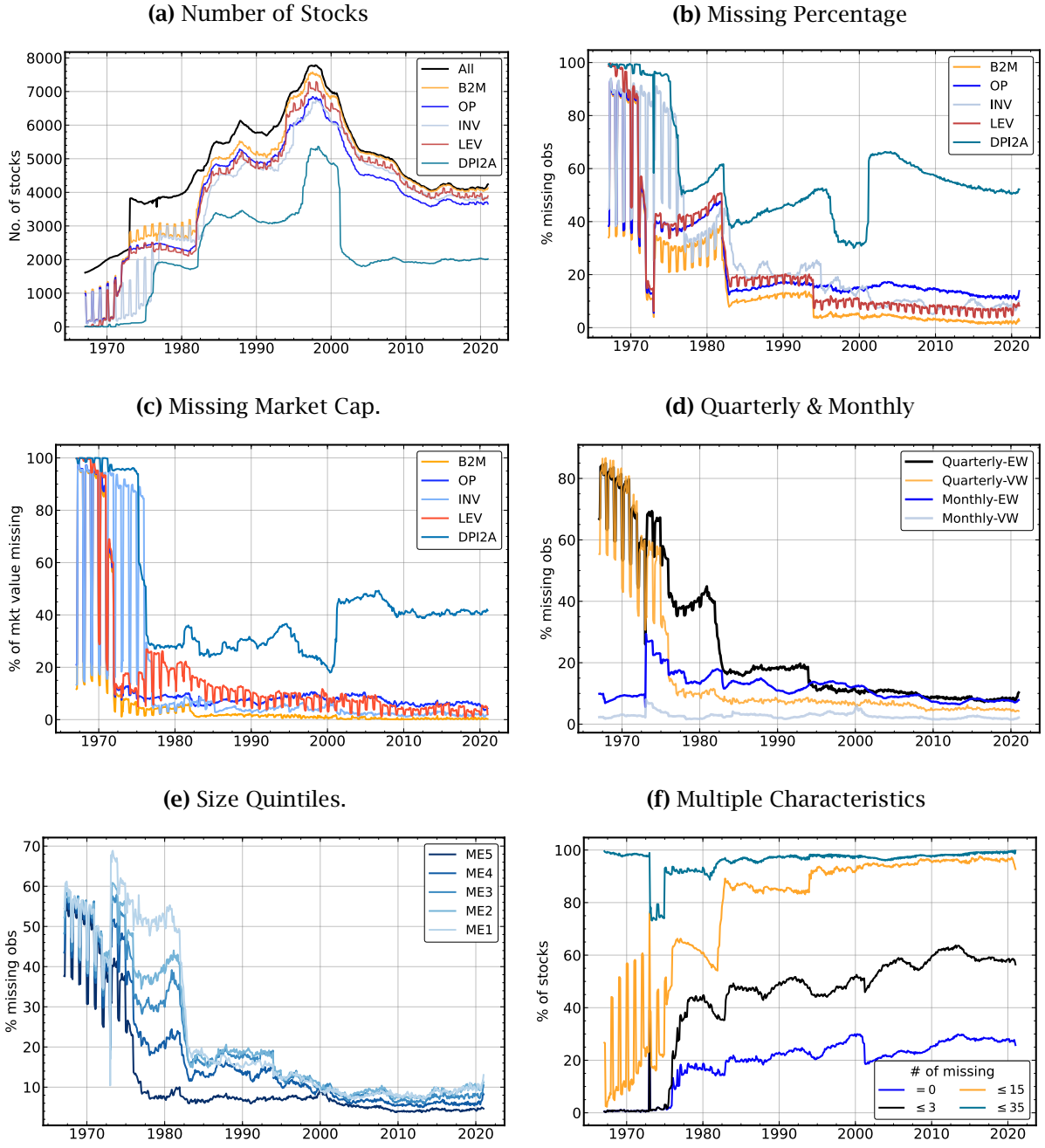
The pattern of missing values of DPI2A differs substantially from those of the other four variables. Previous to 1975, few firms have real investment observations in COMPUSTAT, so DPI2A is virtually completely missing. In contrast to the other variables, the share of missing observations remains above 35% over the rest of the sample. In 2004, 67% firm observations were missing, and more than half were missing in 2020. Although DPI2A has the most missing observations, there are several other variables with more than 20% of missing data in 2020: accruals (AC), fixed-costs-to-sales (FC2Y), operating accruals (OA), and SGA-to-sales (SGA2S).

Figure 1(c) shows the time-series of the share of missing values averaged across all characteristics. We form two groups of characteristics that are updated either monthly or quarterly. Price or return-based characteristics are available at a high frequency, while accounting variables are (at most) available quarterly. Consider first the equal-weighted averages in Panel (c). The time-series of missingness of quarterly characteristics (black line) is similar to those found for B2M, OP, INV, and LEV in Panel (b). Before 1982, more than 40% of observations are missing; between 1982 and 1992, about 20%, and between 8% and 14% after 1992. Since the CRSP database has an (almost) complete record of prices and returns, there are, on average, fewer missing values for characteristics that are updated monthly. However, many monthly characteristics require lags of prices or returns, and thus some observations are missing mechanically. For example, reversals require a return history of 60 months, meaning that newly listed firms do not have any observations for the first five years. As a result, between 10% and 20% of monthly characteristics are missing throughout the sample. The exception is the period from 1973 through 1975, when the inclusion of the NASDAQ added many firms without a history of prices and returns.

Figure D.1 in the Appendix shows the share of missing values of all characteristics over time in the form of heatmaps. Lighter (darker) shades correspond to lower (higher) shares of missing observations. The heatmaps reveal time-series variation as well as heterogeneity across characteristics. The frequency of missing data of most quarterly characteristics, shown in Panel (b), decreases substantially in the early 1980s and again in the mid-1990s. There are several characteristics with many missing values throughout the sample: AC, DPI2A, FC2Y, OA, OP, and SGA2S. The frequency of missing values in monthly variables is directly linked to the number of lagged values that are required. The exceptions are SUV and TURN, which are based on trading volume; however, volume

(PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, PS is redemption value (item PSTKRV), liquidating value (item PSTKL), or par value (item PSTK). The market value of equity (PRC*SHROUT) is as of the current month.

Figure 1: Missing Values over Time



This figure summarizes missing values over time. Subfigure (a) shows the total number of stocks and those that have observed values for our five example characteristics book-to-market (B2M), operating profitability (OP), investment (INV, growth in total assets), leverage (LEV) and real investment (defined as the change in property, plants, equipment and inventory) over lagged total assets (DPI2A). Subfigures (b) and (c) show the percentage and share of the total market value of missing observations for the five example characteristics. Subfigure (d) plots the percentage of missing observations for quarterly and monthly updated characteristics based on equal and market capitalization-weighted averages. Subfigure (e) shows the percentage of missing observations by market capitalization quintiles. Subfigure (f) displays the proportion of missing stocks that have no missing observations or at most 3, 15 or 35 missing characteristics at a given point in time.

for many NASDAQ stocks is missing from CRSP between 1973 and 1983. Thus, the share of stocks with missing values of SUV and TURN is particular to this period, which is shown in the heatmap in Panel (b).

The evidence thus far was based on firm counts without taking firm size into account. Figure 1 Panel (d) also shows the value-weighted percentage of missing observations for monthly (light blue) and quarterly (orange) characteristics. Although the value-weighted percentage is lower than its equal-weighted counterpart, it is still substantial. In particular, quarterly updated characteristics are missing for more than 10% of the market capitalization after 1977.

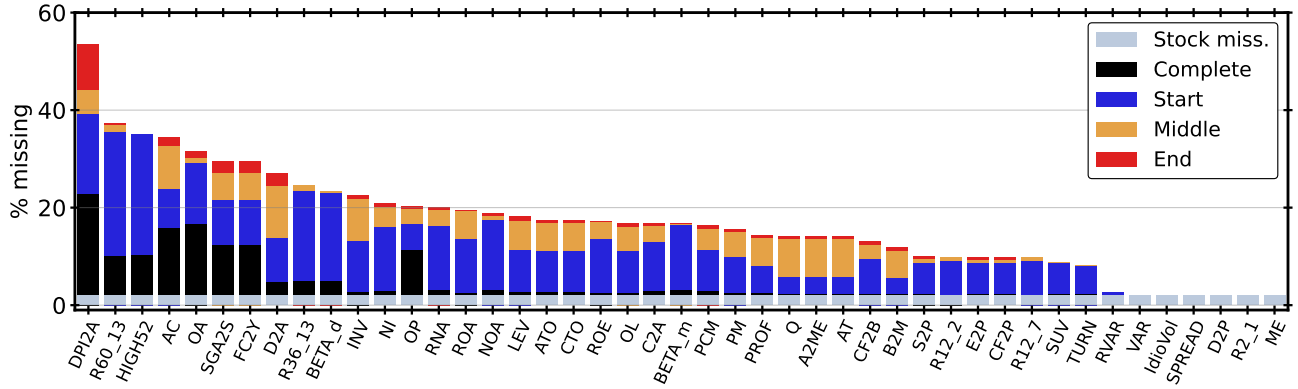
Figure 1 Panel (e) reports the percentage of missing observations for quintiles of market capitalization of companies. We observe that historically, smaller companies had worse data coverage. However, in the last 20 years, small and large companies have shown similar degrees of missingness. Importantly, at no point in time is missing data due only to small-cap companies.

Missing data is a paramount problem when multiple characteristics are required. The missingness in individual characteristics largely underrepresents the severity of the problem. Figure 1 Panel (f) shows the percentage of stocks that have no characteristics, fewer than three, fewer than 15, or fewer than 35 of the 45 characteristics missing. The results are striking. More than 70% of firms are missing at least some popular characteristic at any point in time. The total market cap corresponds to 48%. In other words, an application that requires all 45 characteristics to be observed neglects half of the market capitalization and 70% of the companies at any point in time. As we will show, using a fully observed panel of data may lead to severe sample selection. This can affect all applications that require a full panel of characteristics, which includes characteristic panel models, conditional latent factor models, and machine learning applications.

2.3. What is the structure of missingness?

In order to understand the structure of missingness, we study when, which, and for what values firm fundamentals are missing. Figure 2 displays the percentage of missing observations for each characteristic. We report whether characteristics are missing at the beginning, at the end, or in the middle. Recall that we only include a stock in the sample when we observe its returns and at least one entry in Compustat in a given month. Missingness in the middle implies that we observe some previous and future values. Missingness at the beginning mechanically appears for younger firms, whereas missingness at the end can occur at the end of a company's life. We see that many

Figure 2: Missing Observations by Characteristic



This figure shows the average percentage of missing observations for each characteristic. We decompose the missing values in those missing at the start (no previous observations), the middle (some previous and future observations), the end (no further observations) and completely missing.

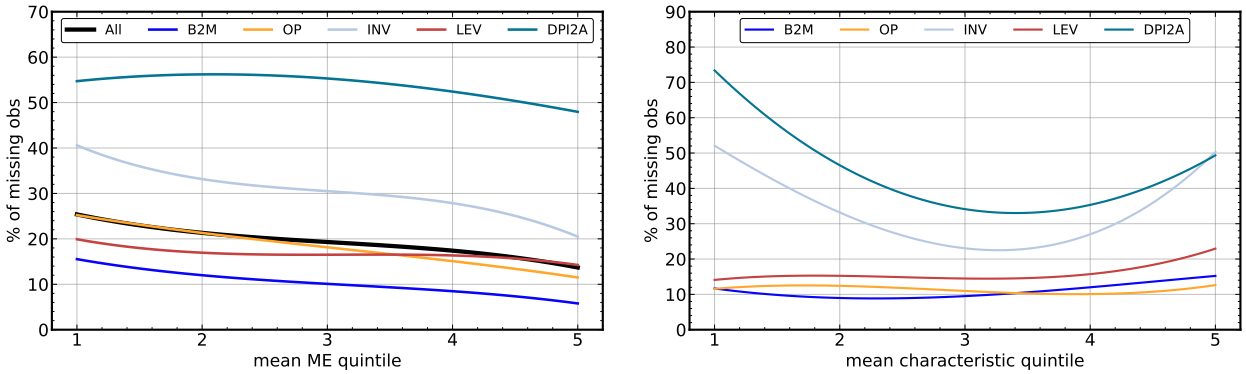
accounting-based variables are missing after having been previously observed, which often occurs in missing time blocks. Overall, we confirm that missingness is a problem for all characteristics.

Some of the missingness patterns are purely mechanical and expected. For example, long-term reversal and momentum have by construction missing observations for a new firm without prior history. At the other extreme, market capitalization is always observed when there is a return in the prior month. Figure D.2 in the Appendix provides missing observations by characteristic pooled by stocks, which can be different from the overall averages if there is heterogeneity in the patterns for individual stocks. Although the overall percentage and relative ranking seem to be quite similar, there are notable differences. Missing in the middle is less pronounced in the pooled averages, which implies that there is a smaller subset of stocks for which observations are primarily missing in the middle. Characteristics that are based on past return observations also constitute a larger percentage for the pooled averages. The lower panel in Figure D.2 in the Appendix shows the value-weighted pooled averages with similar findings.

We investigate how values of characteristics interact with the frequency of missing values. We sort stocks into quintiles of a characteristic and compute the share of missing values among stocks in each quintile. Figure 3 Panel (a) shows the percentage of missing observations by size quintiles. The black line shows the average share of missing values across all 45 characteristics and shows that smaller stocks have more missing values than large stocks; however, the difference is modest. Even in the largest size quintile, 15% of the characteristics are missing. The downward slope is present in most characteristics, but the dependency on size is heterogeneous. The size effect on leverage and

Figure 3: Missing Observations by Characteristic Quintiles

(a) Mean ME Quintiles **(b)** Mean Char. Quintiles



This figure shows the percentage of missing observations for different characteristic quintiles. The left subfigure displays the missing observations for all characteristics and the example characteristics book-to-market, operating profitability, investment, leverage and change in property, plants, equipment and inventory over lagged total assets for the five size quintiles of stocks. The right subfigure presents the proportion of missing values for the five example characteristics for their corresponding characteristic quintile. The characteristic quintiles are based on the average observed characteristic value of the corresponding stock.

DPI2A is almost flat, while it is more pronounced for investment.

We then compute how the frequency of missing values of a characteristic depends on characteristic values themselves. Obviously, we do not observe the actual characteristic realizations when they are missing. Hence, we study the patterns of missingness for firms that are on average in a certain characteristic quintile. In more detail, for each characteristic, we sort stocks into quintiles based on their observed values and compute the proportion of missing values of the characteristic of the stocks in each quintile. The results are shown in Panel (b) of Figure 3. The convex shape of the lines implies that stocks with low and high characteristics have more missing values than stocks with average characteristics. The difference is economically large; missing values of stocks at the extreme of the characteristic distributions are twice as frequent as for stocks at the center (29% vs. 14%, respectively). This U-shaped pattern is true for the majority of individual characteristics, as shown for DPI2A, INV, and to a lesser extent, B2M in Panel (b), and summarized for the other characteristics in Table C.2 in the Appendix. These results suggest that missing values can depend on characteristic realizations themselves. In this sense, missingness is endogenous.

In order to better understand the structure of missingness, we predict missingness of individual firm characteristics with logistic regressions. Table 1 shows the results for different sets of explanatory variables. We report separate regressions for characteristics missing at the beginning, in the

Table 1: Logistic Regressions Explaining Missingess

D2P	IdioVol	ME	R2_1	SPREAD	TURN	VAR	FE	Last Val	Missing Gap	train AUC	test AUC
Missing at the beginning											
1.79*** [232.54]	2.8*** [40.96]	-1.07*** [-121.28]	0.04*** [7.42]	0.69*** [72.77]	0.65*** [108.77]	-2.23*** [-33.02]	F	F	F	0.49	0.50
1.91*** [176.91]	0.97*** [12.52]	-0.45*** [-42.98]	-0.09*** [-13.36]	0.68*** [59.35]	0.85*** [120.59]	-0.68*** [-8.91]	F	F	0.06*** [452.08]	0.61	0.63
							T	F	F	0.69	0.73
							T	F	0.01*** [153.85]	0.69	0.72
0.46*** [36.6]	-1.29*** [-13.72]	-0.65*** [-51.05]	0.11*** [13.66]	-0.02*** [-1.38]	-0.10*** [-11.69]	0.9*** [9.71]	T	F	0.01*** [145.96]	0.71	0.74
Missing in the middle											
0.71*** [318.01]	-0.48*** [-21.43]	-0.91*** [-282.26]	0.05*** [26.54]	0.38*** [109.67]	0.4*** [173.84]	0.15*** [6.62]	F	F	F	0.55	0.51
							T	F	F	0.78	0.82
							T	5.37*** [961.19]	F	0.92	0.96
							T	0.06*** [137.87]	-4.74*** [-279.65]	0.93	0.96
0.29*** [26.41]	-0.34*** [-2.79]	-0.64*** [-40.54]	0.07*** [7.02]	0.40*** [25.63]	-0.28*** [-25.87]	0.42*** [3.53]	T	0.06*** [139.6]	-4.90*** [-270.39]	0.94	0.97
Missing at the end											
0.73*** [454.13]	-0.48*** [-30.33]	-0.98*** [-428.69]	0.04*** [29.06]	0.32*** [129.03]	0.16*** [100.59]	0.22*** [13.6]	F	F	F	0.61	0.55
							T	F	F	0.80	0.83
1.53*** [462.23]	0.92*** [26.04]	-0.92*** [-201.06]	0.06*** [19.98]	0.44*** [90.52]	-0.17*** [-52.93]	-1.03*** [-29.38]	T	F	F	0.82	0.83

This table shows the results of logistic regressions to predict the missingness of individual stock characteristics. We report the results for different sets of explanatory variables for characteristics missing at the beginning, in the middle and at the end. The values of the seven characteristics D2P, IdioVol, ME, R2_1, SPREAD, TURN and VAR are always observed and hence can be included in the regressions. We also include characteristic fixed effects (FE), an indictor variable if the last characteristic value was observed, and the length of a missingness if the last value was not observed. The area under the curve (AUC) measures the accuracy of the logistic regression. The regression is pooled over time, stocks, and characteristics. The model is estimated on the training data (1988-1998) and evaluated out-of-sample on the test data (1999-2020). We also include the z-scores of the regression coefficients in brackets. Stars indicate the statistical significance, where *** corresponds to 1% significance.

middle, and at the end, as, for example, missingness in the middle can be more related to firm fundamentals than mechanical missingness for new firms.[10] We explain missingness with the seven characteristics that are always observed, an indicator if the last observation was missing, and the

[10]For missing at the beginning, we consider the set of all characteristic observations that are missing at the beginning of the sample and include the first time a characteristic is observed. Hence, the results for missing at the beginning essentially predict the change from missing at the beginning to being observed for the first time. For missing at the end, we include only the set of characteristic observations that end in terminal missingness. In more detail, we include the set of only observed values (after potentially missing values) and the first terminal missing value. Thus, the results for missing at the end predict the change from being observed to be missing completely. Missing in the middle excludes the subset of characteristic observations that are missing at the beginning (no prior observations) and at the end (no further observations after missingness). Note that this means that the same stock for the same characteristic can have part of its time-series included in missing at the beginning (first set of observations), missing in the middle (all observed and missing values in the middle), and missing at the end (last block of observations).

length of the missingness if the last observation was missing. We also allow for characteristics' fixed effects. The missing in the middle category is the most important for our analysis and represents the largest part of this sample. It contains all observed and missing characteristic values that have at least one prior observation and an end observation. The area under the curve (AUC) measures the accuracy of the prediction. Our best models achieve an out-of-sample AUC of 0.97, which means that we explain a large part of the missing pattern and that the logistic regression captures important features.

First, characteristic fixed effects are crucial in the prediction, confirming our previous finding that missingness is heterogeneous. Second, the realization of contemporaneous characteristics is highly significant in predicting missingness. As we will show, characteristics are cross-sectionally correlated, which confirms some form of endogeneity in missingness. Finally, missingness is correlated over time. The negative sign on the length of a missing gap indicates that missing data is likely to appear in blocks. Table C.3 in the Appendix reports the number of missing blocks and their mean and median length. Indeed, most missing values cluster together and have an average length of approximately one to two years.

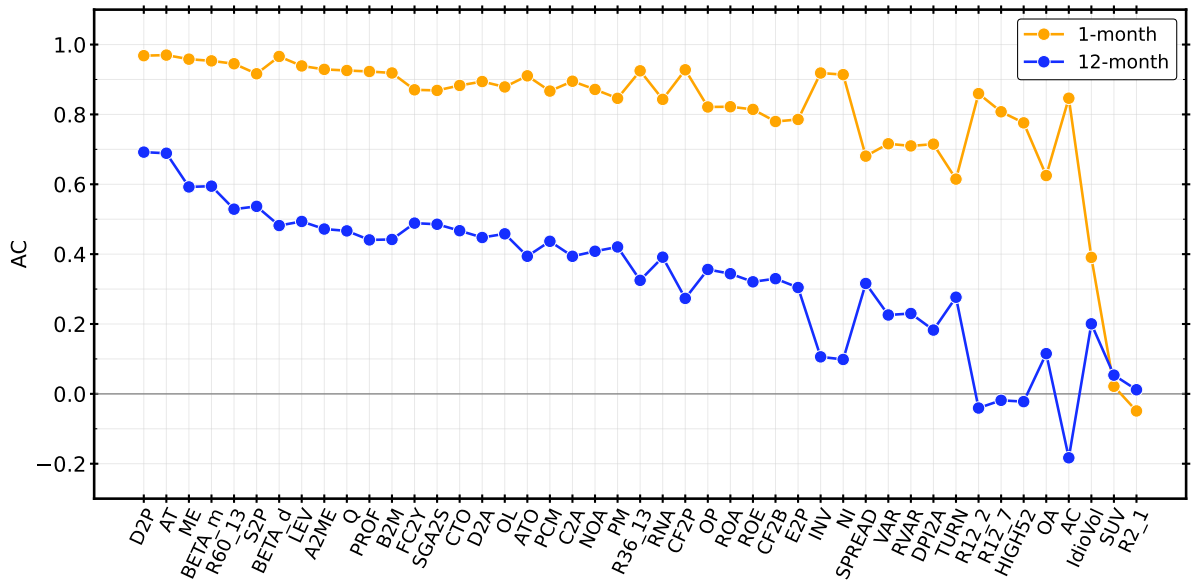
The structure of missingness also has important implications for how to impute missing values. First, imputation methods should allow for different information sets. If no prior values are observed, it is obviously not possible to condition on prior observations in the imputation method. Second, stocks with different fundamentals can be more likely to have missing values. Hence, an imputation method should allow the probability of missingness to be heterogenous and depend on fundamentals. Modeling characteristics with a factor model should allow the joint distribution of missingness to depend on the factor model itself.

2.4. Characteristics Dependency

Characteristics are dependent over time and cross-sectionally on other characteristics. This dependency establishes the foundation of any method that tries to model or predict characteristics. It implies that observing the realizations of other characteristics or prior values allows us to infer the realizations of unobserved characteristics.

Many characteristics are quite persistent. Figures 4 and D.4 show the 45 characteristics sorted by their autocorrelation and standard deviation. As expected, many characteristics, for example, market capitalization and total assets, are rather slow-moving and highly serially correlated. This

Figure 4: Autocorrelation of Characteristic Ranks

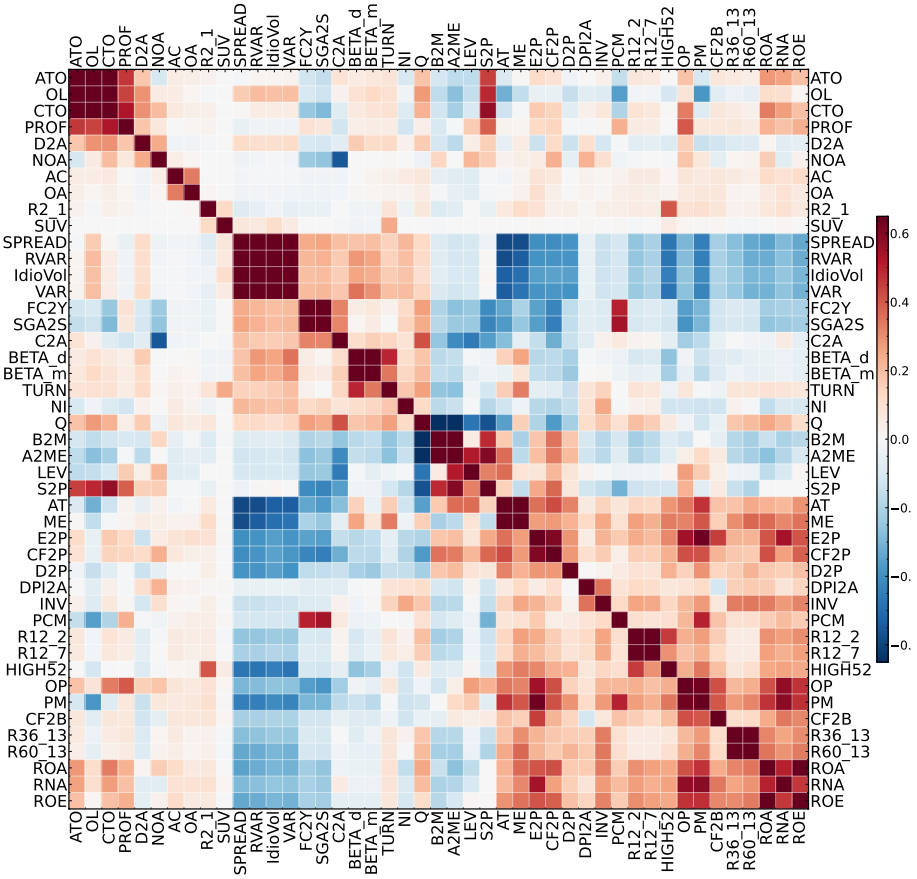


This figure presents the dependency of characteristic ranks and summarizes the 1-month and 12-months autocorrelation coefficients for each characteristic.

implies that the previous values of these persistent characteristics have information for their future realizations. In fact, the autocorrelation of several characteristics is close to one, implying that their previous value would be a good predictor. This predictability persists over longer horizons. Indeed, the 12-month autocorrelation is still more than 0.4 for around half of the characteristics. However, we also find that a number of characteristics, primarily based on prior returns like short-term momentum or idiosyncratic volatility, are highly volatile and seem to show negligible time-series predictability. Hence, the persistence is quite heterogenous. Overall, we conjecture that disregarding time dependency when imputing missing values might lead to an omitted variable bias.

Characteristics are cross-sectionally correlated. Figure 5 shows pairwise correlations in characteristics averaged over time and stocks. We observe obvious clusters of correlations. These could be interpreted as exposure to common characteristic factors. Hence, disregarding observed values of other characteristics when imputing missing values could lead to an additional omitted variable bias. For example, small stocks are more likely to be growth stocks. Therefore, imputing a missing book-to-market value of a small company with a market median, would inherently lead to a bias. The clusters of cross-sectional dependency seem to form around different groups of characteristics. Not surprisingly, characteristics based on past returns exhibit correlations. Similarly, we observe a dependency cluster among trading friction or value characteristics. However, the dependency is

Figure 5: Heatmap of Pairwise Correlation



This figure shows the pairwise correlations across time and stocks for each characteristic. The time period is the sample from 1977 to 2020.

complex and requires a sophisticated tool to capture it from the data.

The general dependency patterns between characteristics seem to be stable over time. Figure 1 shows that the frequency of missing characteristics changes drastically around the year 1977. Figure D.3 in the Appendix shows the pairwise correlations in characteristics averaged over time and stocks from 1967 through 1976, whereas Figure 5 is based on the period from 1977 to 2020. Although the strength of the dependency seems to vary, the location of correlation clusters stays the same. This would be consistent with a factor model in the characteristic space, where the factors stay the same, but the scale of the exposure to those factors can vary.

3. Model

The estimation of a model for the imputation of missing values faces two fundamental challenges. First, it should take advantage of all available information. An ad-hoc imputation method, such as the cross-sectional median, would incur an omitted variable bias. Omitting relevant latent information also leads to an omitted variable bias, even if observations are missing-completely-at-random. Our solution to the problem is to extract all latent cross-sectional information from the data rather than pre-specifying a set of covariates. In other words, we let the data speak about what contemporaneous information can best explain a given characteristic. Second, the model for characteristics, which is estimated on the observed data, should be valid on the unobserved data as well. This is another key feature of our approach. Even when the missingness depends in a complex way on latent information extracted with our model, we provide correct imputed values for the unobserved entries. Flexible methods that are estimated on the observed data and do not account for the dependency between missingness and the information that predicts characteristics can be subject to a selection bias.

Our dataset of month/stock/characteristic observations forms the following three-dimensional vector space:

$$C_{i,t,l} \qquad \text{with } i = 1, ..., N_t, t = 1, ..., T \text{ and } l = 1, ..., L.$$

The data has a cross-sectional dimension of N_t stocks, a time-series dimension T, and the number of different characteristics L. The typical dimensions are around $N_t = 6,000$, $T = 600$, and $L = 45$. The notation of an upper index selects a matrix of this three-dimensional array. We denote by

$$C^t_{i,l} \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L$$

the $N_t \times L$ matrix of characteristics at time t.

Based on our above-mentioned empirical findings, we use the time-series and cross-sectional dependencies in characteristics to infer missing values. The fundamental problem is estimating a low-dimensional model to infer a characteristic value with contemporaneous observed cross-sectional, past and (possibly) future information. The model-implied values are used to impute missing values. We use an estimation approach that allows us to estimate the parameters of the model in the presence of missing values.

3.1. Cross-Sectional Information

An essential building block for our model is based on a cross-sectional factor model. We begin by estimating a low-dimensional cross-sectional factor model with a generalization of PCA for each month t as follows:

$$C_{i,l}^t = F_i^t {\Lambda_l^t}^\top + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

The upper index t indicates that we can have separate factor models for each time t. We assume a K-factor model, that is, $F^t \in \mathbb{R}^{N_t \times K}$ and $\Lambda^t \in \mathbb{R}^{L \times K}$. Without missing values, we can estimate F^t and Λ^t as the singular values of C^t, that is, we apply a simple PCA to ${C^t}^\top C^t$. More specifically, we can obtain $\Lambda^t \in \mathbb{R}^{L \times K}$ as the eigenvectors of the K largest eigenvalues of the $L \times L$ matrix,

$$\frac{1}{N_t} \sum_{i=1}^{N_t} C_i^t {C_i^t}^\top.$$

The different entries in this "characteristic covariance" matrix indicate how close two different characteristics are to each other. In the presence of missing values, we can use the approach of Xiong and Pelger (2023) and estimate the loadings Λ^t as the scaled eigenvectors of the estimated characteristic covariance matrix

$$\hat{\Sigma}_{l,p}^{\text{XS},t} = \frac{1}{Q_{l,p}^t} \sum_{i \in Q_{l,p}^t} C_{i,l}^t C_{i,p}^t,$$

where $Q_{l,p}^t$ is the set of all stocks that are observed for the two characteristics l and p at time t. By construction, $|Q_{l,p}^t| \leq N_t$. The characteristic factors can be estimated by a regression on the estimated loadings $\hat{\Lambda}^t$, as follows:

$$\hat{F}_i^{t,0} = \left(\frac{1}{L} W_{i,l}^t \hat{\Lambda}_l^t (\hat{\Lambda}_l^t)^\top \right)^{-1} \left(\frac{1}{L} \sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t C_{i,l}^t \right),$$

where $W_{i,l}^t = 1$ if characteristic l is observed for stock i at time t and $W_{i,l}^t = 0$ otherwise. Hence, this is a linear regression using only observed values. Xiong and Pelger (2023) provide the formal theory and show that this estimator is consistent under general assumptions on the approximate factor model and the missing pattern. The setup is a large-dimensional panel; that is, both N_t and L are large, but can grow at general and possibly different rates. An approximate factor model assumes that asymptotically most of the dependency is captured by the factors, while the "idiosyncratic"

characteristic residuals $e_{i,l}^t$ are only weakly dependent. This setup allows for a different factor model at each time t and, hence, is a *local* model.

Imputing missing values with the common component of the factor model can be interpreted as forming "characteristic mimicking portfolios" of observed characteristics that have the highest correlation with the target characteristic that should be imputed. The loadings identify which characteristics are "similar', that is, different loadings can capture different correlation clusters shown in Figure 5. The common component of stock i is a weighted average of observed characteristics of this stock, where the weights are learned from the correlation pattern in the cross-section of stocks. A factor model can be interpreted as a generalization of using an industry average for imputing missing characteristics, but instead of defining similarity ad hoc, we learn it from the data.

We generalize the estimator of Xiong and Pelger (2023) by including a ridge regularization in the factor estimation. Empirically, this regularization will lead to substantial improvements for the imputation of firm characteristics. This is because the regularization can reduce the variance of the estimated factors, which can be large for some stocks due to missing patterns in firm characteristics. Our novel robust estimator can be viewed as a combination of the factor model estimator of Xiong and Pelger (2023) for missing data with the robust regularized factor model estimator of Bai and Ng (2019). Our benchmark local cross-sectional estimator (local XS) is defined as follows:

Definition 1 (Local XS Factor Model Estimation). *The cross-sectional factor model F_i^t and Λ_l^t at time t is estimated as follows:*

(a) Estimate the $L \times L$ "characteristic covariance matrix" that captures the cross-sectional dependencies as

$$\hat{\Sigma}_{l,p}^{XS,t} = \frac{1}{Q_{l,p}^t} \sum_{i \in Q_{l,p}^t} C_{i,l}^t C_{i,p}^t,$$

where $Q_{l,p}^t$ is the set of all stocks that have characteristic l and p observed at time t.

(b) The characteristic loadings are estimated as the scaled eigenvectors of the largest K eigenvalues of $\hat{\Sigma}_{l,p}^{XS,t}$. In more detail, denote the diagonal matrix of the largest K eigenvalues of $\hat{\Sigma}^{XS,t}$ by $\hat{D}^t$ and the corresponding eigenvectors by $\hat{V}^t$. The loading estimator is given by

$$\hat{\Lambda}^t = \hat{V}^t (\hat{D}^t)^{1/2}.$$

(c) Estimate the characteristic factors from a regularized ridge regression on the loadings:

$$\hat{F}_i^{t,\gamma} = \left(\frac{1}{L} \sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t (\hat{\Lambda}_l^t)^\top + \gamma I_K \right)^{-1} \left(\frac{1}{L} \sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t C_{i,l}^t \right),$$

where $\gamma \geq 0$ is the regularization parameter and $W_{i,l}^t = 1$ if stock i has characteristic l observed at time t, and $W_{i,l}^t = 0$ otherwise.[11]

(d) We impute missing values with the estimated common component given by $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$.

The estimation is quite simple, and requires only a PCA and ridge regression. Appendix A.1 provides the general inferential theory. The cross-sectional factor model is consistent and the imputed values have an asymptotic normal distribution under general assumptions. Appendix A.2 illustrates the econometric theory in analytical form for a simplified model. We show that the variance of the common component can be large if either little data is observed, or the characteristics with larger loadings are more likely to be missing. This is the reason why we advocate for a regularized model.

The regularized factor estimator is a scaled version of the unregularized factor estimator

$$\hat{F}_i^{t,\gamma} = \Delta^{\gamma,t}\hat{F}^{t,0}, \qquad \text{with } \hat{\Delta}^{\gamma,t} = \left(\frac{1}{L}\sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t (\hat{\Lambda}_l^t)^\top + \gamma I_K\right)^{-1} \left(\frac{1}{L}\sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t (\hat{\Lambda}_l^t)^\top\right),$$

where the limit $\Delta^{\gamma,t} = \text{plim}_{N_t,L\to\infty}\hat{\Delta}^{\gamma,t}$ has eigenvalues ≤ 1.

Shrinkage has two effects: First, the asymptotic covariance of the factors is multiplied by $\Delta^{\gamma,t}$, which reduces the overall variance. Second, the shrinkage leads to a bias in the common component of the form $F_i^t\left(\Delta^{\gamma,t} - I\right)\Lambda_l^{t\top}$. These two effects lead to a bias-variance trade-off, and an optimally selected shrinkage γ can (substantially) reduce the asymptotic mean-squared error (MSE). The MSE is the sum of the asymptotic variance (Var) and the squared asymptotic bias (Bias). The unregularized estimator is unbiased, which means that asymptotically the mean-squared error equals the variance of the common component, that is $\text{Bias}(\hat{C}_{i,l}^{t,0}) = 0$ and MSE=$\text{Var}(\hat{C}_{i,l}^{t,0})$.

For notational convenience, we consider the case of $K = 1$ factor, which simplifies the term $\Delta^{\gamma,t} = \frac{d_1^t}{d_1^t+\gamma}$, where $d_1^t = D_{1,1}^t$ equals the largest population eigenvalue of the systematic component. Asymptotically, the variance of the regularized estimator equals $\text{Var}(\hat{C}_{i,l}^{t,\gamma}) = \left(\Delta^{\gamma,t}\right)^2 \text{Var}(\hat{C}_{i,l}^{t,0})$,

[11] The imputation $C_{i,l}^t$ with the common component is a weighted average of all characteristics of stock i. As by construction for any out-of-sample imputation we do not observe $C_{i,l}^t$, this weighted average does not include $C_{i,l}^t$ in an out-of-sample evaluation. We advocate to also remove $C_{i,l}^t$ in the estimation of the common component for $C_{i,l}^t$. This has two desirable effects. First, the in-sample estimation results are a better reflection of the out-of-sample results. (The out-of-sample results are by construction not affected by this choice.) Second, while in the limit the idiosyncratic term $e_{i,l}^t$ will be orthogonal to the estimated common component, we will have this orthogonality now also in finite samples. This can be beneficial when the common component is used as an input for a follow-up regression as in our combined cross-sectional and time-series model. Importantly, none of the asymptotic results is affected by leaving out one observations, as asymptotically a single observation has a measure of zero, and hence the theory does not require any modifications. Leaving out the target characteristic in the estimation of the latent factor reflects that our goal is not solely to estimate a low rank model, but to use it for the purpose of imputation.

whereas the bias is $\text{Bias}(\hat{C}_{i,l}^{t,\gamma}) = (\Delta^{\gamma,t} - 1)\left(F_i^t(\Lambda_l^t)^\top\right)$. As a result, the ratio of MSEs between the regularized and unregularized estimators can be expressed as follows and, importantly, can be smaller than one:

$$\frac{\text{MSE}(\hat{C}_{i,l}^{t,\gamma})}{\text{MSE}(\hat{C}_{i,l}^{t,0})} = (\Delta^{\gamma} - 1)^2 \frac{(F_i^t(\Lambda_l^t)^\top)^2}{\text{Var}(\hat{C}_{i,l}^{t,0})} + (\Delta^{\gamma})^2.$$

As missingness increases the asymptotic variance for characteristics with many missing values, $\text{Var}(\hat{C}_{i,l}^{t,0})$ can be relatively large, and hence shrinkage becomes particularly beneficial in the presence of missing data. The extreme case of $\gamma \to \infty$ leads to median imputation with the largest bias but also the smallest variance. The optimal value of the tuning parameter γ is based on the bias-variance trade-off to minimize the mean-squared error and selected optimally from the data.

There are three different interpretations of the benefits of including shrinkage. First, as explained above, the optimal bias-variance trade-off improves the precision of imputed values. Second, our shrinkage is adaptive and only has an effect for factors that lead to smaller eigenvalues, but is negligible for dominant factors that lead to large eigenvalues. In this sense, shrinkage can help to (partly) recover weaker factors, which in the presence of missing data can have a large estimation variance. As we see empirically, the largest benefits from shrinkage arise when including a large number of factors K, that is, with shrinkage we can leverage the information in weaker factors. Third, shrinkage also has an interpretation of combining a weighted average of a median imputation and an unregularized factor model imputation. The factors correspond to weighted averages of the observed units. The loadings can be interpreted as the weights in this average. Even in the case when the estimate of the loadings is very precise, the weighted average of the observed characteristics can be noisy if the characteristics with the largest loadings are not observed. Shrinkage will then pull the remaining small loadings toward zero, which otherwise would obtain a disproportionally large weight relative to the case of fully observed data. Intuitively, when a factor cannot be reliably constructed from the observed characteristics, we shrink it toward the median rather than creating a noisy version of it.

The shrinkage parameter γ is selected optimally from the data, and the case of no shrinkage is included as a special case. Empirically, we show that for our data the optimal γ is around $0.01/L$, which corresponds exactly to the empirical magnitude of the average non-systematic variance, that is, the bulk spectrum of the non-systematic eigenvalues of $\frac{1}{L}\hat{\Sigma}^{\text{XS},t}$. This is consistent with assuming

that $\gamma \leq O\left(\frac{1}{\sqrt{\delta}}\right)$, where $\sqrt{\delta} = \sqrt{\min(L, N_t)}$ is the convergence rate of the factor model. This implies that the convergence rate of the regularized estimator is the same as the unregularized estimator. In other words, we maintain the same consistency and convergence rate, but in a finite sample benefit from the lower variance of the estimation.

Based on our empirical findings, the "loadings" Λ are close to constant over time, which results in the model

$$C_{i,l}^t = F_i^t {\Lambda_l}^\top + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

Under the assumption of constant loadings, we can estimate Λ from an average "characteristic covariance matrix". That is, in step 1 of Definition 1, we replace the local characteristic covariance matrix with

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\Sigma}_{l,p}^{\text{XS},t}.$$

As this estimation uses the full data, it represents a *global* model, which we label *global XS* model. If the loadings are constant over time, the global model can be more precise, because it uses substantially more data.[12]

3.2. Time-Series Information

We combine the XS (cross-sectional) information with TS (time-series) information. Given an estimate of the contemporaneous systematic XS factor component, we combine those with past and (possibly) future time-series information. The information based on correlation with other observed characteristics can be efficiently learned with the XS model. However, the component of a characteristic, which is only weakly correlated with other characteristics, can be only learned from past (or future) information, but not from other observed characteristics. We consider a backward cross-sectional model (B-XS) with only the past observed information and a backward-forward cross-sectional model (BF-XS), which combines past and future information.

We use as a motivating example a data-generating process with an AR(1) structure in the system-

[12] If the loadings span the same factor space over time, then the global model provides a reference "rotation" of the factors. Such a reference rotation is related to the idea of "glueing together" locally estimated latent factors as in Zaffaroni (2022) and Pelger (2020).

atic and non-systematic components:

$$C_{i,l}^t = F_i^t (\Lambda_l^t)^\top + e_{i,l}^t, \qquad F_{i,k}^t = \rho_k^F F_i^{t-1} + \epsilon_i^{F,t}, \qquad e_{i,l}^t = \rho_l^e e_{i,l}^{t-1} + \epsilon_{i,l}^{e,t}.$$

The formal treatment of this model is presented in Appendix A.3, and we focus on the intuition here. Under the assumptions of the XS factor model, we can identify the systematic component and the non-systematic residual component at each point in time. This separation allows us to estimate a time-series model for the residual component. Hence, in a first step we can estimate $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$ for each t, and for all observed characteristics the residuals $\hat{e}_{i,l}^{t-1} = C_{i,l}^{t-1} - \hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$. In a second step, we can use a cross-sectional regression to estimate the AR(1) coefficients for residuals, and then combine the XS with the TS model to obtain the imputed value $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top + \hat{\rho}_{e,l}\hat{e}_{i,l}^{t-1}$. We can further reduce the variance of the estimation by leveraging the potential AR(1) structure in the common component.

We propose a simple estimator to combine the contemporaneous systematic correlation and the information in the time-series of the systematic and non-systematic component. For this purpose, we use a three-dimensional vector of covariates $\left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad \hat{F}_i^{t-1,\gamma}(\hat{\Lambda}_l^{t-1})^\top \quad \hat{e}_{i,l}^{t-1}\right)$, and model the characteristic as linear function in these covariates. Note that including $C_{i,l}^{t-1}$ and $\hat{e}_{i,l}^{t-1}$ is equivalent to including $\hat{F}_i^{t-1,\gamma}(\hat{\Lambda}_l^{t-1})^\top$ and $\hat{e}_{i,l}^{t-1}$ in a linear model, but it gives a more interpretable interpretation and more transparently includes conventional models as special cases. Hence, we define the following covariates:

$$X_{i,l}^{t,\text{B-XS}} = \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right),$$

and estimate the model

$$\hat{C}_{i,t}^{l,\text{B-XS}} = (\beta^{l,t,\text{B-XS}})^\top \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$$

with a weighted cross-sectional regression on the partially observed data at time t. We define this model as a *local B-XS model* and summarize the estimation steps in the following definition:

Definition 2 (Local B-XS Factor Model Estimation). *The B-XS (backward-cross-sectional) model is estimated as follows:*

(a) Estimate the XS model $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$ for each time t. If the characteristic $C_{i,l}^{t-1}$ is observed, estimate

the residual $\hat{e}_{i,l}^{t-1} = C_{i,l}^{t-1} - \hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$. *We collect the explanatory covariates in*

$$X_{i,l}^{t,B\text{-}XS} = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right).$$

(b) The coefficients $\beta^{l,t,B\text{-}XS}$ *are estimated with a cross-sectional regression on the partially observed data*

$$\hat{\beta}^{l,t,B\text{-}XS} = \left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^{t,B\text{-}XS}\left(X_{i,l}^{t,B\text{-}XS}\right)^\top\right)^{-1}\left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^{t,B\text{-}XS} C_{i,t}^t\right), \tag{1}$$

where $W_{i,l}^t = 1$ *if* $X_{i,l}^t$ *and* $C_{i,t}^t$ *are observed and 0 otherwise. The imputed values are* $\hat{C}_{i,l}^{t,B\text{-}XS} = \hat{\beta}^{l,t,B\text{-}XS} X_{i,l}^t$.

We can easily generalize this model to include future information by expanding the set of covariates. The *local BF-XS model* replaces the covariates in step 2 of Definition 2 with

$$X_{i,l}^{t,\text{BF-XS}} = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1} \quad C_{i,l}^{t+1} \quad \hat{e}_{i,l}^{t+1}\right).$$

The framework includes several important special cases, as follows:

(a) Time-series AR(1) model (B): $\beta^{l,t,\text{B-XS}} = \begin{pmatrix} 0 & \beta^{\text{l,t,B}} & 0 \end{pmatrix}$.

(b) Last observed value (PV): $\beta^{l,t,\text{B-XS}} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$.

(c) Cross-sectional median: $\beta^{l,t,\text{B-XS}} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$ (as we have centered the rank quantiles at 0).

We estimate the β vectors in a cross-sectional regression using the stacked observed values. In the local model, we use the local XS common component and the observed characteristics for the time t to obtain the local $\hat{\beta}^t$, whereas for the global model we use globally estimated common components in a regression that stacks all characteristics over time. For a given cross-sectional and time-series information set in the vector $X_t^{i,l}$, we obtain the global model from a global regression

$$\hat{\beta}^l = \left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t (X_{i,l}^t)^\top\right)\right)^{-1}\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t C_{i,l}^t\right)\right).$$

Table 2 summarizes the different estimation approaches. For each estimator we have a local version that uses information only at time t and a global version that uses information from the full time-series.

Table 2: Different Imputation Methods

Method	Estimation
Backward-Forward-XS (BF-XS)	$\hat{C}_{i,l}^{t,\text{BF-XS}} = (\hat{\beta}^{\text{BF-XS}})^\top \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1} \quad C_{i,l}^{t+1} \quad \hat{e}_{i,l}^{t+1}\right)$
Backward-XS (B-XS)	$\hat{C}_{i,l}^{t,\text{B-XS}} = (\hat{\beta}^{l,\text{B-XS}})^\top \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$
Forward-XS (F-XS)	$\hat{C}_{i,l}^{t,\text{F-XS}} = (\hat{\beta}^{l,\text{F-XS}})^\top \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t+1} \quad \hat{e}_{i,l}^{t+1}\right)$
Cross-sectional (XS)	$\hat{C}_{i,t}^{\text{XS}} = \hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$
Time-series (B)	$\hat{C}_{i,l}^{t,\text{B}} = \hat{\beta}^{l,\text{B}} C_{i,l}^{t-1}$
Previous value (PV)	$\hat{C}_{i,l}^{t,\text{PV}} = C_{i,l}^{t-1}$
Cross-sectional median	$\hat{C}_{i,l}^{t,\text{median}} = 0$

This table summarizes the different estimation approaches. Each estimation approach has a local and global version.

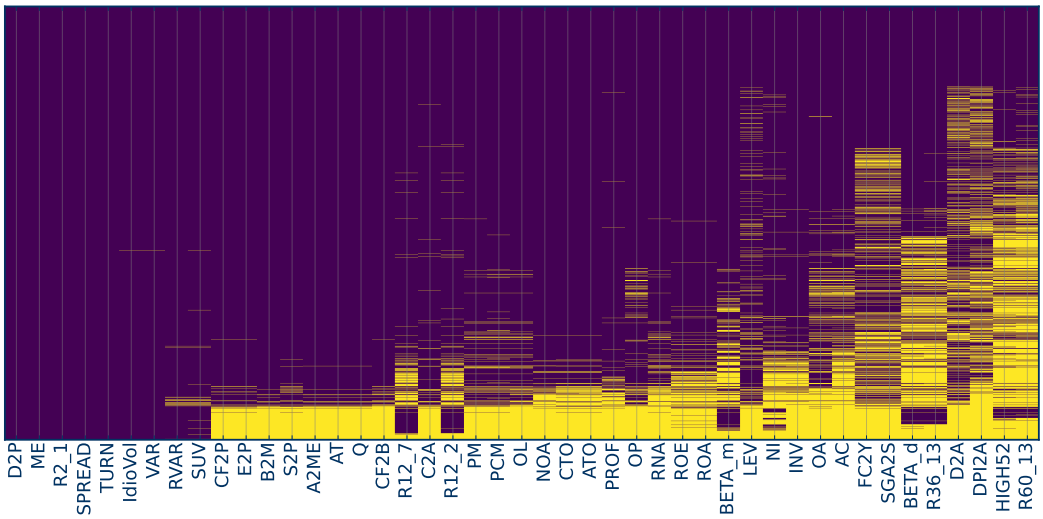
3.3. Model Assumptions and Distribution of Missingness

Our estimation method allows for general missing patterns in firm characteristics. We discuss the formal assumptions on the missingness and the model itself. As we have established in the first part of our empirical analysis, firm characteristics are not missing-completely-at-random (MCAR). This has implications for how to correctly impute missing values.

First, we formalize different probabilistic models for missingness. The simplest case is that data is MCAR. Formally, in this case the probability of missingness takes the form $\mathbb{P}(W_{i,l}^t) = p$ for a constant probability p and hence the missingness can, for example, not depend on the specific characteristics of a company. This assumption is often assumed in the literature on matrix completion as discussed in Chen et al. (2019). However, this assumption is certainly not appropriate for firm characteristics. The missingess in characteristics is complex, as illustrated in Figure 6. We show the joint distribution of missing patterns on a representative example month. The plot shows the missing entries for each firm, where the characteristics are sorted by their missing percentages. Obviously, the MCAR assumption is clearly violated, and the missingness is heterogenous and dependent among characteristics. This dependency is also expected, as many characteristics depend on similar CRSP or Compustat variables in their construction, as explained in more detail in the Internet Appendix. Similarly, we also observe heterogeneous missingess and dependency in the time dimension.

A more general notation is conditional-missing-at-random. Under this notion, characteristics are missing-at-random once we condition on a given a set of covariates $\tilde{X}_{i,l}^t$. Formally, the probability of missingness can be expressed as $\mathbb{P}(W_{i,l}^t) = f(\tilde{X}_{i,l}^t)$ for a function $f(.)$ and a set of observed conditioning variables $\tilde{X}_{i,l}^t$. This more general model can allow for heterogeneous missingess and

Figure 6: Joint Distribution of Missing Patterns



This figure shows the heatmaps of missing data for each stock for a representative example day. Both axis are sorted by the missing percentage, where we first order by firms and then characteristics. Missing data is indicated in yellow. We show the data for April 1986. However, this pattern is consistent across time.

dependency in the cross-sectional and time dimension for an appropriate set of conditioning variables. Our model does not require us to directly model the probability of missingness, and the assumptions are not stated in terms of the probability of missingness. However, in order to provide intuition, we clarify which conditional-missing-at-random distributions are allowed in our model.

We provide the assumptions for the two components of our model. First, we explain the necessary conditions for the cross-sectional regressions that underlie the B-XS model. Then we discuss the assumptions of the latent XS factor model. As the local B-XS model is our benchmark, we focus the analysis on this model. The arguments for the BF-XS, the pure time-series and global models are analogous and straightforward to adopt, given this discussion.

We only need mild assumptions for estimating the coefficients β in a cross-sectional regression with partially observed data. Assume that characteristics can be described by a linear model of the form

$$C^t_{i,l} = \beta X^t_{i,l} + \epsilon^t_{i,l}. \tag{2}$$

This is also the form of the local B-XS model, when we take the common components and residuals as given. A cross-sectional regression with partially observed data as described in equation 2 only requires that $\frac{1}{N_t}\sum_{i=1}^{N_t} W^t_{i,l} X^t_{i,l} (X^t_{i,l})^\top$ converges to a full rank matrix, while $\frac{1}{N_t}\sum_{i=1}^{N_t} W^t_{i,l} X^t_{i,l} \epsilon^t_{i,l}$ goes to

0 for $N_t \to \infty$. Hence, we only require the simple moment conditions of a regression adopted to our setup but do not model the missingness directly. The moment assumptions that we impose correspond to the "ignorability assumption" of the Rubin causal inference model (Rubin (1976, 1978)) and are closely related to the "unconfoundedness assumption" in Rubin and Rosenbaum (1983)) applied to a regression framework. In addition to the standard regularity assumption in a regression with fully observed data, a sufficient condition would be that $\epsilon_{i,l}^t$ is independent of $W_{i,l}^t$ after conditioning on $X_{i,l}^t$. An example for a model that satisfies this assumption is conditional-missing-at-random of the form $\mathbb{P}(W_{i,l}^t) = f_{i,t}(F_i^t(\Lambda_l^t)^\top, C_{i,t-1}^l, W_{i,l}^{t-1})$ for a set of functions $f_{i,t}(.)$ that can vary over time and stocks.

The general assumptions for the XS latent factors are formalized in Xiong and Pelger (2023), whereas Appendix A.2 presents a simplified model to provide intuition for the setup. First, we assume a general approximate factor model in the spirit of Bai (2003) and consider the asymptotics for N_t and L going to infinity. Here, we focus on the assumptions for the missingness. In the case of fully observed data, the PCA estimation of a factor model is symmetric in N_t and L, that is, we obtain the same model if we first estimate the loadings or factors as eigenvectors, and in a second step the other quantity with a regression on those eigenvectors. In the case of missing data, the order of estimation is no longer symmetric, and impacts the assumptions on the generality of missing patterns in one of the two dimensions. Essentially, we need to assume that asymptotically the eigenvectors of the largest eigenvalues of the "characteristic covariance" $\tilde{\Sigma}^{\text{XS},t}$ are the same if they are estimated either on the partially observed data or on the infeasible complete data. A sufficient condition is that, given the model $C_{i,l}^t = F_i^t \Lambda_l^{t\top} + e_{i,l}^t$, the random variable $W_{i,l}^t$ is independent of F_i^t and $e_{i,l}^t$, but can depend in general terms on Λ_l^t and other stock specific information S_i^t. If we switch the estimation order, then a sufficient condition is that the random variable $W_{i,l}^t$ is independent of Λ_l^t and $e_{i,l}^t$, but can have general dependence on F_i^t and other stock specific information S_i^t. An example for a model that satisfies this assumption is conditional-missing-at-random of the form $\mathbb{P}(W_{i,l}^t) = f_{i,t}(F_i^t, S_i^t, W_{i,l}^{t-1})$, where $f_{i,t}$ can be different functions for different stocks and time periods, and the missingness can depend on the latent factors, other stock specific information that are independent of $e_{i,l}^t$, and prior missingness. Empirically, we have implemented both estimation orders, and we obtain the same results. Our benchmark model in Definition 1 uses the more convenient and interpretable order, where the eigenvector decomposition is applied to the smaller $L \times L$ matrix. Appendix A.4 provides a detailed discussion on alternative implementations that can

have theoretical advantages, but, as we show empirically, give the same results as our simple model. Combing the two sets of assumptions, we obtain a consistent estimator for $\beta X_{i,l}^t$, which we use for imputation in the B-XS model.

Our imputation method is particularly well-suited for firm characteristics, as we allow the missingness to be heterogenous, time-varying, stock-specific and dependent on the latent factor model. We will discuss examples of the probability of missingness $\mathbb{P}\left(W_{i,l}^t = 0\right) =: p_{i,l}^t$ that is allowed in our model. The probability can depend on the specific stock i, the characteristic l, and the time t. First, note that our setup allows for different factor and regression models at each time t, and, hence, imposes no assumptions on the temporal structure of $p_{i,l}^t$. This means that the missingness can vary in a completely general way over time, which includes periods of more unobserved data, such as at the beginning of our sample, block-missing patterns, mixed-frequency observations, or missingness because prior values are unobserved. The probability of missingness is also be notably general in the characteristic dimension and can be different for each characteristic. This allows for characteristic-specific heterogeneity; for example, DPI2A has a higher probability of being unobserved than book-to-market ratios. Another case is group-specific heterogeneity, where, for example, there are fewer observations when characteristics are updated quarterly or when a group of characteristics relies on the same accounting variable as an input. Finally, the probability of missingness can in a very general way depend on the features of each stock. More precisely, the probability can be a general, time-varying, and characteristic-specific function of a vector of stock-specific information $S_i^t \in \mathbb{R}^r$, the stock-specific factors F_i^t and past values. For example, the characteristics of small stocks or more extreme characteristic realizations are more likely to be unobserved, which we can account for if it is captured by the latent factors or past observations.

3.4. Discussion

3.4.1. Alternative Imputation Approaches

In this section we discuss related models for characteristic imputation. The covariates $X_{i,l}^t$ in the regression model in Equation 2 can also be only the subset of characteristics that are observed for all stocks at the same time. This is main model in Freyberger et al. (2022). In our data set, this subset of characteristics corresponds to the seven characteristics in Table 1. The advantage of our latent factor model is that we use all partially observed characteristics for the imputation, rather than only a small subset. If some characteristic values are observed for a specific stock and they are correlated

with the missing characteristics, then they should contain useful information. Our approach takes advantage of all the data by leveraging the correlations between all observed characteristics, which can be different for different stocks. Furthermore, a larger information set included in $X_{i,l}^t$ also allows for more general missing patterns, as specified by our ignorability assumption. We interpret using the subset of fully observed covariates for $X_{i,l}^t$ as a special case of our model.

A well-known approach in the literature is the Expectation Maximization (EM) algorithm, which requires specifying the distribution of the characteristics. Chen and McCoy (2022) use the EM algorithm for characteristic imputation and require that the characteristics are jointly normally distributed. In the case of the EM algorithm the ignorability assumption is in terms of the likelihood rather than moment conditions. On an intuitive level, the observed characteristics that they condition on have to be sufficient to estimate a likelihood model only with the observed data without modeling the missingness. The normality assumption implies an underlying linear model between characteristics. Similar to our XS model, their EM implementation leverages the correlation with observed contemporaneous characteristics. The difference is that we use a low rank model for the correlation, whereas they use the correlations with all characteristics under the normality assumption. As stated in Cahan et al. (2023), EM-algorithms are usually considered for low dimensional data, while high dimensional data sets benefit from regularization. The low rank formulation in the form of factor models is the standard for imputation in large matrices as discussed in Chen et al. (2019). An EM-algorithm for large dimensional factor models is introduced in Jin et al. (2021). Xiong and Pelger (2023) show that under appropriate assumptions, this is a special case of our more general XS model. Importantly, the EM algorithm of Chen and McCoy (2022) only includes contemporaneous XS information. We show in our empirical analysis that it is crucial to include the time-series information (TS), whenever it is available, for a reliable imputation.

In fact, the EM algorithm under a normality assumption can be interpreted as a special case of our XS factor model, when setting the number of latent factors to $K = L$. We discuss this perspective in Appendix A.5, where we show that the cross-sectional updating step of the EM algorithm is similar to our cross-sectional regression. The regularized XS factor model relates to an EM algorithm with an underlying factor structure and diagonal matrix for the idiosyncratic component. We show in a simulation in Appendix A.6 that the XS factor model and EM algorithm can perform similarly when the number of latent factors is large while the dimension L is moderate. However, for large dimensions L, the factor model approach substantially outperforms the EM algorithm. In our empir-

ical comparison analysis, we also include the alternative estimation approaches of Freyberger et al. (2022) and Chen and McCoy (2022) and demonstrate that they perform worse than our benchmark B-XS model.

An alternative data-driven approach is to use a machine-learning method to predict characteristics either based on their own past and/or given the contemporaneous realizations of other characteristics. As an example, Beckmeyer and Wiedemann (2022) use a machine-learning algorithm from natural language processing for imputation. Due to the complexity of machine-learning models, there needs to be tuning parameter selection, which usually is obtained by masking characteristics. If such a model is estimated by masking characteristics completely at random, then it would only be appropriate to impute characteristics, which are missing-completely-at-random. Therefore, a machine-learning application with random masking on the training data could lead to a bias in imputed characteristic values. In general, it it is unclear what exactly the required assumptions and theoretical properties are of such an imputation.

3.4.2. Look-Ahead Bias

The choice of imputation method has implications for the follow-up application. Using more data, either in the form of a global model or by incorporating future information, generally improves the quality of the imputation. However, some of the most important use cases of the characteristics data, including out-of-sample asset pricing and investment, should avoid a look-ahead-bias. This means that future information cannot be used in the imputation, as it could make the performance of an investment strategy appear to be better than what it is actually achievable. Blanchet et al. (2022) discuss the trade-off between look-ahead bias and the precision of the imputation.

In our empirical study, the model that uses the most information while avoiding any look-ahead bias is the local B-XS model. The model that uses the most information overall, but also "peeks" into the future, is the global BF-XS model. These two benchmark models allow us to study the trade-off between using more data and using future information. There are other modifications of our models that could avoid a look-ahead bias while using more data. Rather than using only the current month for the local B-XS model, we could use a rolling window for a "locally" global version of the B-XS model. However, as we will show in our analysis below, the factor structure of the cross-sectional factor model is very stable over time. Hence, the global XS and B-XS models are very close to a rolling window look-ahead bias-free version. The more serious look-ahead bias can arise from directly using

future information as an input for imputation, that is, in the forward models.

3.5. Rank Normalization vs. Raw Characteristics

We model rank-normalized data, which can easily be mapped back into raw characteristics. In order to obtain a statistical model for characteristics, we need to appropriately normalize them. Fundamentally, this relates to the conceptual question about how we model dependency. Centered rank-normalized characteristics are the natural choice. By using ranks, we handle the outliers in the raw characteristics and also achieve stationarity in the cross-section and over time.

We show in Appendix B.1 that our main results carry over to raw characteristics. There is a simple mapping between the rank quantiles and raw values through the empirical density function of each characteristic. Therefore, after estimating the density functions, the imputed rank quantiles also provide imputed values for the raw characteristics. We estimate the density function non-parametrically and also parametrically assuming a normal distribution. In both cases, we do not assume that there is a linear dependency between raw characteristics, but only between their relative ranks. As a further robustness result, we also include the results for a factor model that is directly applied to the characteristic space. This requires us to normalize the raw characteristic values by their cross-sectional median and cross-sectional standard deviation after winsorizing the extreme outliers.

We center our ranks at zero; that is, we report characteristic quantiles between $[-0.5, 0.5]$, which is without loss of generality. Hence, the cross-sectional median corresponds to the value of zero. Using uncentered rank quantiles between $[0, 1]$ adds an additional latent cross-sectional factor, one that captures the median and is similar to a "market" or "level" factor.

3.6. Evaluation Metrics

We evaluate the different models based on their RMSE (root-mean-squared errors). The aggregated RMSE for the model implied characteristic $\hat{C}_{i,t,l}$ is averaged over all stocks, time periods, and characteristics, as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

We also consider the RMSE for each characteristic separately, and over time as follows:

$$\text{RMSE}_l = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}, \qquad \text{RMSE}_t = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

All our results are reported in-sample and out-of-sample. The in-sample results evaluate how well a low-dimensional model can approximate the characteristics. Because these results can be biased upwards due to overfitting, we also need to conduct an out-of-sample analysis (OOS). The OOS analysis masks observed entries before we estimate the model on the remaining data. The OOS RMSE compares the masked observed entries with the model-implied values. We consider three different missing patterns for the out-of-sample analysis.

(a) *Missing-Completely-at-Random (MCAR)*:

We mask 10% of the observed characteristics completely randomly.

(b) *Block-Masking over Time (Block)*:

The second case is OOS block-missing, where we mask 10% of characteristics to capture time-series patterns in missingness. Empirically we observe that 40% of the characteristics are missing-at-the-beginning, that is, have no prior observations, and for missing-not-at-the-beginning, the average length of missing blocks is roughly one year, as shown in Table C.3. In order to account for the empirically observed temporal dependency in missing patterns, we mask characteristics randomly in blocks of one year, where 40% of the blocks are at the beginning. It is important to include this case, as for very persistent characteristics the last observed value can provide a very good prediction, but empirically it is often not available.

(c) *Empirical Missingness Distribution (Logit)*:

The third case uses the logistic regression model from Table 1, with all covariates and fixed effects, to mask entries. In more detail, we first use the logistic regression model that explains missingness-at-the-beginning, then conditional on having any previous observation, we use the logistic regression model that explains missing-in-the-middle. This is important as missing patterns at the beginning are systematically different from missingness once a characteristic is observed. The propensity of the logistic regression captures important features of missing patterns. In particular, the probability of missingness is heterogeneous, appears in blocks over time and in the cross-section and depends on the realization of observed characteristics. We mask 10% of the entries with the logistic regression propensity.

As we work with rank quantiles, the characteristics are normalized and the RMSE provides an interpretable measure of the deviation from the true value. In addition, we report the R^2 that measures the explained variation relative to the cross-sectional median imputation. The R^2 is a transformation of the RMSE that includes the total variation in the denominator.

4. Factor Structure in Characteristics

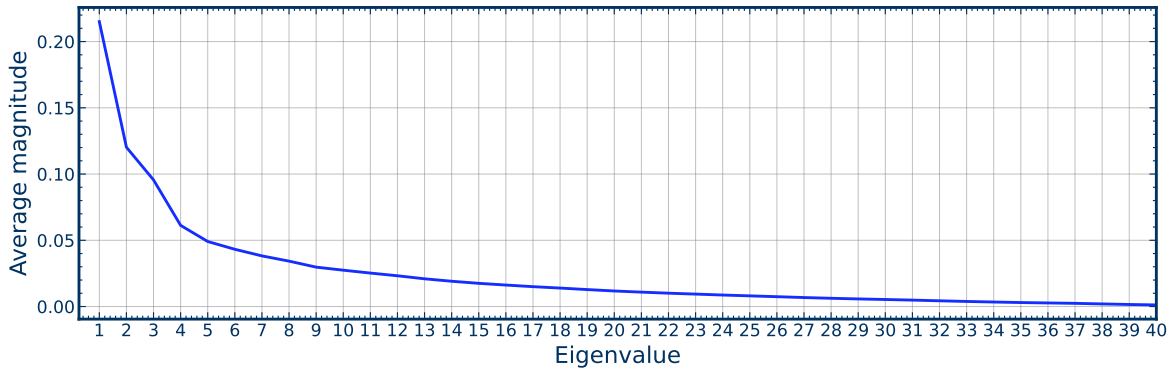
Empirically, firm characteristics are well-described by a factor model. Before conducting an extensive comparison between different imputation methods, we study the properties of a cross-sectional latent factor model. We discuss the choice of the number of factors, their economic interpretation, and their variation over time. Estimating a cross-sectional latent factor model requires observing at least some characteristics for each stock. Moving forward, at each point in time, we only include a stock if it has at least one characteristic observed. This imposes no restriction, as for all stocks some characteristics, like size, are always available. We focus on the data after 1977, which is more homogenous and more widely used in empirical applications. We have confirmed that our general results are robust to this choice.

4.1. Number of Factors

The number of systematic cross-sectional characteristic factors is directly linked to the eigenvalues of the characteristic covariance matrix $\tilde{\Sigma}^{\text{XS},t}_{l,p}$. Figure 7 plots the magnitude of eigenvalues of $\tilde{\Sigma}^{\text{XS},t}$ relative to the sum of all eigenvalues averaged over time. These eigenvalues can be interpreted as the amount of variation explained for different number of factors. The "L"-shaped curve is typical for a dataset with a factor structure. The first four factors explain the most variation in the data, and could be considered as "strong factors". It seems that the factors five to 20 might also contribute to explaining variation the data, but to a much smaller amount. These might be possible weaker factors. Overall, we find strong evidence for a factor structure.

We study the out-of-sample errors as a function of the number of factors and the regularization parameter γ. Our local XS factor model only depends on these two choice parameters K and γ. We report the R^2 for monthly, quarterly, and all characteristics. A median imputation corresponds to a 0-factor model with an R^2 equal to zero. For clarity and brevity, we present the out-of-sample results averaged over the full data. However, we obtain essentially identical results for each month, and the

Figure 7: Eigenvalues of $\Sigma^{XS,t}$



This figure shows the magnitude of eigenvalues of the characteristic covariance matrix $\hat{\Sigma}^{XS,t}$ relative to the sum of all eigenvalues averaged over time.

optimal choices are very stable over time. Hence, we can use past data as validation data to select the tuning parameters K and γ in real-time for future out-of-sample evaluation.

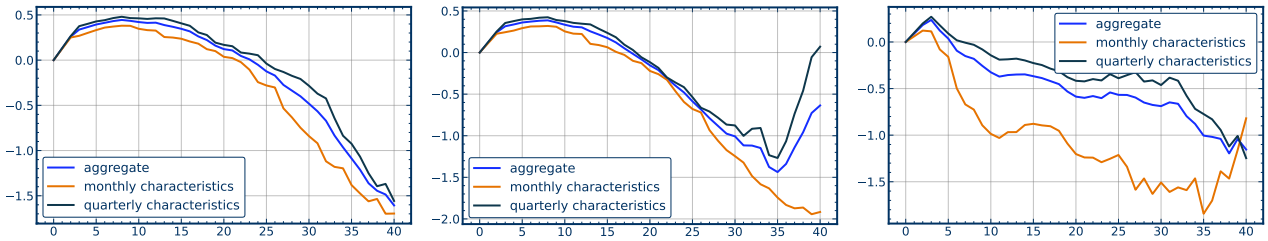
First, we determine the number of factors for a XS model without regularization, that is, $\gamma = 0$. Panel A in Figure 8 shows the out-of-sample R^2 under the three masking schemes MCAR, Block and Logit. As expected, only the dominant strong factors are helpful to explain the variation out-of-sample. In more detail, for MCAR and Block, seven factors seem to be optimal with an R^2 of around 0.4. For the more complex missingness under Logit masking, the first three factors provide an optimal performance. Panel B in Figure 8 shows that the incremental increase in out-of-sample $R2$ for all three masking schemes is the largest for the first three factors. Adding factors beyond the seventh one lowers the explained variation for each of the masking schemes, that is, leads to overfitting.

Second, we study the number of factors for the regularized XS factor model. We start with the value of $\gamma = 0.01/L$, which, based on Figure 7, corresponds to the average noise level for $\frac{1}{L}\hat{\Sigma}^{XS,t}$. We will discuss this choice in more detail below. Panel A in Figure 8 shows that regularization allows us to leverage the information in weaker factors. For all masking schemes we can increase the number of factors to 20 without losing out-of-sample performance. Importantly, as shown in Panel B, the weaker factors can have a positive incremental effect on the out-of-sample R2. In almost all cases, there is no negative effect of including "too many" factors with sufficient regularization.

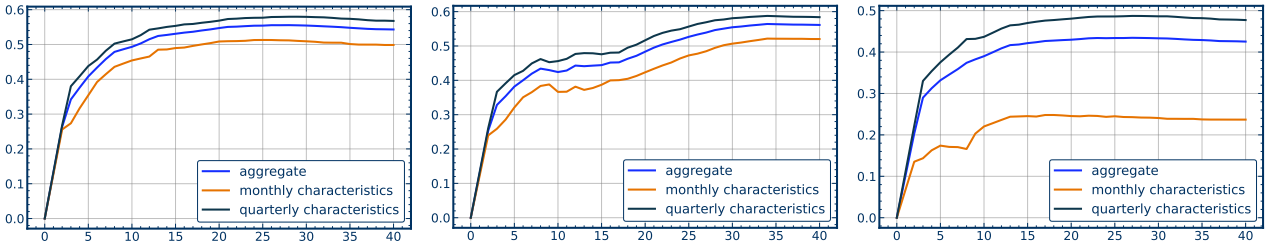
Panel A in Figure 8 suggests that the optimal number of factors for MCAR and Logit is 20, while Block might benefit from including even more factors. The out-of-sample R^2 increase to around 0.55 for MCAR and Block by using regularization and leveraging the information in weaker factors. For

Figure 8: Number of Factors and Regularization

Panel A: Out-of-sample R^2 as Function of Number of Factors

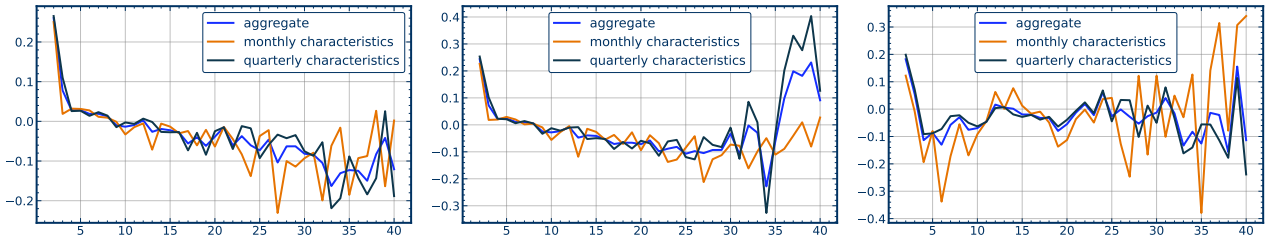


(a) $\gamma = 0$, MCAR **(b)** $\gamma = 0$, Block **(c)** $\gamma = 0$, Logit

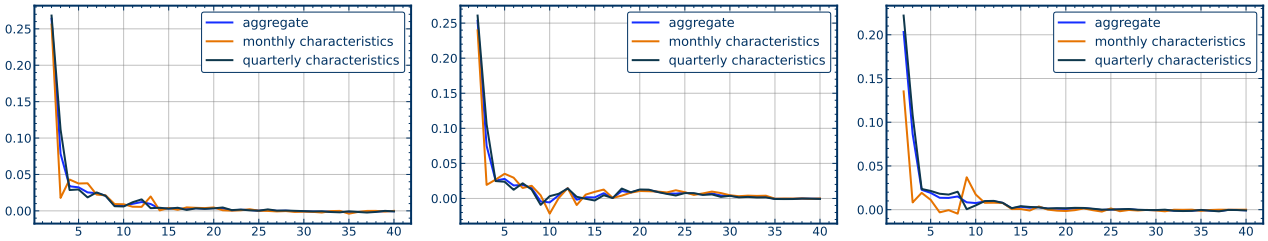


(d) $\gamma = 0.01/L$, MCAR **(e)** $\gamma = 0.01/L$, Block **(f)** $\gamma = 0.01/L$, Logit

Panel B: Incremental Out-of-Sample R^2 as Function of Number of Factors



(g) $\gamma = 0$, MCAR **(h)** $\gamma = 0$, Block **(i)** $\gamma = 0$, Logit

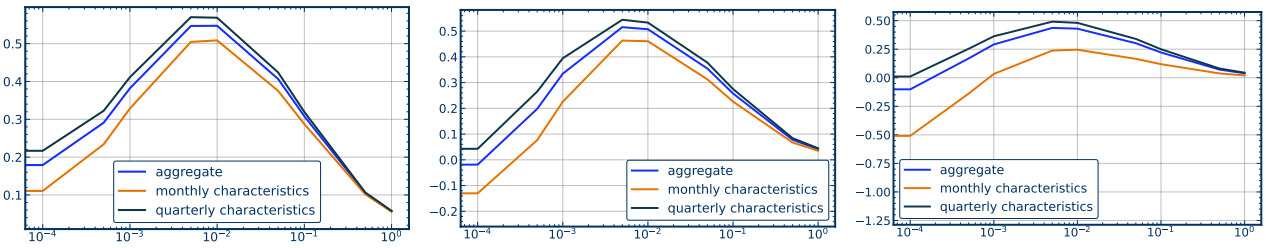


(j) $\gamma = 0.01/L$, MCAR **(k)** $\gamma = 0.01/L$, Block **(l)** $\gamma = 0.01/L$, Logit

The top panel shows the out-of-sample R^2 of the local XS model for different number of latent factors and regularization for three masking schemes missing-completely-at-random, block-missing and logit-missing. The bottom panel shows the incremental change in out-of-sample R^2 for adding factors. The R^2 is the explained variation relative to a cross-sectional median imputation.

the masking with Logit we can push the out-of-sample R^2 from 0.3 to more than 0.4. The results are even stronger for quarterly updated characteristics.

Figure 9: Optimal Regularization



(a) MAR **(b)** Block **(c)** Logit

This figure shows the out-of-sample R^2 for the local XS model with 20 factors for different regularization parameters γ. The R^2 is the explained variation relative to a cross-sectional median imputation. The plots report the scaling constant $\tilde{\gamma}$ for $\gamma = \tilde{\gamma}/L$.

It is important to highlight how factor modeling for missing data imputation differs from other applications of factor models in finance, for example, return modeling for fully observed panels. These differences are the reason why regularization and including weaker factors are important. First, the problem is not only to have a good model to describe characteristics, but also to apply it to partially observed data. Consider the top row of Panel A in Figure 8. The underlying factor model for the data should be the same but when applied to different masking schemes the performance can vary substantially. Even with perfect knowledge of the loadings Λ^t, which correspond to the factor portfolio weights, the construction of the factors depends on which observations are missing. If all characteristics, that have large weights for a particular factor, are missing at the same time, then this factor will be poorly estimated from the partially observed characteristics. The regularization directly addresses this additional challenge. Second, imputation benefits from using all correlations in the data, even if one characteristic is only correlated with a few others, which can be interpreted as a weak factor. Our goal is to leverage all information, which is different from a return asset pricing perspective, which might only focus on strong factors as non-diversifiable risk.

The choice of $\gamma = 0.01/L$ is actually close to optimal for our data. Figure 9 shows the out-of-sample R^2 for 20 factors and different choices of the regularization parameter γ. No regularization, that is, $\gamma = 0$, obviously results in an inferior and possible negative performance due to overfitting. A regularization that is too large pushes the cross-sectional factor model towards median imputation with an R^2 of zero. The choice of $\gamma = 0.01/L$ seems to be universally optimal among all masking schemes and types of characteristics. It corresponds to the average noise level, which we suggest as a starting point for optimizing the regularization. The Internet Appendix collects results for more

combinations of K and γ, which confirms the robustness of our results.

For the remainder of the paper, our baseline XS model has 20 factors with $\gamma = 0.01/L$.

4.2. Local vs. Global Factors

The loading structure of the cross-sectional factor model is relatively stable over time. A global factor model assumes a constant loading matrix Λ, whereas a local factor model allows for time-varying loadings Λ^t. We show that the loading structure is relatively stable over time and, hence, justifies the use of constant loadings. Figure D.5 in the Appendix plots the generalized correlations between the global loadings Λ and local loadings Λ^t for the first 20 factors over time. A generalized correlation equal to 20 would imply that the two loading matrices span the same space. Although there is some variation, the generalized correlation is close to the maximum. We conclude that it is meaningful to analyze the composition of the global factors, as it provides a close proxy for the local factors.

4.3. Structure of Factors

The characteristic factors have a meaningful economic interpretation. The loadings Λ can be interpreted as weights to construct the characteristic factors. We focus on the global model, as it is described by only one set of weights that are closely related to the local weights. The Internet Appendix shows the full factor weights grouped by characteristic categories. For better interpretability, we generalize the approach of Pelger and Xiong (2021) to create sparse group proxy factors. In more detail, given the estimated factor loadings $\hat{\Lambda}$, we use a of group lasso estimator to obtain a sparse set of loadings that provides a common component that is close to the non-sparse factor model. The groups are based on the characteristics categories.[13] Although we observe some decline in the

[13] Consider a low-rank factorization of a matrix $C^t = \hat{F}^t \hat{\Lambda}^T$. We want to estimate a sparse approximation of this model where the loadings $\tilde{\Lambda}$ are interpretable from an economic perspective, or, in other words, group-sparse for predefined groups G. Our sparse factor model is estimated iteratively by repeating the following steps until convergence:

$$\tilde{\Lambda}^{j+1} = \arg\min_{\Lambda} \|\hat{F}^t \hat{\Lambda}^\top - \tilde{F}^{t,j}(\Lambda)^\top\|_F + \sum_{g\in G}\sum_{k=1}^{K} \|\Lambda_{g,k}\|_2 \alpha_g$$
$$\tilde{F}^{t,j+1} = \arg\min_{F} \|\hat{F}\hat{\Lambda}^T - F(\tilde{\Lambda}^{j+1})^\top\|_F.$$

For each $g \in G$ we set the all the characteristics outside of this group to zero. The weights α_g account for the different numbers of elements in each group, which is needed to balance out the groups. If we set the groups equal to each individual element, we would obtain the usual lasso penalty. The group lasso sets groups instead of individual elements to zero.

performance of the group sparse factor model relative to our full factor model, it still substantially outperforms the median imputation, and hence it is reasonable to use it for interpretation.

Figures D.6 and D.7 show the sparse factor weights for the first ten factors. We focus on these dominating factors because Figure 8 has demonstrated that they explain most of the variation in characteristics. The first factor loads on profitability as well as value and trading frictions, and can be interpreted as capturing the correlations between these groups. Factor 2 is a value factor. Factors 3, 5 and 7 capture different types of dependencies in the profitability cluster. They represent different forms of long-short weights between profitability characteristics. Factors 4, 6, and 10 are based on trading frictions. While Factor 4 is long only and captures a general positive correlation between trading frictions, factor 6 has long-short weights and captures the negative dependencies between two groups of characteristics within the trading friction cluster. Factor 8 loads on past returns, whereas factor 9 puts most weight on intangibles.

5. Imputation

5.1. Aggregate Comparison between Methods

In an extensive comparison study we compare the quality of different imputation approaches. We include the different variations of our model framework and the most widely used conventional ways to treat missing data. The baseline model without look-ahead bias (i.e., without using future information) is the local B-XS. The baseline model that uses as much information as possible is the global BF-XS. All cross-sectional models use 20 latent factors with a regularization of $\gamma = 0.01/L$ based on the analysis in the previous section. We consider the global and local versions of our models and different subsets of time-series information, that is backward, forward, or none. Another special case would be not to use the cross-sectional model, but only the backward (AR(1) type) model. The popular conventional approaches encompass using only the previous value, a cross-sectional median, or the industry-specific median for imputation. In total, we examine the following 11 models: global BF-XS, global B-XS, global F-XS, global XS, global B, local B-XS, local XS, local B, previous value (PV), XS median, and industry median.

Importantly, some models require more input information, for example, past observed values, and hence are not applicable to missing data when this information is not available. The practically relevant problem is to impute characteristics for the full dataset, and compare different approaches for the same data. Hence, our main analysis uses the natural fallback approach, where we replace a

Table 3: Imputation Error for Different Imputation Methods

	In-Sample			OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
	Imputation RMSE											
local B-XS (XS)	**0.13**	**0.14**	**0.12**	**0.14**	**0.14**	**0.13**	**0.18**	**0.18**	**0.19**	**0.18**	**0.16**	**0.23**
local XS	0.19	0.18	0.20	0.19	0.19	0.21	0.20	0.19	0.21	0.22	0.21	0.25
local B (median)	0.15	0.15	0.14	0.15	0.15	0.15	0.22	0.22	0.23	0.21	0.19	0.26
prev val (median)	0.17	0.16	0.18	0.17	0.16	0.18	0.23	0.22	0.25	0.22	0.19	0.27
median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.30
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.30
global BF-XS (B-XS, XS)	**0.09**	**0.09**	**0.11**	**0.13**	**0.13**	**0.12**	**0.17**	**0.16**	**0.18**	**0.18**	**0.16**	**0.23**
global B-XS (XS)	0.13	0.14	0.12	0.14	0.14	0.13	0.18	0.18	0.19	0.19	0.16	0.24
global XS	0.20	0.19	0.21	0.20	0.19	0.22	0.20	0.19	0.22	0.23	0.21	0.26
global B (median)	0.15	0.15	0.14	0.15	0.15	0.15	0.22	0.21	0.23	0.21	0.19	0.26
	Explained Variation R^2											
local B-XS (XS)	**0.82**	**0.78**	**0.83**	**0.80**	**0.76**	**0.81**	**0.62**	**0.63**	**0.60**	**0.58**	**0.74**	**0.23**
local XS	0.55	0.58	0.54	0.53	0.56	0.51	0.54	0.56	0.52	0.42	0.51	0.22
local B (median)	0.75	0.75	0.74	0.74	0.73	0.73	0.40	0.47	0.37	0.45	0.65	0.02
prev val (median)	0.63	0.72	0.60	0.62	0.71	0.59	0.29	0.44	0.23	0.44	0.64	0.00
median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ind-median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
global BF-XS (B-XS, XS)	**0.87**	**0.92**	**0.85**	**0.83**	**0.81**	**0.83**	**0.63**	**0.70**	**0.60**	**0.58**	**0.76**	**0.16**
global B-XS (XS)	0.81	0.78	0.82	0.79	0.77	0.80	0.59	0.63	0.57	0.55	0.72	0.15
global XS	0.49	0.54	0.47	0.48	0.53	0.45	0.49	0.54	0.47	0.53	0.54	0.22
global B (median)	0.74	0.75	0.74	0.73	0.74	0.73	0.40	0.47	0.37	0.45	0.65	0.02

This table shows imputation RMSE and R^2 by imputation method averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from all observed data for the three masking schemes MCAR, Block, and Logit. We report the fallback method in brackets, which is used when a method is not applicable. For example, for missing characteristics of stocks without any prior values, B-XS will be replaced by XS. The R^2 is the explained variation relative to a cross-sectional median imputation.

method that requires time-series information by a pure cross-sectional model when the time-series information is not available. Our benchmark local B-XS is hence replaced by a local XS, when data is missing at the beginning. We also provide a complete comparison analysis for the subsets of the data, when specific time-series information is available.

The main results are summarized in Table 3, which shows the imputation errors for these different imputation methods. We report the in-sample, OOS missing-completely-at-random, OOS block-missing, and OOS logit-missing results for all characteristics and separated by their updating frequency. The first striking observation is that the cross-sectional median or industry median results in roughly twice as large imputation errors compared to our baseline models local B-XS and global BF-XS. These results are robust to the updating frequency and the in- or out-of-sample analysis. We conclude that the current standard of ignoring the time-series and cross-sectional dependency is

strongly suboptimal. The local and global versions of our model are relatively close. This is expected for the cross-sectional models as the local and global loadings are approximately the same as shown before. We will revisit the comparison between local and global models in more detail in Section 5.3.

Our baseline local B-XS is the best look-ahead-bias free model. Both a backward model (B) and pure cross-sectional model (XS) outperform the median imputation. However, combining the time-series and cross-sectional information results in the best model, which increases the Block and Logit R^2 by 10 to 20 points. The previous value performs worse than using an AR(1) time-series model, as characteristics are usually not stale but rather only autocorrelated. The MCAR masking inflates the performance of pure time-series models like B and the previous value. Although many characteristics are persistent, for around 40% of the characteristics there are no prior values, and if they are available, they tend to miss in blocks. Hence, the Block and Logit masking schemes provide the more realistic comparison. Within the global models the global BF-XS dominates the alternative approaches. This is not surprising, as using future information should be beneficial. However, the difference between the global BF-XS and local B-XS for the out-of-sample data is relatively small. This suggests that a look-ahead-bias-free local B-XS performs close to a model that uses all available information.

The general ordering of imputation methods holds among all masking mechanisms. The outperformance of the baseline models, local B-XS, and global BF-XS, is even more pronounced for the logit-masking. In this case, most masking occurs for quarterly characteristics. When monthly characteristics are masked, it is likely that a large number of characteristics is masked simultaneously and/or the block of missing data is rather long, which presents a challenge for all imputation methods. The in-sample results can be interpreted as an evaluation of the parsimonious characteristic model, whereas the out-of-sample results also test how well the parsimonious model can be estimated from the partially observed data. The fact that the in-sample and out-of-sample MCAR results are extremely close is evidence that our characteristic models do not overfit but rather provide a good description for characteristics.

The lower part of Table 3 reports the R^2, which measures the explained variation relative to a cross-sectional median imputation. It clarifies how substantial the improvements are for our baseline models. The local B-XS and global BF-XS can achieve an impressive out-of-sample R^2 of 0.58 for logit-masking. The median imputation has, by definition, an R^2 of zero. Interestingly, for monthly updated characteristics, which tend to be less persistent, a pure time-series has an R^2 close to zero

Table 4: Imputation Error for Extreme Characteristic Quintiles

	In-Sample			OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
First characteristic quintile												
local B-XS (XS)	**0.17**	**0.17**	**0.15**	**0.17**	**0.18**	**0.16**	**0.22**	**0.23**	**0.22**	**0.23**	**0.20**	**0.29**
local XS	0.24	0.24	0.24	0.24	0.24	0.25	0.25	0.24	0.25	0.28	0.26	0.31
local B (median)	0.19	0.19	0.18	0.19	0.19	0.18	0.29	0.29	0.29	0.28	0.25	0.35
prev val (median)	0.20	0.20	0.20	0.20	0.20	0.21	0.30	0.29	0.31	0.28	0.25	0.36
median	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
global BF-XS (B-XS, XS)	**0.12**	**0.11**	**0.13**	**0.16**	**0.17**	**0.15**	**0.21**	**0.21**	**0.22**	**0.23**	**0.20**	**0.30**
global B-XS (XS)	0.16	0.17	0.15	0.17	0.18	0.16	0.23	0.23	0.23	0.24	0.21	0.30
global XS	0.25	0.25	0.25	0.25	0.25	0.26	0.26	0.25	0.27	0.29	0.27	0.33
global B (median)	0.18	0.19	0.18	0.19	0.19	0.18	0.29	0.29	0.29	0.28	0.25	0.35
Fifth characteristic quintile												
local B-XS (XS)	**0.17**	**0.17**	**0.15**	**0.17**	**0.18**	**0.16**	**0.22**	**0.21**	**0.22**	**0.23**	**0.21**	**0.28**
local XS	0.22	0.22	0.23	0.23	0.22	0.24	0.23	0.22	0.25	0.27	0.26	0.31
local B (median)	0.19	0.19	0.18	0.19	0.19	0.19	0.28	0.27	0.30	0.28	0.24	0.35
prev val (median)	0.20	0.20	0.21	0.21	0.20	0.22	0.29	0.28	0.31	0.28	0.25	0.35
median	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
global BF-XS (B-XS, XS)	**0.11**	**0.10**	**0.13**	**0.16**	**0.16**	**0.15**	**0.21**	**0.20**	**0.22**	**0.23**	**0.21**	**0.29**
global B-XS (XS)	0.16	0.17	0.15	0.17	0.18	0.16	0.22	0.21	0.23	0.24	0.22	0.29
global XS	0.24	0.23	0.25	0.24	0.23	0.26	0.24	0.23	0.26	0.28	0.27	0.32
global B (median)	0.19	0.19	0.19	0.19	0.19	0.19	0.28	0.27	0.30	0.28	0.24	0.35

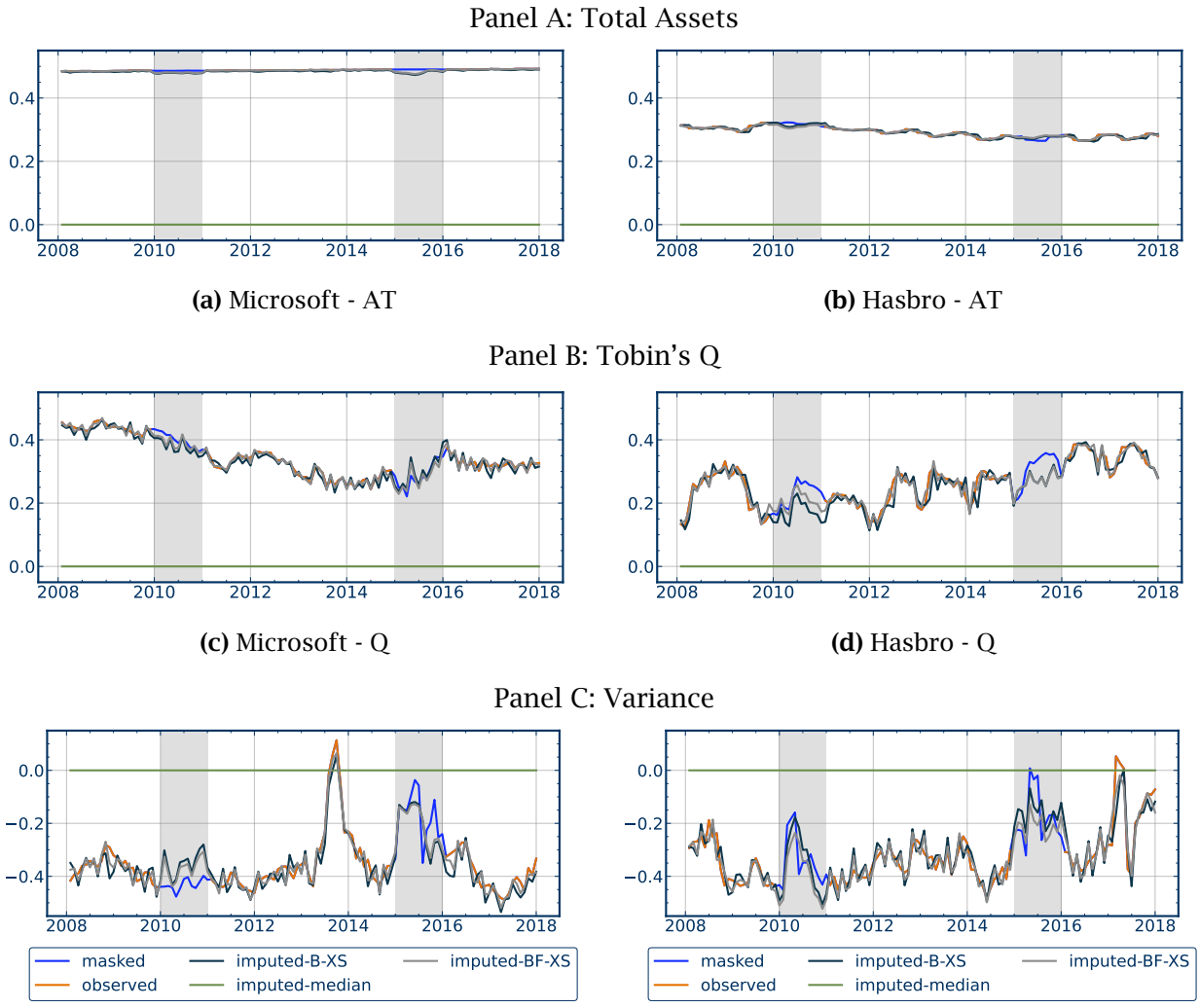
This table shows imputation RMSE by imputation method for different types of missingness for the subset of masked values that are in the first or fifth characteristic quintile. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from all observed data for the three masking schemes MCAR, Block, and Logit. We report the fallback method in brackets, which is used when a method is not applicable. For example, for missing characteristics of stocks without any prior values B-XS will be replaced by XS.

for logit-masking. Hence, under the most realistic masking scheme, it is even more crucial to optimally leverage the cross-sectional and time-series information. It is again important to emphasize that any comparison study has to be precise about the data that is used for comparison. In the Internet Appendix we collect the results for different subsets of the data. When we study the 60% of the data without missing at the beginning, our local B-XS achieves a remarkable out-of-sample R^2 of 0.86 for logit-masking.

Many applications use only the subset of largest or smallest characteristic values. One prominent example is using portfolio sorting strategies based on the extreme quantiles of characteristics. These applications depend on a precise imputation of the extreme characteristic quantiles, but they are less affected by the imputation quality in the center of the distribution. The outperformance of our baseline models relative to naive imputation is even more pronounced for these values.

Table 4 reports the RMSE for the masked characteristic values that are in the first or fifth characteristic quintile. By construction the median imputation performs particularly poorly. The local B-XS

Figure 10: Illustrative Model-Implied and Imputed Time-Series



(a) Microsoft - AT **(b)** Hasbro - AT

(c) Microsoft - Q **(d)** Hasbro - Q

(e) Microsoft - VAR **(f)** Hasbro - VAR

This figure shows illustrative realized and model-implied characteristic time-series for Microsoft and Hasbro. We plot the realized characteristic rank over time, and the model-implied values with the B-XS, BF-XS, and median model. The gray-shaded areas indicate missing blocks of one-year which are not part of the estimation, and, hence, serve as out-of-sample evaluation. We consider Total Assets, Tobin's Q and Variance, which are three representative characteristics of decreasing persistence.

has around half of the median RMSE. This confirms that our baseline models provide the preferred imputed values even for extreme realizations.

In order to provide intuition, we illustrate the model implied and imputed time-series for representative examples. Figure 10 shows characteristic time-series for Microsoft and Hasbro, two representative companies in different industries and hence with different fundamentals. We show their characteristic time-series for three characteristics with different levels of persistence. The most per-

sistent is Total Assets (AT). Tobin's Q has a medium level of persistence, whereas the local variance is a fast-fluctuating characteristic. These three examples are relatively representative, as they capture stylized features of other characteristics. We show the model-implied values in-sample as well as the imputation results for out-of-sample missing blocks of 12 months.

There are several obvious conclusions that we can draw from these examples. First is that the median value creates considerably large errors in observed and imputed values. Importantly, if we used the median imputed values for the missing blocks, we would also distort the time-series of the characteristics. For example, the centered rank quantile for the Total Assets of Microsoft would jump from about 0.5 to 0 and back to 0.5. In contrast, the imputed values with our methods reflect substantially better the level and dynamics of characteristics. Second, our two baseline models are very exact on the in-sample data. Obviously, the imputation is more challenging on the out-of-sample data. Third, our models reflect dynamic changes in the out-of-sample data, which are captured by the cross-sectional factor component. As we will see in Section 5.3, this cross-sectional component is more relevant for fast-changing characteristics such as the variance. Finally, the BF-XS seems to "connect" the two endpoints of the missing data, whereas the B-XS model is, for obvious reasons, "anchored" at the starting point of the missing block.

The aggregated comparison results are robust over time and with respect to the market capitalization of the stocks. In the Internet Appendix we show the RMSE results for each month. The relative ordering of the different methods is stable over time. We also report the RMSE for deciles of different sizes. Although the errors are larger in magnitude among smaller stocks, the relative comparison between the models remains the same. Importantly, even the largest decile accounts for a substantial part of the imputation errors, and, hence, the results are not driven by fitting only small-cap stocks.[14]

5.2. Imputation Results for Different Types of Missingness

In the next step we aim to understand how the imputation results are affected by the type of missingness. Hence, we show all the results of the previous subsection for data missing at the beginning, the middle, and the end of the sample. Table 5 shows the in-sample and out-of-sample RMSE results. Note that the type of missingness determines which models can be used. For example,

[14] Figures IA.8, IA.10, and IA.12 in the Internet Appendix show the RMSE for each month. Table IA.3 in the Internet Appendix reports the RMSE for deciles of different sizes.

Table 5: Imputation Error for Types of Missingness

	In-Sample			OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
Beginning of the sample												
local B-XS	-	-	-	-	-	-	-	-	-	-	-	-
local XS	**0.23**	**0.22**	**0.25**	**0.23**	**0.21**	**0.25**	**0.22**	**0.20**	**0.23**	**0.26**	**0.26**	**0.27**
local B	-	-	-	-	-	-	-	-	-	-	-	-
prev	-	-	-	-	-	-	-	-	-	-	-	-
XS-median	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.31	0.31	0.31	0.31	0.31
ind-median	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.31	0.31	0.31	0.31	0.31
global BF-XS	-	-	-	-	-	-	-	-	-	-	-	-
global F-XS	**0.11**	**0.06**	**0.16**	**0.17**	**0.17**	**0.17**	**0.20**	**0.19**	**0.21**	**0.21**	**0.20**	**0.25**
global B-XS	-	-	-	-	-	-	-	-	-	-	-	-
global XS	0.24	0.23	0.27	0.24	0.22	0.27	0.22	0.21	0.25	0.27	0.26	0.28
global B	-	-	-	-	-	-	-	-	-	-	-	-
Middle of the sample												
local B-XS	**0.13**	**0.14**	**0.12**	**0.14**	**0.14**	**0.12**	**0.17**	**0.17**	**0.17**	**0.12**	**0.12**	**0.13**
local XS	0.19	0.18	0.20	0.19	0.18	0.21	0.19	0.18	0.20	0.19	0.19	0.20
local B	0.14	0.15	0.14	0.15	0.15	0.14	0.19	0.18	0.21	0.14	0.13	0.15
prev	0.16	0.16	0.18	0.16	0.16	0.18	0.21	0.19	0.24	0.15	0.14	0.19
XS-median	0.29	0.28	0.29	0.28	0.28	0.29	0.28	0.28	0.29	0.29	0.29	0.29
ind-median	0.29	0.28	0.29	0.28	0.28	0.29	0.28	0.28	0.29	0.29	0.29	0.29
global BF-XS	**0.09**	**0.08**	**0.10**	**0.12**	**0.13**	**0.11**	**0.15**	**0.14**	**0.16**	**0.11**	**0.10**	**0.12**
global F-XS	0.08	0.05	0.12	0.14	0.14	0.13	0.16	0.16	0.18	0.13	0.12	0.14
global B-XS	0.13	0.14	0.12	0.13	0.14	0.13	0.17	0.17	0.18	0.13	0.12	0.13
global XS	0.20	0.19	0.21	0.20	0.19	0.22	0.20	0.19	0.21	0.20	0.20	0.21
global B	0.14	0.14	0.14	0.14	0.15	0.14	0.19	0.18	0.21	0.14	0.13	0.15
End of the sample												
local B-XS	**0.17**	**0.17**	**0.16**	**0.17**	**0.17**	**0.16**	**0.20**	**0.20**	**0.20**	**0.12**	**0.12**	**0.12**
local XS	0.25	0.25	0.25	0.24	0.23	0.24	0.22	0.22	0.23	0.23	0.23	0.23
local B	0.18	0.19	0.18	0.18	0.18	0.18	0.23	0.22	0.24	0.13	0.13	0.13
prev	0.21	0.20	0.21	0.20	0.20	0.22	0.25	0.23	0.27	0.13	0.13	0.15
XS-median	0.35	0.37	0.34	0.33	0.32	0.33	0.32	0.32	0.32	0.32	0.32	0.32
ind-median	0.35	0.37	0.34	0.33	0.32	0.33	0.32	0.32	0.32	0.32	0.32	0.32
global BF-XS	-	-	-	-	-	-	-	-	-	-	-	-
global F-XS	-	-	-	-	-	-	-	-	-	-	-	-
global B-XS	**0.17**	**0.17**	**0.17**	**0.17**	**0.18**	**0.16**	**0.20**	**0.20**	**0.20**	**0.12**	**0.12**	**0.13**
global XS	0.26	0.26	0.26	0.24	0.24	0.25	0.23	0.22	0.23	0.23	0.23	0.24
global B	0.19	0.19	0.18	0.18	0.18	0.18	0.23	0.22	0.24	0.13	0.13	0.13

This table shows imputation RMSE by imputation method for different types of missingness. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from all observed data for the three masking schemes MCAR, Block, and Logit.

when observations are missing at the beginning of the sample, we obviously cannot use any of the models that require prior observations. Similarly, for observations at the end, the forward models are excluded. Only missingness in the middle of the sample allows us to use all models.

The best models for missing observations at the beginning of the sample are the global F-XS, when using all possible information, and the local XS, when avoiding a look-ahead bias. These are the special cases of our baseline models that exclude the prior information. Therefore, we recommend

to use these two baseline models for imputing the missing values at the beginning.

The best model for missing observations in the middle are the global BF-XS, for full observations, and the local B-XS, among the look-ahead bias-free models. Overall our baseline models dominate the other approaches. Finally, we show that the global B-XS and local B-XS are the best model for missingness at the end of the sample. We conclude that the best model which still avoids future information is the local B-XS, and, if data is missing at the beginning, we replace it with the local XS.

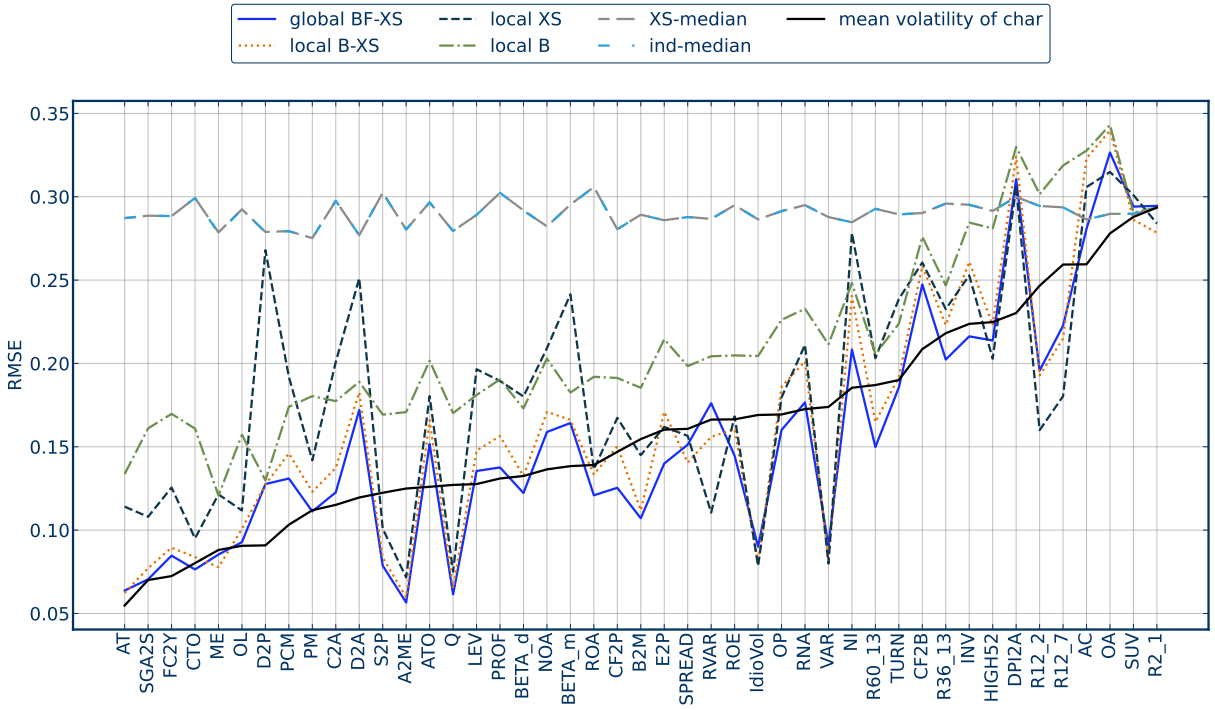
5.3. What information matters?

Which characteristics are hard to predict and what information is the most useful? In order to answer these questions, we compare the imputation errors for each characteristic. In the main text we focus on the out-of-sample results with block-missing patterns, while the Internet Appendix collects the in-sample and out-of-sample missing-completely-at-random and logit-masking results. Figure 11 plots the out-of-sample block-missing imputation errors for individual characteristics sorted in ascending order based on their time-series volatility. Characteristics on the right, for example, short-term momentum, fluctuate the most and, hence, might be harder to predict from the time-series, while the characteristics on the left, for example, total assets, are more persistent.

The median or industry median are in almost all cases the worst possible models. The pure cross-sectional model, which includes the median as a special case for a zero-factor model, strictly dominates the median imputation. The imputation of more volatile characteristics seems to benefit more from cross-sectional information. On the other hand, the more persistent characteristics seem to rely more on time-series information. A pure time-series or pure cross-sectional model is not uniformly better, and, in almost all cases, a combination of both types of information leads to superior results. The global BF-XS model has the smallest errors, except for idiosyncratic volatility (IdioVol), variance (VAR), closeness to last year high (HIGH52) and momentum (R12_2), where the XS and B-XS perform slightly better. The local B-XS is, for almost all characteristics, the best local model. The results are comparable for the logit masking.

The results are qualitatively similar for missing-completely-at-random. Overall, the benefit of cross-sectional information for more persistent information seems to shrink. This is expected, as there are only a few missing points in a row, and, hence, the last observed values can be very informative. However, the relative ranking remains the same. The results are comparable for the in-sample analysis.

Figure 11: Imputation Error for Individual Characteristics

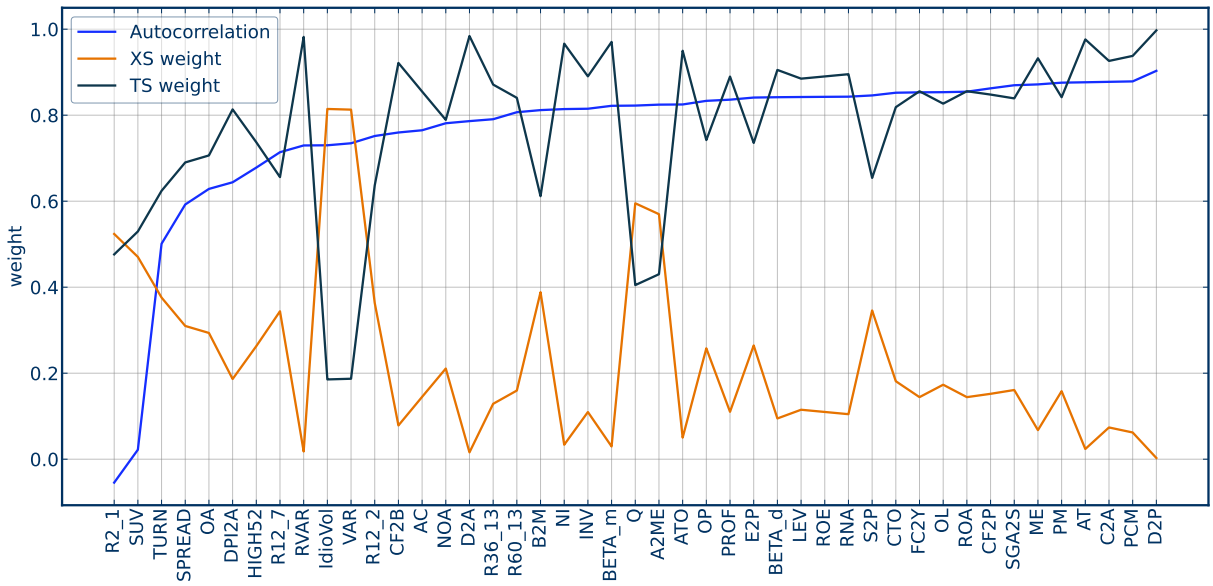


This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error out-of-sample for masked characteristics from all observed data for the block-masking scheme. We use the fallback method as indicated in Table 3, when a method is not applicable.

In order to assess the relative importance of the time-series and cross-sectional types of information we compare the relative weights in the regressions of the B-XS model. Figure 12 shows the absolute values of the regression coefficients on the cross-sectional factor model and the time-series information for the B-XS model. The characteristics are sorted in ascending order based on their autocorrelation. As expected, the time-series weight increases with the autocorrelation. This means that the more persistent characteristics use more time-series information for the imputation. In contrast, highly volatile and only weakly serially correlated characteristics tend to put larger weights on the cross-sectional factor model. Most importantly, all characteristics put some weight on the TS and XS information.

Finally, we compare the global and local models in more detail. Figure D.8 shows a comparison of imputation RMSE for local and global methods across individual characteristics. As previously, the characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. As expected by the aggregate statistics, the global models are very similar to their local counterparts. However, highly volatile characteristics can benefit from local models. This is,

Figure 12: Information Used for Imputation



This figure shows the absolute values of the regression coefficients on the cross-sectional factor model and the time-series information for the global B-XS model. The characteristics are sorted in ascending order based on their autocorrelations.

for example, visible for the pure cross-sectional models and implies that the models are relatively stable over time for most characteristics, but there can be some time variation among the more volatile characteristics.

5.4. Alterative Imputations Methods

Our imputation method outperforms alternative imputation methods. In this section we compare our local B-XS model with the approaches suggested by Freyberger et al. (2022) and Chen and McCoy (2022). The out-of-sample RMSE are shown in Table 6.

We start by comparing the method of Freyberger et al. (2022) to the median and local XS imputation. Freyberger et al. (2022) suggest a cross-sectional regression on the subset of fully observed characteristics. This is a special case of our cross-sectional regression, where we set X in equation 1 equal to the seven always-observed characteristics listed in Table 1. First, conditioning on these characteristics performs better than using the median. However, our local XS performs substantially better, that is, there is useful cross-sectional information in partially observed characteristics. Our XS factor model can be interpreted as a cross-sectional regression model that uses the latent factors rather than the seven always observed characteristics. There are certainly variations for the cross-sectional regression models based on fully observed characteristics. For example, the data can be

Table 6: Imputation Error for Alternative Methods

	OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
local B-XS	**0.14**	**0.14**	**0.13**	**0.17**	**0.17**	**0.18**	**0.12**	**0.12**	**0.13**
local XS	0.19	0.19	0.21	0.20	0.19	0.21	0.22	0.21	0.25
XS reg. fully obs.	0.26	0.26	0.25	0.26	0.26	0.26	0.27	0.27	0.27
EM	0.17	0.16	0.18	0.17	0.17	0.19	0.20	0.18	0.23
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.30

This table compares the out-of-sample imputation RMSE for alternative imputation methods averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the errors on the subset of data that are not missing at the beginning, that is, this data has some prior values of the characteristics observed. We compare our benchmark model, local B-XS, with the Expectation Maximization (EM) algorithm suggested by Chen and McCoy (2022) and a cross-sectional regression on the subset of fully observed characteristics (XS reg. fully obs.) suggested by Freyberger et al. (2022).

split into different blocks that have more fully observed characteristics. However, this would always use less information than our local XS model, which takes advantages of all partially observed data. For the same reason, even when including past values in the cross-sectional model regression model, it will perform worse than the B-XS.

Next, we compare the EM algorithm suggested by Chen and McCoy (2022) with our methods. As illustrated in Appendix A.5, the EM algorithm could be interpreted as an XS model with many factors. We observe that the EM-imputed values seem to be better than the pure XS model. Our simulation study in Appendix A.6 suggests that our cross-sectional dimension with $L = 45$ is on the lower side, but we could expect the XS model to have a stronger relative performance for datasets with more characteristics. Once we include the time-series information, the B-XS strongly outperforms the EM algorithm, particularly for the most practically relevant case of logit-masking.

We want to emphasize that the papers of Freyberger et al. (2022) and Chen and McCoy (2022) have different goals and their contributions are complementary to our work. The main focus of Freyberger et al. (2022) is on the efficient weighting of imputed values in GMM applications, which can be applied to any of the methods that provide asymptotic standard errors. The main point of Chen and McCoy (2022) is to study the imputation implications for investments with prediction based long-short portfolios, but not the quality of the imputation.

6. Asset Pricing Implications

Firm characteristics are the most widespread source of conditioning information for expected returns. As a result, missing values of firm characteristics have two fundamental effects on asset pricing: (1) sample selection of firms based on observability of their characteristics (Section 6.1), and (2) imputation bias when a wrong method is used to create a complete panel (Section 6.2). Naturally, the magnitudes of both effects depend on both the specific pattern of missing data and the actual application and/or estimation technique. In all the follow-up analyses, in order to ensure that the characteristic information is available to an investor in real time, we use the values of observed or imputed characteristics lagged by six months.[15]

6.1. Sample Selection on Observables: Firms with missing characteristics are different.

Investment outcomes and asset pricing tests depend on the universe of companies one considers. We show that firms with missing characteristics are different from those with observed values. Hence, using only the subsample of stocks with fully observed fundamentals leads to a selection bias in measuring investment performance or interpreting the outcomes of asset pricing tests. In this section, we highlight the impact of such a sample selection on three key objects of interest: a) measuring the market risk premia, b) portfolio performance and risk factor recovery in IPCA, a popular conditional asset pricing model pioneered by Kelly et al. (2019), and c) measuring the risk-return trade-off of cross-sectional strategies via simple sorted portfolios.

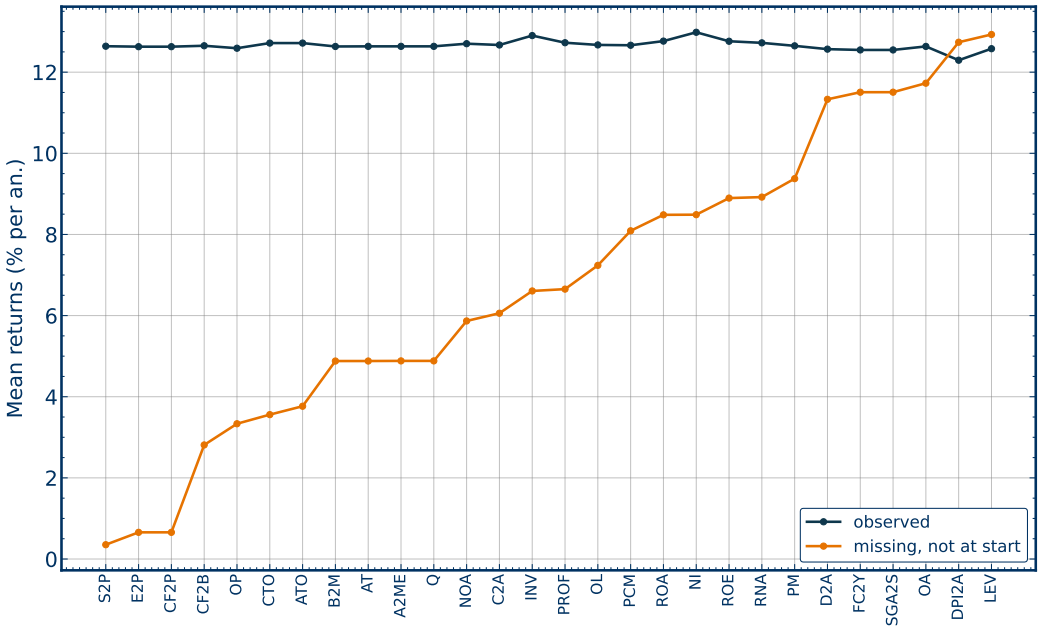
6.1.1. Market Premia with Observables

We begin this section by documenting a simple empirical fact: Even the average return on a market-style long-only portfolio of available stocks depends on whether the portfolio is constructed of companies that have a particular observed characteristic. In other words, even having – or not having – observable values for popular firm fundamentals such as book-to-market ratio, or Tobin's Q, on its own has an impact on asset returns, separate from its actual value.

Figure 13 shows average returns of stocks with observed or missing corresponding characteristics. In each month, we compute mean returns of all stocks with observed data of a particular

[15] Our results are very similar when a smaller or larger number of lags is being used for investment. Note that our focus is not on the optimal horizon of the information, but rather on the impact of missing some of the data and/or a choice of the imputation method. Investors could also use different lag horizons for different characteristics, and yet, even in that setting our results largely remain unchanged.

Figure 13: Market Premium Conditional on Observing a Firm Characteristic



This figure depicts the average annual return of stocks with observed or missing not-at-the-beginning (that is missing in middle or end of a stock sample) characteristics. The portfolios are value-weighted.

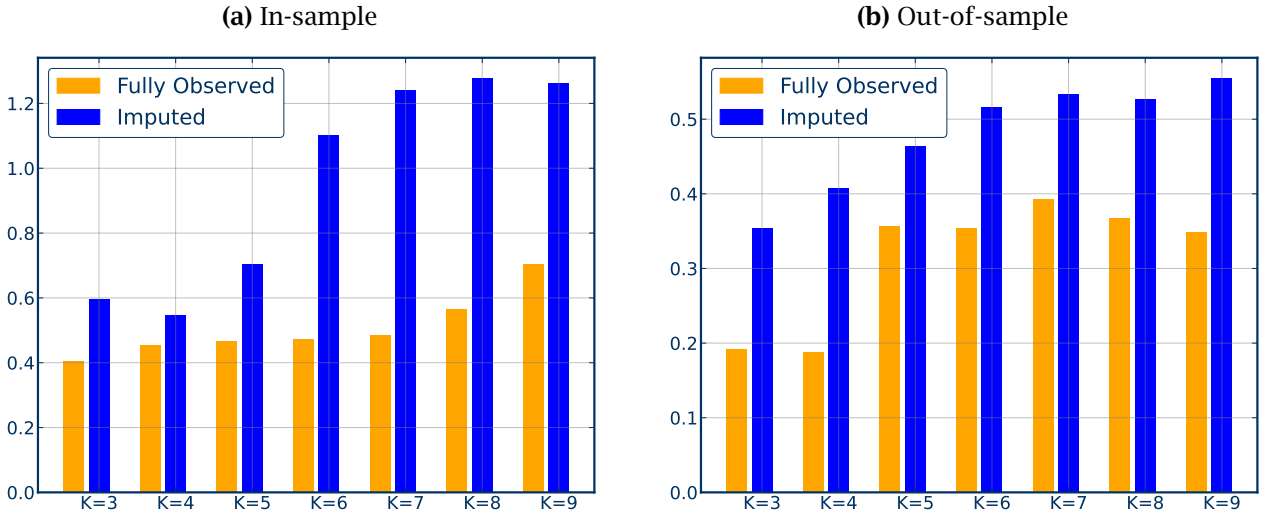
characteristics as well as mean returns of stocks for which the characteristic is missing in the middle or the end of the stock sample. Note that trading based on missingness not-at-the-beginning is an investable strategy. The presence of many firm-specific fundamentals seems to have an impact on asset returns due to the selection of companies into the observable set. Note that the selection bias impact is not uniform across different characteristics. In particular, companies with missing price-based ratios (e.g., sales-to-price, cash-flow-to-price, and earnings-to-price ratios) tend to have lower average returns than stocks with observed data. This effect is smaller for investment-related characteristics. The differences in market premia are economically large with up to 12% (p.a.) for the sales-to-price ratio.

It is clear that even estimating the market risk premium on stocks with missing observations can be negatively affected by a selection bias.

6.1.2. Omitted Firms and Investment Opportunity Set

We now show that neglecting firms with missing firm fundamentals generally leads to suboptimal investment decisions and subpar portfolio performance. To show this, we estimate the conditional latent factor model of Kelly et al. (2019) on the subset of stocks with fully observed characteristics and the larger set of stocks with imputed characteristics using the local B-XS imputation (with the

Figure 14: Sharpe Ratios with IPCA Factors



This figure shows the in- and out-of-sample Sharpe ratios of mean-variance efficient combination for different numbers of IPCA factors. We estimate a conditional latent factor model with the Instrumented Principal Component Analysis of Kelly et al. (2019). The estimation is either on the small subset of fully observed or the large set of all imputed stocks. The in-sample analysis is estimated on the full time period, while the out-of-sample analysis estimates the loadings and mean-variance efficient weights on the first half of the time-series and evaluates the portfolios on the second half.

local XS as fallback).

The Instrumented Principal Component Analysis (IPCA) models the exposure of individual stocks to latent risk factors as a function of firm characteristics. Intuitively, the IPCA factors are obtained as PCA factors of characteristic managed portfolios. Importantly, this analysis requires stocks to have a complete set of observable characteristics. Hence, to estimate the model we have to either take a small subset of fully observed data or work with the whole sample of firms and impute missing characteristics' values.

We evaluate the performance of the IPCA factors based on the Sharpe ratio of the implied pricing kernel (optimal factor portfolio). We first obtain the mean-variance efficient combination of the latent IPCA factors and then report the Sharpe ratio of this investment strategy. Importantly, we report these results for both in-sample and out-of-sample evaluations of different numbers of latent factors. The out-of-sample analysis estimates the IPCA model and mean-variance efficient factor combination on the first half of the sample and reports its out-of-sample performance on the second half of the data.

Figure 14 reports our findings. Estimating latent risk factors and their optimal combinations on the fully observed dataset leads to subpar investment performance both in- and out-of-sample.

This result holds for any number of latent factors. In fact, even a considerably parsimonious four-factor IPCA model based on all the available stocks outperforms a nine-factor model based on the fully observed data. Figure D.10 in the Appendix generalizes IPCA to a nonlinear conditional factor model. For each characteristic we consider 10 basis functions based on indicator functions for deciles. This corresponds to a kernel approximation of a non-linear loading function. Intuitively, we apply PCA to a collection of decile-sorted portfolios rather than only one linear long-short portfolio for each characteristic. The qualitative results are the same, but we observe that the differences between fully observed and imputed data are more pronounced. This suggests that the imputation bias might be amplified in more complex models.

These findings indicate several important implications. First, the stochastic discount factor (SDF) estimated on all stocks seems to be closer to the true SDF than the one estimated on only the subset of stocks with fully observed characteristics. Second, an investment strategy based on the fully observable (and a nonrepresentative) subset of firms, achieves a subpar performance, leaving money on the table. Note that these conclusions do not depend on the method used to extract the SDF. In fact, we obtain similar results for characteristic-mimicking factors built via simple cross-sectional regressions, as well as machine-learning prediction of returns with neural networks.[16] In other words, the issue lies in the sample selection based on characteristic observability, not the method used for building the SDF.

In order to further understand the sources of the risk-return trade-off and its relation to missing firm characteristics, we now turn to the simple, yet very popular, way of constructing cross-sectional strategies based on characteristics, namely decile-sorted portfolios.

6.1.3. Conditional Sorts

In this section, we show how the selection bias with missing data affects conditional expected returns. We focus specifically on the simplest asset pricing application to dissect the implications for different characteristics and for conditional means and variances.

Most multivariate asset pricing applications, including multiple sorts, panel regressions on multiple characteristics, and IPCA, from the previous subsection, require the presence of multiple characteristics. In order to illustrate the effect of requiring the presence of multiple characteristics, we focus on the properties of the most basic investment strategies, deciles sorts, and study how they are

[16] The results are available upon request.

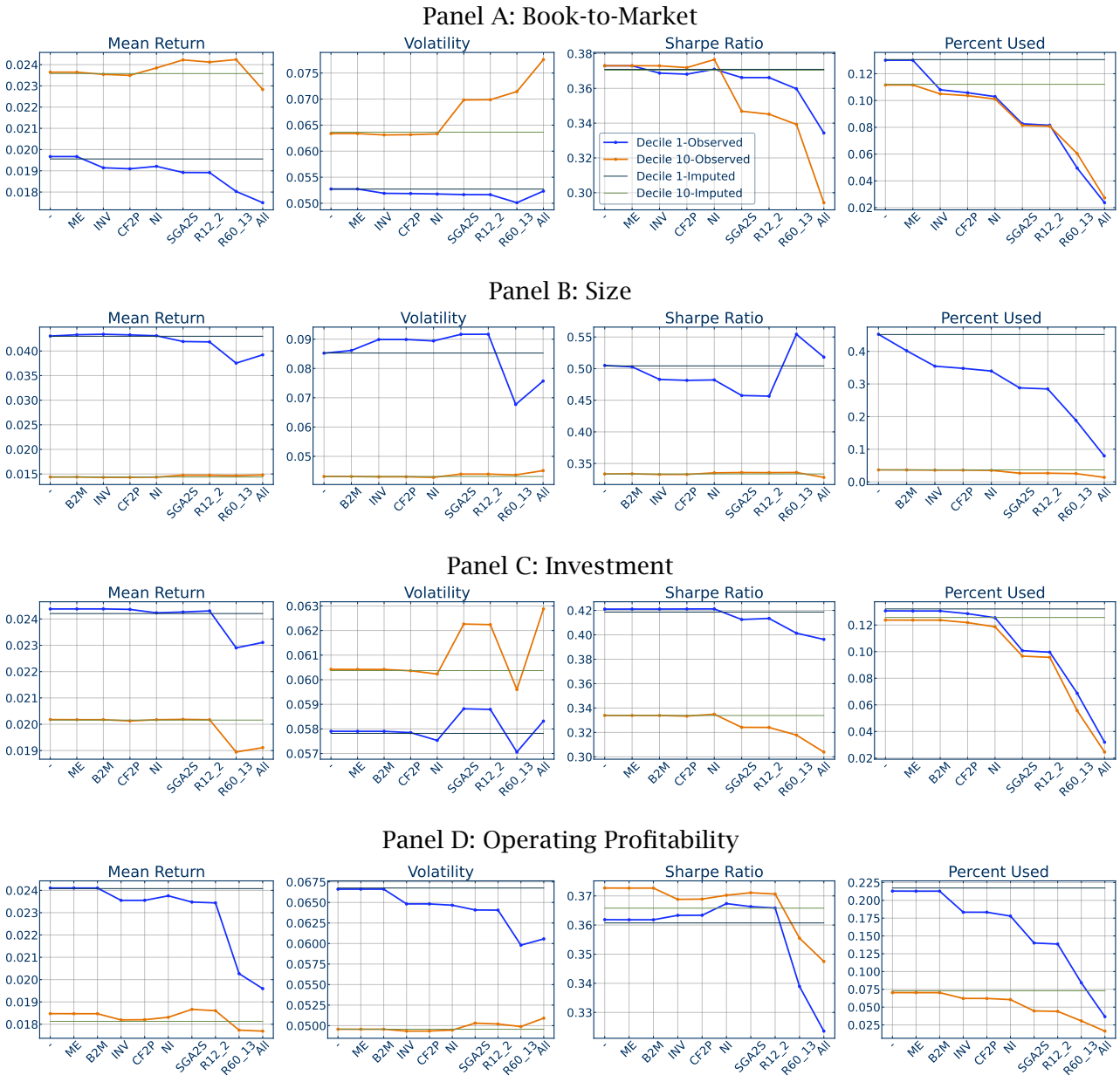
affected by the requirements of observing data for additional characteristics. Following the usual convention, the decile cutoff values are based on NYSE breakpoints, similar to Fama and French (1993).

First, we study the empirical effect of data selection and imputation on conditional returns for some of the most widely used characteristics, size (ME), book-to-market ratio (B2M), investment (INV), operating profitability (OP), momentum (R12_2) and long-term reversal (R60_12). In addition, we consider the accounting-based characteristics net share issues (NI) and expenses to sales (SGA2S), since those seem to be strongly affected by missing values. We construct value-weighted decile-sorted portfolios for the main characteristics, size, value, investment, and operating profitability. In order to understand the effect of requiring the presence of multiple characteristics, we study the asset pricing implications for the first and last deciles of these four characteristic sorts when requiring additional characteristics to be observed. In more detail, we first include only stocks that have the sorting characteristic available. We then take the subset of stocks for which size is also available. We continue stepwise, by incrementally requiring that, in addition, INV, OP, NI, SGA2S, R12_2, R60_13 or all 45 characteristics be available. The decile cutoff points remain the same NYSE breakpoints.

Figure 15 shows the Sharpe ratio, mean return, standard deviation, and percentage of stocks used in the first and 10th deciles. First, we use the least restrictive sample of stocks that requires only that leading characteristic be observed, progressively requiring more additional fundamentals to be observed. We also include the strategies with characteristic values imputed with the local B-XS model (and local XS as fallback), which is free of the look-ahead bias and could be easily used by investors in real time. The first obvious observation is that using all the stocks with imputed values, or all the stocks for which we require the availability of a single sorting variable, leads to essentially the same means and Sharpe ratios. This is reassuring, since it further confirms the validity of our imputation approach even for the firms that have fairly extreme values of the fundamentals. This result is in direct contrast with the situation in which we require additional characteristics to be observed, which changes the composition of the decile portfolios, its rate of return and its Sharpe ratio.

Requiring more characteristics drastically reduces the number of stocks that are included in portfolio sorts. In the case of size, the number of small stocks (the 10th decile, based on NYSE breakpoints) drops from almost 50% of all the tradable companies to less than 10%, when all the

Figure 15: Univariate Sorts with and without Missing Values



This figure shows the Sharpe ratio, average return, standard deviation and percentage of stocks for the univariate first and tenth value-weighted characteristic-sorted deciles for different subset of stock with and without imputation. We sequentially restrict the set of stocks to those for which multiple characteristics available. First, we include all stocks for which only the sorting characteristic is available, then in addition we require in addition the availability of size (ME). In the next step, the sorting characteristics, size, and investment (INV) need to be observed. We continue with operating profitability (OP), Net Share Issues (NI), Selling, general and administrative expenses to sales (SGA2S, momentum (R12_2) and long-term reversal (R60_13). We sort based on book-to-market, size, investment and operating profitability. We impute missing values with our baseline local BW-XS model.

characteristics are required to be observed. Restricting stocks to having contemporaneous observations for the book-to-market (B2M), investment (INV), and operating profitability (OP), removes 15%

of the overall sample. These results are even more extreme for portfolios sorted by the book-to-market ratio, see Panel A in Figure 15. In this case, the number of available stocks for the extreme growth and value deciles drops from more than 10% to approximately 2% of the sample, when all the characteristics are required to be observed. The requirement to observe ME, INV, and OP (in addition to B2M) already leads to a relative reduction of 10%-20% of the initial number of firms available for this strategy. The smaller number of stocks has an expected effect on the volatilities of the portfolio sorts, because one would expect that having fewer stocks would lead to less-diversified portfolios, and, hence, higher overall volatility. Indeed, we observe that in most cases volatility increases. Note, however, that in general this does not have to yield a monotonic effect: Because characteristics have complex missing patterns, both lower degree of diversification and firm selection contribute to the overall effect on volatility, making it difficult to predict the overall sign of the effect.

Importantly, the systematic structure in missing data creates a selection bias in mean returns. The mean returns of extreme deciles on investment, size, value, and operating profitability are already affected by requiring the presence of only three additional characteristics. Again, due to the complex nature of missingness, it can have an ambiguous effect on the risk premia. In all four cases, requiring the presence of all the characteristics leads overall to lower average returns. As the average returns tend to decrease in the more restrictive subsample of stocks, while the volatility effect increases in many cases, the Sharpe ratios tend to decrease as well. However, the exact effect on the Sharpe ratio and the corresponding t-statistics can be fairly complex.

Our results extend to the conditional mean based on the majority of characteristics. Figure D.9 shows the Sharpe ratios and mean returns for the top and bottom deciles of stocks, sorted by a given characteristic for two types of samples: first, requiring only that a single characteristic is observed, and second, requiring all 45 characteristics to be observed at the same time. We group firm-specific variables by their type and report both the Sharpe ratio and average return of the corresponding deciles.

For most characteristics, the Sharpe ratios on the fully observed panel are lower than those on the larger panel of firms with missing information. Consider, for example, the case of sorting based on operating leverage (characteristic OL in the intangible category in Figure D.9). In a fully observed panel, the Sharpe ratio of the bottom decile based on OL is 25% lower compared to the case of a simple univariate sort that requires only a single observed characteristic. Similar patterns can be observed for dividend-to-price (D2P), momentum (R12_7), expenses-to-assets (DPI2A), spread (SPREAD) and

return on assets and equity (ROA/ROE) among others. Hence, the combination of possible lower expected returns and/or higher volatility on a restricted sample can create a negative selection bias for simple asset pricing statistics. The directional effect on mean returns is more complex than Sharpe ratios, emphasizing again the complex interaction between the sorting characteristics and missingness. It seems that in many cases, when mean returns are larger on the restricted sample, the increase in volatility dominates, thus resulting in a lower Sharpe ratio. The corresponding Sharpe ratios and mean returns of deciles with imputed data are close to the sorts that require only a single characteristic to be observed.

The systematic selection bias in the expected returns of decile-sorted portfolios carries over to univariate long-short factors. Table IA.4 in the Internet Appendix reports the mean, standard deviation, Sharpe ratio, percentage, and market value of missing characteristics for univariate long-short decile factors. As in the case of case of decile sorts, these factors are constructed with NYSE breakpoints. We compare the results when using (1) only stocks with fully observed characteristics (i.e., 45 observed characteristics), (2) stocks with at least 10 characteristics observed and imputed data, (3) only the specific sorting characteristic observed, the combination of (2) and (3), and the difference between (2) and (3). Stock selection obviously has a strong effect on risk premia and Sharpe ratios even for simple univariate long-short factors. As a long-short factor combines the impact of selection and imputation in the two separate legs, the effects can be complex and more or less pronounced than for the individual legs.

Our findings on the selection bias in measuring risk premia on the market and conditional sorts are crucial, not only for reduced-form asset pricing models, but also for the majority of structural models. Since most of them are calibrated on observed moments of market dynamics, yet limit the sample of firms to those that have specific observed characteristics (e.g., investment, profitability and PPE/assets), selection bias may lead to a substantial impact on the estimated structural parameters and the magnitude of established economic channels.

6.2. Imputation Bias: Median Imputation Distorts Asset Pricing Tests

Having established that researchers should use all available data to avoid selection bias, we now compare the implications of different imputation methods. We focus on one of the most widely used tests in asset pricing: cross-sectional regressions of excess returns on lagged characteristics. Naturally, they require a complete set of characteristics for every company observed during a partic-

ular time period, and, hence, are ideally positioned to highlight the impact of different imputation methods on the estimation of risk premia.

Cross-sectional regressions on characteristics produce slope coefficients, which represent returns on the factor-mimicking pure-play portfolios, see Fama and MacBeth (1973), Fama (1976), and Back et al. (2013). Comparing risk premia estimates based on historical data and different imputation approaches, however, does not indicate which method is better. To measure a potential bias caused by different imputation techniques, one needs to define the true properties of these strategies. Therefore, we first take the whole available dataset with genuinely missing values for some of the characteristics and use the available sample to build pure-play portfolios. This gives us a benchmark for both factor dynamics and risk premia. We then mask some of the additional observed characteristic values using a realistic, data-driven procedure: the logistic regression used in Table 1. Note that this logistic regression propensity is well-suited to describe the empirical patterns of "missingness" and creates a realistic reference dataset.[17] Using this dataset, we impute masked values using the local B-XS model or cross-sectional median approach and examine the resulting pure-play portfolios on the reconstructed datasets by comparing them with those observed in the full sample.

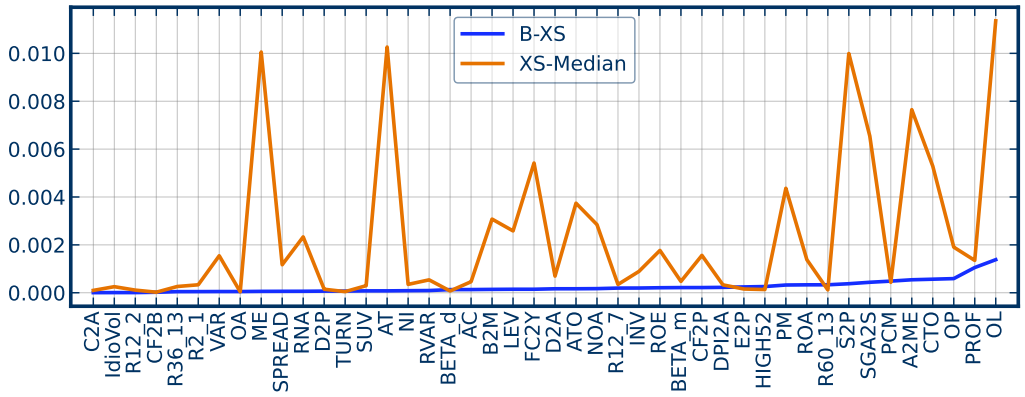
Note that pure-play portfolio weights depend on the joint cross-sectional behavior of firm characteristics. Hence, the imputation of some characteristic values may affect the risk premia of the factors that are based on fully observed characteristics. For example, the risk premia associated with size could be affected by the imputation of book-to-market ratio, even though market capitalization of firms is always observed.

Figure 16 illustrates imputation bias in factor-mimicking portfolios caused by different approaches to data completion, relative to the true pure-play portfolio constructed from the full initial dataset. Panel (a) reports the absolute error in the resulting factor risk premia. The B-XS imputed values yield uniformly and substantially smaller risk premia errors compared to the median imputation. For some characteristics (e.g., Total Assets (AT) and Operating Leverage (OL)), the imputation error achieved using B-XS leads to a risk premia bias, which is four to five times smaller than that achieved with median imputation,which is popular in the current literature. Overall, the bias in the risk premia estimates, based on the data imputed using B-XS, is small, not only in relative terms (compared to the median) but also economically.

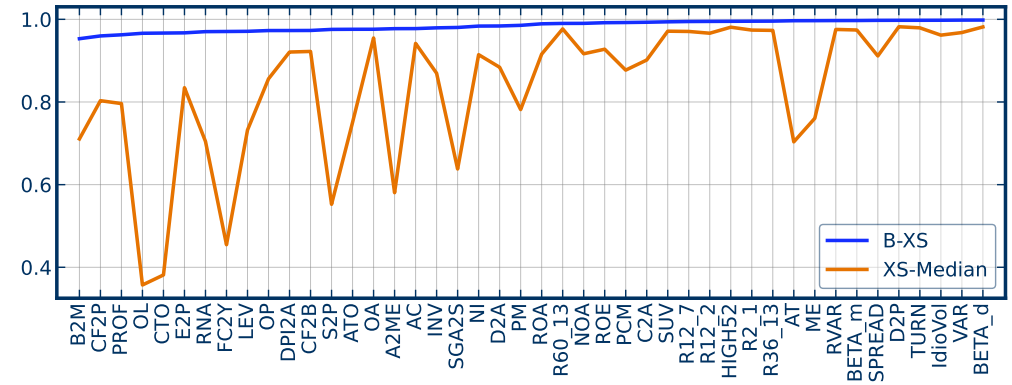
[17] We obtain qualitatively similar results for masking block-missing data, which are available upon request.

Figure 16: Imputation Bias in Pure-Play Mimicking Portfolios

(a) Absolute Error in Risk Premia Estimation



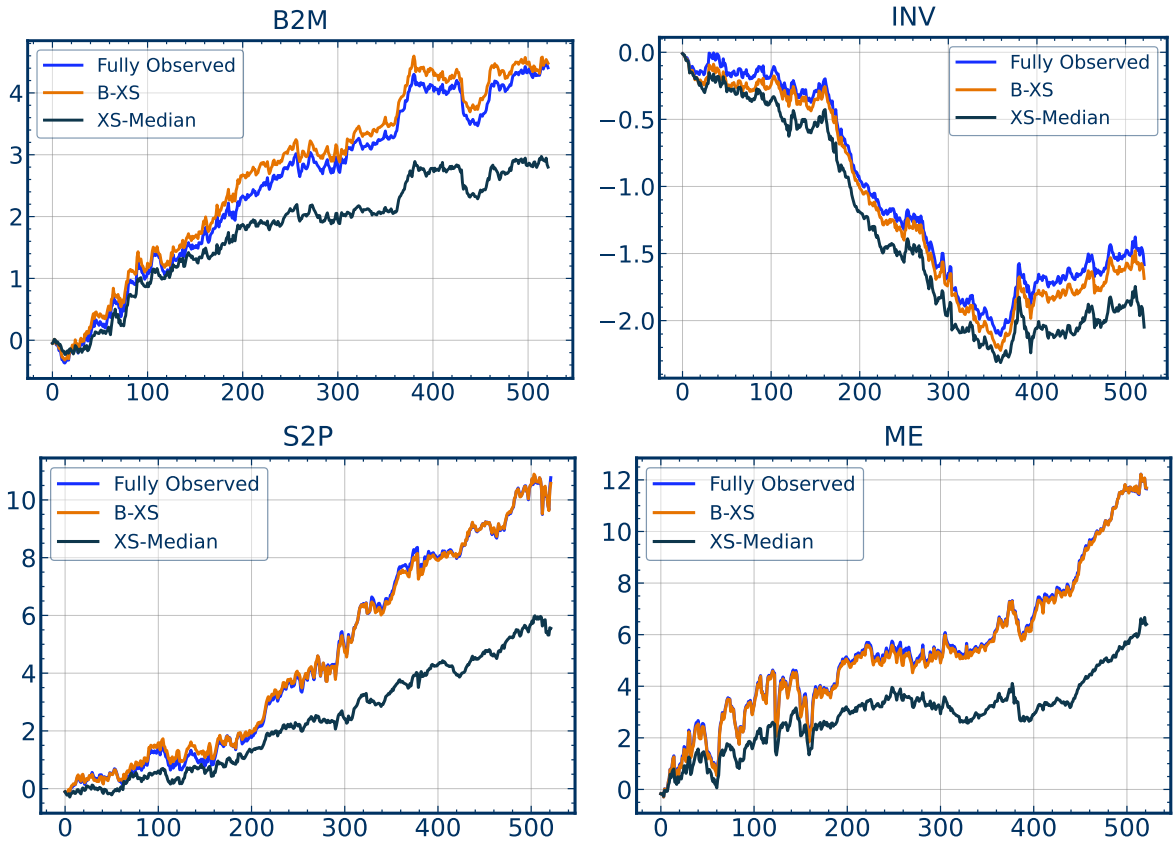
(b) Correlation with the True Factor-Mimicking Portfolio



This figure presents properties of the pure-play mimicking portfolios constructed with imputed characteristics values. Factor portfolios are built via cross-sectional regressions of stock excess returns on characteristics. Some of the characteristic values are masked based on the empirical pattern with the logistic regression propensity, and imputed via the local B-XS or median value. Panel (a) displays the absolute error in monthly portfolio risk premia based on imputed data relative to true observed values, measured on the full sample. Panel (b) displays correlations between the time-series of the factor-mimicking portfolios without masking and those with imputed values.

Figure 16, Panel (b), presents correlations between the pure-play mimicking portfolios constructed using imputed data and the original sample. The differences between the B-XS and median imputation approaches are even more pronounced. The mimicking portfolio time-series are rather close to the reference value for B-XS with correlations higher than 92% for all the characteristics. In contrast, factor returns based on the median imputation provide a substantially worse approximation to the true portfolio strategy, which is uniform for all the characteristics. In some cases, the correlation to the true factor is below 40%, see Operating Leverage. As a result, median-based imputation fails to accurately recover not only the average returns of factor-mimicking portfolios but also their dynamic

Figure 17: Characteristic Mimicking Factor Portfolios



This figure shows the time-series of cumulative excess returns of characteristic-mimicking factor portfolios with and without imputation. We estimate characteristic-mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference.

properties, for example, volatility.

Figure 17 further illustrates the cumulative performance of pure-play strategies constructed on the original dataset and the dataset masked data but completed with imputed values. We highlight results for the book-to-market ratio (B2M), profitability (PROF), and sales-to-price ratio (S2P). Portfolio performance, achieved using B-XS imputation, is remarkably stable and provides a precise approximation to the true behavior of the factor-mimicking portfolios. In contrast, we observe a substantial bias in the time-series for the factors using median imputation. As a result, this bias may affect not only expected returns on the factors but also their volatility and co-movement with other portfolios, invalidating results in many empirical applications. The Internet Appendix presents the results for the remaining characteristics, which reveal the same patterns.

It is important to clarify the difference between evaluating the imputation bias based on parameter estimates or with some prediction metric. In many economic studies, the objects of interest are the parameters or the structure of a model. Our empirical study is an example that shows that the median imputation can severely bias the parameter estimates of an asset pricing model. In our case, the returns of the factor portfolios are the "slope parameters" of the characteristic regression, and as such depend on the precision of all imputed values. It is possible that some prediction based metrics, for example, a prediction R^2 or Sharpe ratio from trading based on forecasting stock returns with characteristics is less affected by the median imputation. This can occur for example if the extreme prediction deciles are primarily driven by a subset of mostly observed characteristics, like price-trend based characteristics. However, it would be misleading to conclude from a high prediction metric that the median imputation does not affect the model structure, for example in terms of coefficients and variable importance.[18]

7. Conclusion

This paper focuses on a widespread yet rarely recognized issue of missing data in firm-specific characteristics. First, we document the systematic feature of missing data: It is pervasive and widespread among the majority of firms. In our representative dataset of the 45 most often used characteristics, more than 70% of firms are missing at least one characteristic at any given point in time. We show that firm fundamentals are not missing-completely-at-random but rather display complex systematic patterns. We leverage the complicated cross-sectional and time-series dependency

[18] Our empirical findings are complementary to Freyberger et al. (2022), who also confirm that the median imputation leads to biased parameter estimates. They estimate the coefficients on characteristics in a return prediction to study the incremental value of characteristics after publication. Chen and McCoy (2022) study return forecasting and compare a prediction metric for imputation with the median and the EM-algorithm. Their dataset and imputation method are different from our study. There are several reasons why a long-short portfolio based on return prediction deciles might perform similar for the EM-algorithm and median imputed values. First, extreme prediction deciles are often driven by a subset of mostly observed characteristics, like price-trend based characteristics. Hence, although the conditional expected return for less extreme returns might not be well-explained with median imputed values, this might not be reflected in the performance of the extreme long-short portfolio. Second, if a specific characteristic data set includes many characteristics that are not relevant for extreme returns, then setting their values to zero in a regression of returns on characteristics is similar to using a regularized ridge regression, and hence might be even beneficial for this specific application. Third, long-short prediction portfolios only account for a relative relationship but not for the level in the risk-premia, which might be affected by the imputation method. Fourth, their dataset seems to include a larger number of characteristics that seem to be more "idiosyncratic", i.e. less commonly used characteristics, that are not correlated with common firm fundamentals. By construction, the imputed characteristic values with cross-sectional models are expected to be close to the median for characteristics that are close to uncorrelated with other characteristics. In summary, the effect of the imputation method depends on the application, metric and dataset.

in firm characteristics to propose a new imputation method that is easy to use and substantially outperforms existing alternatives.

Our findings are relevant for numerous applications in asset pricing, since, as we demonstrated, asset returns are affected by missing observations of the firms' characteristics. The effects are particularly pronounced when requiring a large set of characteristics to be observed. While, for the sake of clarity, we demonstrate our findings with widely used portfolio sorts, cross-sectional regressions, and conditional latent factor models, we suspect that it has a first-order effect in return predictability regressions of more complex models (including machine learning), as well as all the recently proposed advanced frameworks of stock returns that typically require a large balanced panel of stock characteristics. Furthermore, because most of the structural models in asset pricing are evaluated on a subset of firms with fully observed characteristics (e.g., investment, profitability, and Tobin's Q), we expect the issue of missing data to have an impact on models in macrofinance, especially those based on production.

The problem of missing data is not limited to firm characteristics, and is encountered universally in various applications in finance: I/B/E/S forecast data, ESG ratings of firms, and many others. It is also likely to be more severe in the international context. Given the growth in Big Data applications and new sources of information being available at an increasing speed, we suspect that the issue of missing data will become even more paramount going forward. We are confident that our paper lays out foundations and general guidelines for imputing missing data and can be applied in numerous different settings in the follow-up research.

References

Abrevaya, J., and S. G. Donald, 2017, A GMM approach for dealing with missing data on regressors, *Review of Economics and Statistics* 99, 657–662.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, 2022, Matrix completion methods for causal panel data models, *Journal of the American Statistical Association, forthcoming* forthcoming.

Back, K., N. Kapadia, and B. Ostdiek, 2013, Slopes as factors: Characteristic pure plays, working paper, available at SSRN: https://papers.ssrn.com/abstract_id=2295993.

Bai, J., 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135–171.

Bai, J., and S. Ng, 2019, Principal components and regularized estimation of factor models, *Journal of Econometrics* 212, 78–96.

Bai, J., and S. Ng, 2021, Matrix completion, counterfactuals, and factor analysis of missing data, *Journal of the American Statistical Association* 1746–1763.

Bai, Jushan, and Serena Ng, 2023, Approximate factor models with weaker loadings, *Journal of Econometrics, forthcoming* .

Beckmeyer, H., and T. Wiedemann, 2022, Recovering missing firm characteristics with attention-based machine learning, available on SSRN: https://papers.ssrn.com/abstract_id=4003455.

Blanchet, J., F. Hernandez, V. A. Nguyen, M. Pelger, and X. Zhang, 2022, Bayesian imputation of missing data with optimal look-ahead-bias and variance tradeoff, *Working paper* .

Bryzgalova, S., M. Pelger, and J. Zhu, 2019, Forest through the trees: Building cross-sections of stock returns, *Journal of Finance, forthcoming* .

Cahan, E., J. Bai, and Serena Ng, 2023, Factor-based imputation of missing values and covariances in panel data of large dimensions, *Journal of Econometrics, forthcoming* .

Chen, A., and J. McCoy, 2022, Missing values and the dimensionality of expected returns, *Working paper* .

Chen, L., M. Pelger, and J. Zhu, 2022, Deep learning in asset pricing, *Management Science, forthcoming* .

Chen, X, J. Fan, C. Ma, and Y. Yan, 2019, Inference and uncertainty quantification for noisy matrix completion, *The Proceedings of the National Academy of Sciences* 116, 22931–22937.

Cochrane, J. H., 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Connor, G., and R. Korajczyk, 1988, Risk and return in an equilibrium apt: Application to a new test methodology, *Journal of Financial Economics* 21, 255–289.

Dagenais, M.G., 1973, The use of incomplete observations in multiple regression analysis: A generalized least squares approach, *Journal of Econometrics* 1, 317–328.

Dello Preite, M., R. Uppal, P. Zaffaroni, and I. Zviadadze, 2022, What is missing in asset pricing factor models?, *Working paper* .

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *Working paper* .

Duan, Junting, Markus Pelger, and Ruoxuan Xiong, 2023, Target PCA: Transfer learning large dimensional panel data, *Journal of Econometrics, forthcoming* .

Emmanuel, T., T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, 2021, A survey on missing data in machine learning, *Journal of Big Data* 8, 1–37.

Fama, E. F., 1976, *Foundations of Finance* (Basic Books).

Fama, E. F, and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, E. F., and J. D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Fan, J., Y. Liao, and W. Wang, 2016, Projected principal component analysis in factor models, *Annals*

of Statistics 44, 219–254.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber, 2022, Missing data in asset pricing panels, *Working paper* .

Freyberger, J., A. Neuhierl, and M. Weber, 2020, Dissecting characteristics nonparametrically, *Review of Financial Studies* 33, 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet, 2016, Time-varying risk premium in large cross-sectional equity data sets, *Econometrica* 84, 985–1046.

Giglio, S., and D. Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.

Green, J., J.R. Hand, and X. F. Zhang, 2017, The characteristics that provide independent information about average u.s. monthly stock returns, *The Review of Financial Studies* 30, 4389–4436.

Gu, S., B. T. Kelly, and D. Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

Jin, S., K. Miao, and L. Su, 2021, On factor models with random missing: EM estimation, inference, and cross validation, *Journal of Econometrics* 222, 745–777.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh, 2023, Machine-learning the skill of mutual fund managers, *Journal of Financial Economics* 150, 94–138.

Kelly, B.T., S. Pruitt, and Y. Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Koh, P. S., and D. M. Reeb, 2015, Missing R&D, *Journal of Accounting and Economics* 60, 73–94.

Kozak, S., S. Nagel, and S. Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Lettau, M., 2022, High-dimensional factor models with an application to mutual fund characteristics, *Working paper* .

Lettau, Martin, and Markus Pelger, 2020a, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.

Lettau, Martin, and Markus Pelger, 2020b, Factors That Fit the Time Series and Cross-Section of Stock Returns, *The Review of Financial Studies* 33, 2274–2325.

Lewellen, J., 2015, The Cross-section of Expected Stock Returns, *Critical Finance Review* 4, 1–44.

Light, N., D. Maslov, and O. Rytchkov, 2017, Aggregation of information about the cross section of stock returns: A latent variable approach, *The Review of Financial Studies* 30, 1339–1381.

Little, R. J. A., 1992, Regression with missing X's: A review, *Journal of the American Statistical Association* 87, 1227–1237.

Little, R. J. A., and D. B. Rubin, 2020, *Statistical Analysis with Missing Data* (John Wiley & Sons, Inc.).

Lyandres, Evgeny, Le Sun, and Lu Zhang, 2008, The new issues puzzle: Testing the investment-based explanation, *The Review of Financial Studies* 21, 2825–2855.

Pelger, M., 2020, Understanding systematic risk: A high-frequency approach, *Journal of Finance* 75,

2179–2220.

Pelger, M., and R. Xiong, 2021, Interpretable sparse proximate factors for large dimensions, *Journal of Business & Economic Statistics* 1–23.

Raja, P. S., and K. Thangavel, 2020, Missing value imputation using unsupervised machine learning technique, *Soft Computing* 24, 4361–4392.

Rao, C. R., and H. Toutenburg, 1999, *Linear Models: Least Squares and Alternatives* (Springer).

Robins, J. M., A. Rotnitzky, and L. P. Zhao, 1994, Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* 89, 846–866.

Rosenbaum, P. R., and D. B. Rubin, 1983, The central role of the propensity score in observational studies for causal effects, *Biometrika* 70, 41–55.

Rubin, D. B., 1976, Inference and missing data, *Biometrika* 63, 581–592.

Rubin, D. B., 1978, Bayesian inference for causal effects: The role of randomization, *The Annals of Statistics* 6, 34–58.

Rubin, D. B., and P. R. Rosenbaum, 1983, The central role of the propensity score in observational studies for causal effects, *Biometrika* 70, 41–55.

Wooldridge, J.M., 2007, Inverse probability weighted estimation for general missing data problems, *Journal of Econometrics* 141, 1281–1301.

Xiong, R., and M. Pelger, 2023, Large dimensional latent factor modeling with missing observations and applications to causal inference, *Journal of Econometrics* 233, 271–301.

Yates, F., 1933, The analysis of replicated experiments when the field results are incomplete, *Empire Journal of Experimental Agriculture* 1, 129–142.

Zaffaroni, P, 2022, Factor models for conditional asset pricing, *Working paper* .

Appendix A. Model

Appendix A.1. General Results for XS Factor Model

This section summarizes the general distribution results for the local XS factor model. As the model is estimated independently at each time period t, it does not impose any assumptions on the time-series structure and all quantities can vary with t. In order to keep the notation simple, we only add an index t to the main objects of the factor model, but in principle, all objects can be indexed by t and vary over time.

The distribution results are derived under the general assumptions for approximate factor models stated in Xiong and Pelger (2023). They are on a similar level of generality as the approximate factor model in Bai (2003). In order to provide intuition, we present a simplified model in Appendix A.2, which is a special case of our general framework. The distribution results depend on the moments of the factors and loadings. We assume that the factors and loadings are systematic such that there exist positive definite matrices Σ_F^t, Σ_Λ^t, and $\Sigma_{\Lambda,i}^t$ that satisfy for any t and i the following:

$$\frac{1}{N_t}\sum_{i=1}^{N_t} F_i^t (F_i^t)^\top \xrightarrow{p} \Sigma_F^t, \qquad \frac{1}{L}\sum_{l=1}^{L} \Lambda_l^t (\Lambda_l^t)^\top \xrightarrow{p} \Sigma_\Lambda^t, \qquad \frac{1}{L}\sum_{l=1}^{L} \Lambda_l^t (\Lambda_l^t)^\top W_{i,l}^t \xrightarrow{p} \Sigma_{\Lambda,i}^t.$$

The largest sample eigenvalues of $\hat{\Sigma}^{\text{XS},t}$ are denoted by $\hat{D}^t$, and D^t denotes the population eigenvalues of the matrix $\Sigma_F^t \Sigma_\Lambda^t$. We denote by $\delta = \min(L, N_t)$ the convergence rate of the factor model.

The order of the estimation is a choice, which affects the underlying assumptions. The estimation can either be first applied in the stock dimension N_t for obtaining the loadings and the characteristic dimension L to obtain the factors, or the other way around. In turns out that the empirical results are essentially the same for either order. We discuss this aspect in more detail in Appendix A.4. Theorem 1 provides the distribution theory without regularization.

Theorem 1 (XS Factor Model Estimator without Regularization). *Assume that the assumptions in Theorem 2 of Xiong and Pelger (2023) hold and that $N_t, L \to \infty$. We define $\delta = \min(N_t, L)$, and denote by $\hat{H}^t = \frac{1}{N_t L}\hat{D}^{t\,-1}(\hat{\Lambda}^t)^\top \Lambda^t (F^t)^\top F^t$ a $K \times K$ rotation matrix. Then, for $\gamma = 0$ the following results hold for each i, l and t:*

(a) **Auxiliary limits***: The largest sample eigenvalues converge to the population eigenvalues, that is, $\hat{D}^t \xrightarrow{p} D^t$, where $\Sigma_F^{t\,1/2}\Sigma_\Lambda^t\Sigma_F^{t\,1/2} = \Upsilon^t D^t \Upsilon^{t\top}$ with diagonal matrix D^t and eigenvectors Υ^t. The rotation matrix has the limit $\hat{H}^t \xrightarrow{p} H^t = \Upsilon^t \Sigma_F^{t\,1/2}$.*

(b) **Loadings***: For $\sqrt{N_t}/L \to 0$, the asymptotic distribution of the loadings is*

$$\sqrt{N_t}\left(\Sigma_{\Lambda,l}^{XS,t}\right)^{-1/2}\left(\hat{\Lambda}_l^t - \hat{H}^t\Lambda_l^t\right) \xrightarrow{d} N(0, I_K), \tag{A.1}$$

where $\Sigma_{\Lambda,l}^{XS,t} = (D^t)^{1/2} H^t \tilde{\Sigma}_{\Lambda,l}^t (H^t)^\top (D^t)^{1/2}$ and $\tilde{\Sigma}_{\Lambda,l}^t$ is the asymptotic variance of the loadings defined in Theorem 2.1 in Xiong and Pelger (2023).

(c) **Factors**: *For $\sqrt{L}/N_t \to 0$ and $\sqrt{N_t}/L \to 0$, the asymptotic distribution of the factors is*

$$\sqrt{\delta}\left(\Sigma_{F,i}^{XS,t,0}\right)^{-1/2}\left(\hat{F}_i^{t,0} - \left((\hat{H}^t)^\top\right)^{-1} F_i^t\right) \xrightarrow{d} N(0, I_K), \tag{A.2}$$

where $\Sigma_{F,i}^{XS,t,0} = (D^t)^{-1/2}((H^t)^\top)^{-1}\tilde{\Sigma}_{F,i}^{t,0}(H^t)^{-1}(D^t)^{-1/2}$ and $\tilde{\Sigma}_{F,i}^{t,0}$ is the asymptotic variance of the factors defined in Theorem 2.2 in Xiong and Pelger (2023).

(d) **Common component**: *The asymptotic distribution of the common component is*

$$\sqrt{\delta}\left(\Sigma_{C,i,l}^{XS,t,0}\right)^{-1/2}\left(\hat{F}_i^{t,0}(\hat{\Lambda}_l^t)^\top - F_i^t(\Lambda_l^t)^\top\right) \xrightarrow{d} N(0,1) \tag{A.3}$$

where $\Sigma_{C,i,l}^{XS,t,0} = \frac{\delta}{L}(\Lambda_l^t)^\top\Sigma_{F,i}^{t,0}\Lambda_l^t + \frac{\delta}{N_t}(F_i^t)^\top\Sigma_{\Lambda,l}^t F_i^t - 2\frac{\delta}{N_t}(\Lambda_l^t)^\top\Sigma_{F,\Lambda,i,l}^{Cov,t}F_i^t$ and $\Sigma_{F,\Lambda,i,l}^{Cov,t}$ is given in Theorem 2.3 in Xiong and Pelger (2023).

Theorem 1 corresponds to Theorem 2 in Xiong and Pelger (2023), with the only difference being that we use a different identification assumption, where the eigenvalues are assigned to the loadings instead of the factors. This is why we multiply the asymptotic covariance matrix of the loading estimator with the eigenvalues, while we divide the asymptotic covariance matrix of the factors by them. This is without loss of generality, when we do not apply regularization. As discussed in Bai and Ng (2019) it is beneficial to assign the eigenvalues to the object that is regularized.

Theorem 2 (XS Factor Model Estimator with Regularization). *Assume that the assumptions in Theorem 2 of Xiong and Pelger (2023) hold and that $N_t, L \to \infty$. The regularized factor estimator is given by $\hat{F}_i^{t,\gamma} = \hat{\Delta}^{\gamma,t}\hat{F}_i^{t,0}$ with $\hat{\Delta}^\gamma = \left(\frac{1}{L}\sum_{l=1}^L W_{i,l}^t\hat{\Lambda}_l^t(\hat{\Lambda}_l^t)^\top + \gamma I_K\right)^{-1}\left(\frac{1}{L}\sum_{l=1}^L W_{i,l}^t\hat{\Lambda}_l^t(\hat{\Lambda}_l^t)^\top\right)$. Then, for $\gamma > 0$ the following results hold for each i, l and t (using the notation of Theorem 1):*

(a) **Auxiliary limits**: *The shrinkage matrix converges to*

$$\hat{\Delta}^{\gamma,t} \xrightarrow{p} \left(H^t\Sigma_{\Lambda,i}^t H^{t\top} + \gamma I_K\right)^{-1}\left(H^t\Sigma_{\Lambda,i}^t H^{t\top}\right) =: \Delta^{\gamma,t}.$$

(b) **Factors**: *For $\sqrt{L}/N_t \to 0$ and $\sqrt{N_t}/L \to 0$, the asymptotic distribution of the factors is*

$$\sqrt{\delta}(\Delta^{\gamma,t}\Sigma_{F,i}^{XS,t,0}\Delta^{\gamma,t})^{-1/2}\left(\hat{F}_i^{t,\gamma} - \hat{\Delta}^{\gamma,t}\left((\hat{H}^t)^\top\right)^{-1}F_i^t\right) \xrightarrow{d} N(0, I_K).$$

(c) **Common component**: *The asymptotic distribution of the common component is*

$$\sqrt{\delta}(\Sigma_{C,i,l}^{XS,t,\gamma})^{-1/2}\left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top - F_i^t(\Lambda_l^t)^\top - F_i^t\left(\Delta^{\gamma,t} - I_K\right)\left(\Lambda_l^t\right)^\top\right) \xrightarrow{d} N(0,1),$$

where $\Sigma_{C,i,l}^{XS,t,\gamma} = \frac{\delta}{L}(\Lambda_l^t)^\top\Delta^{\gamma,t}\Sigma_{F,i}^{t,0}\Delta^{\gamma,t}\Lambda_l^t + \frac{\delta}{N_t}(F_i^t)^\top\Delta^{\gamma,t}\Sigma_{\Lambda,l}^t\Delta^{\gamma,t}F_i^t - 2\frac{\delta}{N_t}\Lambda_l^\top\Delta^{\gamma,t}\Sigma_{F,\Lambda,i,l}^{Cov,t}\Delta^{\gamma,t}F_i^t$.

The proof is based on the following identity, which relates the regularized estimation to the unregularized estimation:

$$\hat{F}_i^{t,\gamma} = \left(\frac{1}{L}\sum_{l=1}^L W_{i,l}^t\hat{\Lambda}_l^t(\hat{\Lambda}_l^t)^\top + \gamma I_K\right)\left(\frac{1}{L}\sum_{l=1}^L W_{i,l}^t(\hat{\Lambda}_l^t)^\top C_{i,l}^t\right) = \hat{\Delta}^{\gamma,t}\hat{F}^{t,0}.$$

This implies that $\sqrt{\delta}\left(\hat{F}_i^{t,\gamma} - \Delta^{\gamma,t}\left((\hat{H}^t)^\top\right)^{-1} F_i^t\right) = \sqrt{\delta}\hat{\Delta}^{\gamma,t}\left(\hat{F}_i^{t,0} - \left((\hat{H}^t)^\top\right)^{-1} F_i^2\right)$. Theorem 2(a) follows from the proof of Lemma 8.4 on page 71 in the Internet Appendix of Xiong and Pelger (2023). Combining these two points allows us to use Theorem 1(c) to prove Theorem 2(b). The last statement follows from Theorem 1(c) and the following expansion:

$$\begin{aligned}&\sqrt{\delta}\left(\hat{C}_{i,l}^{\gamma,t} - C_{i,l}^t - F_i^t\left(\Delta^{\gamma,t} - I\right)\Lambda_l^{t\top}\right)\\ =&\sqrt{\delta}\left((\hat{H}^t)^{-1}\hat{\Lambda}_l^t - \Lambda_l^t\right)\Delta^{\gamma,t}F_i^t + \sqrt{\delta}\left((\hat{H}^t)^\top\hat{F}_i^{t,0} - F_i^t\right)\Delta^{\gamma,t}\Lambda_l^{t\top} + o_p(1).\end{aligned}$$

This expansion follows from the same arguments as in the unregularized case, which underlies the proof of Theorem 1.(d).

Shrinkage has two effects: First, the variance terms are multiplied by $\Delta^{\gamma,t}$, which reduces the variance. Second, the shrinkage leads to a bias in the common component of the form $F_i^t\left(\Delta^{\gamma,t} - I\right)\Lambda_l^{t\top}$. These two effects lead to a bias-variance trade-off, and an optimally selected shrinkage γ can (substantially) reduce the asymptotic mean-squared error (MSE). The MSE is the sum of the asymptotic variance (Var) and the squared asymptotic bias (Bias). As shown in Xiong and Pelger (2023), the unregularized estimator is unbiased, which means that asymptotically the mean-squared error equals the variance, that is $\text{Var}(\hat{C}_{i,l}^{t,0}) = \frac{1}{\delta}\Sigma_{C,i,l}^{t,0}$ and $\text{Bias}(\hat{C}_{i,l}^{t,0}) = 0$.

For notational convenience, we consider the case of $K = 1$ factor, which simplifies $\Delta^{\gamma,t} = \frac{d_1^t}{d_1^t+\gamma}$, where $d_1^t = D_{1,1}^t$ equals the largest population eigenvalue of the systematic component. Asymptotically, the variance of the regularized estimator equals $\text{Var}(\hat{C}_{i,l}^{t,\gamma}) = (\Delta^\gamma)^2\text{Var}(\hat{C}_{i,l}^{t,0}) = \frac{1}{\delta}\Sigma_{C,i,l}^{t,0}(\Delta^\gamma)^2$, whereas the bias is $\text{Bias}(\hat{C}_{i,l}^{t,\gamma}) = (\Delta^\gamma - 1)\left(F_i^t(\Lambda_l^t)^\top\right)$. As a result, the ratio of MSEs between the two estimators can be expressed as follows and can be smaller than one:

$$\frac{\text{MSE}(\hat{C}_{i,l}^{t,\gamma})}{\text{MSE}(\hat{C}_{i,l}^{t,0})} = (\Delta^\gamma - 1)^2\frac{(F_i^t(\Lambda_l^t)^\top)^2}{\text{Var}(\hat{C}_{i,l}^{t,0})} + (\Delta^\gamma)^2.$$

Missingness increases the asymptotic variance, and for characteristics with many missing values, $\Sigma_{C,i,l}^{t,0}$ can be relatively large, and hence shrinkage can be very beneficial. The extreme case of $\gamma \to \infty$ leads to median imputation with the largest bias, but also the smallest variance. The optimal value of the tuning parameter γ is based on the bias-variance trade-off to minimize the mean-squared error and selected optimally from the data via validation.

The convergence rate of the unregularized estimator is $O\left(\frac{1}{\sqrt{\delta}}\right)$. The bias term does not affect the asymptotic convergence rate if the regularization is set to $\gamma \leq O\left(\frac{1}{\sqrt{\delta}}\right)$. Empirically, we confirm that the optimal γ is around $0.01/L$ and of similar magnitude as the noise. This suggests that it is reasonable to assume that the regularization satisfies $\gamma \leq O\left(\frac{1}{\sqrt{\delta}}\right)$. This has the consequence that in the limit for $\delta \to \infty$ and $\gamma \to 0$, the asymptotic variance is the same as without shrinkage. However, shrinkage still has finite sample effects, as we confirm empirically and in simulations. The reason is that in a finite sample, the higher order effects of the asymptotic expansion can play a role. It is possible to explicitly model the asymptotic effect on the higher-order terms for shrinkage

with $\gamma = O\left(\frac{1}{\sqrt{\delta}}\right)$, but this is beyond the scope of this paper. These higher-order effects are more relevant in the setup of missing data, which increases the variance, and for our only moderately large dimension $L = 45$. We also expect the bias-variance trade-off to become asymptotically a first-order aspect for weak factors. More specifically, under certain assumptions, weak factors can be estimated consistently but at a lower convergence rate (see among others Bai and Ng (2023), Xiong and Pelger (2023) and Duan et al. (2023) for the effect on convergence rates and the impact of missingness). Hence, it is possible that the asymptotic variance of weak factors is of a similar order as the bias term even for small regularization values. Empirically, we observe that the benefit is the largest for weaker factors. While it is beyond the scope of this paper to provide a complete analysis for weak factors, the bias-variance tradeoff would follow the same arguments as outlined above.

Moving forward, we assume that $\gamma \leq O\left(\frac{1}{\sqrt{\delta}}\right)$, and hence we have consistent estimation of the imputed values at the rate of $\frac{1}{\sqrt{\delta}}$ as follows:

$$\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top = F_i^t(\Lambda_l^t)^\top + O_p\left(\frac{1}{\sqrt{\delta}}\right).$$

Appendix A.2. A Simplified XS Factor Model

In this section, we present a simplified factor model with the stronger Assumptions 2 and 3, which substantially simplifies the notation but conveys the main conceptual insights of the general model. It allows us to highlight the effect of missing observations, and how they can increase the variance of the estimation. As previously, in principle all variables can be indexed by time t, as we estimate separate models for each time period t. For notational ease, we use the index t only for the main objects.

We allow for general patterns in the missing observations. $Q_{l,p}^t = \{i : W_{il}^t = 1 \text{ and } W_{ip}^t = 1\}$ denotes the set of all stocks for which characteristics l and p are observed at time t. $|Q_{l,p}^t|$ is the cardinality of the set $Q_{l,p}^t$. Assumption 1 states the conditions on the observation pattern.

Assumption 1 (Observational Pattern).

(a) W^t *is independent of* F^t *and* e^t.

(b) For a given observation matrix W^t, $\frac{|Q_{l,p}^t|}{N_t} \geq \underline{q} > 0$ *and there exist constants* q_{lp} *and* $q_{lp,rs}$ *for all* i, l, r, s *such that* $q_{lp} = \lim_{N_t \to \infty} \frac{|Q_{l,p}^t|}{N_t}$ *and* $q_{lp,rs} = \lim_{N_t \to \infty} \frac{|Q_{lp}^t \cap Q_{rs}^t|}{N_t}$.

Assumption 1 allows very general observation patterns that can vary over time and depend on unit-specific features. In particular, the observation pattern can depend on the factor loadings that capture cross-sectional information. For the purpose of identification, we assume that the observation pattern is independent of the factors. Note that the estimator of the common components is "symmetric" in N_t and L, and therefore we could switch the roles of N_t and L in the assumptions above. In that case, the observation pattern would be independent of the loadings but can depend on the factors. The assumption that the observation pattern is independent of the errors is

closely related to the unconfoundedness assumption in Rosenbaum and Rubin (1983). Assumption 1 implicitly assumes that for any two units, the number of time periods when both are observed is proportional to N_t. This simplifies the presentation of our results and is sufficient for most empirically relevant cases, but this assumption can be relaxed as discussed in Xiong and Pelger (2023).

The general case assumes an approximate factor structure at the same level of generality as in Bai (2003). The factors and loadings have nontrivial time-series and cross-sectional dependency. We allow the errors to be weakly correlated in the time-series and cross-sectional dimensions. The asymptotic distributions are based on general martingale central limit theorems. The general assumptions are stated in Xiong and Pelger (2023).

The consistency results are based on Assumption 2 for the simplified model. The key elements are that the factors and loadings are systematic in the sense that they lead to exploding eigenvalues, whereas the error terms are non-systematic with bounded eigenvalues in the second momdent matrix of C^t. These are standard factor model assumptions. The asymptotic distribution results require additional restrictions on the missing patterns, as stated in Assumption 3.

Assumption 2 (Simplified Factor Model).
There exists a positive constant $c < \infty$ such that:

(a) Factors: $F_i^t \overset{\text{i.i.d.}}{\sim} (0, \Sigma_F)$, $\mathbf{E}[\|F_i^t\|^4] \leq c$, *and* $\mathbf{E}\|F_i^t F_i^{t\top} - \Sigma_F\|^{2+\epsilon} \leq c$ *for some* $\epsilon \in (0,1)$.

(b) Factor loadings: $\Lambda_l^t \overset{\text{i.i.d.}}{\sim} (0, \Sigma_\Lambda)$ *and* $\mathbf{E}[\|\Lambda_i^t\|^4] \leq c$.

(c) Errors: $e_{il}^t \overset{\text{i.i.d.}}{\sim} (0, \sigma_e^2)$, $\mathbf{E}[(e_{il}^t)^8] \leq c$.

(d) Independence: F^t, Λ^t and e^t are independent.

(e) Eigenvalues: The eigenvalues of $\Sigma_\Lambda \Sigma_F$ are distinct.

Assumption 3 (Moments of Simplified Factor Model).

(a) Systematic loadings: $\frac{1}{L}\sum_{l=1}^L \Lambda_l^t (\Lambda_l^t)^\top W_{il}^t \xrightarrow{P} \Sigma_{\Lambda,t}$ *for some positive definite matrix $\Sigma_{\Lambda,t}$ for any t.*

(b) Dependency in missing pattern: $\frac{1}{L^2}\sum_{r=1}^L \sum_{s=1}^L \frac{q_{rl,sl}}{q_{rl}q_{sl}} \xrightarrow{P} \omega_{ll}$, $\lim_{L\to\infty} \frac{1}{L^3}\sum_{p=1}^L \sum_{r=1}^L \sum_{s=1}^L \frac{q_{pr,sl}}{q_{pr}q_{sl}} \xrightarrow{P} \omega_l$ *and* $\lim_{L\to\infty} \frac{1}{L^4}\sum_{l=1}^L \sum_{p=1}^L \sum_{j=r}^L \sum_{j=1}^L \frac{q_{lp,rs}}{q_{lp}q_{rj}} \xrightarrow{P} \omega$ *for all l and some constants ω_{ll}, ω_l, and ω.*

Assumption 3 has two key elements. First, the full rank assumption of $\Sigma_{\Lambda,t}$ captures that the factor loadings are systematic for the observed entries. Second, the number of observed units at every time period t is proportional to L and different units share a number of observed entries that is proportional to N_t. The impact of the missing pattern on the asymptotic variances of the estimators is captured by the three key parameters ω, ω_l and ω_{ll}. Note that by construction these constants satisfy $\omega_{ll}, \omega_l, \omega \geq 1$. These parameters tend to increase with a larger fraction of missing observations, or with stronger dependency between the rows of the observation patterns. If observations are missing-completely-at-random with probability $1-p$, then $\omega_{jj} = \frac{1}{1-p}$, $\omega_j = 1$ and $\omega = 1$. Intuitively, given the same proportion of missing entries, these key parameters increase the more

the missing pattern deviates from missing-completely-at-random, and hence can be interpreted as a measure of complexity in the missing pattern. In fact, these parameters encapsulate all information about the missing patterns for the simplified model. Two missing patterns that have the same parameters $\omega_{ll}, \omega_l, \omega$ result in the same asymptotic distribution for the estimator. We show that larger values of these parameters will increase the asymptotic variance for the estimators of the latent factor model.

As stated in Xiong and Pelger (2023), the simplified model is only a special case of the general approximate factor model. All the theorems are derived under the general assumptions. Corollary 1 shows the asymptotic distribution of the latent factor model for the simplified model, and highlights the role of the missing pattern parameters ω, ω_l and ω_{ll}. The distribution results of Theorem 1 and 2 simplify under Assumptions 2 and 3, and we can provide explicit expressions for the asymptotic variances. If we assume in addition that the proportions of observed time-series (q_{lp} and $q_{lp,rs}$) are independent of the second moment of the loadings $\Lambda_l^t \Lambda_l^{t\top}$, we can further separate the effect of missing patterns from the properties of the factor model. The following corollary is Corollary 1 of Xiong and Pelger (2023), which we restate here in our notation for the readers' convenience.

Corollary 1. *Suppose Assumptions 1, 2 and 3 hold and $N_t, L \to \infty$. Then Theorem 1 and 2 hold. If in addition, q_{lp} and $q_{lp,rs}$ are independent of $\Lambda_m \Lambda_m^\top$ for all l, p, r, s, m, then the asymptotic variances simplify as follows with the weights ω, ω_l and ω_{ll} defined in Assumption 3:*

(a) The asymptotic variance term for the loadings in formula (A.1) *simplifies to*

$$\tilde{\Sigma}^t_{\Lambda,l} = \omega_{ll} \cdot \Sigma^{\text{obs}}_{\Lambda} + (\omega_{ll} - 1)\Sigma^{\text{miss}}_{\Lambda,j},$$

where

$$\Sigma^{\text{obs}}_{\Lambda} = \Sigma_F^{-1}\sigma_e^2, \qquad \Sigma^{\text{miss}}_{\Lambda,l} = \Sigma_F^{-1}\Sigma_\Lambda^{-1}(\Lambda_j^\top \otimes \Sigma_\Lambda)\Xi_F(\Lambda_j \otimes \Sigma_\Lambda)\Sigma_\Lambda^{-1}\Sigma_F^{-1},$$

and $\mathbf{E}[\text{vec}(F_i^t F_i^{t\top} - \Sigma_F)\text{vec}(F_i^t F_i^{t\top} - \Sigma_F)^\top] = \Xi_F$.

(b) The asymptotic variance term for the factors in formula (A.2) *simplifies to*

$$\tilde{\Sigma}^{t,0}_{F,i} = \frac{\delta}{N_t}\Sigma^{\text{obs}}_{F,i} + \frac{\delta}{N_t}(\omega - 1)\Sigma^{\text{miss}}_{F,i},$$

where

$$\Sigma^{\text{obs}}_{F,i} = \Sigma^{-1}_{\Lambda,i}\sigma_e^2, \qquad \Sigma^{\text{miss}}_{F,t} = {\Sigma^t_{\Lambda,i}}^{-1}(I_K \otimes (F_i^{t\top}\Sigma_F^{-1}\Sigma_\Lambda^{-1}))(\Sigma^t_{\Lambda,i} \otimes \Sigma_\Lambda)\Xi_F(\Sigma^t_{\Lambda,i} \otimes \Sigma_\Lambda)(I_K \otimes (\Sigma_\Lambda^{-1}\Sigma_F^{-1}F_i^t)){\Sigma^t_{\Lambda,i}}^{-1}.$$

(c) The asymptotic variance term for the common component in formula (A.3) *simplifies to*

$$\Sigma^{XS,t,0}_{C,i,l} = \frac{\delta}{N_t}\left[F_i^{t\top}(\omega_{ll} \cdot \Sigma^{\text{obs}}_{\Lambda} + (\omega_{ll} - 1) \cdot \Sigma^{\text{miss}}_{\Lambda,l})F_i^t + (\omega - 1)\Lambda_l^{t\top}\Sigma^{\text{miss}}_{F,i}\Lambda_l^t - 2(\omega_l - 1)\Lambda_l^{t\top}\Sigma^{\text{miss, cov}}_{\Lambda,F,i,l}F_i^t\right] + \frac{\delta}{L}\Lambda_l^{t\top}\Sigma^{\text{obs}}_{F,i}\Lambda_l^t,$$

where

$$\Sigma^{Cov,t}_{F,\Lambda,i,l} = {\Sigma^t_{\Lambda,i}}^{-1}(I_K \otimes (F_i^{t\top}\Sigma_F^{-1}\Sigma_\Lambda^{-1}))(\Sigma_{\Lambda,i} \otimes \Sigma_\Lambda)\Xi_F(\Lambda_l^t \otimes \Sigma_\Lambda)\Sigma_\Lambda^{-1}\Sigma_F^{-1}.$$

The simplified model provides a clear interpretation of the effect of missing data. Without missing data, the distribution results correspond to the familiar form established in Bai (2003). The missing data leads to correction terms, which we indicate with the label "miss". In the presence of missing values, these correction terms are necessary to capture the additional uncertainty. Importantly, the parameters ω, ω_l, and ω_{ll}, which depend only on the missing pattern, but not on the factor model, determine the weights of correction terms. The asymptotic covariance of the loadings is a weighted combination of the variance of an OLS regression of the population factors F^t on C_l^t and the correction term. The weight $\omega_{ll} \geq 1$ depends on the number of the observed entries and the similarities in observation patterns for different units. Without missing data, $\omega_{ll} = 1$ and the correction term disappears. If the data is observed uniformly at random with probability p, the weight equals $\omega_{ll} = 1/p$ which is increasing in the proportion of missing observations.

Similarly, the asymptotic variance of the factors has two components: the variance of an OLS regression of the population loadings on C_i^t using only observed entries, and the correction term. The weight $\omega \geq 1$ increases the scale of the correction term. When all entries are observed, or all entries are observed missing-completely-at-random (with either the same or different probabilities), then $\omega = 1$, the correction term vanishes, and the asymptotic variance depends only on $\Sigma_{F,i}^{\text{obs}}$. If the missing pattern does not depend on the loadings, then $\Sigma_{\Lambda,i}^t = p_i \Sigma_\Lambda$ and $\Sigma_{F,t}^{\text{obs}}$ simplifies to $\frac{1}{p_i}\Sigma_\Lambda^{-1}\sigma_e^2$ which is the variance of an OLS regression of the population loadings on C_i^t scaled by the inverse proportion of observed entries for stock i.

The distribution of the common component depends on all three parameters ω, ω_l, and ω_{ll}. If all entries are observed at random, then $\omega_l = 1$ and the contribution of the loading and factor distribution to the common component are separated similar to the conventional PCA setup in Bai (2003). In this case, only the two terms $\omega_{ll}(F_i^t)^\top \Sigma_\Lambda^{\text{obs}} F_i^t$ and $(\Lambda_l^t)^\top \Sigma_{F,i}^{\text{obs}} \Lambda_l^t$ remain in the asymptotic variance.

Appendix A.3. The B-XS Model

The B-XS model combines time-series information and cross-sectional information. In order to study its properties, we need to impose assumptions on the time-series properties of the data-generating process. We use as a motivating example an AR(1) structure in the systematic and the non-systematic components, and show that our estimator can estimate such a model.

Assume that the data generating process is given by

$$\begin{aligned} C_{i,l}^t &= F_i^t \Lambda_l^t + e_{i,l}^t, \\ F_{i,k}^t &= \rho_{F,k} F_i^{t-1} + \epsilon_i^{F,t}, \\ e_{i,l}^t &= \rho_e e_{i,l}^{t-1} + \epsilon_{i,l}^{e,t}. \end{aligned}$$

Obviously, we need to impose identifying assumptions. First, we assume that the assumptions of the XS-model hold for each time t. This means in particular that the factors are systematic among

the stock (i) and characteristic (l) dimension, whereas the residual component $e_{i,l}^t$ is non-systematic in i and l. Furthermore, the systematic factors F_i^t are orthogonal to the non-systematic component, that is., $\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^t F_i^t e_{i,l}^t \xrightarrow{p} 0$ and $\frac{1}{N_t}\sum_{i=1}^{N_t} F_i^t e_{i,l}^t \xrightarrow{p} 0$. This gives us the identification and separation of $F_i^t\Lambda_l$ from the residual $e_{i,l}^t$. Note that the cross-sectional factor model does not impose any assumptions in the time-dimension. The additional time-series structure is that both the factors $F_{i,k}^t$ and the residuals $e_{i,l}^t$ follow simple AR(1) processes, and hence, the innovations $\epsilon_i^{F,t}$ and $\epsilon_{i,l}^{e,t}$ are uncorrelated over time.

We assume that Λ_l^t has the same span over time, that is, it holds that $\Lambda_l^t = \Lambda_l^{t-1}A^t$ up to some invertible (and potentially time-varying) $K \times K$ matrix A^t. We have confirmed empirically that his condition holds. This implies an AR(1) model in the systematic component $F_i^t\Lambda_l^\top = \rho_C F_i^{t-1}\Lambda_l^\top + \epsilon_{i,l}^{C,t}$. In our estimation, we will leverage the AR(1) structure in the common component and the residuals.

As a starting point, we show formally how to combine the cross-sectional factor model and the time-series dependence in the residuals. First, we estimate the common component $\hat{F}_i^t\hat{\Lambda}_l^t$ for all time periods t and stocks i. The cross-sectional information is sufficient to estimate the common component at time t, but cannot estimate the component that is only weakly contemporaneously correlated with other characteristics. Hence, if prior residuals are observed, they can help to predict the non-systematic component for missing data. Therefore, as a second model we estimate the parameters of the time-series model for the residuals. Note that we can use the stock dimension to estimate the AR(1) coefficient, that is, we are not running a regression in the time-series dimension. Because the common component and the residuals are orthogonal, we can combine the predictors of the two components.

In more detail, our identification assumptions allow us to proceed as follows:

(a) For all entries we can estimate the systematic component $\hat{F}_i^t(\hat{\Lambda}^t)^\top$. For observed entries we can also estimate the non-systematic residual component $\hat{e}_{i,l}^t = C_{i,l}^t - \hat{F}_i^T(\hat{\Lambda}^t)^\top$.

(b) Using the residuals, we estimate $\hat{\rho}^e$ with a cross-sectional regression of $\hat{e}_{i,l}^t$ on $\hat{e}_{i,l}^{t-1}$ on the partially observed data:

$$\hat{\rho}^{e,l} = \left(\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^{t-1}\left(\hat{e}_{i,l}^{t-1}\right)^2\right)^{-1}\left(\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\hat{e}_{i,l}^{t}\right).$$

(c) We combine the cross-sectional and time-series models:

$$\hat{C}_{i,l}^t = \hat{F}_i^t\Lambda_l^t + \hat{\rho}^{e,l}\hat{e}_{i,l}^{t-1}.$$

The estimation of the AR(1) model on the partially observed data allows for general missing patterns. The underlying assumption is that $\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^t\left(e_{i,l}^{t-1}\right)^2$ converges to a full rank matrix and that $\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\epsilon_{i,l}^{e,t} \xrightarrow{p} 0$. This essentially only limits the dependency of the missingness $W_{i,l}^{t-1}$ on the idiosyncratic shock $\epsilon_{i,l}^{e,t}$, but otherwise keeps the structure very general.

The estimator can be algebraically decomposed into the following elements:

$$
\begin{aligned}
&\hat{F}_i^t \hat{\Lambda}_l^t + \hat{\rho}\hat{e}_{i,l}^{t-1} - \left(F_i^t \Lambda_l^t + \rho e_{i,l}^{t-1}\right) \\
&= \left(\hat{F}_i^t \hat{\Lambda}_l^t - F_i^t \Lambda_l^t\right) + \rho\left(\hat{e}_{i,l}^{t-1} - e_{i,l}^{t-1}\right) + (\hat{\rho} - \rho)\,\hat{e}_{i,l}^{t-1} \\
&= \left(\hat{F}_i^t \hat{\Lambda}_l^t - F_i^t \Lambda_l^t\right) + \rho\left(\hat{F}_i^{t-1} \hat{\Lambda}_l^{t-1} - F_i^{t-1} \Lambda_l^{t-1}\right) \\
&\quad + \hat{e}_{i,l}^{t-1}\left(\left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}(\hat{e}_{i,l}^{t-1})^2\right)^{-1}\left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\left(\rho\hat{e}_{i,l}^{t-1} + \left(\hat{e}_{i,l}^{t} - \rho\hat{e}_{i,l}^{t-1}\right)\right)\right) - \rho\right) \\
&= \left(\hat{F}_i^t \hat{\Lambda}_l^t - F_i^t \Lambda_l^t\right) + \rho\left(\hat{F}_i^{t-1} \hat{\Lambda}_l^{t-1} - F_i^{t-1} \Lambda_l^{t-1}\right) + \hat{e}_{i,l}^{t-1}\left(\left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}(\hat{e}_{i,l}^{t-1})^2\right)^{-1}\left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\epsilon_{i,l}^{e,t}\right)\right. \\
&\quad \left. + \left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}(\hat{e}_{i,l}^{t-1})^2\right)^{-1}\left(\frac{1}{N_t}\sum_{i=1}^{N_t} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\left(\left(\hat{F}_i^t \hat{\Lambda}_l^t - F_i^t \Lambda_l^t\right) + \rho\left(\hat{F}_i^{t-1} \hat{\Lambda}_l^{t-1} - F_i^{t-1} \Lambda_l^{t-1}\right)\right)\right)\right).
\end{aligned}
$$

Using the result that $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top - F_i^t(\Lambda_l^t)^\top = O_p\left(\frac{1}{\sqrt{\delta}}\right)$, and the standard regression assumptions that $\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^{t-1}\hat{e}_{i,l}^{t-1}\epsilon_{i,l}^{e,t} = O_p\left(\frac{1}{\sqrt{\delta}}\right)$ and that $\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}} W_{i,l}^{t}\left(e_{i,l}^{t-1}\right)^2$ converges to a full rank matrix, together with standard moment bounds, we can show the consistency of the estimator,

$$
\hat{F}_i^t \hat{\Lambda}_l^t + \hat{\rho}\hat{e}_{i,l}^{t-1} - \left(F_i^t \Lambda_l^t + \rho e_{i,l}^{t-1}\right) = O_p\left(\frac{1}{\sqrt{\delta}}\right).
$$

The decomposition above also provides the foundation to determine the asymptotic distribution, which depends on the asymptotic distributions of $\sqrt{\delta}\left(F_i^t \hat{\Lambda}_l^t - F_i^t \Lambda_l^t\right)$, $\sqrt{\delta}\left(F_i^{t-1} \hat{\Lambda}_l^{t-1} - F_i^{t-1} \Lambda_l^{t-1}\right)$, and $\frac{1}{\sqrt{N_t}}\sum_{i=1}^{N_t} W_{i,l}^{t-1} e_{i,l}^{t-1}\epsilon_{i,l}^{e,t}$. Note that these three terms are in general not independent. Hence, the asymptotic variance of the overall estimator is complex, but it is feasible to obtain it explicitly.

We consider this particular estimator, which combines a cross-sectional factor model with a time-series model in the residuals in Appendix A.4. For this estimator we have a formal econometric theory. If uncertainty quantification is relevant for the researcher, then this simple estimator, which already gives a substantial improvement over the local XS model, provides an alternative for our baseline estimator.

The time t information is sufficient to obtain a consistent estimator of the common component $F_i^t(\Lambda_l^t)^\top$. However, for persistent common components, including the information in $\hat{F}_i^{t-1}(\hat{\Lambda}_l^{t-1})^\top$ can further reduce the variance of the common component estimator for time t. The idea is to estimate the common component at time t as a weighted average of the common component estimator at time t and its prediction based on time $t-1$ information. A natural estimator would run an AR(1) estimation in common components and combine it with the pure cross-sectional model and the AR(1) model in the residuals.

We propose a simple estimator that combines these elements in one regression. This is a generalization of the model that uses only the contemporaneous common component and the time-series

model in the residuals. We define a three-dimensional vector of covariates as

$$\left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad \hat{F}_i^{t-1}(\hat{\Lambda}_l^{t-1})^\top \quad \hat{e}_{i,l}^{t-1}\right).$$

Note that including $C_{i,l}^{t-1}$ and $\hat{e}_{i,l}^{t-1}$ is equivalent to including $\hat{F}_i^t(\hat{\Lambda}_l^t)^\top$ and $\hat{e}_{i,l}^{t-1}$ in a linear model, but it gives a more direct interpretation, that more clearly includes conventional models as special cases. Hence, we use the following covariates:

$$X_i^{l,t} = \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$$

We estimate the model

$$\hat{C}_{i,t}^{l,\text{B-XS}} = (\beta^{l,\text{B-XS}})^\top \left(\hat{F}_i^t(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$$

with a weighted regression on the partially observed data.

$$\hat{\beta}^{l,\text{B-XS}} = \left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t X_{i,l}^{t\top}\right)^{-1} \left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t C_{i,t}^t\right).$$

Appendix A.4. Alternative Model Estimation and Assumptions on Missingness

We show that alternative implementations of our method to combine contemporaneous systematic and past information lead to essentially the same results. The implementation of our benchmark B-XS model is guided by the goal to have a simple and easy-to-implement method with formal theoretical justification. There are alternative implementations that help to understand the impact of the different elements and assumptions on the information set.

Our benchmark model B-XS conditions on the contemporaneous common component and the past value of the common component and residual. Formally, the information set in the cross-sectional regression is given by $X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$, and the weights are obtained from the following cross-sectional regression:

$$\hat{\beta}^{l,t} = \left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t X_{i,l}^{t\top}\right)^{-1} \left(\sum_{i=1}^{N_t} W_{i,l}^t X_{i,l}^t C_{i,t}^t\right). \tag{A.4}$$

We consider four alternative specifications for this cross-sectional regression. In one alternative model we use only the past residuals in addition to the contemporaneous common component, as motivated by the illustrative model studied in Appendix A.3. For this alternative, the cross-sectional regression A.4 uses $X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad \hat{e}_{i,l}^{t-1}\right)$, that is, we drop the past common component or equivalently the past observed entry. This means that the cross-sectional regression has only two covariates. A special case of this model does not run the cross-sectional regression to determine the weights for the XS and TS components, but relies on the data-generating process of the illustrative model being correctly specified, that is, it uses $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top + \hat{\rho}^{e,l}\hat{e}_{i,l}^{t-1}$, as discussed in Appendix A.3. A

Table A.1: Imputation Error for Alternative Implementations

	OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
$X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$	**0.14**	**0.14**	**0.13**	**0.17**	**0.17**	**0.18**	**0.12**	**0.12**	**0.13**
$X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1}\right)$	0.14	0.14	0.13	0.18	0.17	0.18	0.13	0.12	0.14
$X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad \hat{e}_{i,l}^{t-1}\right)$	0.15	0.16	0.15	0.18	0.18	0.18	0.14	0.13	0.15
$X_{i,l}^t = \left(\hat{F}_i^{t,\gamma} \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$	0.13	0.14	0.12	0.17	0.17	0.18	0.12	0.12	0.13
$X_{i,l}^t = \left(\hat{F}_i^{t,\gamma} \quad C_{i,l}^{t-1}\right)$	0.14	0.14	0.13	0.17	0.17	0.18	0.13	0.12	0.13
$\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top + \hat{\rho}^{e,l}\hat{e}_{i,l}^{t-1}$	0.17	0.16	0.18	0.18	0.17	0.19	0.16	0.15	0.17
$\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top$	0.19	0.19	0.21	0.20	0.19	0.21	0.22	0.21	0.25

This table shows the imputation RMSE by imputation method averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the imputation error out-of-sample for masked characteristics for the three masking schemes MCAR, Block, and Logit. We report the errors on the subset of data that are not missing at the beginning, that is, these data have some prior values of the characteristics observed. Our benchmark B-XS model corresponds to the information set $X_{i,l}^t = \left(\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$.

second alternative model uses only the past observed value, but not the residual. Such a formulation would implicitly assume that the serial correlation in the common component and the residual is the same.

The third alternative model uses the latent characteristic factors, the past common component and the residual as covariates. This specification is similar to a model that we used in a previous version of this paper. For this specification, we use $X_{i,l}^t = \left(\hat{F}_i^{t,\gamma} \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right)$ in regression A.4, which corresponds to $K + 2$ covariates. The advantage of this specification is that missingness can depend in a general form on the covariates in this cross-sectional regression, that is, the latent factors and past values. In this sense, this regression allows for more general missing patterns as the baseline model defined in Definition 2. This relates to our discussion on the assumptions on the missing pattern in Section 3.3, where we can also allow the missing pattern to depend on the factors themselves. As a fourth alternative, we only include the latent factors and past values.

We compare the five models in Table A.1, which shows the out-of-sample RMSE. We report the results for the subset of the data, where the previous values are available. For entries for which we do not observe any previous values, we would use the XS model in all the cases and hence obtain the same relative performance ranking.

We have three findings. First, the results for the five specifications that run a cross-sectional regression on a combination of cross-sectional and time-series information are very close. Second, the past residual values seem to already include most of the time-series information. Including past common components can additionally reduce the variance, but only leads to minor improvements. Third, a model with a loading estimation that allows for more general missing patterns has no impact on the results. In addition to these alternative models, we have also analyzed a model,

where we switch the order in which we estimate loadings and factors. This has implications for the assumptions on the missing pattern. However, the out-of-sample results are also not affected.

Table A.1 also includes the results for the estimator $\hat{F}_i^{t,\gamma}(\hat{\Lambda}_l^t)^\top + \hat{\rho}^{e,l}\hat{e}_{i,l}^{t-1}$, which depends on the assumption that the data-generating process of the illustrative model is correctly specified. This estimator provides a substantial improvement over the pure cross-sectional model, which indicates that the illustrative model captures relevant information. Our cross-sectional regressions on cross-sectional and time-series information are more general models and do not rely on the correct specification of the illustrative model. They provide further out-of-sample improvements, which suggests that their additional flexibility is beneficial.

In summary, our B-XS model in Definition 2 is a reasonable benchmark. While alternative implementations lead to essentially the same out-of-sample results, our benchmark specification has the following advantages. It is simple and transparent. The formulation maps easily into econometric theory. Lastly, it is fast to implement as the spectral decomposition is applied to an $L \times L$ matrix rather than an $N_t \times N_t$ matrix.

Appendix A.5. XS Factor Model and EM Algorithm

We compare how our XS factor model and the EM algorithm leverage the correlation in the data. In order to provide some intuition, we focus on the expectation step of the EM algorithm and the cross-sectional regression step of the XS model. We emphasize that we do not model the complete EM algorithm and XS estimation, but only focus on this one step. Jin et al. (2021) and Proposition 1 in Xiong and Pelger (2023) provide a formal treatment of the EM algorithm under additional assumptions.

The EM algorithm iterates an E-step (expectation) with an M-step (maximization). As in Chen and McCoy (2022), we assume for the EM algorithm that the data is normally distributed. In the following, we take the covariance matrix of the data as given. The covariance matrix corresponds to the parameters that are estimated in the M-step of the EM algorithm. Both the EM algorithm and our XS model have to estimate the characteristic covariance matrix from the partially observed data, and those estimates are, in general, not the same. Here, we take the same characteristic covariance matrix as given and show how the two methods use it to impute missing data. Note that the N_t dimension of the number of stocks is substantially larger than the dimension L for the number of characteristics. Hence, it is reasonable to assume that estimation errors are dominated from the estimation step in the L dimension, which we can approximate by taking the characteristic covariance matrix as given.

We model the L characteristics at time t and assume that for each stock i the L dimensional characteristic vector follows

$$C_i \overset{i.i.d.}{\sim} N(0, \Sigma).$$

For convenience, we drop the index t, and set the mean of our centered characteristics to 0. It is

straightforward to extend the argument to non-zero means. Without loss of generality, we can order the characteristics for stock i such that the first block $C_{i,1}^{\text{obs}}$ is observed, whereas the second block $C_{i,2}^{\text{miss}}$ is missing, which leads to the decomposition

$$C_i = \begin{pmatrix} C_{i,1}^{\text{obs}} \\ C_{i,2}^{\text{miss}} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}.$$

We want to impute the missing $C_{i,2}^{\text{miss}}$ using the observed $C_{i,1}^{\text{obs}}$ and the covariance matrix Σ. We define the spectral decomposition of the characteristic covariance matrix as $\Sigma = VDV^\top$ for a diagonal matrix D and orthonormal matrices V, which yields the following decomposition of Σ in terms of eigenvectors:

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} V_1 D V_1^\top & V_1 D V_2^\top \\ V_2 D V_1^\top & V_2 D V_2^\top \end{pmatrix}.$$

The expectation step of the EM algorithm uses the conditional expectation of a normal distribution resulting in

$$\hat{C}_{i,2}^{\text{EM}} = \Sigma_{2,1}\Sigma_{1,1}^{-1} C_{i,1}^{\text{obs}} = V_2 D V_1^\top \left(V_1 D V_1^\top\right)^{-1} C_{i,1}^{\text{obs}}.$$

How does this compare to the XS model? We set the loadings to the first K largest eigenvalues scaled by the eigenvectors, that is, $\Lambda_{l,k} = V_{l,k}\sqrt{D_{k,k}}$ for $k = 1, ..., K$. We split the loadings into the observed and missing block, $\Lambda^\top = \begin{pmatrix} \Lambda_1^\top & \Lambda_2^\top \end{pmatrix}^\top$. The cross-sectional regression using only observed data imputes the missing data with

$$\begin{aligned} \hat{C}_{i,2}^{\text{XS}} =& \hat{F}_i \Lambda_2^\top C_{i,1}^{\text{obs}} = \Lambda_2 \left(\Lambda_1^\top \Lambda_1\right)^{-1} \Lambda_1^\top C_{i,1}^{\text{obs}} \\ =& V_{2,1:K}\sqrt{D_{1:K}} \left(V_{1,1:K} D_{1:K} V_{1,1:K}^\top\right)^{-1} \sqrt{D_{1:K}} V_{1,1:K}^\top C_{i,1}^{\text{obs}} \end{aligned}$$

where $V_{1,1:K}$ and $V_{2,1:K}$ are the first K columns of the eigenvector blocks, and $D_{1:K}$ is the submatrix of the first K eigenvalues. By simple algebra and using the generalized inverse denoted by $(.)^\dagger$, the XS estimate can be written as

$$\begin{aligned} \hat{C}_{i,2}^{\text{XS}} =& \Lambda_2 \Lambda_1^\top \left(\Lambda_1 \Lambda_1^\top\right)^\dagger C_{i,1}^{\text{obs}} \\ =& V_{2,1:K} D_{1:K} V_{1,1:K}^\top \left(V_{1,:K} D_{1:K} V_{1:K}^\top\right)^\dagger C_{i,1}^{\text{obs}}. \end{aligned}$$

Hence, the prediction given the covariance matrix is the same for EM and XS for $K = L$. The EM algorithm and XS use a similar update step with the difference being that the XS model uses only the covariance captured by the first K principal components. If we assume a low rank factor structure of the form $\Sigma = \Lambda\Lambda^\top + \sigma_e^2 I_L$, then the E-step of the EM algorithm maps into our regularized cross-sectional regression for a specific choice of γ.

Appendix A.6. Simulation

We compare in a simulation the XS factor model with the expectation maximization (EM) algorithm. As both methods leverage the cross-sectional dependency for a given t, we omit the time dimension and focus on a matrix estimation problem.

In this section we compare our robust cross-sectional factor model (XS) against the EM algorithm. We consider data generated from an approximate K-factor model, $C_{i,l} = F_i\Lambda_l + \epsilon_{i,l}$, where the factors follow $F_i \overset{iid}{\sim} N(0, I_K)$, the loadings follow $\Lambda \overset{iid}{\sim} N(0, I_K)$, and the idiosyncratic residual follows $\epsilon_{i,l} \overset{iid}{\sim} \sigma^{\epsilon} N(0, I)$.

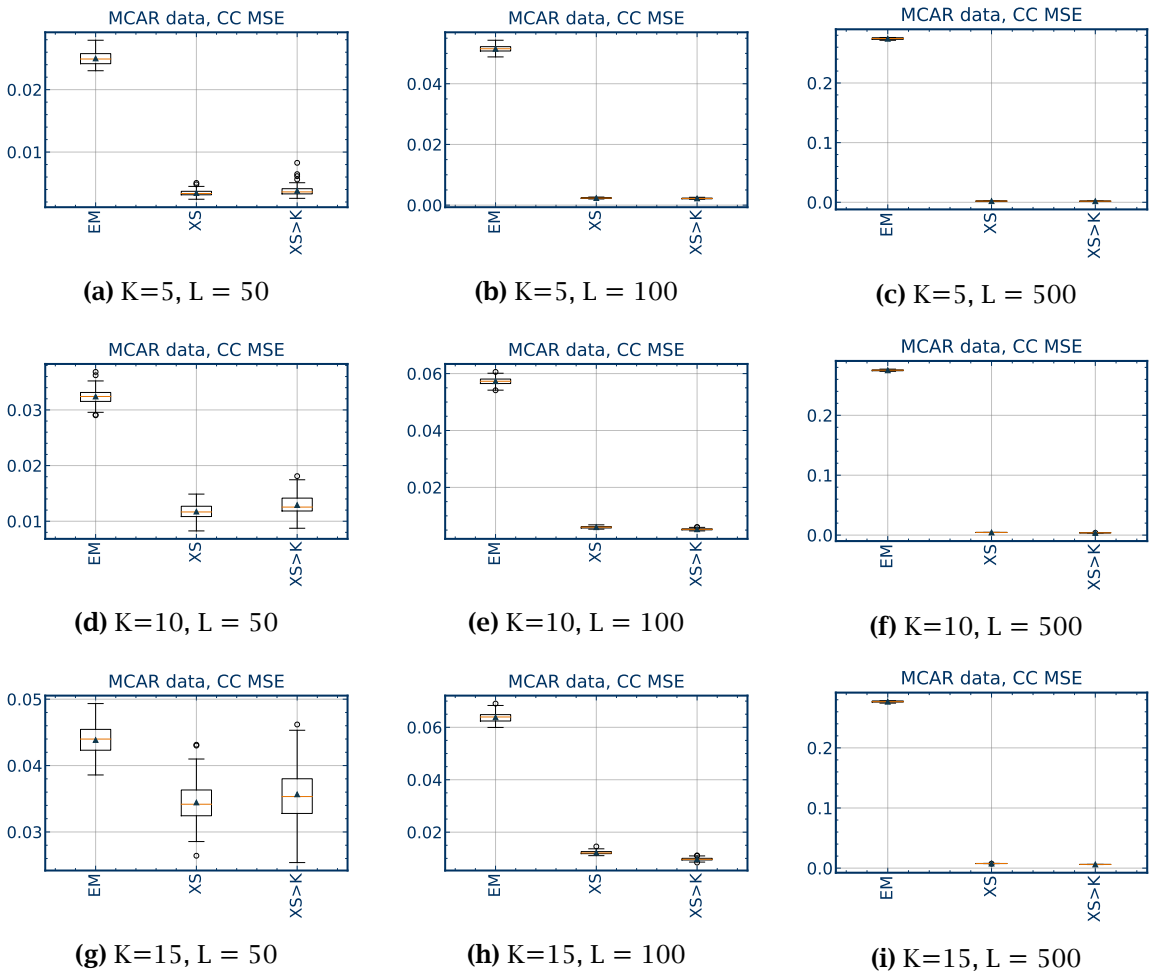
We mask 30% of the data according to two different schemes: missing-completely-at-random (MCAR) and missing-conditionally-at-random. In the MCAR scheme, the masking follows an i.i.d Bernoulli distribution $P(W_{i,l} = 1) = 0.3$. In the case of conditional masking, the probability of missingness depends on the loadings, where entries with larger loading values are more likely to be missing. This mirrors the stylized empirical observation that more extreme realizations are more likely to be missing. Note, that this is also a more challenging setup, as entries with larger loading values are more informative about the latent factor model. We mask loadings based on a logistic type model. In more detail, we mask characteristics with a probability $P(W_{i,l} = 1) \propto 1 - 1/(1 + \exp(\|\Lambda_l\|))$, where we scale the probability to obtain 30% missing observations.

We report the imputation error relative to the true common component, that is, we calculate $(\hat{C}_{i,l} - F_i\Lambda_l^{\top})^2$. Note that the errors $\epsilon_{i,l}$ are i.i.d. and without additional time-series information and the assumption of persistence cannot be learned by either of the cross-sectional models. Hence, the errors relative to the entries including the errors would mechanically be scaled up without affecting the relative ranking. With our metric, a model, that perfectly recovers the cross-sectional dependency structure modeled by the factors, would have an error of zero.

We consider two specifications of the XS model. In the first infeasible version, we use the correct number of latent factors K. This provides a reasonable benchmark to understand how well the factor model can be recovered. The second feasible version determines the number of latent factors $\hat{K}$ using cross-validation on the masked data. Specifically, we take the originally masked data, mask another 10 % of the observed entries, and then select the number of factors used for imputation as the number that minimizes the imputation error over the cross-validation masked entries. This mirrors how we select the optimal number of factors in the empirical analysis, and is therefore a relevant baseline to consider. The regularization γ is set to the estimated average idiosyncratic variance, that is, the average of the non-systematic eigenvalues of the sample characteristic covariance matrix. In the case where we estimate the number of factors via cross-validation, we use the starting value of $\gamma = 10^{-4}$ for the first iteration.

Figures A.1 and A.2 show the results for varying the numbers of factors $K = 5, 10, 15$ and the number of characteristics over $L = 50, 100, 500$. The number of stocks equals $N = 1,000$, and we run each simulation 100 times.

Figure A.1: Errors with Missing-Completely-at-Random



(a) K=5, L = 50 **(b)** K=5, L = 100 **(c)** K=5, L = 500

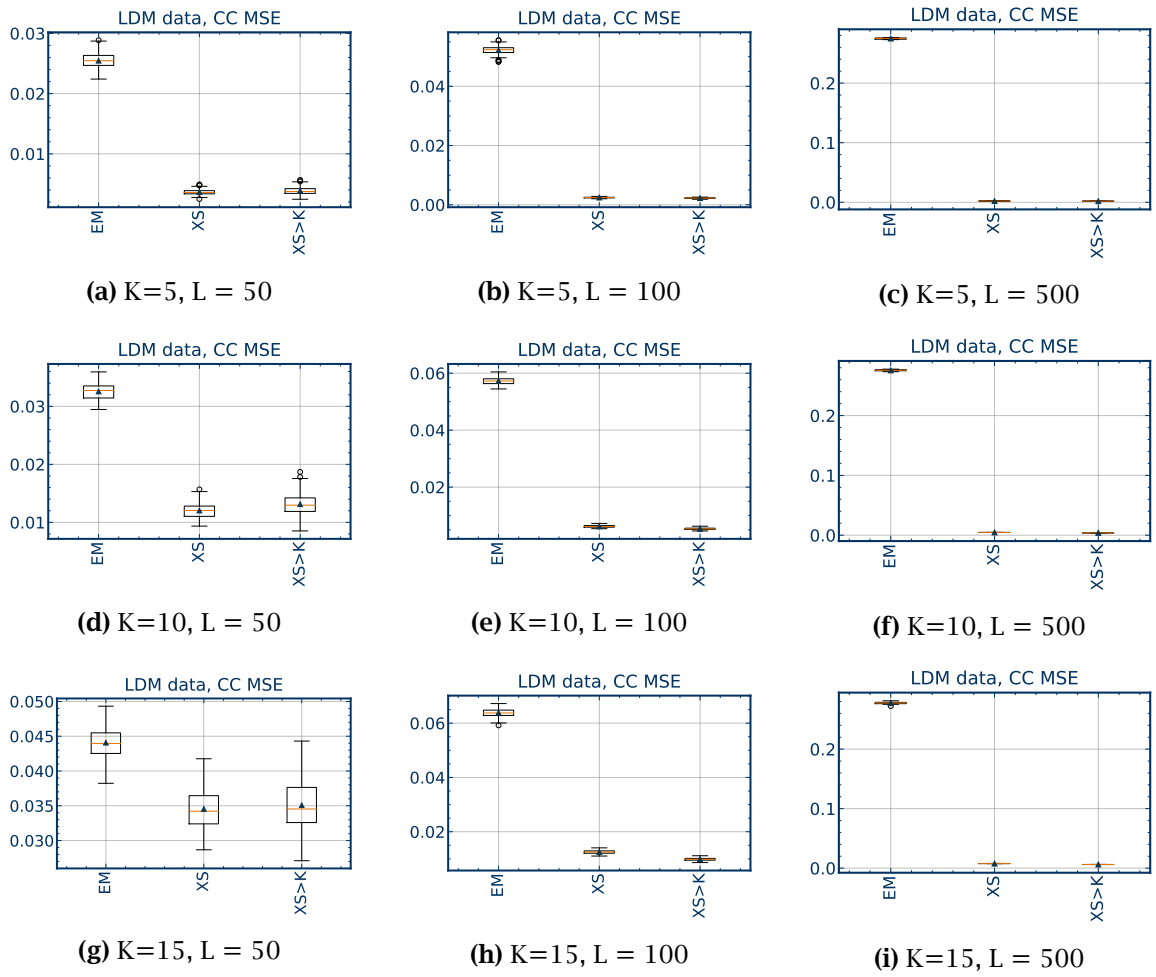
(d) K=10, L = 50 **(e)** K=10, L = 100 **(f)** K=10, L = 500

(g) K=15, L = 50 **(h)** K=15, L = 100 **(i)** K=15, L = 500

This figure shows the error for different number of factors and different number of characteristics L. We mask 30% of the entries completely-at-random. We compare the EM algorithm, the XS factor model with the correct number of factors K and the XS factor model where we select the number of factors on the validation data. The errors are relative to the true common component with significance levels (candlestick=5%, dashed line=1%, box=10%). The number of stocks equals $N = 1,000$, and each simulation is run 100 times. The number of factors is selected on a validation dataset of 10% of masked observed entries. The regularization γ is set to the average idiosyncratic variance from an iterative estimation with a starting value of $\gamma = 10^{-4}$.

First, we find that the XS model with the correct number of factors outperforms all other configurations. Second, we can reliably estimate the number of latent factors as a tuning parameter via cross-validation. The feasible XS factor model performs quite similarly to the infeasible XS model with the correct number of factors. Of course, as we estimate additional parameters, the feasible model has a larger variance. However, for a large amount of data, the difference becomes negligible. Third, the EM algorithm always performs worse than the XS model. The differences to the EM algorithm are more pronounced if the data is generated by a small number of factors. This makes sense, as the XS factor model leverages the low rank structure of the covariance matrix, while the

Figure A.2: Imputation Errors with Missing-Conditionally-at-Random



(a) K=5, L = 50 **(b)** K=5, L = 100 **(c)** K=5, L = 500

(d) K=10, L = 50 **(e)** K=10, L = 100 **(f)** K=10, L = 500

(g) K=15, L = 50 **(h)** K=15, L = 100 **(i)** K=15, L = 500

This figure shows the error for different number of factors and different number of characteristics L. We mask 30% of conditional on the loadings, where observations with larger loadings are more likely to be missing. We compare the EM algorithm, the XS factor model with the correct number of factors K and the XS factor model, where we select the number of factors on the validation data. The errors are relative to the true common component with significance levels (candlestick=5%, dashed line=1%, box=10%). The number of stocks equals $N = 1,000$, and each simulation is run 100 times. The number of factors is selected on a validation dataset of 10% of masked observed entries. The regularization γ is set to the average idiosyncratic variance from iterative estimation with a starting value of $\gamma = 10^{-4}$.

EM algorithm does not impose this structure. However, for a large number of factors K and low dimension of L, the performance between the XS factor models and EM becomes close. This should not be surprising as there is a very close connection between the conditional expectation of the missing characteristics conditional on a known covariance matrix and the imputation that the XS model makes as discussed in Appendix A.5. Fourth, the EM algorithm becomes unreliable for a large number of characteristics L. This is also expected, as the reliable estimation of large dimensional covariance matrices requires some form of regularization, which is achieved by our factor model. More specifically, the number of parameters estimated for the EM model is growing at a quadratic

rate in the number of characteristics, whereas this number grows linearly in the XS model. Fifth, the EM model is more sensitive to the characteristic-dependent missingness for a smaller number of characteristics. As expected the errors of all models are larger for this type of more complicated missing pattern.

The simulation provides two key takeaways. (1) Overall, we see that the XS model is robust to varying configurations and provides precise imputed values. (2) The EM approach works quite well when the number of characteristics is relatively small, but runs into problems when the number of characteristics increases. This highlights an important limitation of the the EM approach in that it becomes intractable as the number of covariates increases. While there are certainly ways to mitigate this, for example, by considering EM applied to a factor-model for the covariates, the algorithm would then begin to look more similar to the XS algorithm but with slightly different updates for the loadings.

Appendix B. Empirical Results for Factor Model

Appendix B.1. Rank Normalization vs. Raw Characteristics

Our main analysis reports the results for rank quantiles, but the results carry over to raw characteristics. The Internet Appendix shows the out-of-sample imputation RMSE in the original characteristic space without transforming characteristics into ranks. We consider OOS block-missing for different number of cross-sectional factors. The raw characteristics are normalized by their cross-sectional mean and variance.[19] The RMSE are further normalized by the RMSE of a simple median imputation. The first model is our baseline factor model estimated on ranks and transformed back into the characteristic space with the empirically estimated density function of each characteristic. We estimate the density function with the machine-learning method of k-nearest neighbors.

The second and third models estimate the factor model directly on the characteristics. In the fourth and fifth cases, we estimate the factor model in the kernel transformed space with a Gaussian kernel and revert it back to the raw characteristics.

We observe that a factor model estimated on rank quantiles and inverted back to raw characteristics outperforms a factor model directly applied to raw characteristics. If we use a normal distribution rather than of a non-parametric density function to invert the model into the raw characteristic space, we get a slightly worse result, but one that is still substantially better than directly estimating a factor model in the raw characteristic space. We conclude that the rank quantile space

[19] Because of the outliers we need to winsorize the data. In more detail, we first estimate the cross-sectional mean and standard deviation of each raw characteristics for each day. Then, we winsorize the values that deviate more than five standard deviations from the cross-sectional mean. After winsorizing, we reestimate the mean and standard deviation, which we use to finalize the normalization of the raw characteristics. The results are in Table IA.1 in the Internet Appendix.

is appropriate for the latent factor model and provides better results than a factor model in the raw characteristic space.[20]

Appendix B.2. Factors in Returns vs. Characteristics

Factor modeling for characteristics and returns are related but distinct problems. If conditional expected returns are a linear function in characteristics, and the characteristics themselves follow a low rank factor model, then the factor structure in characteristic implies a factor model in returns. However, a factor model that is optimal for explaining variation in characteristics does not necessarily explain the most variation in returns and the other way around. In other words, PCA applied to a panel of returns or to a characteristic covariance matrix solve different objectives, and hence can result in different factors.

We compare the latent factors obtained from characteristics and projected returns. In more detail, we first extract the loadings $\hat{\Lambda}$ of our global cross-sectional factor model as the eigenvectors of the average characteristic covariance matrix $\hat{\Sigma}^{\text{XS}}$. They correspond to the portfolio weights for constructing the characteristic factors. Second, we consider the projected stock returns $R^{\text{managed}}_{t,l} = \frac{1}{N_t} R_{t,i} C^{t-1}_{i,l}$, which correspond to L managed long-short portfolios sorted on past characteristics, and can be interpreted as L univariate factor portfolios.[21] We then apply a PCA to the return covariance matrix of the managed portfolios R^{managed}_t, which results in portfolio weights for return factors.

Figure IA.16 in the Internet Appendix reports the average generalized correlation between the factor weights obtained from characteristics and returns. A value of one implies that the weights are the same up to a rotation. It is apparent that, although there is some overlap in the factor structure, characteristic and return factors are not the same. Next, Figure IA.17 in the Internet Appendix studies the investment implications for return and characteristic factors. We apply the two sets of factors weights to the L managed portfolios R^{managed}_t and form the mean-variance efficient portfolio for different numbers of factors. We report the results in-sample and out-of-sample, where we use the first half of the data for estimation, and the second half for out-of-sample evaluation. Interestingly, the Sharpe ratios of return and characteristic factors are quite similar.

This discussion illustrates that different factor models solve different objectives. If the goal is to estimate latent factors that explain the risk premia and span the pricing kernel, then this can be achieved by using a method like Risk-Premium-PCA applied to the returns of managed portfolios, which includes the mean return as part of the objective function. On the other hand, if the goal is to capture the correlation in characteristics for imputation, this naturally implies a PCA in the characteristic space. Finally, the correlation for managed portfolios can be explained by a PCA applied to a return covariance matrix of those managed portfolios.

[20] This result extends to models that include time-series information as the persistence for characteristic rank quantiles and characteristics normalized by their cross-sectional mean and standard deviation is very similar.

[21] The projected portfolios are similar to the construction in Kelly et al. (2019) and Kozak et al. (2020).

Appendix C. Tables

Table C.1: Firm Characteristics by Category

	Past Returns		
(1)	r2_1	Short-term momentum	Monthly
(2)	r12_2	Momentum	Monthly
(3)	r12_7	Intermediate momentum	Monthly
(4)	r36_13	Long-term momentum	Monthly
(5)	LT_Rev	Long-term reversal	Monthly
	Investment		
(6)	Investment	Investment	Quarterly
(7)	NOA	Net operating assets	Quarterly
(8)	DPI2A	Change in property, plants, equipment and inventory over assets	Quarterly
(9)	NI	Net Share Issues	Quarterly
	Profitability		
(10)	PROF	Profitability	Mixed Quart. & Yearly
(11)	ATO	Net sales over lagged net operating assets	Quarterly
(12)	CTO	Capital turnover	Quarterly
(13)	FC2Y	Fixed costs to sales	Mixed Quart. & Yearly
(14)	OP	Operating profitability	Quarterly
(15)	PM	Profit margin	Quarterly
(16)	RNA	Return on net operating assets	Quarterly
(17)	ROA	Return on assets	Quarterly
(18)	ROE	Return on equity	Quarterly
(19)	SGA2S	Selling, general and administrative expenses to sales	Quarterly
(20)	D2A	Capital intensity	Quarterly
	Intangibles		
(21)	AC	Accrual	Quarterly
(22)	OA	Operating accruals	Quarterly
(23)	OL	Operating leverage	Quarterly
(24)	PCM	Price to cost margin	Quarterly
	Value		
(25)	A2ME	Assets to market cap	Quarterly
(26)	BEME	Book to Market Ratio	Quarterly
(27)	C	Ratio of cash and short-term investments to total assets	Quarterly
(28)	CF	Free Cash Flow to Book Value	Quarterly
(29)	CF2P	Cashflow to price	Quarterly
(30)	D2P	Dividend Yield	Monthly
(31)	E2P	Earnings to price	Mixed Quart. & Monthly
(32)	Q	Tobin's Q	Mixed Quart. & Monthly
(33)	S2P	Sales to price	Mixed Quart. & Monthly
(34)	Lev	Leverage	Quarterly
	Trading Frictions		
(35)	AT	Total Assets	Quarterly
(36)	Beta	CAPM Beta	Monthly
(37)	IdioVol	Idiosyncratic volatility	Monthly
(38)	LME	Size	Monthly
(39)	LTurnover	Turnover	Monthly
(40)	MktBeta	Market Beta	Monthly
(41)	High52	Closeness to past year high	
(42)	Resid_Var	Residual Variance	Monthly
(43)	Spread	Bid-ask spread	Monthly
(44)	SUV	Standard unexplained volume	Monthly
(45)	Variance	Variance	Montly

This table shows the 45 firm-specific characteristics sorted into six categories. More details on the construction are in Table IA.5 in the Internet Appendix.

Table C.2: Missing by Characteristic Quintiles

	All	ME Quintile				Characteristic Quintile			
		[1-2]	(2-3]	(3-4]	(4-5]	[1-2]	(2-3]	(3-4]	(4-5]
A2ME	12.43%	13.44%	10.51%	10.23%	9.93%	8.50%	9.56%	11.43%	15.25%
AC	43.20%	39.89%	34.04%	32.28%	26.67%	52.34%	26.01%	23.93%	51.18%
AT	12.43%	13.44%	10.51%	10.23%	9.93%	11.25%	10.20%	9.29%	9.01%
ATO	19.36%	22.33%	17.71%	16.24%	14.06%	19.27%	15.69%	14.11%	14.89%
B2M	10.69%	12.13%	8.67%	7.95%	6.63%	8.53%	7.75%	8.59%	12.31%
BETA_d	46.97%	56.44%	48.95%	44.73%	31.59%	39.19%	29.91%	28.54%	38.25%
BETA_m	35.85%	43.79%	37.57%	33.96%	23.76%	35.33%	22.39%	21.68%	32.85%
C2A	14.54%	15.49%	12.28%	12.10%	12.39%	15.45%	14.34%	12.39%	7.57%
CF2B	11.99%	14.17%	10.00%	8.86%	7.11%	9.73%	10.20%	10.09%	13.30%
CF2P	8.94%	10.81%	7.16%	5.38%	2.86%	8.62%	6.35%	6.36%	5.77%
CTO	19.35%	22.32%	17.70%	16.23%	14.06%	19.37%	15.25%	14.60%	15.24%
D2A	24.79%	25.89%	21.39%	20.77%	19.39%	22.07%	18.57%	18.61%	19.21%
D2P	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
DPI2A	55.95%	51.92%	52.41%	50.37%	44.98%	57.90%	37.42%	33.58%	38.17%
E2P	8.94%	10.81%	7.16%	5.38%	2.86%	8.70%	6.33%	5.94%	9.14%
FC2Y	28.24%	28.17%	24.02%	22.34%	23.87%	15.19%	17.68%	17.27%	20.42%
HIGH52	61.96%	70.83%	64.36%	60.54%	44.51%	83.61%	59.03%	49.68%	78.85%
INV	33.04%	38.42%	32.44%	30.16%	24.25%	43.89%	23.13%	21.88%	37.65%
IdioVol	0.04%	0.09%	0.03%	0.01%	0.00%	0.05%	0.03%	0.03%	0.05%
LEV	16.87%	16.14%	13.46%	14.17%	13.40%	12.68%	12.97%	13.45%	16.62%
ME	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NI	32.01%	39.49%	32.54%	29.44%	22.76%	39.96%	23.09%	24.72%	32.03%
NOA	20.41%	23.11%	19.00%	17.98%	16.52%	17.71%	17.17%	17.08%	15.99%
OA	32.31%	24.86%	20.88%	20.51%	19.30%	40.22%	17.57%	15.58%	42.48%
OL	14.88%	16.34%	12.74%	12.30%	12.36%	15.26%	11.42%	11.68%	13.30%
OP	18.95%	14.32%	10.00%	8.81%	7.08%	10.94%	10.85%	9.61%	8.99%
PCM	17.12%	21.26%	16.81%	13.15%	10.61%	17.53%	14.05%	11.89%	10.13%
PM	13.91%	14.98%	11.53%	10.82%	9.91%	11.82%	11.73%	11.56%	14.21%
PROF	18.24%	21.22%	16.95%	15.13%	11.73%	18.78%	13.32%	12.78%	14.74%
Q	12.43%	13.44%	10.51%	10.23%	9.93%	14.38%	11.61%	9.76%	8.32%
R12_2	20.73%	26.04%	21.98%	19.41%	13.29%	36.47%	14.75%	11.49%	41.87%
R12_7	20.56%	25.75%	21.80%	19.32%	13.23%	39.37%	15.27%	12.00%	45.58%
R2_1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
R36_13	48.09%	58.13%	50.21%	45.42%	32.03%	57.91%	28.88%	22.84%	57.89%
R60_13	63.55%	74.31%	66.17%	60.78%	44.31%	63.36%	36.02%	29.05%	56.02%
RNA	21.66%	24.03%	19.65%	18.63%	17.24%	21.01%	16.50%	15.87%	18.25%
ROA	24.85%	28.86%	23.71%	21.98%	18.45%	25.90%	20.29%	17.08%	20.22%
ROE	23.15%	27.61%	21.93%	19.76%	15.17%	25.53%	17.74%	14.86%	20.98%
RVAR	0.04%	0.07%	0.03%	0.01%	0.03%	0.02%	0.02%	0.03%	0.04%
S2P	9.27%	11.08%	7.26%	5.42%	2.91%	7.87%	6.21%	6.50%	8.21%
SGA2S	28.27%	28.23%	24.03%	22.35%	23.87%	14.81%	17.56%	17.45%	20.65%
SPREAD	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
SUV	7.74%	10.50%	8.07%	6.30%	4.23%	28.97%	6.01%	7.66%	36.33%
TURN	5.55%	7.80%	5.57%	4.30%	2.82%	9.18%	4.80%	3.53%	3.96%
VAR	0.04%	0.07%	0.03%	0.01%	0.03%	0.02%	0.02%	0.03%	0.04%

This table reports the percentage of missing observations for different size and characteristic quintiles. The means are pooled by stocks.

Table C.3: Lengths of Missing Blocks

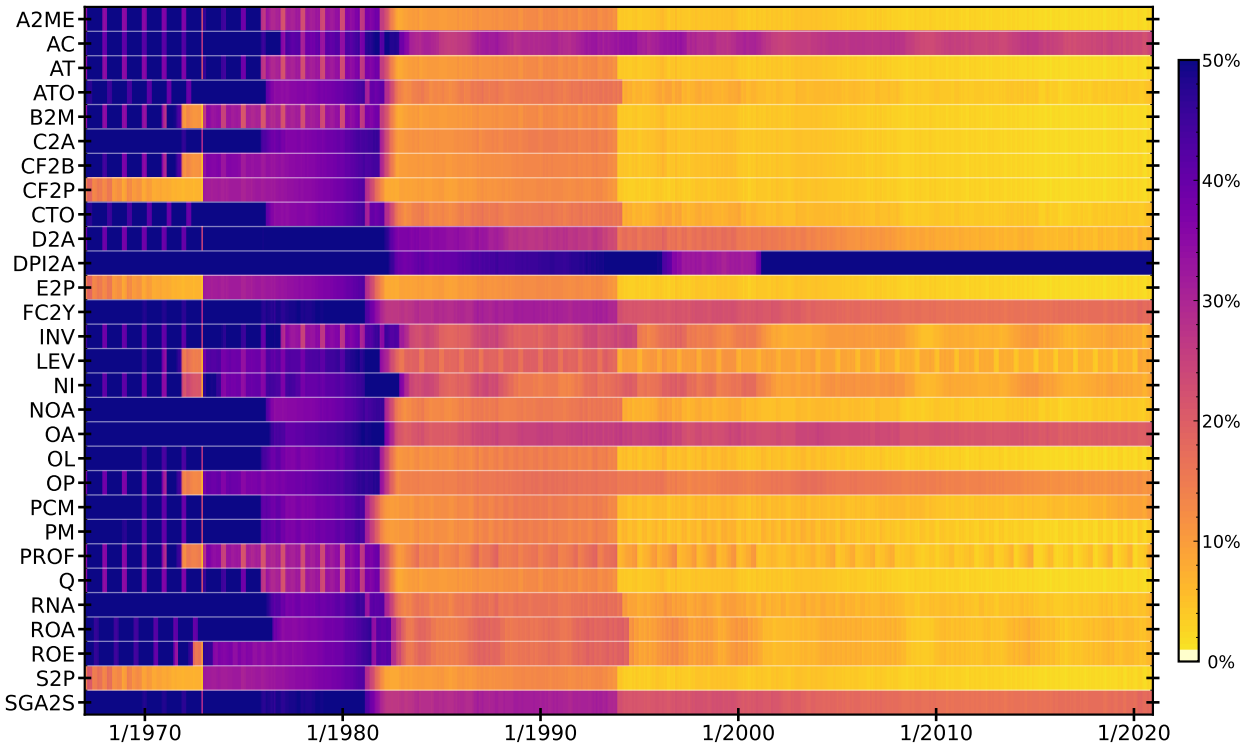
	number of gaps	mean length	median length		number of gaps	mean length	median length
A2ME	11693	11.14	9	OA	3814	18.6	7
AC	11948	12	9	OL	11320	8.85	3
AT	11693	11.14	9	OP	6542	11.75	6
ATO	7550	11.95	9	PCM	10324	9.65	3
B2M	11617	11.16	9	PM	11535	9.54	3
BETA_d	1324	31.46	4	PROF	11595	11.3	9
BETA_m	1556	28.58	5	Q	11693	11.14	9
C2A	6599	12.18	6	R12_2	1406	42.02	23
CF2B	6447	11.93	6	R12_7	2165	26.92	7
CF2P	4770	13.93	6	R2_1	2040	25.54	6
CTO	7458	12.05	9	R36_13	1812	33.59	23
D2A	14002	14.67	9	R60_13	1169	44.34	48
D2P	2040	25.54	6	RNA	12979	9.61	6
DPI2A	5612	29.51	12	ROA	6968	12.42	9
E2P	4770	13.93	6	ROE	6818	12.57	9
FC2Y	7927	15.5	9	RVAR	2019	25.89	7
HIGH52	1137	23.45	4	S2P	5238	13.32	6
INV	13076	11.28	9	SGA2S	7919	15.52	9
IdioVol	2162	24.31	6	SPREAD	2085	25.01	6
LEV	13952	13.64	9	SUV	2129	22.96	4
ME	2040	25.54	6	TURN	2156	22.53	3
NI	8757	12.11	9	VAR	2019	25.89	7
NOA	4071	16.71	7				

This table shows the number of missing blocks and their mean and median length for each characteristic.

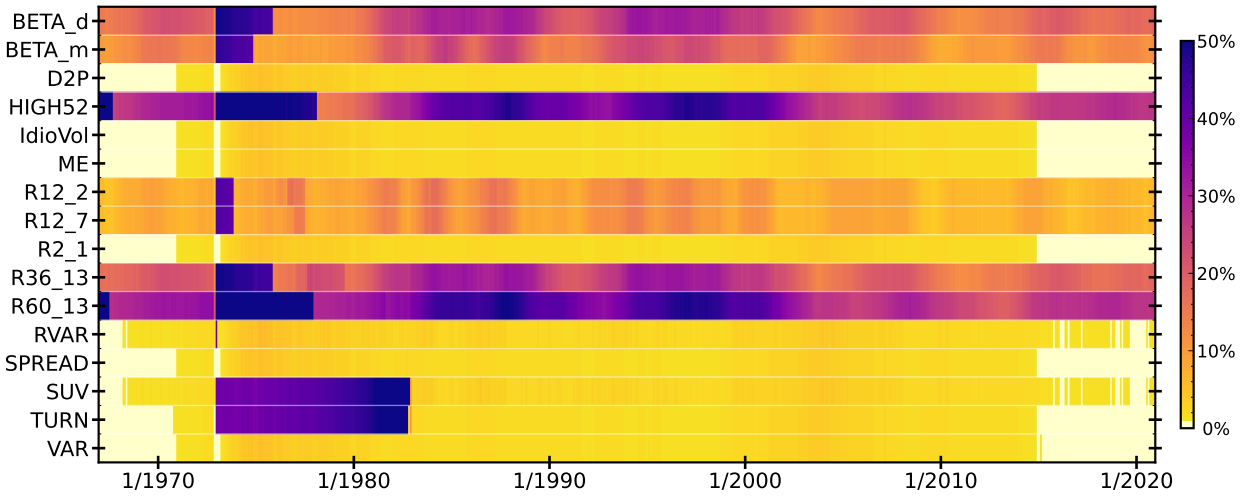
Appendix D. Figures

Figure D.1: Missing Observations over Time By Characteristics

(a) Quarterly Characteristics



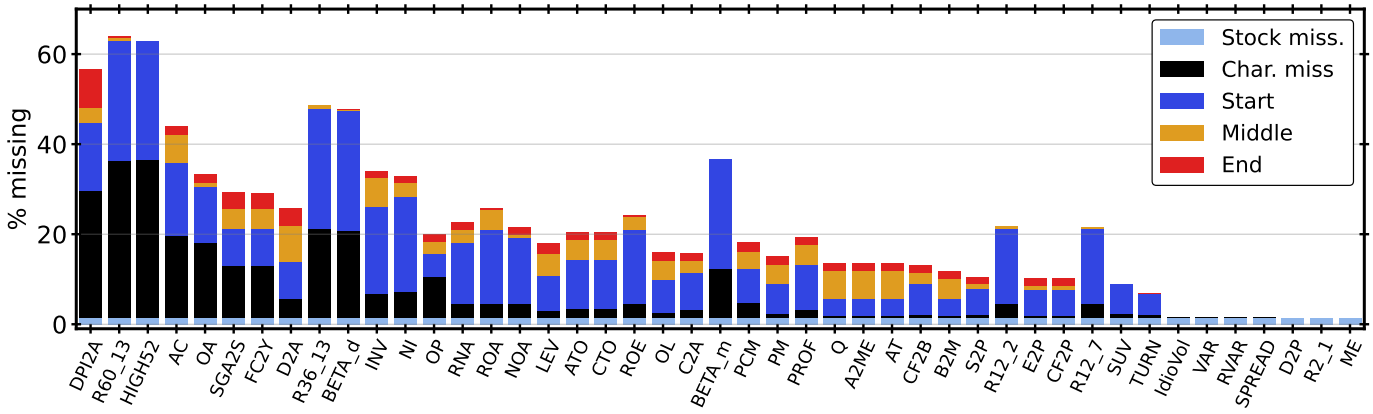
(b) Monthly Characteristics



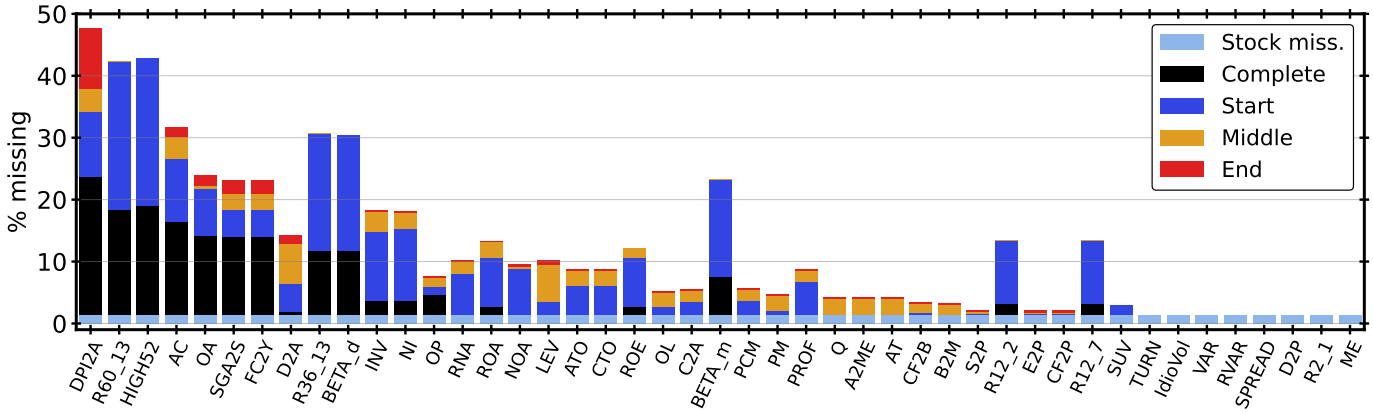
This figure is a heatmap of percentage of missing values for all 45 characteristics over time. Quarterly characteristics collect all characteristics that are updated at a frequency lower than monthly.

Figure D.2: Missing Observations by Characteristic Pooled by Stocks

(a) Pooled Mean across Stocks (Equally-weighted)

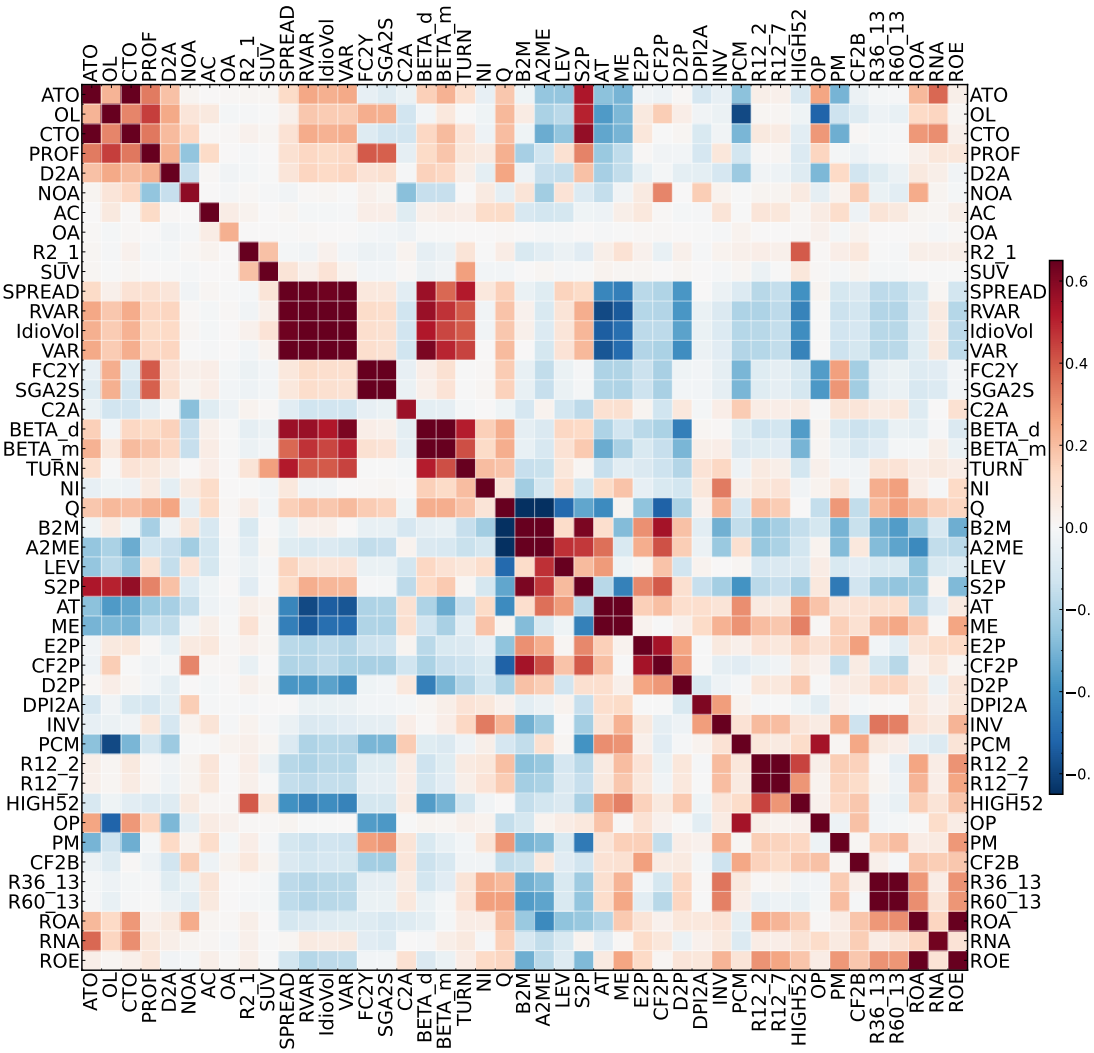


(b) Pooled Mean across Stocks (Value-weighted)



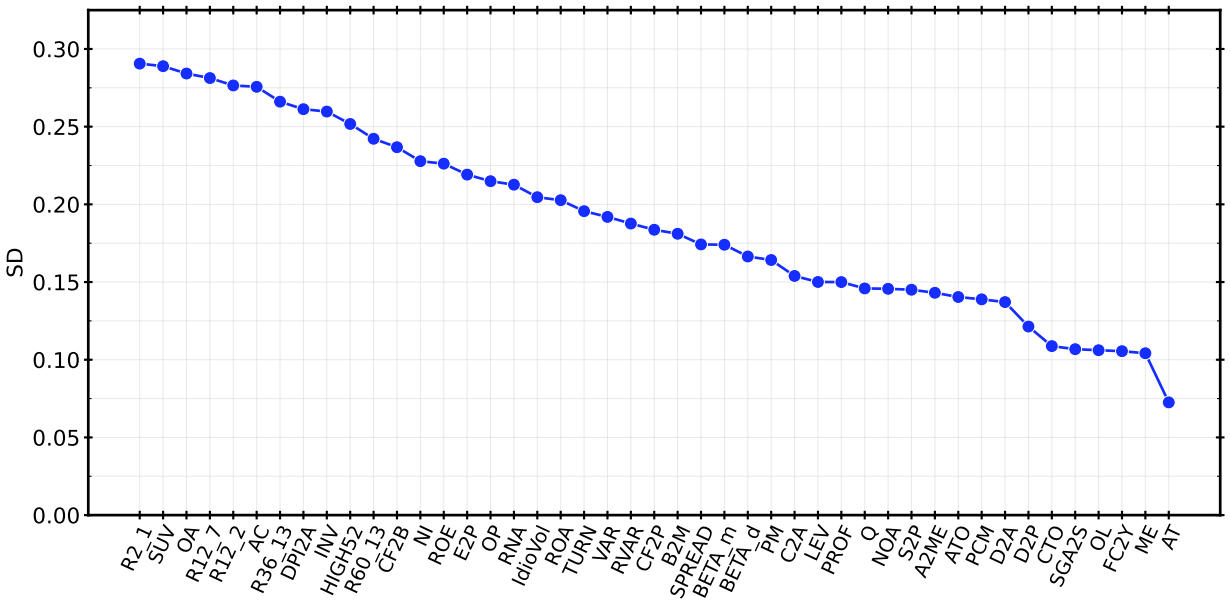
This figure shows the average percentage of missing observations for each characteristic. The means are pooled by stocks, which are equally weighted in the top panel and value-weighted in the bottom panel. We decompose the missing values in those missing at the start (no previous observations), the middle (some previous and future observations), the end (no further observations), and completely missing.

Figure D.3: Heatmap of Pairwise Correlation from 1967–1976



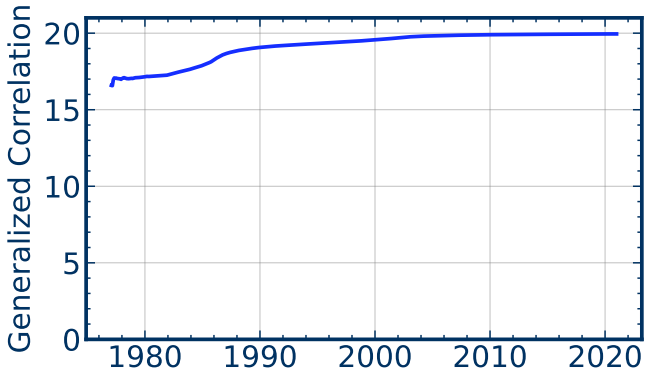
This figure shows the pairwise correlations across time and stocks for each characteristic. The time period is the early sample from 1967-1976.

Figure D.4: Standard Deviation of Characteristic Ranks



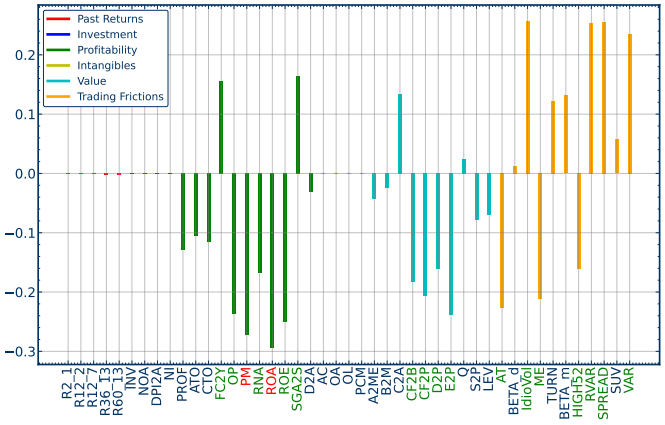
This figure presents the time-series variation of characteristic ranks. It shows the sorted standard deviation over time for each characteristic.

Figure D.5: Generalized Correlation of Global and Local Factor Weights

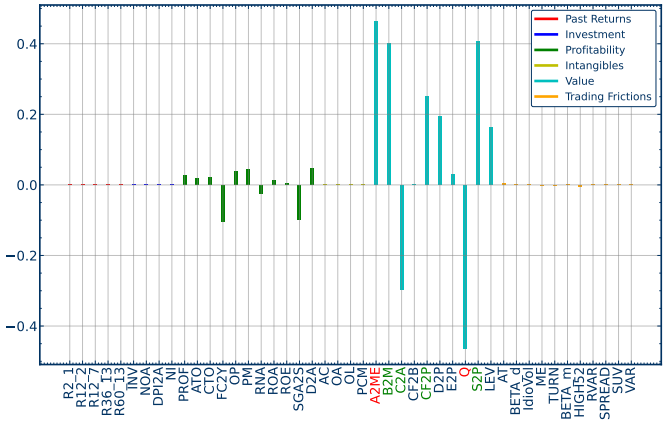


This figure shows the time-series of the generalized correlation of the constant global Λ with the time-varying local Λ^t estimated each month. We consider a 20-factor model.

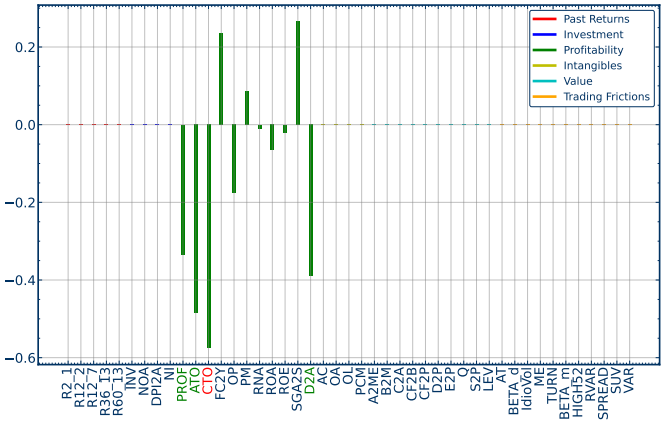
Figure D.6: Composition of Proxy Factors by Characteristic Categories



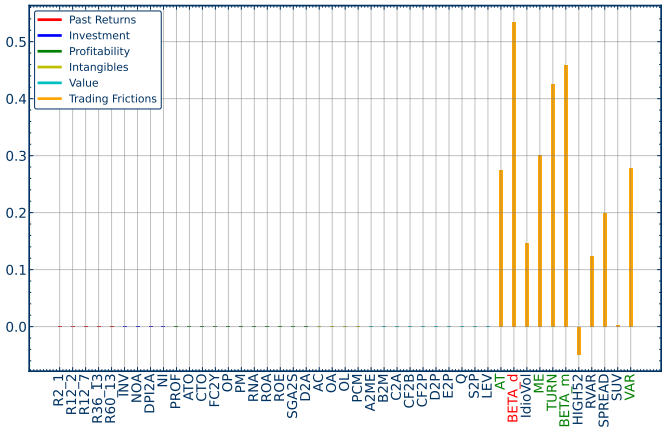
(a) Factor 1



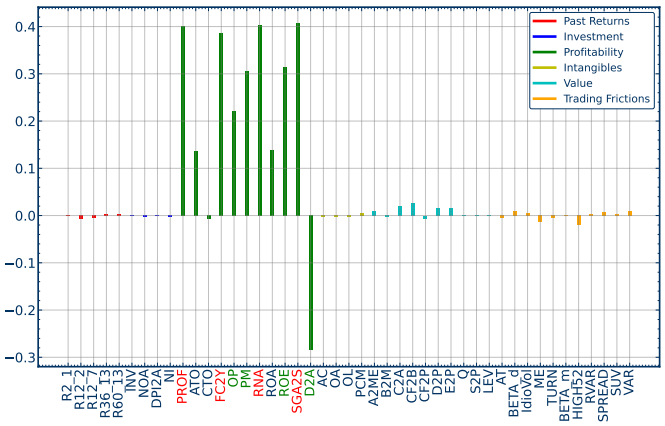
(b) Factor 2



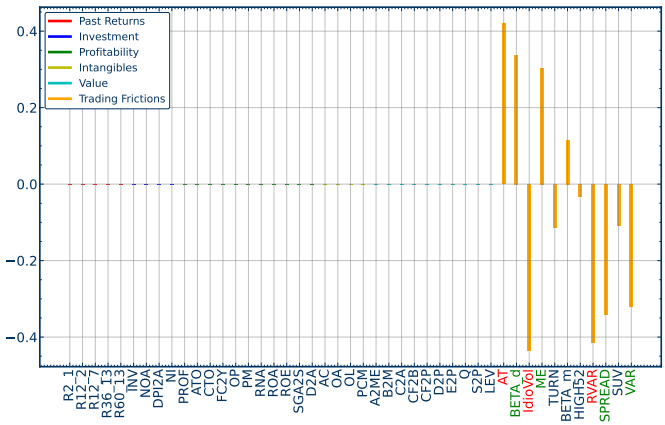
(c) Factor 3



(d) Factor 4



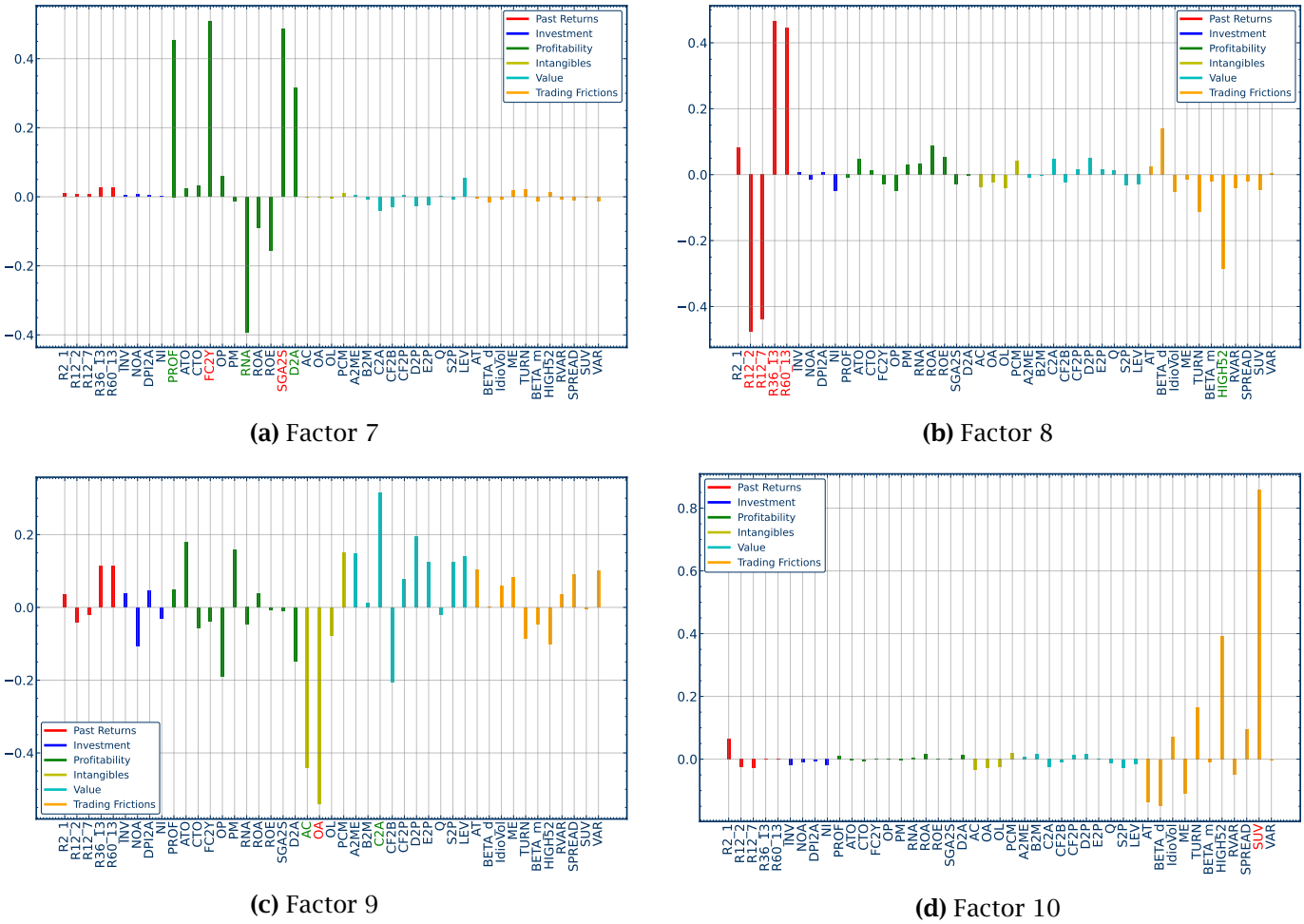
(e) Factor 5



(f) Factor 6

This figure shows the global factor loadings on the characteristics for the first six factors. The loadings are colored by the category to which the characteristic belongs. The sparse approximation is based on a group lasso.

Figure D.7: Composition of Proxy Factors by Characteristic Categories

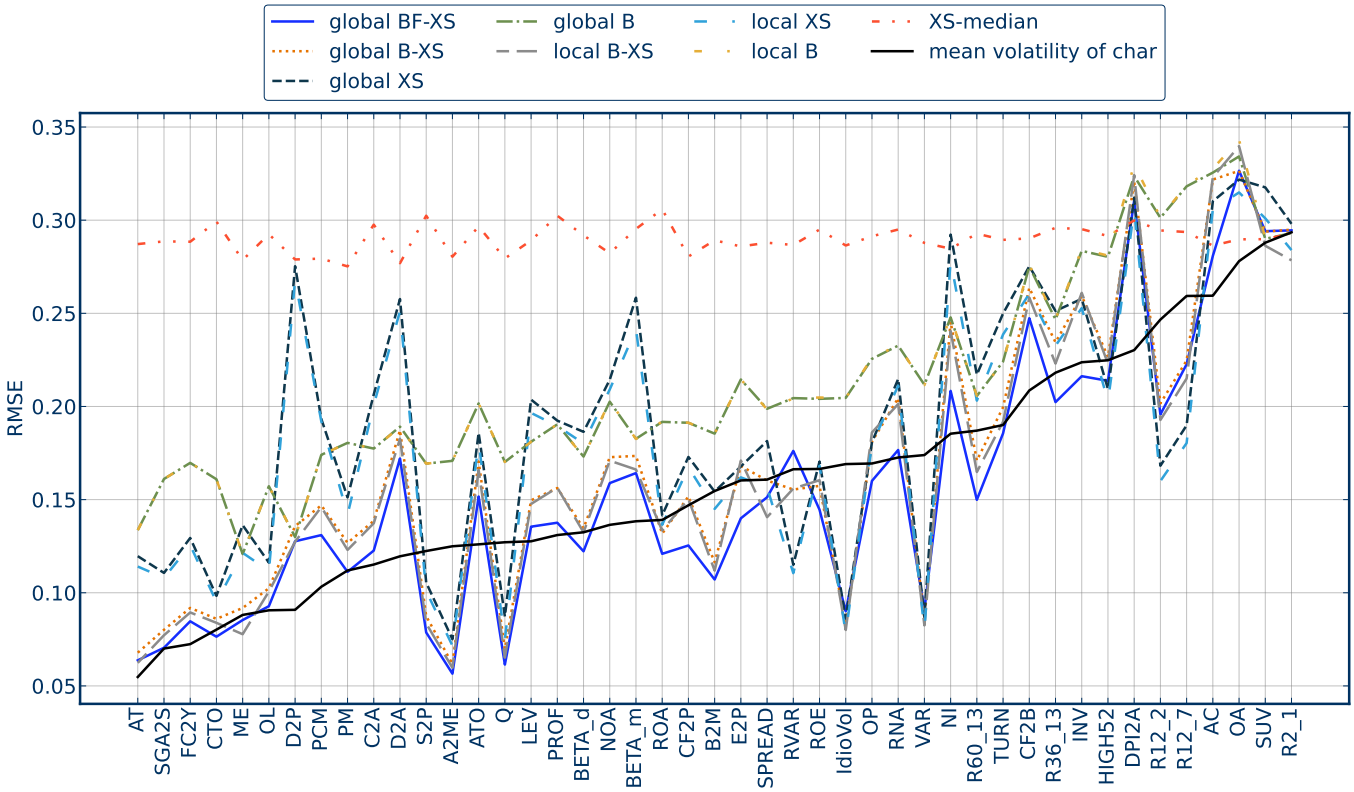


(a) Factor 7 **(b)** Factor 8

(c) Factor 9 **(d)** Factor 10

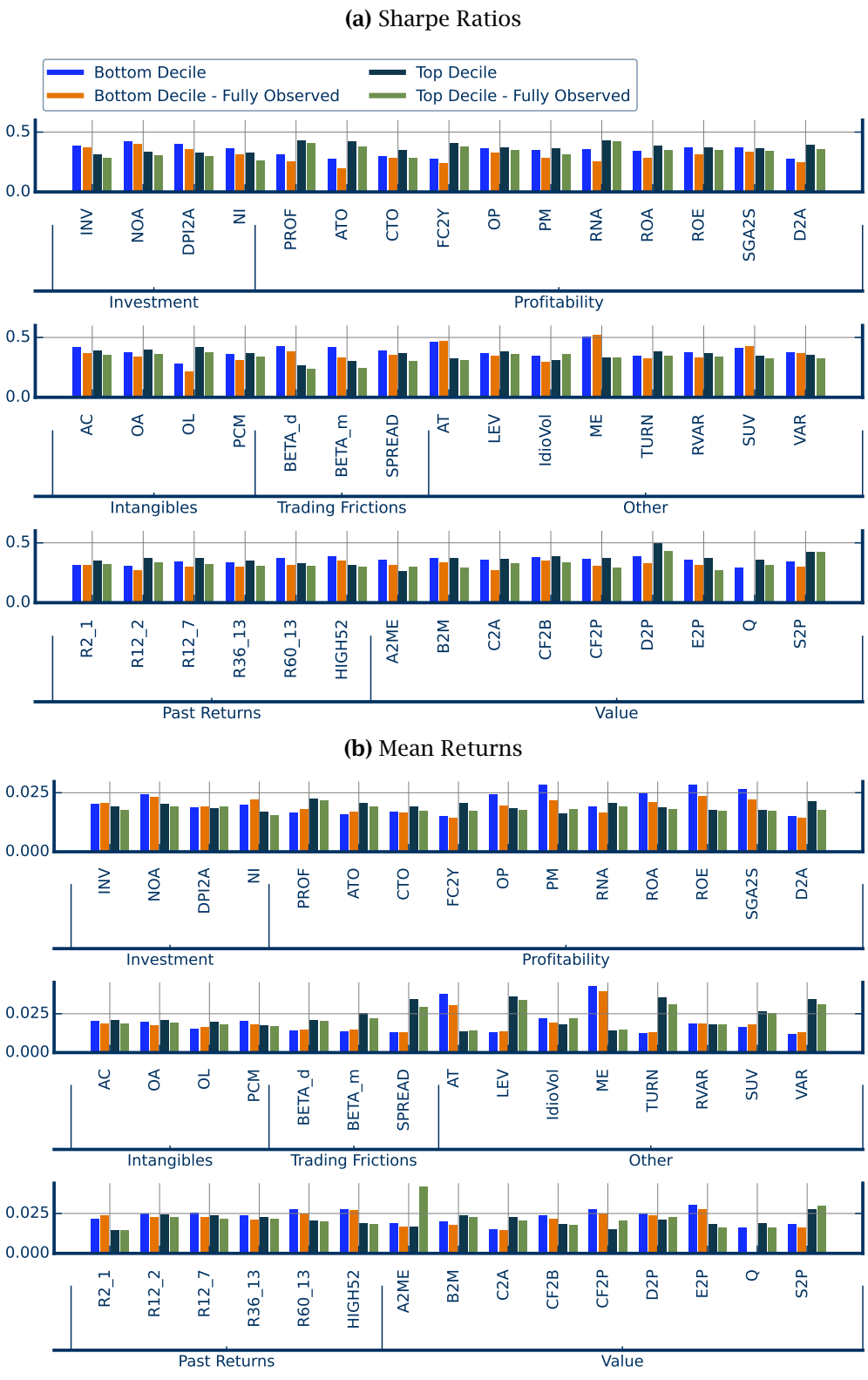
This figure shows the global factor loadings on the characteristics for the 7th to 10th factors. The loadings are colored by the category to which the characteristic belongs. The sparse approximation is based on a group lasso.

Figure D.8: Global and Local Imputation for Individual Characteristics



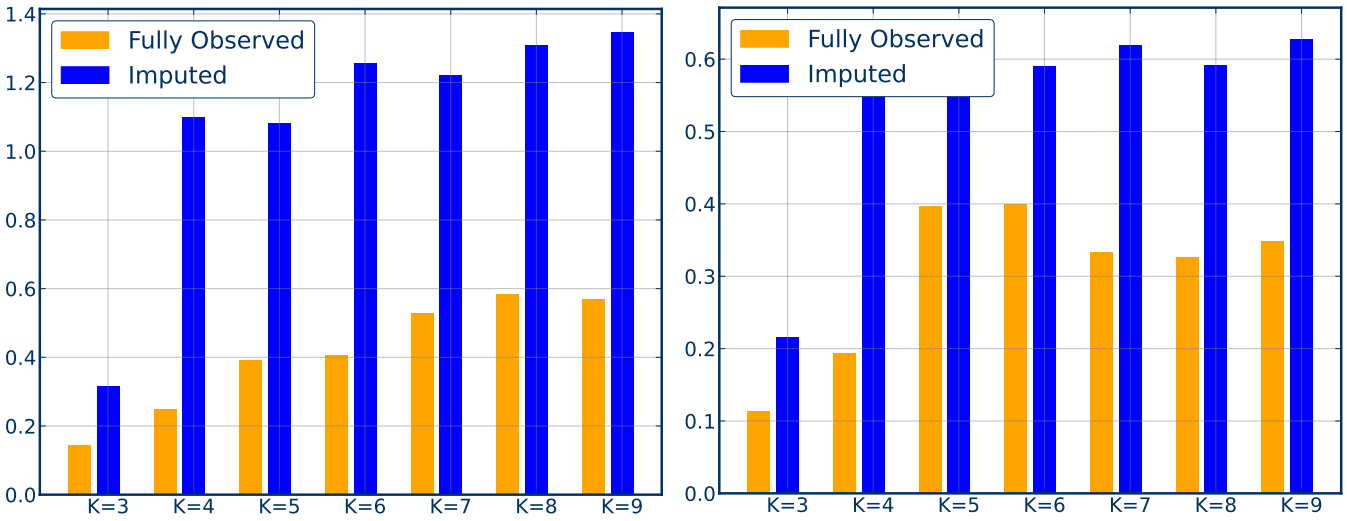
This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error out-of-sample for masked characteristics from all observed data for the block-masking scheme. We use the fallback method as indicated in Table 3, when a method is not applicable.

Figure D.9: Top and Bottom Deciles with and without Missing Values



This figure shows the Sharpe ratios and average returns for value-weighted decile-sorted portfolios formed from stocks with observed single or full panel of characteristics. The left set of plots shows the Sharpe ratios of the top and bottom deciles, while the right set of plots shows the mean returns. The light blue and green bars correspond to the first and last deciles, comprised of a fully observed panel of stocks with all the characteristics. The dark blue and green bars correspond to the return on the extreme deciles formed by stocks required to have only the characteristic available used in sorting.

Figure D.10: Sharpe Ratios with Non-parametric IPCA Factors



(a) In-Sample

(b) Out-of-Sample

This figure shows the in- and out-of-sample Sharpe ratios of mean-variance efficient combination for different number of IPCA factors. We estimate a conditional latent factor model with the Instrumented Principal Component Analysis of Kelly et al. (2019). We generalize IPCA to a nonlinear conditional factor model by considering for each characteristic 10 basis functions based on indicator functions for cross-sectional deciles. This corresponds to a kernel approximation of a non-linear loading function. The estimation is either on the smaller subset of fully observed or the larger set of all imputed stocks. The in-sample analysis is estimated on the full time period, while the out-of-sample analysis estimates the loadings and mean-variance efficient weights on the first half of the time-series and evaluates the portfolios on the second half.

Internet Appendix for Missing Financial Data

Svetlana Bryzgalova

London Business School, CEPR, sbryzgalova@london.edu

Sven Lerner

Stanford University, Institute for Computational and Mathematical Engineering, svenl@stanford.edu

Martin Lettau

University of California at Berkeley, Haas School of Business, NBER, CEPR, lettau@berkeley.edu

Markus Pelger

Stanford University, Department of Management Science & Engineering, mpelger@stanford.edu

Abstract

The Internet Appendix collects additional empirical results supporting the main text.

Keywords: Missing data, firm characteristics, cross-sectional asset pricing, PCA, factor model, big data, asset pricing

JEL classification: C14, C38, C55, G12

This draft: January 20, 2024

Contents

1. Overview

The Internet Appendix collects additional empirical results supporting the main text. Section 2 includes robustness results for the imputation, while Section 3 shows additional asset pricing results. Section 4 provides a detailed description of the data.

Section 2 first shows results for the structure of factors. The characteristic factors have a meaningful economic interpretation. The loadings Λ can be interpreted as weights to construct the characteristic factors. We focus on the global model, as it is described by only one set of weights, which are closely related to the local weights. While the main text shows the factors weights for a sparse proxy model, we include in Figures IA.1 and IA.1 the results for the full model. We can observe that certain categories are dominant for specific factors.

Our main analysis reports the results for rank quantiles, but the results carry over to raw characteristics. Table IA.1 shows the out-of-sample imputation RMSE in the original characteristic space without transforming characteristics into ranks. We consider OOS block-missing for different number of cross-sectional factors. The raw characteristics are normalized by their cross-sectional mean and variance.[1] The RMSE are further normalized by the RMSE of a simple median imputation. The first model is our baseline factor model estimated on ranks and transformed back into the characteristic space with the empirically estimated density function of each characteristic. We estimate the density function with the machine learning method of k-nearest neighbors. The second and third model estimate the factor model directly on the characteristics. In the fourth and fifth cases, we estimate the factor model in the kernel transformed space with a Gaussian kernel and revert it back to the raw characteristics. Furthermore, we also provide additional results for the out-of-sample errors as function of the number of factors and the regularization parameter γ. Figures IA.3 and IA.4 show the out-of-sample R^2 under the three masking schemes MCAR, Block and Logit, which confirm our baseline specification.

Section 2.2 reports additional robustness results on different subsets of the data. Table IA.2 reports the imputation results for the subset of the data where a method is applicable without using a fallback method. The relative ordering of method is the same as for our comparison study in the main text. The aggregated comparison results are robust with respect to the market capitalization of the stocks. Table IA.3 reports the RMSE for deciles of different sizes. Although the errors are larger in magnitude among smaller stocks, the relative comparison between the models stays the same. Importantly, even the largest decile accounts for a substantial part of the imputation errors, and, hence, the results are not driven by fitting only small-cap stocks.

Section 2.3 show the imputation accuracy for individual characteristics for different masking

[1] Because of the outliers we need to winsorize the data. In more detail, we first estimate the cross-sectional mean and standard deviation of each raw characteristics for each day. Then, we winsorize the values that deviate more than five standard deviations from the cross-sectional mean. After winsorizing, we reestimate the mean and standard deviation, which we use to finalize the normalization of the raw characteristics.

mechanisms. The results of the block-missing masking are comparable for the logit masking, as shown in Figure IA.7. The results are qualitatively similar for missing-completely-at-random, as shown in Figure IA.5. Overall, the benefit of cross-sectional information for more persistent information seems to shrink in the case of missing-completely-at-random. This is expected, as there are only a few missing points in a row, and, hence, the last observed values can be very informative. However, the relative ranking remains the same. The results are comparable for the in-sample analysis.

Section 2.4 shows that the imputation results are robust over time. Figures IA.8, IA.10, and IA.12 show the RMSE for each month. The relative ordering of the different methods is quite stable over time.

Section 3 collects asset pricing results for the cross-sectional regressions and univariate factors. In Subsection 3.1 we compare latent factors extracted from characteristics and managed portfolios. In more detail, we first extract the loadings $\hat{\Lambda}$ of our global cross-sectional factor model as the eigenvectors of the average characteristic covariance matrix $\hat{\Sigma}^{\text{XS}}$. They correspond to the portfolio weights for constructing the characteristic factors. Second, we consider the projected stock returns $R_{t,l}^{\text{managed}} = \frac{1}{N_t} R_{t,i} C_{i,l}^{t-1}$, which correspond L managed long-short portfolios sorted on past characteristics, and can be interpreted as L univariate factor portfolios. We then apply a PCA to the return covariance matrix of the managed portfolios R_t^{managed}, which results in portfolio weights for return factors. Figure IA.16 reports the average generalized correlation between the factor weights obtained from characteristics and returns. Figure IA.17 studies the investment implications for return and characteristic factors and reports the Sharpe ratio of the the mean-variance efficient portfolio for different number of factors. Our results illustrates that different factor models solve different objectives.

In Section 3.2 we collect the cumulative performance of pure-play strategies constructed on the original dataset and the dataset masked data but completed with imputed values. Figures IA.18 to IA.21 show the results for all characteristics. Portfolio performance, achieved using B-XS imputation, is remarkably stable and provides a precise approximation to the true behavior of the factor-mimicking portfolio. In contrast, we observe a substantial bias in the time-series for the factor achieved using median imputation. As a result, this bias may affect not only expected returns on the factor but also its volatility and co-movement with other portfolios, invalidating results in many empirical applications.

Section 3.3 shows that the systematic selection bias in the expected returns of decile-sorted portfolios carries over to univariate long-short factors. Table IA.4 reports the mean, standard deviation, Sharpe ratio, percentage, and market value of missing characteristics for univariate long-short decile factors. As in the case of case of decile sorts, these factors are constructed with NYSE breakpoints. We compare the results when using (1) only stocks with fully observed characteristics (i.e. 45 observed characteristics), (2) stocks with at least 10 characteristics observed and imputed

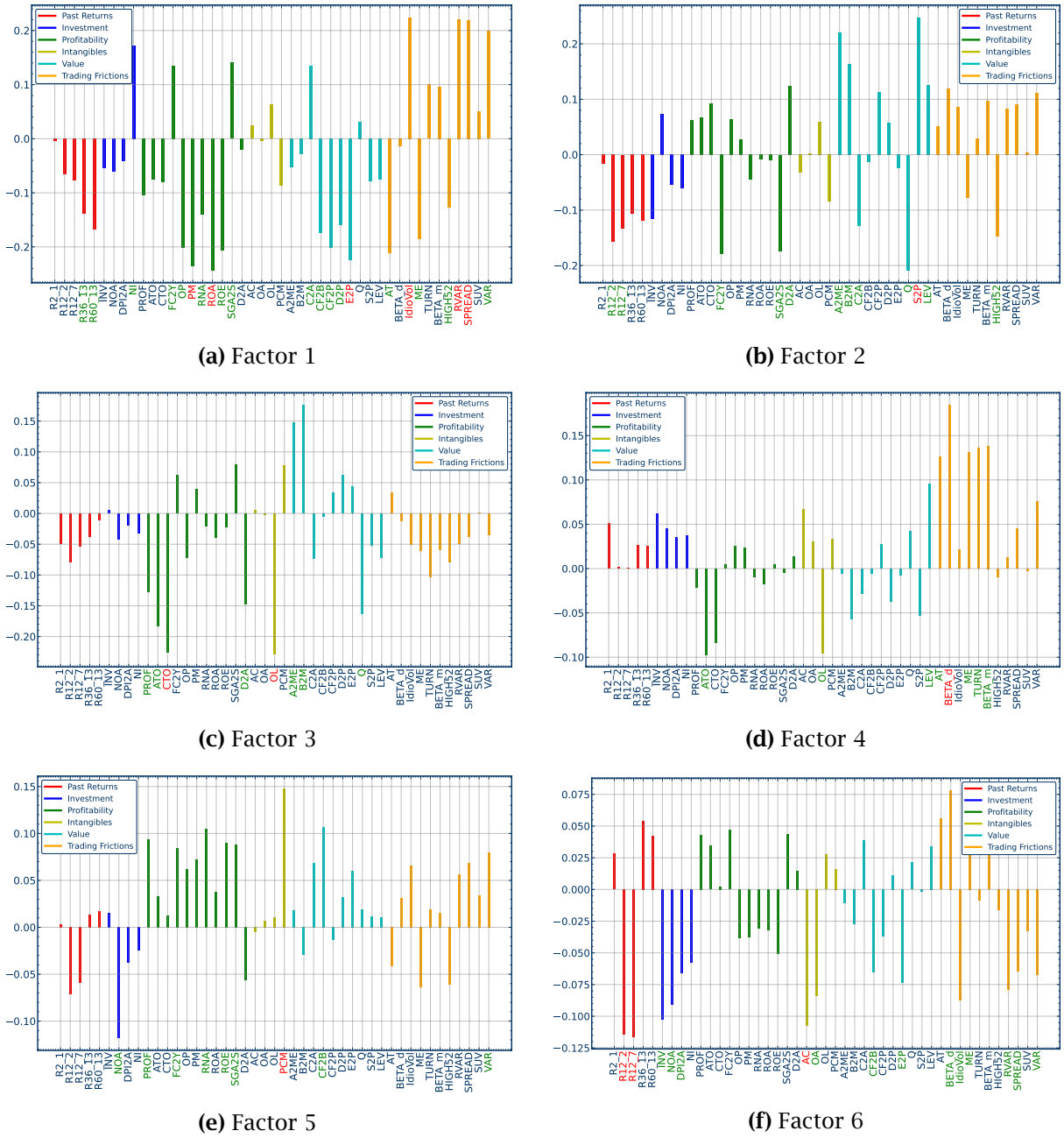
data, (3) only the specific sorting characteristic observed, the combination of (2) and (3), and the difference between (2) and (3). Stock selection obviously has a strong effect on risk premia and Sharpe ratios even for simple univariate long-short factors. As a long-short factor combines the impact of selection and imputation in the two separate legs, the effects can be complex and more or less pronounced than for the individual legs.

Lastly, Section 4 provides a detailed description of the characteristic data. Table IA.5 reports the construction of the 45 characteristics. Table IA.6 shows the dependencies between CRSP and Compustat input variables in the construction of characteristics, which can mechanically lead to dependencies in missiningness.

2. Imputation

2.1. Factor Structure

Figure IA.1: Composition of Latent Factors by Characteristic Categories



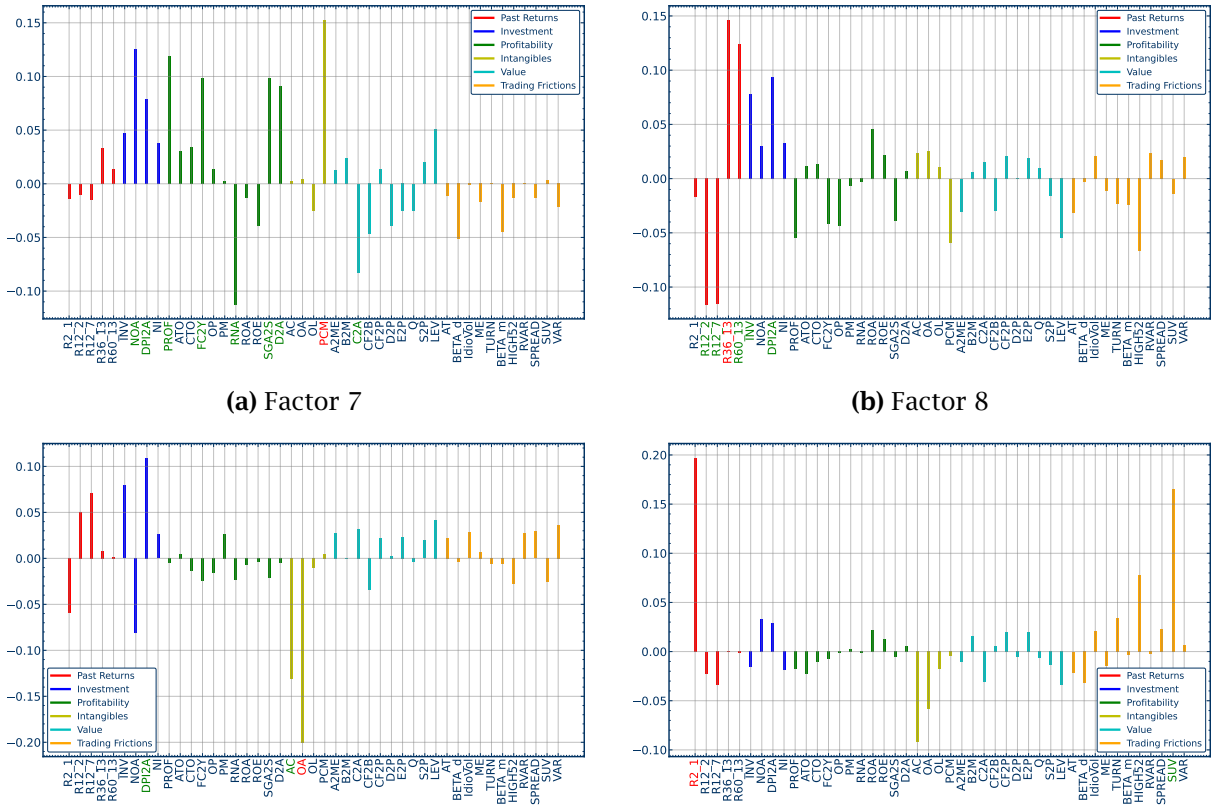
(a) Factor 1 **(b)** Factor 2

(c) Factor 3 **(d)** Factor 4

(e) Factor 5 **(f)** Factor 6

This figure shows the global factor loadings on the characteristics for the first six factors. The loadings are colored by the category to which the characteristic belongs.

Figure IA.2: Composition of Latent Factors by Characteristic Categories



(a) Factor 7 **(b)** Factor 8

(c) Factor 9 **(d)** Factor 10

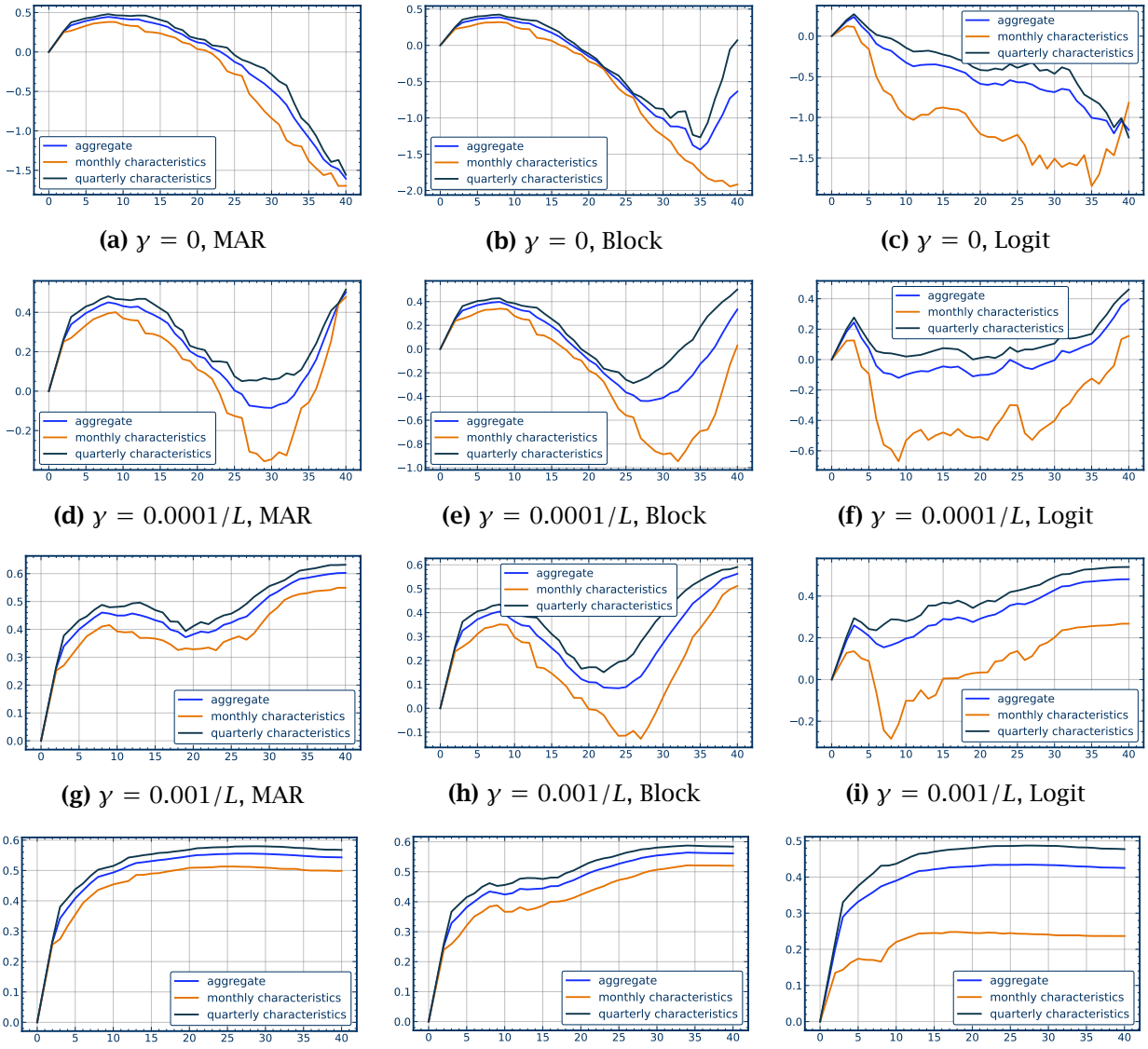
This figure shows the global factor loadings on the characteristics for the 7th to 10th factors. The loadings are colored by the category to which the characteristic belongs.

Table IA.1: OOS RMSE in Characteristic Space for XS Factor Models

Number of factors	all characteristics	quarterly characteristics	monthly characteristics
Constant factor weights on ranks			
20	0.690	0.656	0.744
Factor model on normalized raw characteristics, global fit			
10	0.919	0.924	0.911
11	0.916	0.922	0.909
12	0.914	0.921	0.905
13	0.914	0.920	0.904
14	0.913	0.919	0.903
15	0.913	0.919	0.904
16	0.913	0.919	0.906
17	0.912	0.919	0.904
18	0.913	0.919	0.904
19	0.913	0.919	0.904
20	0.912	0.919	0.903
Factor model on normalized raw characteristics, local fit			
10	0.920	0.926	0.913
11	0.919	0.925	0.911
12	0.917	0.923	0.908
13	0.916	0.922	0.907
14	0.916	0.922	0.908
15	0.916	0.922	0.908
16	0.914	0.920	0.906
17	0.914	0.921	0.905
18	0.913	0.919	0.904
19	0.912	0.918	0.903
20	0.912	0.919	0.903
Factor model on kernel transformation of ranks global fit			
10	0.793	0.768	0.831
11	0.789	0.763	0.828
12	0.786	0.761	0.825
13	0.784	0.760	0.820
14	0.784	0.759	0.821
15	0.784	0.761	0.820
16	0.785	0.761	0.823
17	0.789	0.765	0.826
18	0.790	0.765	0.828
19	0.792	0.768	0.828
20	0.790	0.765	0.827
Factor model on kernel transformation of ranks local fit			
10	0.796	0.772	0.832
11	0.791	0.767	0.829
12	0.788	0.764	0.825
13	0.786	0.763	0.821
14	0.786	0.763	0.820
15	0.786	0.764	0.819
16	0.783	0.760	0.819
17	0.785	0.761	0.820
18	0.785	0.761	0.821
19	0.785	0.762	0.820
20	0.784	0.761	0.819

This table shows the out-of-sample imputation RMSE in the original characteristic space without transforming characteristics into ranks. The characteristics are normalized by their cross-sectional mean and variance. The RMSE are further normalized by the RMSE of a medial that sets imputed values to zero, i.e., a simple median imputation. The first model is our baseline factor model estimated on ranks and transformed back into the characteristic space with the empirically estimated density function of each characteristic. We estimate the density function with the machine-learning method, k-nearest neighbor. The second and third models estimate the factor model directly on the characteristics. In the fourth and fifth cases, we estimate the factor model in the kernel transformed space with a Gaussian kernel and revert it back to the raw characteristics. For the out-of-sample analysis we use the block-masking scheme of 10% of the data.

Figure IA.3: Out-of-Sample R^2 as Function of Number of Factors



(a) $\gamma = 0$, MAR **(b)** $\gamma = 0$, Block **(c)** $\gamma = 0$, Logit

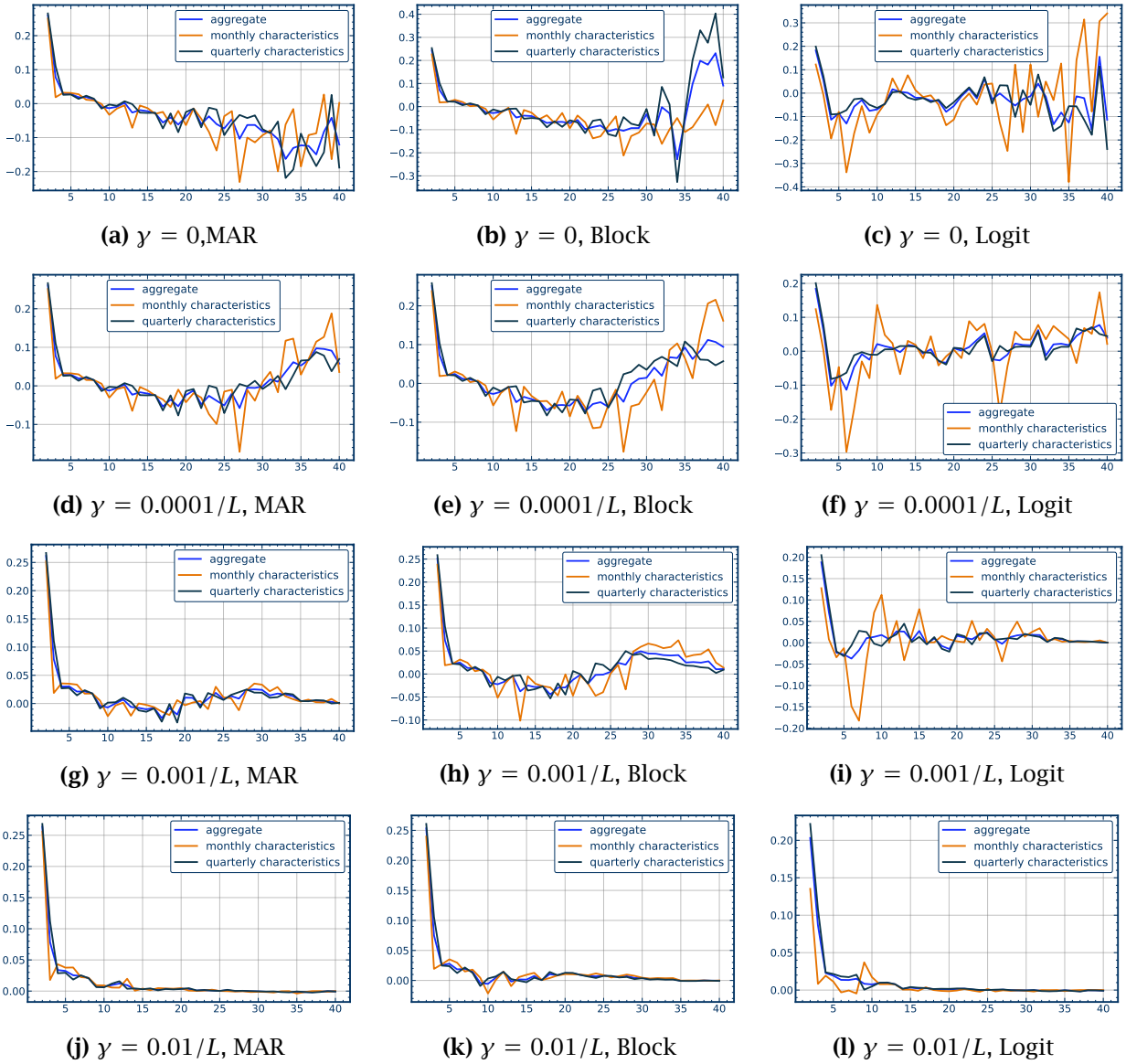
(d) $\gamma = 0.0001/L$, MAR **(e)** $\gamma = 0.0001/L$, Block **(f)** $\gamma = 0.0001/L$, Logit

(g) $\gamma = 0.001/L$, MAR **(h)** $\gamma = 0.001/L$, Block **(i)** $\gamma = 0.001/L$, Logit

(j) $\gamma = 0.01/L$, MAR **(k)** $\gamma = 0.01/L$, Block **(l)** $\gamma = 0.01/L$, Logit

The figure shows the out-of-sample R^2 of the local XS model for different numbers of latent factors and regularization for three masking schemes missing-completely-at-random, block-missing and logit-missing. The R^2 is the explained variation relative to a cross-sectional median imputation.

Figure IA.4: Incremental Out-of-Sample R^2 as Function of Number of Factors



(a) $\gamma = 0$,MAR **(b)** $\gamma = 0$, Block **(c)** $\gamma = 0$, Logit

(d) $\gamma = 0.0001/L$, MAR **(e)** $\gamma = 0.0001/L$, Block **(f)** $\gamma = 0.0001/L$, Logit

(g) $\gamma = 0.001/L$, MAR **(h)** $\gamma = 0.001/L$, Block **(i)** $\gamma = 0.001/L$, Logit

(j) $\gamma = 0.01/L$, MAR **(k)** $\gamma = 0.01/L$, Block **(l)** $\gamma = 0.01/L$, Logit

The figure shows the incremental change in out-of-sample R^2 of the local XS model for different numbers of latent factors and regularization for three masking schemes missing-completely-at-random, block-missing and logit-missing. The R^2 is the explained variation relative to a cross-sectional median imputation.

Table IA.2: Imputation Error for Different Imputation Methods

	In-Sample			OOS MCAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
	Imputation RMSE											
	Missing in the middle											
global BF-XS	0.09	0.08	0.10	0.12	0.13	0.12	0.15	0.14	0.16	0.11	0.10	0.12
global F-XS	0.08	0.05	0.12	0.14	0.14	0.13	0.17	0.16	0.18	0.17	0.15	0.21
	Not missing at the beginning											
global B-XS	0.13	0.14	0.12	0.14	0.14	0.13	0.17	0.17	0.18	0.13	0.12	0.13
local B-XS	0.13	0.14	0.12	0.14	0.14	0.13	0.17	0.17	0.18	0.12	0.12	0.13
global B	0.14	0.14	0.14	0.15	0.15	0.14	0.19	0.18	0.21	0.14	0.13	0.15
local B	0.14	0.15	0.14	0.15	0.15	0.14	0.19	0.18	0.21	0.14	0.13	0.15
prev	0.16	0.16	0.18	0.17	0.16	0.18	0.21	0.19	0.24	0.15	0.14	0.18
	All types of missingness											
global XS	0.20	0.19	0.21	0.20	0.19	0.22	0.20	0.19	0.22	0.23	0.21	0.26
local XS	0.19	0.18	0.20	0.19	0.19	0.21	0.20	0.19	0.21	0.22	0.21	0.25
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.30
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.30
	Explained Variation R^2											
	Missing in the middle											
global BF-XS	0.88	0.93	0.86	0.83	0.83	0.83	0.69	0.77	0.65	0.91	0.91	0.62
global F-XS	0.85	0.98	0.81	0.78	0.76	0.79	0.63	0.69	0.61	0.62	0.78	0.23
	Not missing at the beginning											
global B-XS	0.82	0.79	0.83	0.80	0.78	0.80	0.63	0.66	0.61	0.86	0.87	0.58
local B-XS	0.83	0.79	0.83	0.81	0.77	0.81	0.65	0.65	0.64	0.86	0.87	0.59
global B	0.75	0.77	0.74	0.74	0.76	0.74	0.50	0.61	0.46	0.85	0.85	0.47
local B	0.76	0.77	0.75	0.75	0.76	0.74	0.51	0.61	0.46	0.85	0.85	0.47
prev	0.64	0.74	0.60	0.63	0.73	0.59	0.36	0.57	0.28	0.83	0.84	0.05
	All types of missingness											
global XS	0.49	0.53	0.47	0.47	0.53	0.45	0.48	0.53	0.45	0.38	0.47	0.14
local XS	0.55	0.58	0.54	0.53	0.56	0.51	0.53	0.56	0.51	0.42	0.51	0.22
XS-median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ind-median	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00

This table shows imputation RMSE and R^2 by imputation method averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from all observed data for the three masking schemes MCAR, Block, and Logit. We report the results for the subset of the data where a method is applicable. As different methods require different data availability, we report the results separately for all the data missing in the middle, no missing at the beginning, and any type of missingness. The R^2 is the explained variation relative to a cross-sectional median imputation.

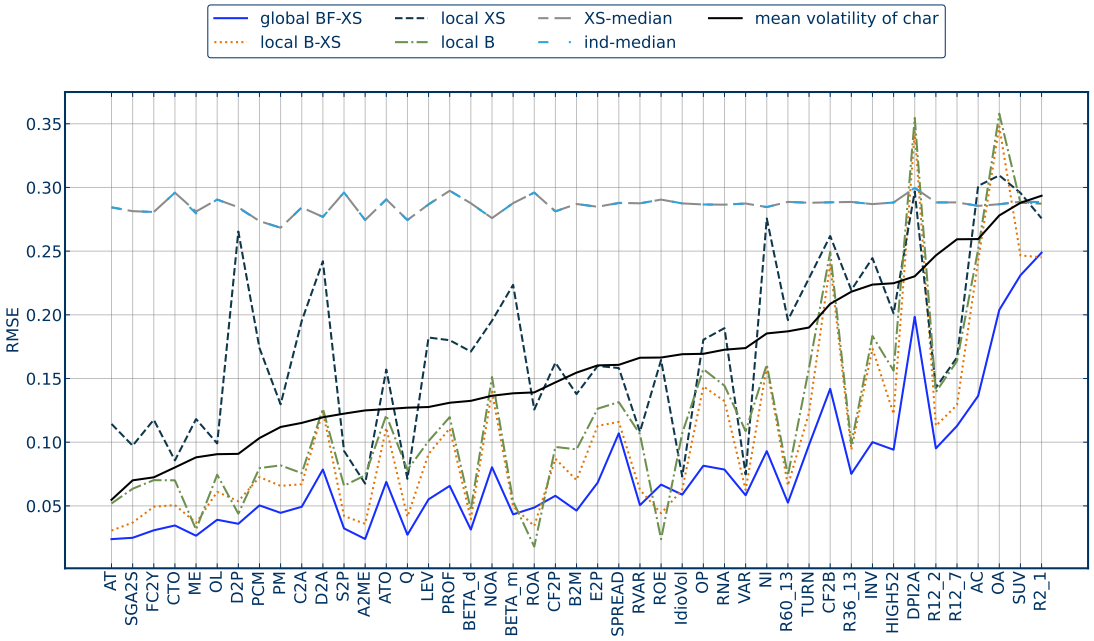
Table IA.3: Imputation Error by Size Deciles

		In-Sample			OOS MCAR			OOS Block			OOS Logit		
size decile	method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
1	local B-XS	0.15	0.16	0.14	0.16	0.16	0.15	0.20	0.20	0.20	0.14	0.14	0.16
	local XS	0.22	0.21	0.23	0.22	0.22	0.23	0.23	0.22	0.24	0.24	0.23	0.26
	local B	0.17	0.17	0.16	0.17	0.17	0.16	0.23	0.22	0.24	0.16	0.15	0.18
2	local B-XS	0.15	0.15	0.13	0.15	0.16	0.14	0.19	0.19	0.19	0.14	0.13	0.15
	local XS	0.21	0.20	0.22	0.21	0.21	0.23	0.22	0.21	0.23	0.23	0.22	0.26
	local B	0.16	0.16	0.15	0.16	0.16	0.16	0.22	0.21	0.23	0.15	0.14	0.17
3	local B-S	0.14	0.15	0.13	0.15	0.15	0.14	0.19	0.18	0.19	0.13	0.13	0.15
	local XS	0.20	0.20	0.22	0.21	0.20	0.22	0.21	0.20	0.22	0.23	0.22	0.26
	local B	0.15	0.16	0.15	0.16	0.16	0.15	0.21	0.20	0.23	0.14	0.14	0.17
4	local B-XS	0.14	0.14	0.12	0.14	0.15	0.13	0.18	0.18	0.19	0.13	0.13	0.14
	local XS	0.20	0.19	0.21	0.20	0.19	0.22	0.20	0.19	0.22	0.23	0.22	0.26
	local B	0.15	0.15	0.15	0.15	0.15	0.15	0.21	0.19	0.22	0.14	0.13	0.17
5	local B-XS	0.13	0.14	0.12	0.14	0.15	0.13	0.18	0.17	0.18	0.13	0.12	0.15
	local XS	0.19	0.18	0.21	0.20	0.19	0.21	0.20	0.19	0.22	0.22	0.21	0.25
	local B)	0.15	0.15	0.14	0.15	0.15	0.15	0.20	0.19	0.22	0.13	0.13	0.16
6	local B-XS	0.13	0.14	0.12	0.13	0.14	0.13	0.17	0.17	0.18	0.12	0.12	0.14
	local XS	0.18	0.18	0.20	0.19	0.18	0.21	0.19	0.18	0.21	0.22	0.20	0.25
	local B	0.14	0.14	0.14	0.14	0.14	0.14	0.19	0.18	0.21	0.13	0.12	0.16
7	local B-XS	0.12	0.13	0.11	0.13	0.13	0.12	0.16	0.16	0.17	0.11	0.11	0.13
	local XS	0.18	0.17	0.19	0.18	0.18	0.20	0.19	0.18	0.20	0.21	0.20	0.25
	local B	0.14	0.14	0.14	0.14	0.14	0.14	0.19	0.18	0.21	0.12	0.12	0.15
8	local B-XS	0.12	0.13	0.11	0.12	0.13	0.12	0.16	0.15	0.16	0.11	0.11	0.13
	local XS	0.17	0.17	0.19	0.18	0.17	0.19	0.18	0.17	0.20	0.20	0.19	0.24
	local B	0.13	0.13	0.13	0.13	0.13	0.13	0.18	0.17	0.20	0.12	0.12	0.15
9	local B-XS	0.12	0.12	0.10	0.12	0.13	0.11	0.15	0.15	0.15	0.11	0.11	0.13
	local XS	0.17	0.16	0.18	0.17	0.17	0.19	0.17	0.17	0.19	0.19	0.18	0.22
	local B	0.13	0.13	0.12	0.13	0.13	0.13	0.17	0.16	0.19	0.12	0.11	0.15
10	local B-XS	0.11	0.12	0.10	0.11	0.12	0.11	0.14	0.14	0.15	0.10	0.10	0.12
	local XS	0.17	0.16	0.17	0.17	0.16	0.18	0.17	0.16	0.18	0.18	0.17	0.21
	local B	0.12	0.12	0.12	0.12	0.12	0.12	0.16	0.15	0.17	0.11	0.10	0.14

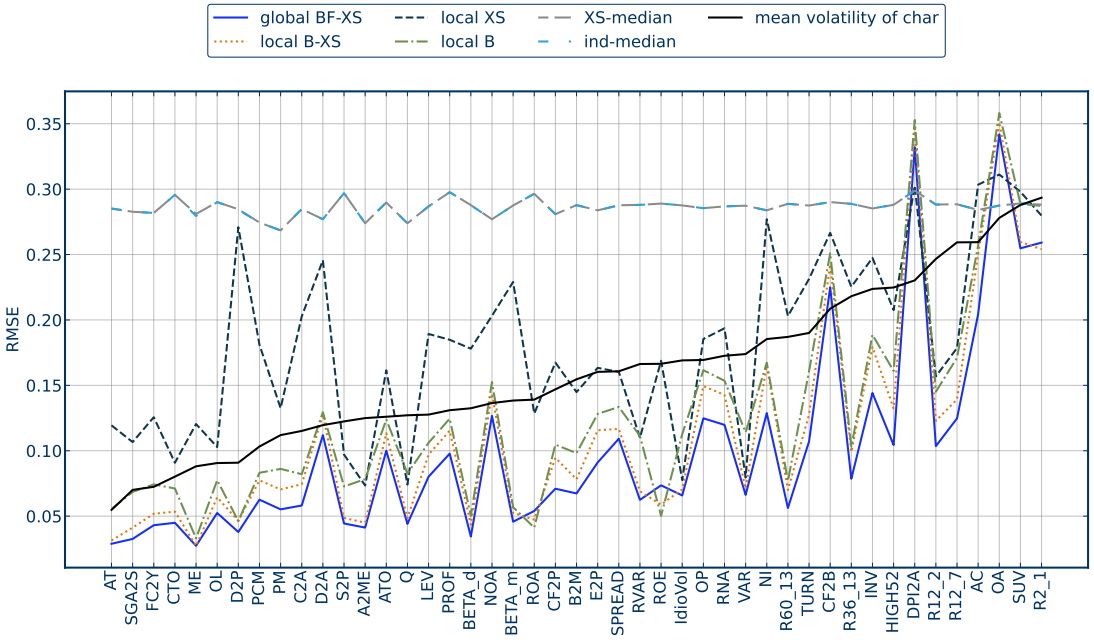
This table shows out-of-sample imputation RMSE by imputation method for each size deciles, overall and also for monthly updated and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from all observed data for the three masking schemes MCAR, Block, and Logit. We report the results for the subset of the data where a method is applicable.

Figure IA.5: Imputation Error for Individual Characteristics

Panel A: In-Sample



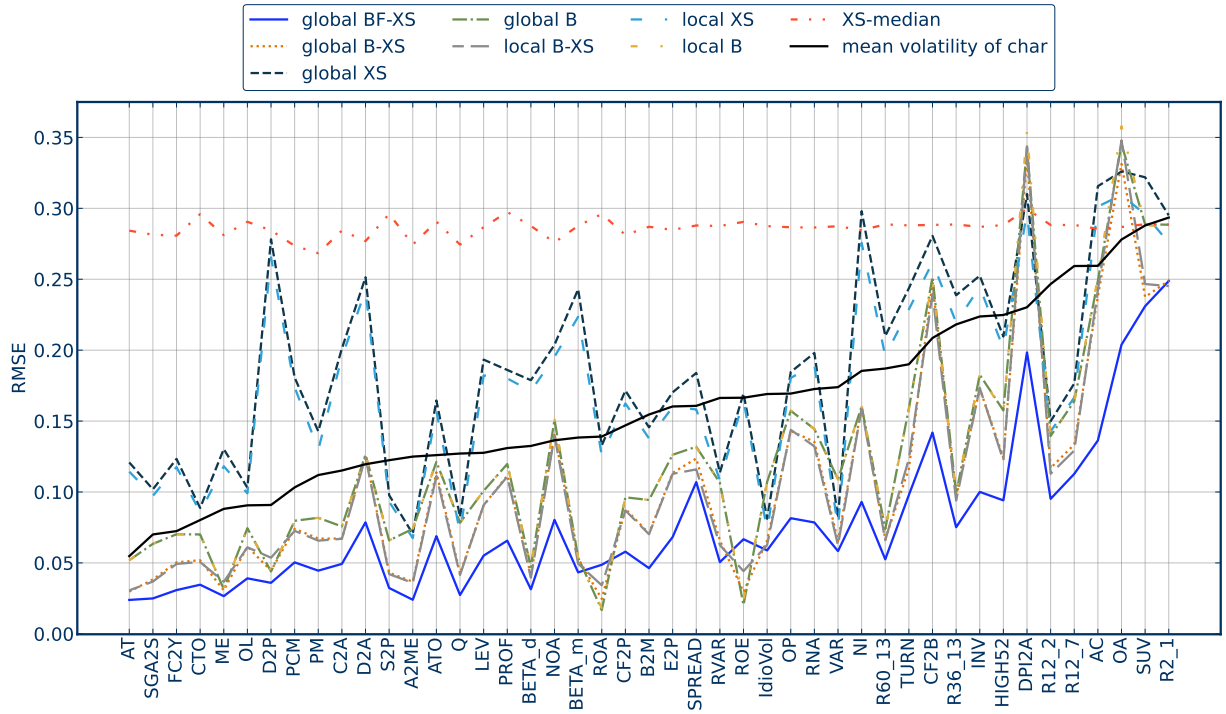
Panel B: Out-of-Sample Missing-Completely-at-Random



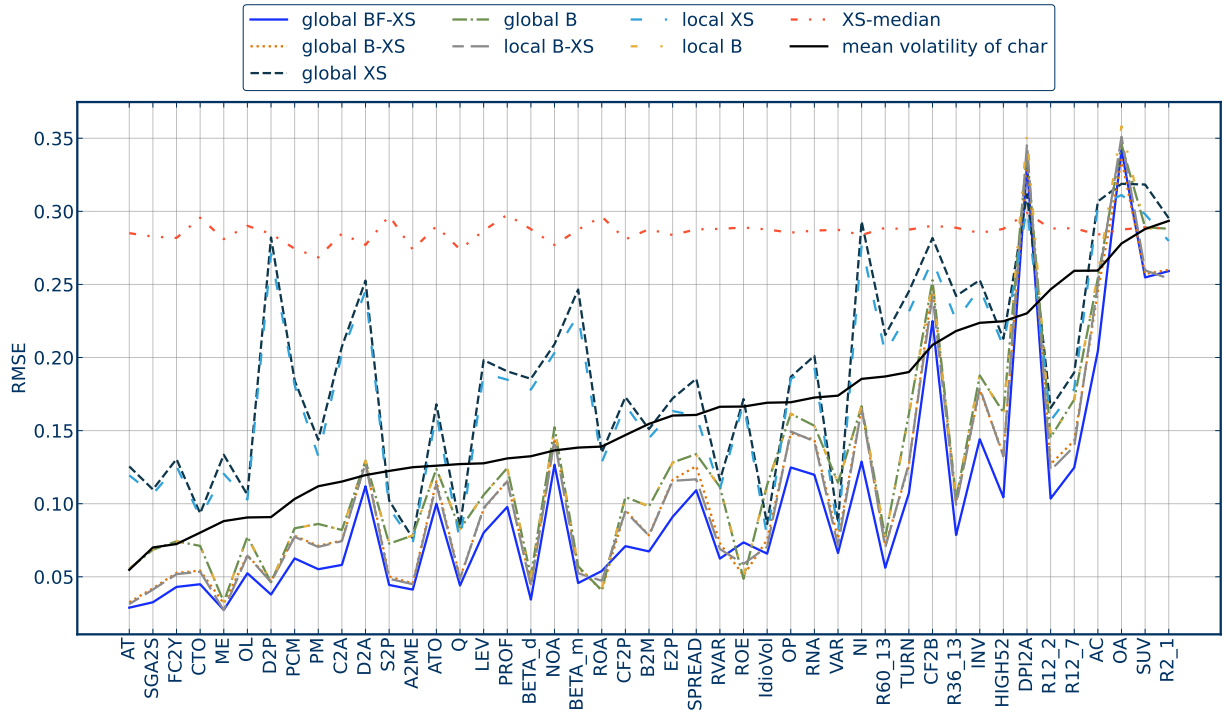
This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for MCAR-masked characteristics from all observed data. We use the fallback method as indicated in Table 3, when a method is not applicable.

Figure IA.6: Global and Local Imputation for Individual Characteristics

Panel A: In-Sample



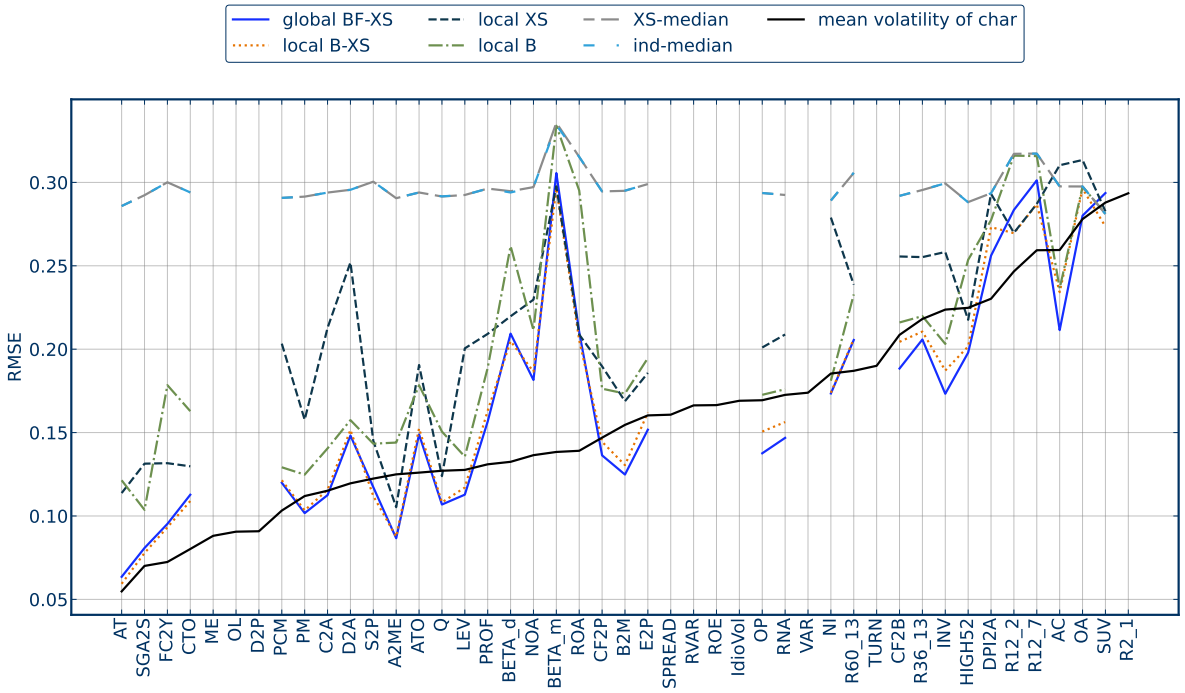
Panel B: Out-of-Sample Missing-Completely-at-Random



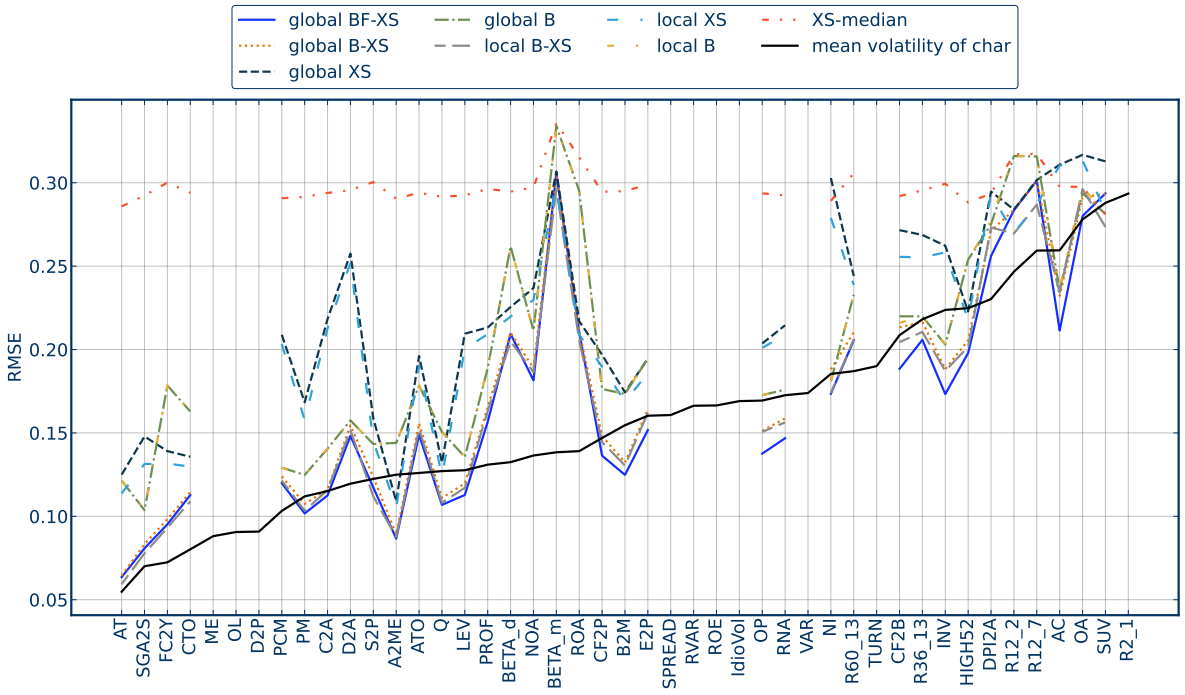
This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for MCAR-masked characteristics from all observed data. We use the fallback method as indicated in Table 3, when a method is not applicable.

Figure IA.7: Imputation Error For Individual Characteristics

Panel A: Out-of-Sample Logit

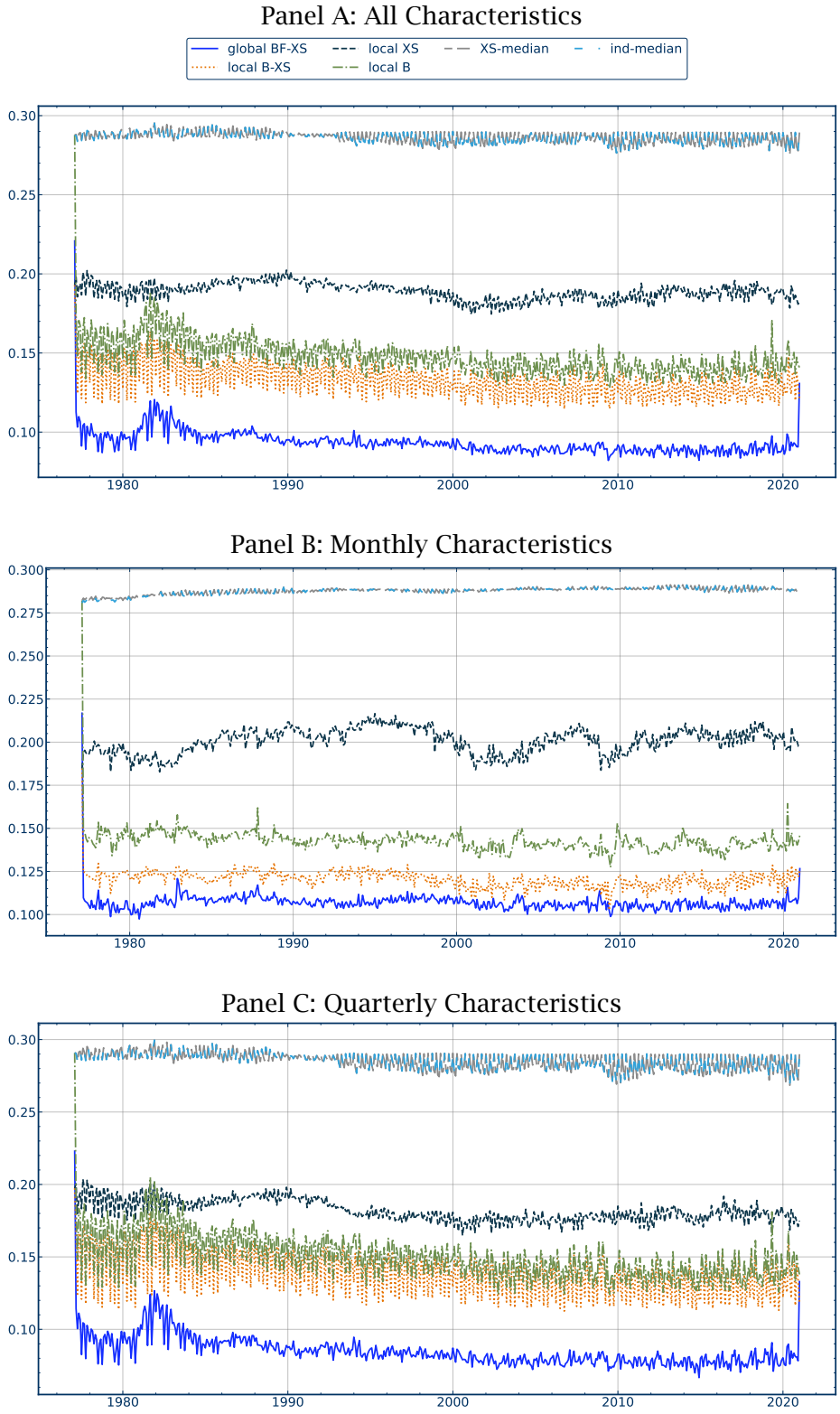


Panel B: Out-of-Sample Logit



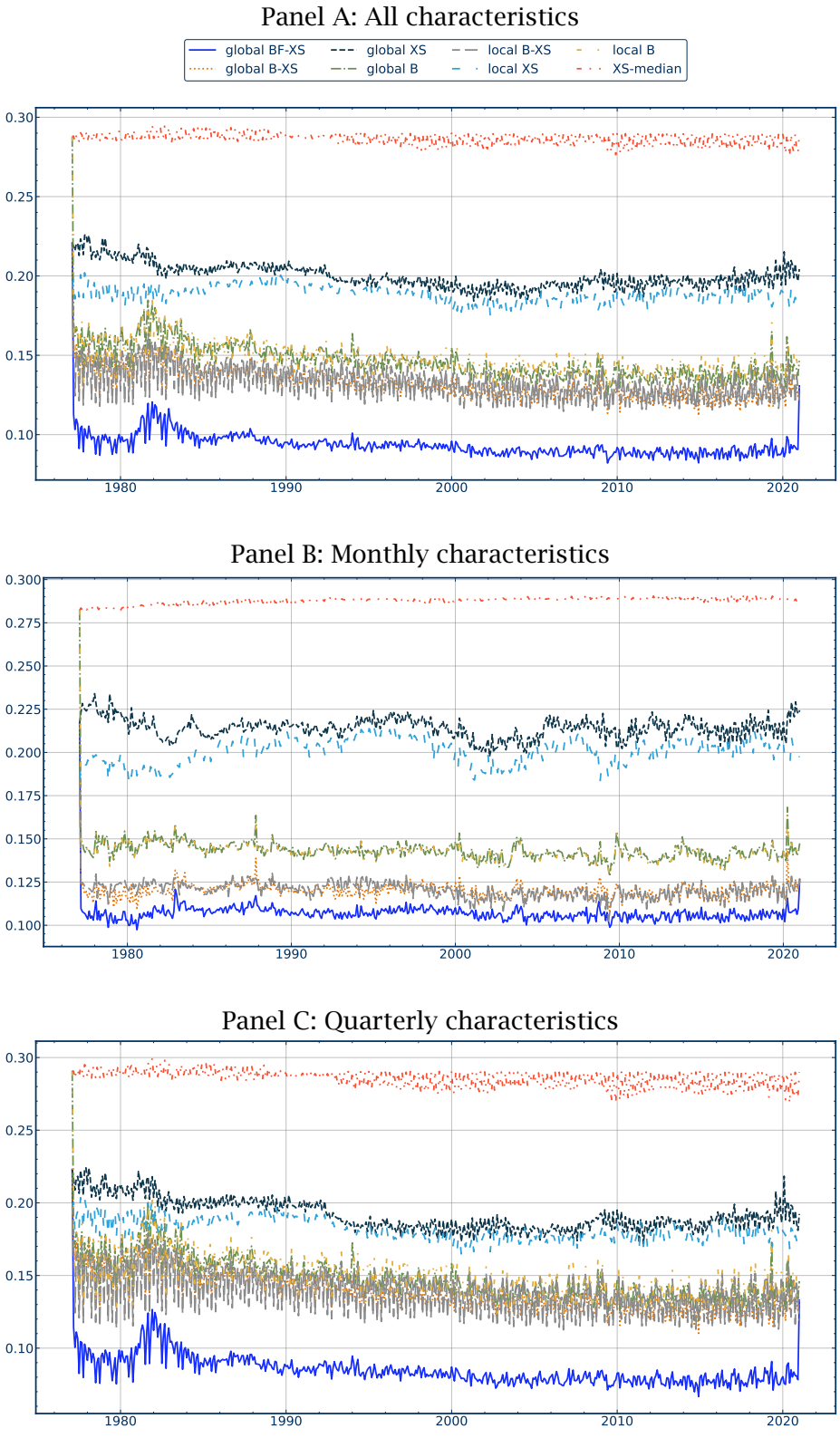
This figure shows the imputation out-of-sample RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error out-of-sample for masked characteristics from all observed data for the logit-masking scheme. We use the fallback method as indicated in Table 3, when a method is not applicable. The lines have empty entries for characteristics that are always observed.

Figure IA.8: In-Sample Imputation Error over Time



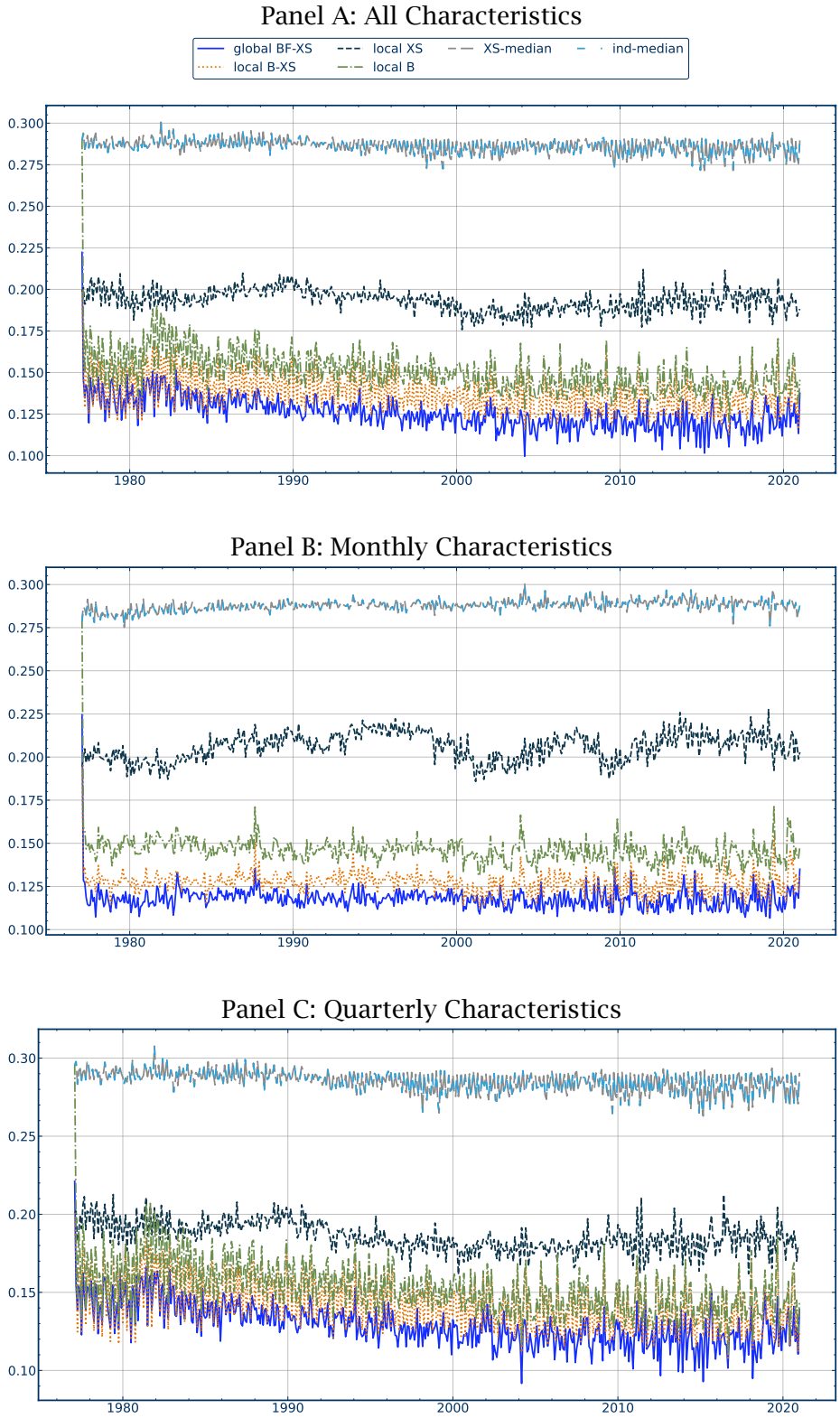
This figure shows in-sample time-series RMSE_t for different imputation methods. This is evaluated over all observed data in the sample.

Figure IA.9: In-Sample Imputation Error over Time



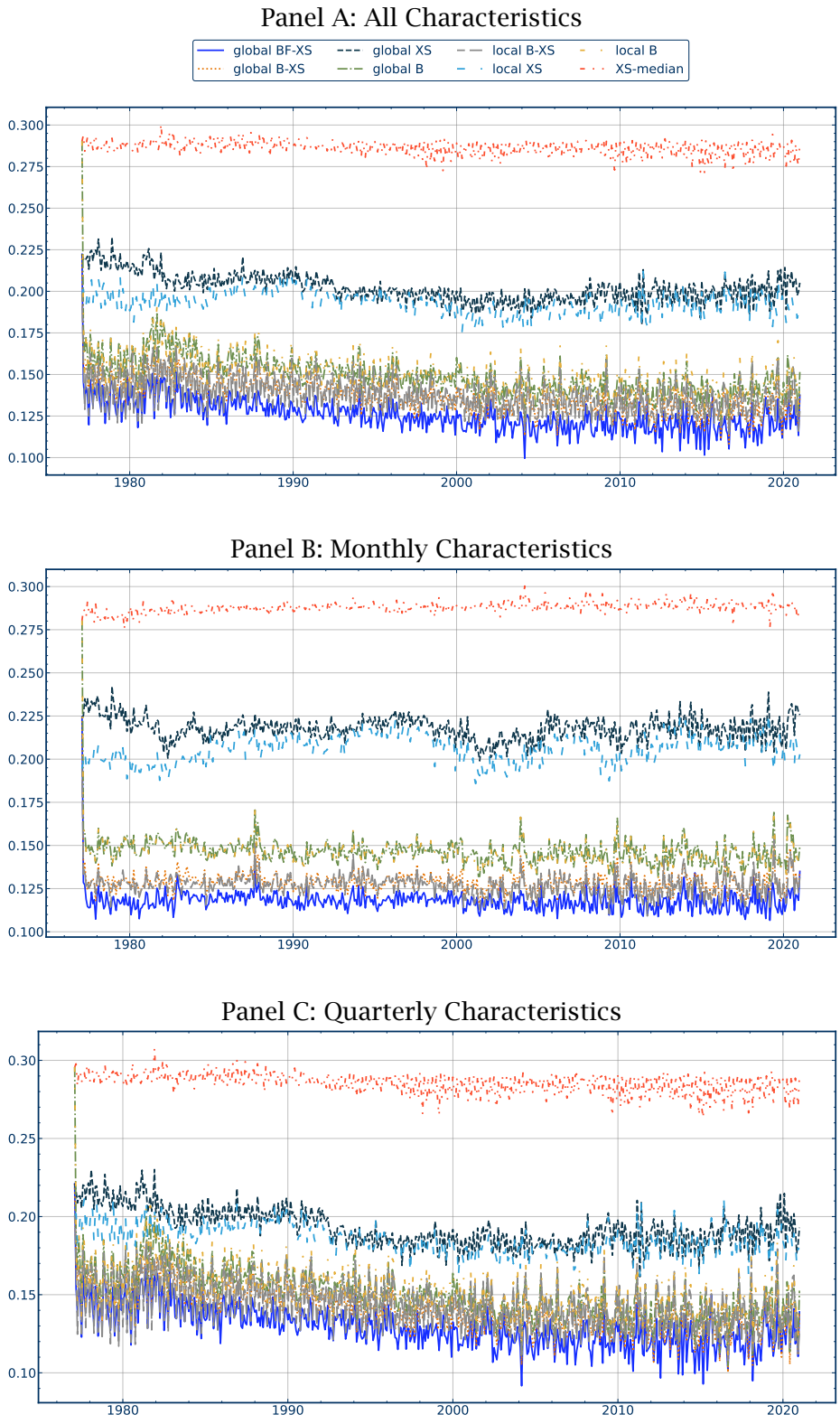
This figure shows in-sample time-series RMSE_t for different imputation methods. This is evaluated over all observed data in the sample.

Figure IA.10: Out-Of-Sample Missing-Completely-at-Random Imputation Error Over Time



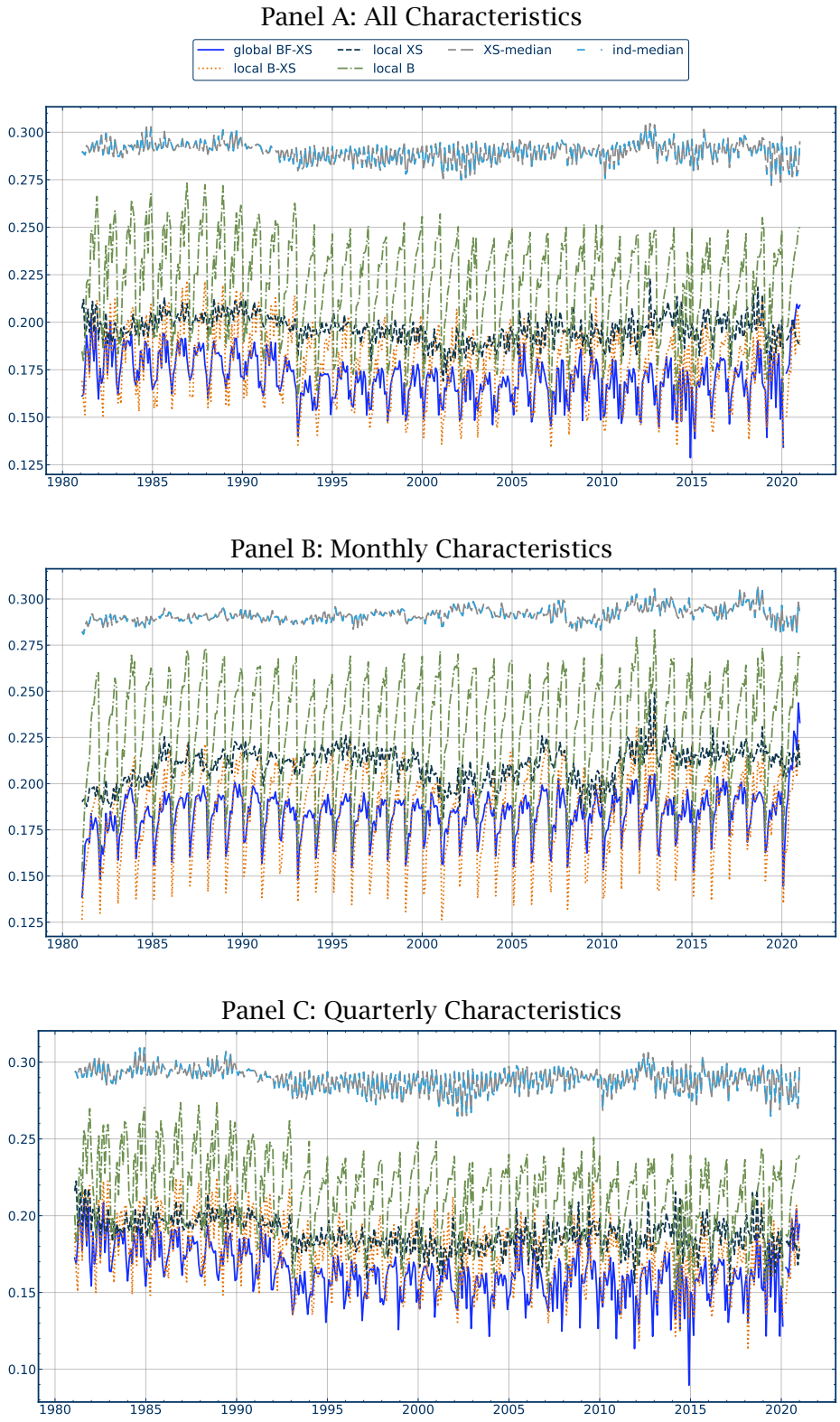
This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure IA.11: Out-of-Sample Missing-Completely-at-Random Imputation Error over Time



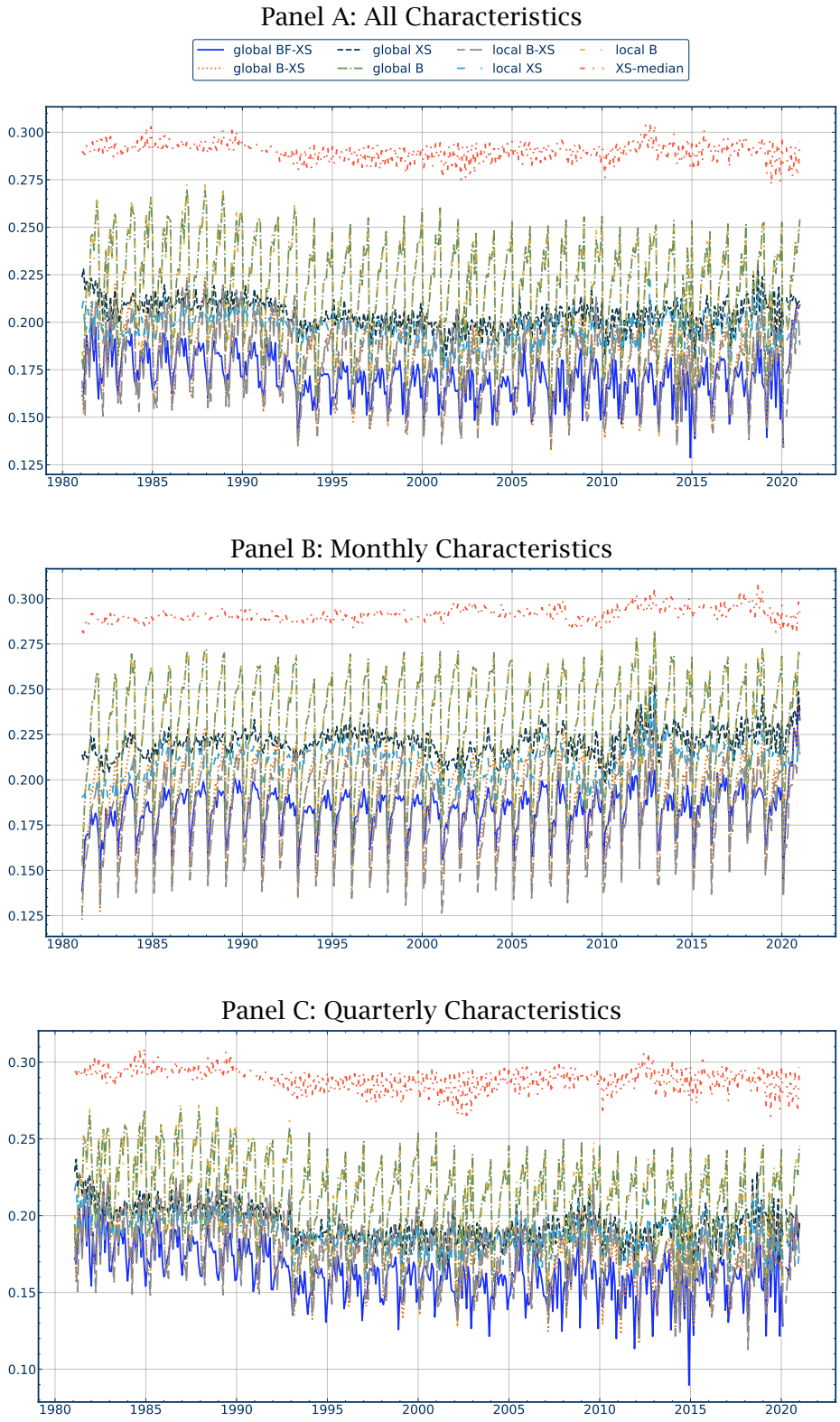
This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure IA.12: Out-of-Sample Block Missing Imputation Error over Time



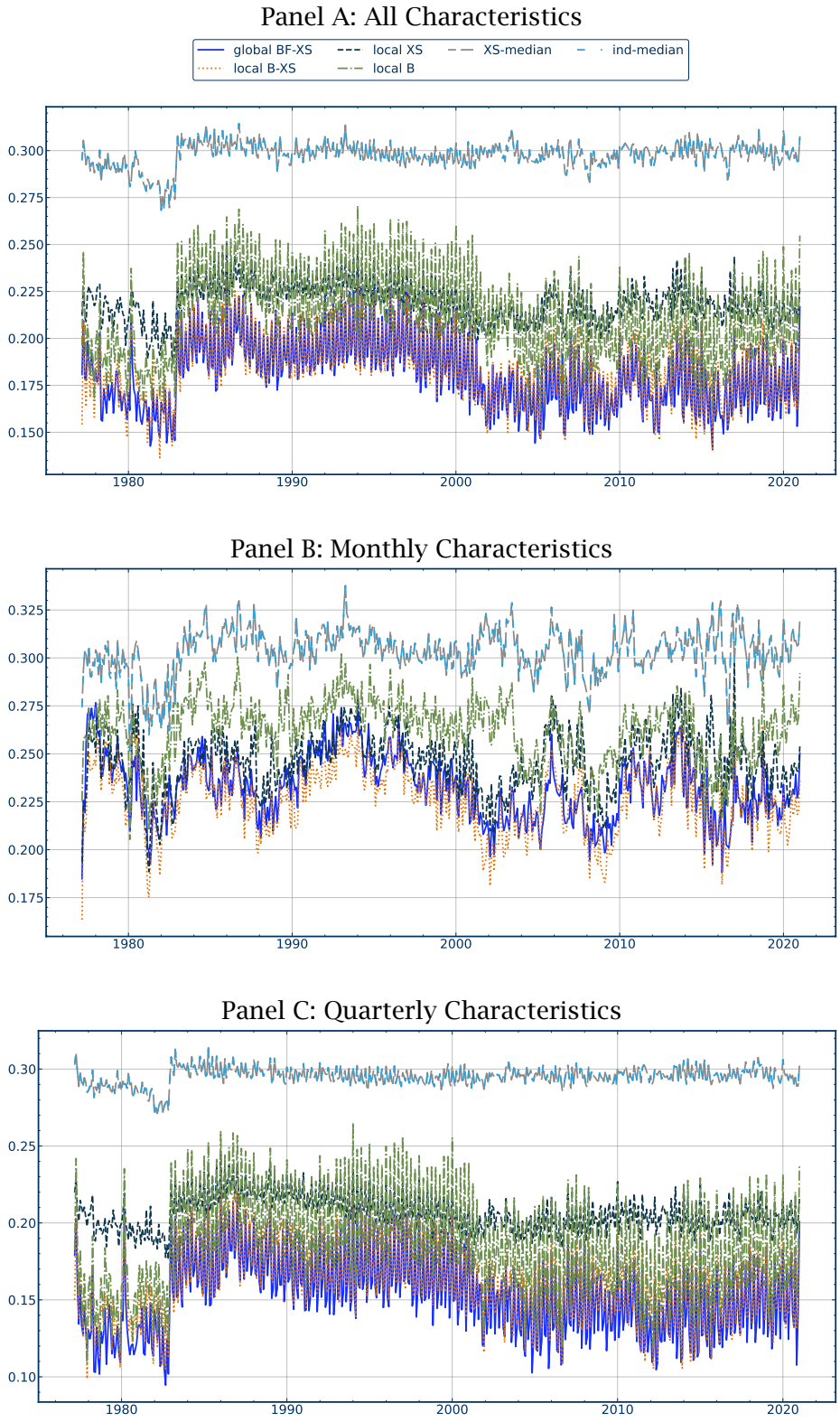
This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure IA.13: Out-of-Sample Block Missing Imputation Error over Time



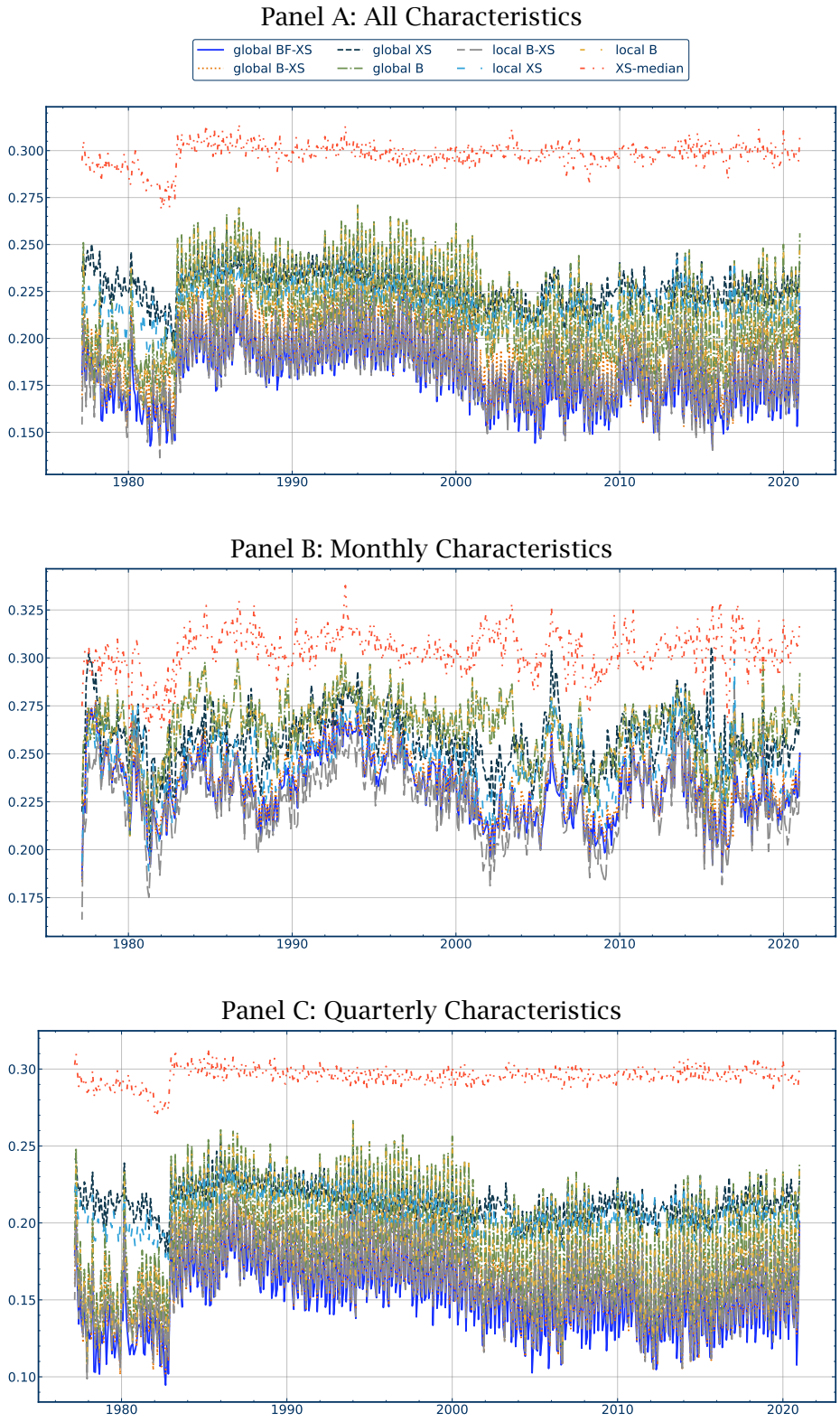
This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure IA.14: Out-of-Sample Logit Missing Imputation Error over Time



This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure IA.15: Out-of-Sample Logit Missing Imputation Error over Time

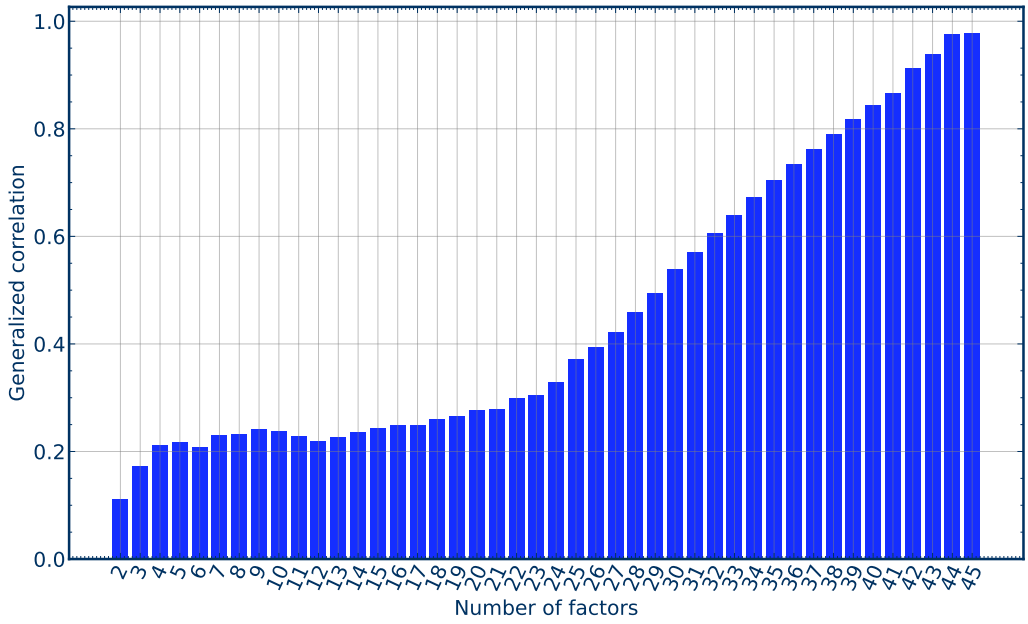


This figure shows out-of-sample time-series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

3. Asset Pricing Implications

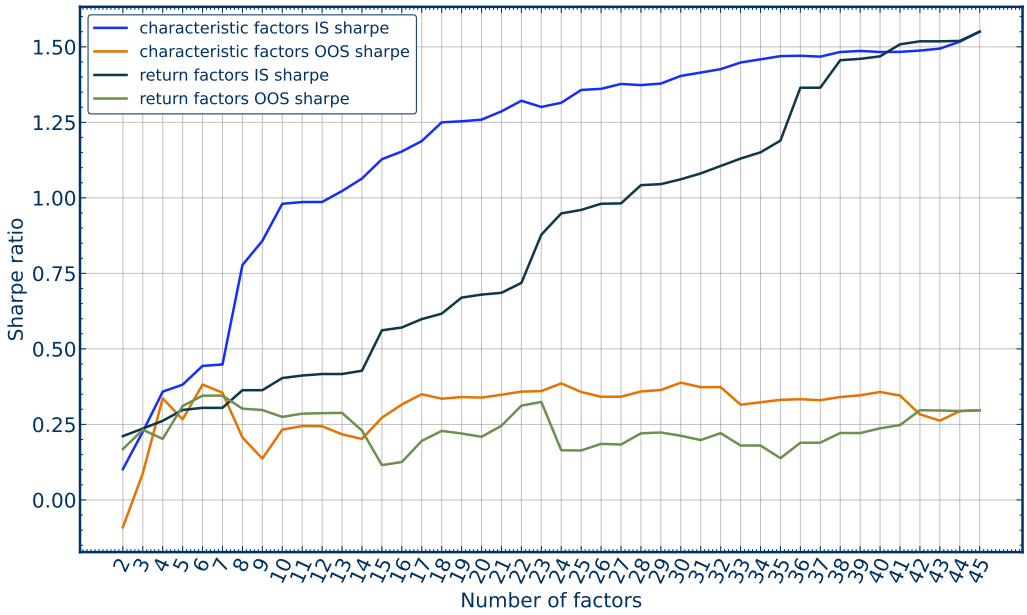
3.1. Factors in Returns vs. Characteristics

Figure IA.16: Generalized Correlations between Characteristics and Return Factors



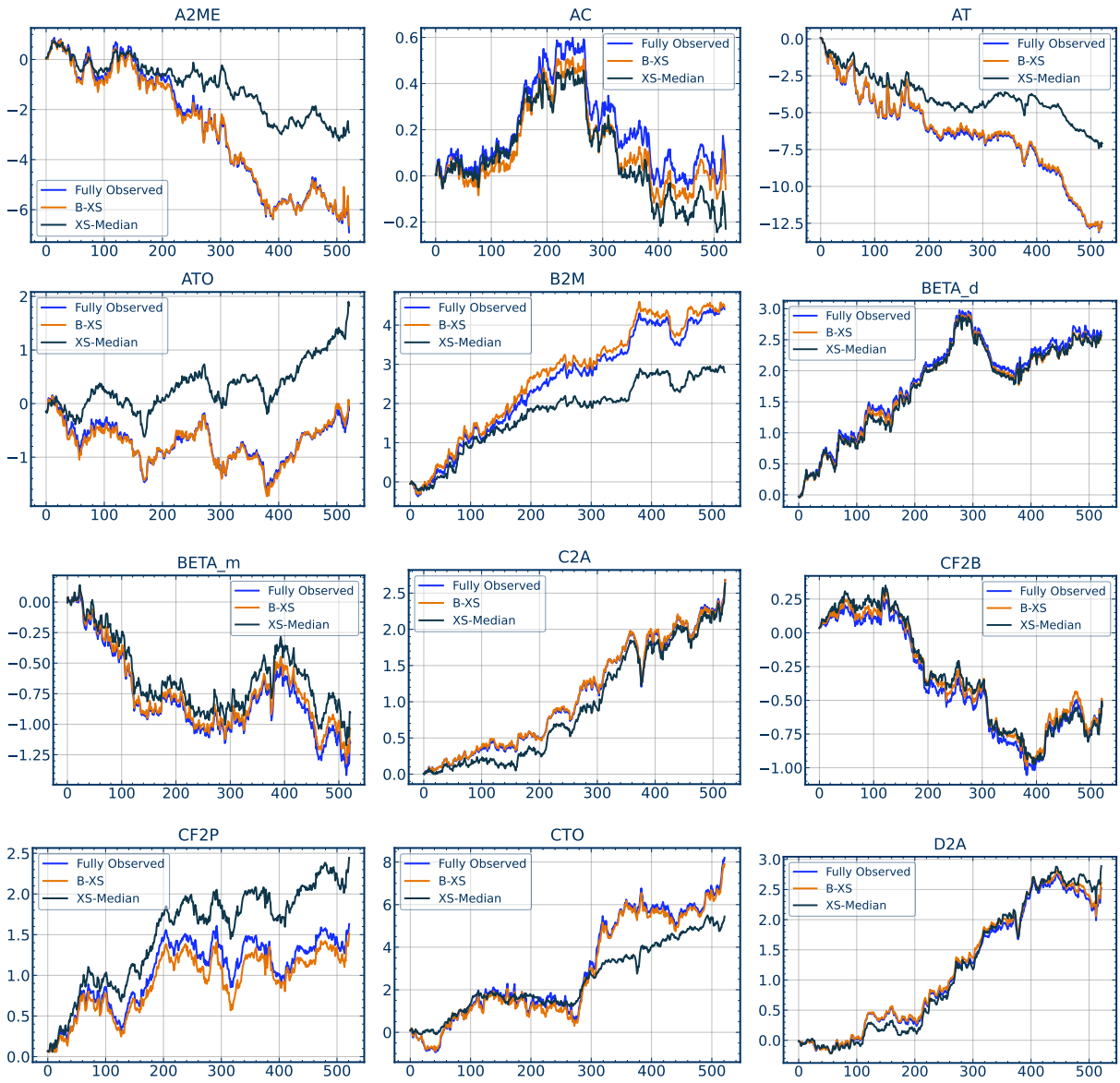
This figure shows the average generalized correlations between characteristic and return factors for different numbers of latent factors. We first extract the loadings $\hat{\Lambda}$ of our global cross-sectional factor model as the eigenvectors of the average characteristic covariance matrix $\hat{\Sigma}^{\text{XS}}$. They correspond to the portfolio weights for constructing the characteristic factors. Second, we consider the projected stock returns $R^{\text{managed}}_{t,l} = \frac{1}{N_t} R_{t,i} C^{t-1}_{i,l}$, which correspond to L managed long-short portfolios sorted on past characteristics, and can be interpreted as L univariate factor portfolios. We then apply a PCA to the return covariance matrix of the managed portfolios R^{managed}_t, which results in portfolio weights for return factors. Finally, we report the average generalized correlation between the factor weights obtained from characteristics and returns. An average generalized correlation of one implies that the same space is spanned.

Figure IA.17: Sharpe Ratios of Characteristics and Return Factors



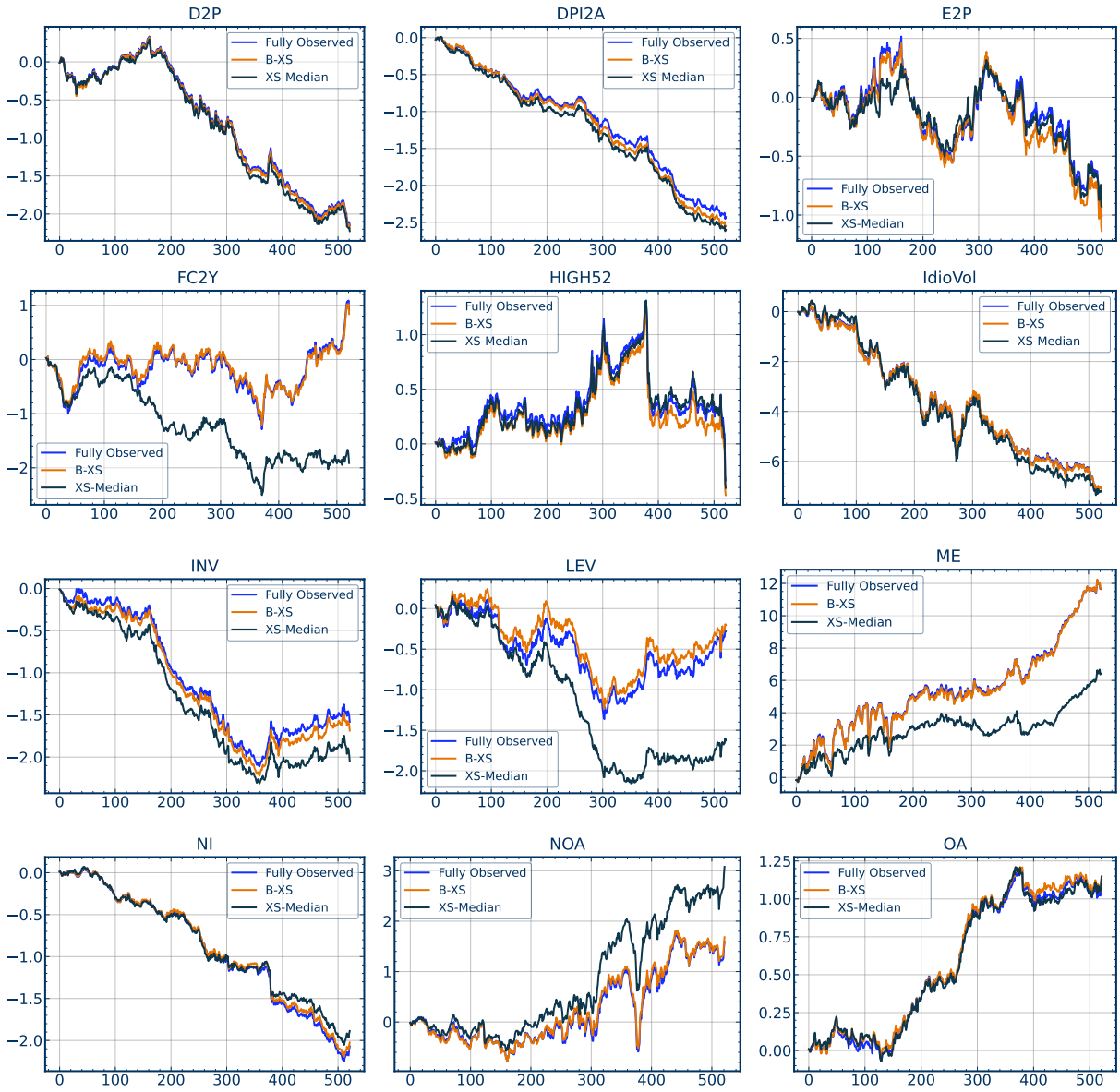
This figure shows the Sharpe ratios of mean-variance efficient portfolios based on characteristic and return factors. We first extract the loadings $\hat{\Lambda}$ of our global cross-sectional factor model as the eigenvectors of the average characteristic covariance matrix $\hat{\Sigma}^{\text{XS}}$. They correspond to the portfolio weights for constructing the characteristic factors. Second, we consider the projected stock returns $R_{t,l}^{\text{managed}} = \frac{1}{N_t} R_{t,i} C_{i,l}^{t-1}$, which correspond to L managed long-short portfolios sorted on past characteristics, and can be interpreted as L univariate factor portfolios. We then apply a PCA to the return covariance matrix of the managed portfolios R_t^{managed}, which results in portfolio weights for return factors. We apply the two sets of factor weights to the L managed portfolios R_t^{managed} and form the mean-variance efficient portfolio for different numbers of factors. We report the results in-sample and out-of-sample, where we use the first half of the data for estimation and the second half for out-of-sample evaluation.

Figure IA.18: Characteristic-Mimicking Factor Portfolios



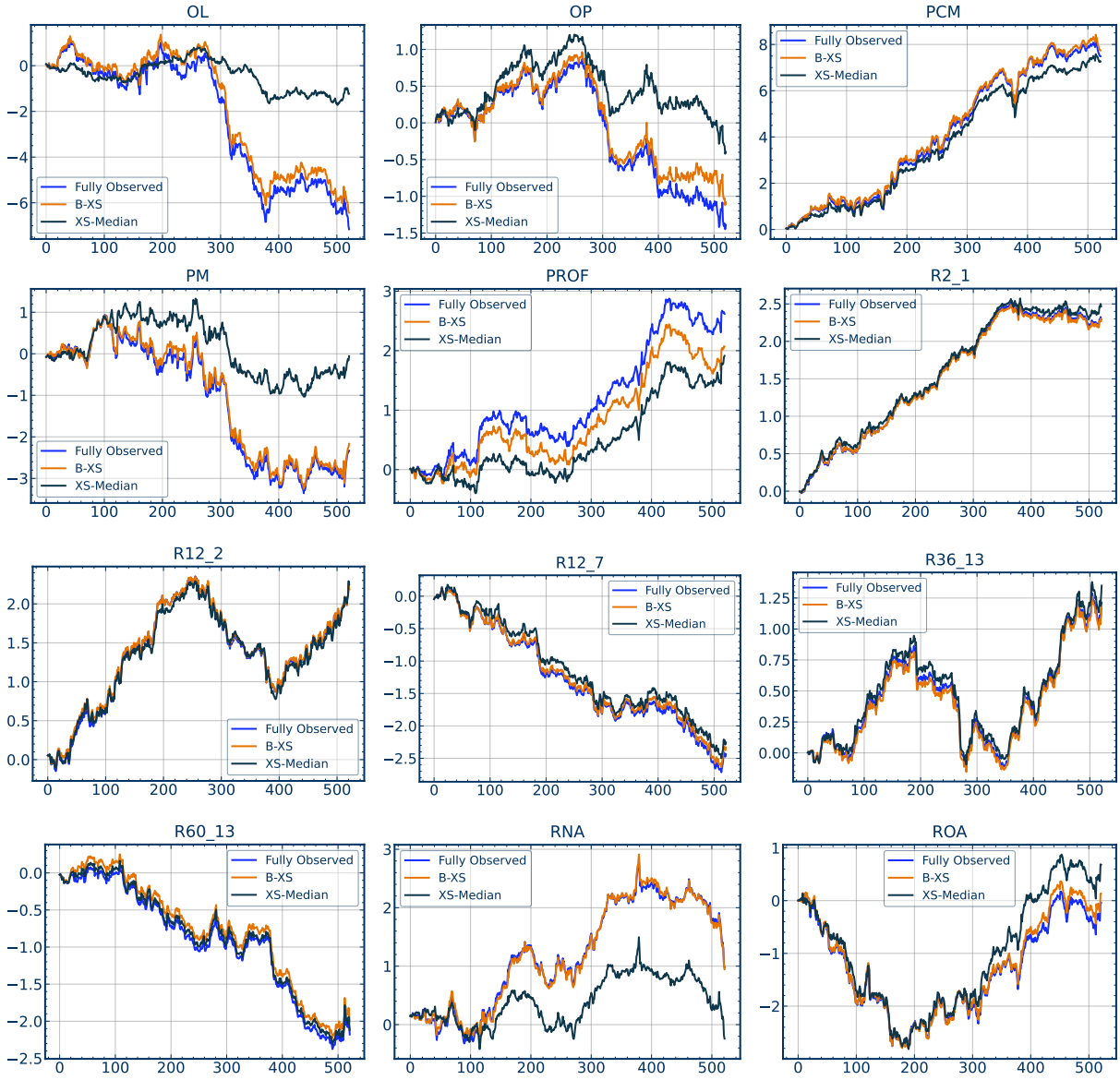
This figure shows the time-series of cumulative excess returns of characteristic-mimicking factor portfolios with and without imputation. We estimate characteristic-mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference.

Figure IA.19: Characteristic-Mimicking Factor Portfolios



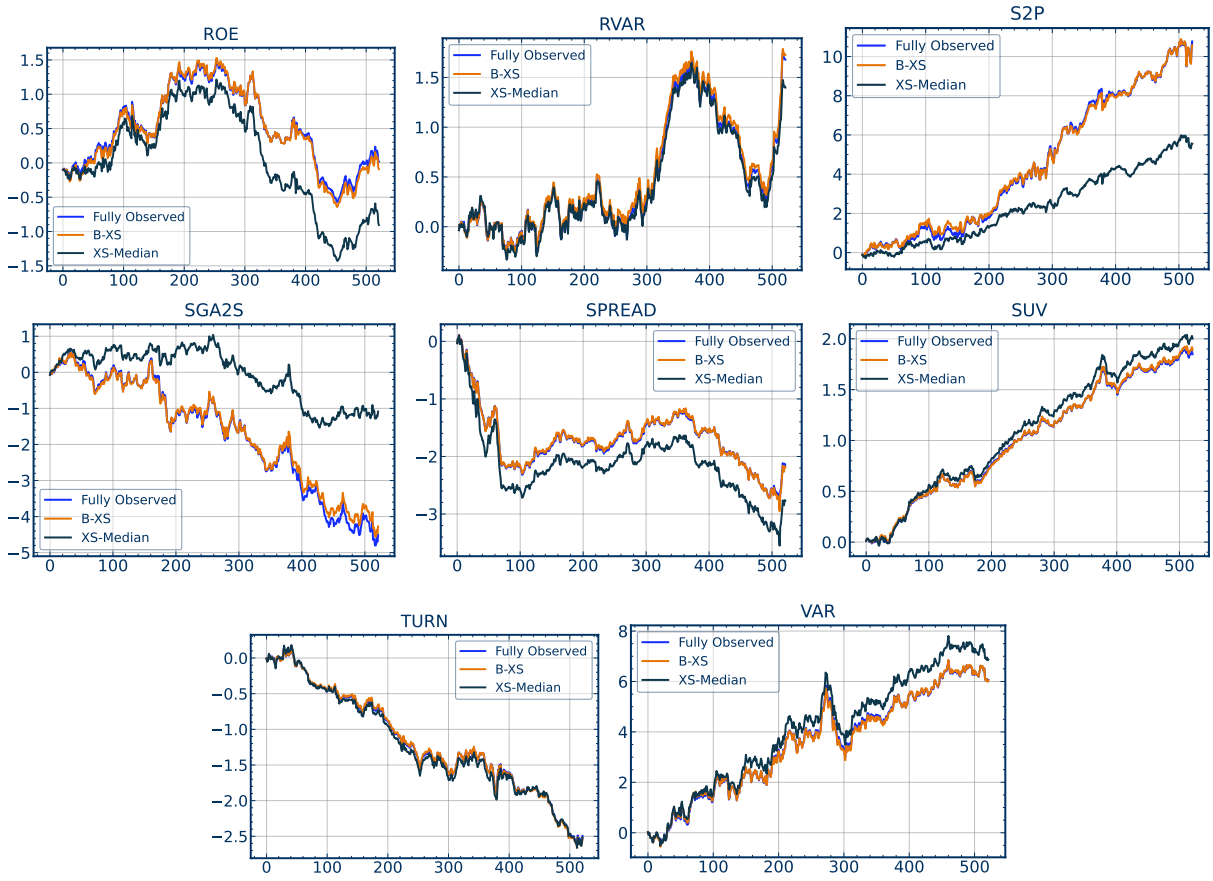
This figure shows the time-series of cumulative excess returns of characteristic-mimicking factor portfolios with and without imputation. We estimate characteristic-mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference.

Figure IA.20: Characteristic-Mimicking Factor Portfolios



This figure shows the time-series of cumulative excess returns of characteristic-mimicking factor portfolios with and without imputation. We estimate characteristic-mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference.

Figure IA.21: Characteristic-Mimicking Factor Portfolios



This figure shows the time-series of cumulative excess returns of characteristic-mimicking factor portfolios with and without imputation. We estimate characteristic-mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference.

3.3. Univariate Long-Short Factors

Table IA.4: Univariate Long-Short Decile Factors with and without Imputation

	(1) Fully Observed					(2) Obs ≥ 10					(3) Specific Char Observed					(2) + (3)					(2) - (3)				
	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%
A2ME	3.52	11.2	0.31	0.22	0.41	2.44	9.06	0.27	0.4	0.63	1.66	6.27	0.26	0.92	0.99	1.66	6.27	0.26	0.93	1.0					
AC	1.79	5.22	0.34	0.22	0.41	1.87	5.0	0.37	0.4	0.63	1.97	5.07	0.39	0.7	0.79	1.98	5.05	0.39	0.72	0.82	2.15	13.98	0.15	0.02	0.03
AT	1.41	4.51	0.31	0.22	0.41	1.35	4.34	0.31	0.4	0.63	1.38	4.24	0.32	0.92	0.99	1.38	4.24	0.32	0.93	1.0					
ATO	2.14	5.29	0.4	0.22	0.41	2.13	5.21	0.41	0.4	0.63	2.25	5.31	0.42	0.89	0.99	2.25	5.31	0.42	0.89	0.99	1.88	18.31	0.1	0.01	0.0
B2M	2.38	7.74	0.31	0.22	0.41	2.53	7.28	0.35	0.4	0.63	2.3	5.71	0.4	0.93	0.99	2.3	5.71	0.4	0.93	1.0	1.69	25.72	0.07	0.0	0.0
BETA_d	2.11	8.7	0.24	0.22	0.41	2.25	8.5	0.27	0.4	0.63	2.18	8.03	0.27	0.78	0.95	2.18	8.03	0.27	0.79	0.95					
BETA_m	2.38	9.21	0.26	0.22	0.41	2.61	8.8	0.3	0.4	0.63	2.65	8.53	0.31	0.85	0.96	2.65	8.53	0.31	0.86	0.96					
C2A	2.09	6.19	0.34	0.22	0.41	2.04	5.98	0.34	0.4	0.63	2.21	6.09	0.36	0.91	0.99	2.22	6.1	0.36	0.91	0.99	4.87	15.3	0.32	0.0	0.0
CF2B	1.79	5.25	0.34	0.22	0.41	1.89	5.02	0.38	0.4	0.63	1.87	4.71	0.4	0.92	0.99	1.88	4.71	0.4	0.92	0.99	1.04	20.62	0.05	0.0	0.0
CF2P	1.89	6.53	0.29	0.22	0.41	2.09	6.65	0.31	0.4	0.63	1.52	3.98	0.38	0.93	0.99	1.52	3.98	0.38	0.93	1.0					
CTO	1.87	4.97	0.38	0.22	0.41	1.95	4.93	0.4	0.4	0.63	2.07	4.96	0.42	0.89	0.99	2.07	4.96	0.42	0.9	0.99	1.2	18.15	0.07	0.01	0.0
D2A	1.67	5.81	0.29	0.22	0.41	1.79	5.25	0.34	0.4	0.63	1.86	5.34	0.35	0.78	0.91	1.83	5.21	0.35	0.83	0.94	1.95	7.2	0.27	0.04	0.03
D2P	2.3	5.1	0.45	0.22	0.41	2.32	4.88	0.48	0.4	0.63	2.12	4.17	0.51	1.0	1.0	2.12	4.17	0.51	1.0	1.0					
DPI2A	1.76	6.17	0.29	0.22	0.41	1.74	5.9	0.3	0.4	0.63	1.85	5.92	0.31	0.5	0.61	1.84	5.91	0.31	0.61	0.77	1.58	9.68	0.16	0.11	0.16
E2P	2.27	6.39	0.36	0.22	0.41	2.23	5.97	0.37	0.4	0.63	2.08	5.23	0.4	0.93	0.99	2.08	5.23	0.4	0.93	1.0	3.25	25.49	0.13	0.0	0.0
FC2Y	1.73	4.65	0.37	0.22	0.41	1.76	4.58	0.38	0.4	0.63	1.93	4.99	0.39	0.75	0.8	1.95	4.99	0.39	0.78	0.83	2.73	7.24	0.38	0.03	0.03
HIGH52	1.19	4.37	0.27	0.22	0.41	1.23	4.23	0.29	0.4	0.63	1.27	3.94	0.32	0.66	0.9	1.27	3.94	0.32	0.66	0.9					
INV	1.78	6.27	0.28	0.22	0.41	1.91	6.07	0.32	0.4	0.63	1.96	6.01	0.33	0.83	0.97	1.96	5.98	0.33	0.84	0.98	0.85	15.38	0.06	0.01	0.0
IdioVol	2.78	8.45	0.33	0.22	0.41	2.91	8.47	0.34	0.4	0.63	3.37	8.98	0.37	1.0	1.0	3.37	8.98	0.38	1.0	1.0					
LEV	2.07	6.06	0.34	0.22	0.41	1.91	6.0	0.32	0.4	0.63	1.75	5.74	0.31	0.86	0.95	1.75	5.72	0.31	0.88	0.97	2.15	12.19	0.18	0.02	0.02
ME	1.48	4.59	0.32	0.22	0.41	1.48	4.46	0.33	0.4	0.63	1.43	4.3	0.33	1.0	1.0	1.43	4.3	0.33	1.0	1.0					
NI	1.91	6.26	0.31	0.22	0.41	1.87	5.99	0.31	0.4	0.63	1.84	5.57	0.33	0.83	0.97	1.84	5.58	0.33	0.84	0.97	2.07	10.69	0.19	0.01	0.0
NOA	1.64	5.97	0.27	0.22	0.41	1.65	5.75	0.29	0.4	0.63	1.64	5.04	0.33	0.89	0.99	1.64	5.04	0.33	0.89	0.99	1.25	30.89	0.04	0.01	0.0
OA	1.8	5.29	0.34	0.22	0.41	1.78	5.09	0.35	0.4	0.63	1.94	5.12	0.38	0.75	0.8	1.93	5.13	0.38	0.76	0.83	1.11	23.06	0.05	0.01	0.03
OL	1.81	4.98	0.36	0.22	0.41	1.94	4.99	0.39	0.4	0.63	2.03	4.92	0.41	0.9	0.99	2.04	4.92	0.42	0.91	0.99	2.63	15.15	0.17	0.01	0.0
OP	1.88	5.28	0.36	0.22	0.41	1.87	5.12	0.36	0.4	0.63	1.97	5.04	0.39	0.83	0.93	1.97	5.04	0.39	0.84	0.93	2.19	23.16	0.09	0.0	0.0
PCM	1.63	4.98	0.33	0.22	0.41	1.61	4.76	0.34	0.4	0.63	1.7	4.81	0.35	0.89	0.99	1.7	4.81	0.35	0.91	0.99	3.96	13.19	0.3	0.01	0.0
PM	1.79	5.71	0.31	0.22	0.41	1.69	5.27	0.32	0.4	0.63	1.55	4.48	0.35	0.9	0.98	1.55	4.48	0.35	0.91	0.99	2.74	19.2	0.14	0.01	0.0
PROF	1.87	4.83	0.39	0.22	0.41	1.92	4.73	0.41	0.4	0.63	1.99	4.79	0.41	0.9	0.99	1.99	4.79	0.42	0.9	0.99	2.06	18.77	0.11	0.0	0.0
Q	1.6	5.13	0.31	0.22	0.41	1.65	4.96	0.33	0.4	0.63	1.78	5.23	0.34	0.92	0.99	1.79	5.23	0.34	0.93	1.0					
R12_2	2.34	6.46	0.36	0.22	0.41	2.41	6.24	0.39	0.4	0.63	2.59	6.34	0.41	0.92	0.98	2.59	6.34	0.41	0.92	0.98					
R12_7	2.26	6.61	0.34	0.22	0.41	2.39	6.48	0.37	0.4	0.63	2.65	6.53	0.41	0.92	0.98	2.66	6.53	0.41	0.92	0.98					
R2_1	1.85	6.17	0.3	0.22	0.41	1.92	5.93	0.32	0.4	0.63	2.07	5.94	0.35	1.0	1.0	2.07	5.94	0.35	1.0	1.0					
R36_13	1.94	6.5	0.3	0.22	0.41	2.06	6.36	0.32	0.4	0.63	2.02	6.05	0.33	0.77	0.95	2.01	6.05	0.33	0.78	0.95	0.49	14.98	0.03	0.01	0.0
R60_13	1.76	6.08	0.29	0.22	0.41	1.87	6.12	0.31	0.4	0.63	1.89	5.87	0.32	0.64	0.91	1.89	5.86	0.32	0.65	0.91	1.77	14.03	0.13	0.01	0.0
RNA	1.85	5.13	0.36	0.22	0.41	1.79	4.88	0.37	0.4	0.63	1.89	4.92	0.39	0.87	0.98	1.89	4.91	0.38	0.88	0.98	2.99	16.75	0.18	0.01	0.0
ROA	1.66	5.07	0.33	0.22	0.41	1.64	4.79	0.34	0.4	0.63	1.75	4.91	0.36	0.87	0.98	1.76	4.9	0.36	0.87	0.98					
ROE	1.7	5.15	0.33	0.22	0.41	1.63	4.76	0.34	0.4	0.63	1.75	4.84	0.36	0.87	0.98	1.76	4.84	0.36	0.88	0.98					
RVAR	2.83	9.11	0.31	0.22	0.41	2.98	8.95	0.33	0.4	0.63	3.5	9.25	0.38	1.0	1.0	3.5	9.25	0.38	1.0	1.0					
S2P	2.81	6.95	0.4	0.22	0.41	2.51	6.41	0.39	0.4	0.63	2.69	6.39	0.42	0.92	1.0	2.69	6.38	0.42	0.93	1.0	-0.06	26.34	-0.0	0.0	0.0
SGA2S	1.78	5.1	0.35	0.22	0.41	1.78	4.85	0.37	0.4	0.63	1.97	5.37	0.37	0.75	0.8	1.99	5.31	0.37	0.78	0.83	2.65	7.38	0.36	0.03	0.03
SPREAD	3.17	10.44	0.3	0.22	0.41	3.32	10.19	0.33	0.4	0.63	3.68	9.59	0.38	1.0	1.0	3.68	9.59	0.38	1.0	1.0					
SUV	1.78	5.24	0.34	0.22	0.41	1.8	5.13	0.35	0.4	0.63	1.96	5.0	0.39	0.93	1.0	1.96	5.0	0.39	0.94	1.0					
TURN	2.49	7.96	0.31	0.22	0.41	2.61	7.75	0.34	0.4	0.63	2.74	7.63	0.36	0.94	1.0	2.74	7.63	0.36	0.94	1.0					
VAR	2.77	9.32	0.3	0.22	0.41	2.92	9.2	0.32	0.4	0.63	3.35	9.39	0.36	1.0	1.0	3.35	9.39	0.36	1.0	1.0					

This table shows the mean, volatility, Sharpe ratio, value percentage and percentage of total stocks used to construct long-short decile factors. The quantile cutoffs are based on deciles of fully present NYSE data. We consider (1) fully observed data, (2) only ≥ 10 observed characteristics, where the rest is imputed, (3) specific characteristic present, other characteristics missing, (4) union of 3 and 2 (in 3 or 2), (5) not in (3) but in (2). Mean and standard deviations are reported as percentages. We use the local B-XS model to impute missing values that have prior observations available and the local XS model for the case without prior observations.

4. Data

Table IA.5: Firm Characteristics

Acronym	Name	Definition	Reference	Freq
A2ME	Assets to market cap	Total assets (AT) over market capitalization (PRC*SHROUT) as of current month	Bhandari (1988)	Q
AC	Accrual	Change in operating working capital per split-adjusted share from the fiscal year end t-2 to t-1 divided by book equity (defined in B2M) per share in t-1. Operating working capital per split-adjusted share is defined as current assets (ACTQ) minus cash and short-term investments (CHEQ) minus current liabilities (LCTQ) minus debt in current liabilities (DLCQ) minus income taxes payable (TXPQ).	Sloan (1996)	Q
AT	Total Assets	Total Assets (ATQ)	Gandhi and Lustig (2015)	Q
ATO	Net sales over lagged net operating assets	Net sales (SALEQ) over lagged net operating assets. Net operating assets are the difference between operating assets and operating liabilities (defined in NOA)	Soliman (2008)	Q
B2M	Book to Market Ratio	Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITCQ), minus preferred stock (PSTKQ). SH is shareholdersâ□□ equity (SEQQ). If missing, SH is the sum of common equity (CEQQ) and preferred stock (PSQ). If missing, SH is the difference between total assets (ATQ) and total liabilities (LTQ). The market value of equity (PRC*SHROUT) is as of the current month.	Fama and French (1992)	Q
Beta_d	CAPM Beta	Product of correlations between the excess return of stock i and the market excess return and the ratio of volatilities. We calculate volatilities from the standard deviations of daily log excess returns over a one-year horizon requiring at least 120 observations. We estimate correlations using overlapping three-day log excess returns over a five-year period requiring at least 750 non-missing observations.	Frazzini and Pedersen (2014)	M
Beta_m	Market Beta	Coefficient of the market excess return from the regression on excess returns in the past 60 months (24 months minimum)	Fama & MacBeth (1973)	M
C2A	Ratio of cash and short-term investments to total assets	Ratio of cash and short-term investments (CHEQ) to total assets (ATQ)	Palazzo (2012)	Q
CF2B	Free Cash Flow to Book Value	Cash flow to book value of equity is the ratio of net income (NIQ), depreciation and amortization (DPQ), less change in working capital (WCAPCH), and capital expenditure (CAPX) over the book-value of equity (defined in B2M)	Hou et al. (2011)	Q
CF2P	Cashflow to price	Cashflow over market capitalization (PRC*SHROUT) as of currrent month. Cashflow is defined as income before extraordinary items (IBQ) plus depreciation and amortization (DPQ) plus deferred taxes (TXDBQ).	Desai, Rajgopal & Venkatachalam (2004)	Q
CTO	Capital turnover	Ratio of net sales (SALEQ) to lagged total assets (ATQ)	Haugen and Baker (1996)	Q
D2A	Capital intensity	Ratio of depreciation and amortization (DPQ) to total assets (ATQ)	Gorodnichenko and Weber (2016)	Q
D2P	Dividend Yield	Total dividends (DIVAMT) paid from July of t-1 to June of t per dollar of equity (ME) in June of t	Litzenberger and Ramaswamy (1979)	M
DPI2A	Change in property, plants, and equipment	Changes in property, plants, and equipment (PPEGTQ) and inventory (INVTQ) over lagged total assets (ATQ)	Lyandres, Sun, and Zhang (2008)	Q
E2P	Earnings to price	The earnings used in months (t, t+1, t+2) are the earning from the quarter reported at time t (IBQ). P (actually ME) is price times shares outstanding at the end of current month.	Basu (1983)	Q
FC2Y	Fixed costs to sales	Ratio of selling, general, and administrative expenses (XSGAQ), research and development expenses (XRDQ), and advertising expenses (XADQ) to net sales (SALEQ)	D'□□Acunto, Liu, Pflueger, and Weber (2016)	Y
HIGH52	Closeness to past year high	The ratio of stock price at the end of the current calendar month and the highest daily price in the past year	George and Hwang (2004)	M
IdioVol	Idiosyncratic volatility	"Standard deviation of the residuals from a regression of excess returns on the Fama and French three-factor model"	Ang, Hodrick, Xing, and Zhang (2006)	M
INV	Investment	Change in total assets (ATQ) from the fiscal quarter ending in month t-12 to the fiscal quarter ending in t, divided by t-12 total assets	Cooper, Gulen, and Schill (2008)	Q
LEV	Leverage	Ratio of long-term debt (DLTTQ) and debt in current liabilities (DLCQ) to the sum of long-term debt, debt in current liabilities, and stockholders'□□ equity (SEQQ)	Lewellen (2015)	Q
ME	Size	Total market capitalization at the end of the current month defined as price times shares outstanding	Fama and French (1992)	M
LT_Rev	Long-term reversal	Cumulative return from 60 months before the return prediction to 13 months before	Jegadeesh and Titman (2001)	M
TURN	Turnover	Turnover is last month'□□s volume (VOL) over shares outstanding (SHROUT)	Datar, Naik, and Radcliffe (1998)	M

Continued on next page.

Acronym	Name	Definition	Reference	Freq
NI	Net Share Issues	The change in the natural log of split-adjusted shares outstanding (CSHO*AJEX) from the fiscal yearend in t-2 to the fiscal yearend in t-1	Pontiff and Woodgate (2008)	M
NOA	Net operating assets	Difference between operating assets minus operating liabilities scaled by lagged total assets (ATQ). Operating assets are total assets (ATQ) minus cash and short-term investments (CHEQ), minus investment and other advances (IVAOQ). Operating liabilities are total assets (ATQ), minus debt in current liabilities (DLCQ), minus long-term debt (DLTTQ), minus minority interest (MIBQ), minus preferred stock (PSTKQ), minus common equity (CEQQ).	Hirshleifer, Hou, Teoh, and Zhang (2004)	Q
OA	Operating accruals	Changes in non-cash working capital minus depreciation (DPQ) scaled by lagged total assets (ATQ). Non-cash working capital is defined in Accrual (AC)	Sloan (1996)	Q
OL	Operating leverage	Sum of cost of goods sold (COGSQ) and selling, general, and administrative expenses (XSGAQ) over total assets (ATQ)	Novy-Marx (2011)	Q
OP	Operating profitability	Annual revenues (REVTQ) minus cost of goods sold (COGSQ), interest expense (TIEQ), and selling, general, and administrative expenses (XSGAQ) divided by book equity (defined in B2M)	Fama and French (2015)	
PCM	Price to cost margin	Difference between net sales (SALEQ) and costs of goods sold (COGSQ) divided by net sales (SALEQ)	Bustamante and Donangelo (2016)	Q
PM	Profit margin	Operating income after depreciation (OIADPQ) over net sales (SALEQ)	Soliman (2008)	Q
PROF	Profitability	Gross profit (GP) divided by the book value of equity (defined in B2M)	Ball, Gerakos, Linnainmaa, and Nikolaev (2015)	Y
Q	Tobin's Q	"Tobinâ□□s Q is total assets (ATQ), the market value of equity (SHROUT times PRC) minus cash and short-term investments (CEQQ), minus deferred taxes (TXDBQ) scaled by total assets (ATQ)"	Kaldor (1966)	Q
R12_2	Momentum	To be included in a portfolio for month t (formed at the end of month t-1), a stock must have a price for the end of month t-13 and a good return for t-2. In addition, any missing returns from t-12 to t-3 must be -99.0, CRSP's code for a missing price. Each included stock also must have ME for the end of month t-1.	Fama and French (1996)	M
R12_7	Intermediate momentum	Cumulative return from 12 months before the return prediction to seven months before	Novy-Marx (2012)	M
R36_13	Long-term reversal	Cumulative return from 36 months before the return prediction to 13 months before	De Bondt and Thaler (1985)	M
R2_1	Short-term reversal	Lagged one-month return	Jegadeesh and Titman (1993)	M
RNA	Return on net operating assets	Ratio of operating income after depreciation (OIADPQ) to lagged net operating assets. Net operating assets are the difference between operating assets minus operating liabilities. (defined in NOA)	Soliman (2008)	Q
ROA	Return on assets	Income before extraordinary items (IBQ) to lagged total assets (ATQ)	Balakrishnan, Bartov, and Faurel (2010)	Q
ROE	Return on equity	Income before extraordinary items (IBQ) to lagged book-value of equity (defined in B2M)	Haugen and Baker (1996)	Q
RVAR	Residual Variance	Variance of the residuals from a regression of excess returns in the past two months on the CAPM model	Ang, Hodrick, Xing, and Zhang (2006)	M
S2P	Sales to price	Ratio of net sales (SALEQ) to the market capitalization (ME)	Lewellen (2015)	Q
SGA2S	Selling, general and administrative expenses to sales	Ratio of selling, general and administrative expenses (XSGAQ) to net sales (SALEQ)	Freyberger, Neuhierl, Weber (2017)	Q
SPREAD	Bid-ask spread	The average daily bid-ask spread in the current month	Chung and Zhang (2014)	M
SUV	Standard unexplained volume	Difference between actual volume and predicted volume in the current month. Predicted volume comes from a regression of daily volume on a constant and the absolute values of positive and negative returns. Unexplained volume is standardized by the standard deviation of the residuals from the regression	Garfinkel (2009)	M
VAR	Variance	Variance of daily returns in the past 60 days	Ang, Hodrick, Xing, and Zhang (2006)	M

This table summarizes the information about the 45 characteristics. We report the abbreviation, name, definition, reference and updating frequency.

Table IA.6: CRSP and Compustat Dependencies in the Construction of Characteristics

Characteristic	CRSP Dependencies		Compustat Dependencies	
	Monthly	Daily	Quarterly	Yearly
A2ME	prc, shrout		atq	
AC			actq, atq, ceqq, cheq, dlcq, lctq, ltq, pstkq, pstkq, seqq, txditcq, txpq	
AT			atq	
ATO			atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, pstkq, saleq	
B2M	prc, shrout		atq, ceqq, ltq, pstkq, pstkq, seqq, txditcq	
BETA_d	ret	ret		
BETA_m	ret			
C2A			atq, cheq	
CF2B			atq, capxy, ceqq, dpq, ltq, niq, pstkq, pstkq, seqq, txditcq, wcapchy	
CF2P	prc, shrout		dpq, ibq, txdbq	
CTO			atq, saleq	
D2A			atq, dpq	
D2P	divamt, prc, shrout			
DPI2A			atq, invtq, ppegtq	
E2P	prc, shrout		ibq	
FC2Y			saleq, xrdq, xsgaq	xad
HIGH52	prc	prc		
INV			atq	
IdioVol	ret	ret		
LEV			dlcq, dlttq, seqq	
ME	prc, shrout			
NI			ajexq, cshoq	
NOA			atq, atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, pstkq	
OA			actq, atq, cheq, dlcq, dpq, lctq, txpq	
OL			atq, cogsq, xsgaq	
OP			atq, ceqq, cogsq, ltq, pstkq, pstkq, revtq, seqq, tieq, txditcq, xsgaq	
PCM			cogsq, saleq	
PM			oiadpq, saleq	
PROF			atq, ceqq, ltq, pstkq, pstkq, seqq, txditcq	gp
Q	prc, shrout		atq, ceqq, txdbq	
R12_2	prc, prc, ret, shrout			
R12_7	ret			
R2_1	ret			
R36_13	ret			
R60_13	ret			
RNA			atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, oiadpq, pstkq	
ROA			atq, ibq	
ROE			atq, ceqq, ibq, ltq, pstkq, pstkq, seqq, txditcq	
RVAR	ret	ret		
S2P	prc, shrout		saleq	
SGA2S			saleq, xsgaq	
SPREAD	ret	askhi, bidlo		
SUV	ret	ret, vol		
TURN	shrout, vol			
VAR	ret	ret		

This table shows the CRSP and Compustat dependencies in the construction of characteristics. We report for each characteristic, which CRSP and Compustat variables are used in the construction, and the corresponding updating frequency.