An Imputed Walk Down Wall Street: Imputation Strategies and Portfolio Performance in Empirical Asset Pricing

Master's Thesis

Université Paris Dauphine - PSL

Master 2 - Research in Finance (104)

Author: Kaoutar DIAN

Supervisor: Amedeo ANDRIOLLO, Assistant Professor of Finance

June 2026

Declaration on the Use of Artificial Intelligence

This thesis was completed with the assistance of Claude, a large language model developed by Anthropic, used primarily as a productivity tool given the time constraints of completing a research project alongside a full-time internship.

- **Writing:** Claude was used to help draft and edit sections of the thesis. The author directed the content, reviewed sections, and worked closely with the thesis supervisor Amedeo Andriollo, who provided rounds of detailed feedback that shaped the final version throughout.

- **Code:** Claude assisted in writing and refining the Python scripts used in the empirical analysis. The author ran all scripts, interpreted all outputs, and made every methodological decision. Due to computational limitations, the supervisor ran the neural network component on his own machine.

- **Learning:** Claude was used as an interactive tool to consolidate understanding of the methods and literature covered in the thesis.

The research question, empirical design, and conclusions are developed under the supervision of Amedeo Andriollo.

Abstract

In this thesis, six different imputation methods for handling missing data in cross-sectional asset pricing are considered, and an evaluation of how the imputation methods impact forecasts of equity returns is presented. The dataset considered is Chen and Zimmermann (2022), which collects most of the signals commonly used in forecasting returns in the empirical asset pricing literature. In particular, 50 signals are selected during the period 2000-2021. Regarding the imputation strategies, the following methods have been compared: mean imputation, previous-value imputation, EM algorithm, XS and B-XS models by Bryzgalova et al. (2025), and conditional mean method by Freyberger, Höppner, Neuhierl, and Weber (2025).

The evaluation is conducted first by a random masking exercise, in which 10 percent of observed values are randomly hidden and the imputed values are compared to the true ones. The conditional mean method gives the least RMSE, being 45 percent lower than the average benchmark. The other two techniques that follow are the previous value and B-XS. All the methods give better results than the naive benchmark. Portfolio performance is analyzed using univariate sorting, bivariate sorting based on both missingness and characteristics, beta sorting, and long-short predictions portfolios.

We document a discrepancy between the accuracy of imputation and the performance of the investment portfolios generated. There is no universal imputation method which is the best with regard to all criteria. Factor models XS and B-XS give the greatest improvement in the PCR portfolio with respect to mean imputation (with $\Delta SR = +0.498 \; and \; +0.489$), even though they do not yield the highest accuracy of the masking test. In case of neural networks, mean imputation is the best baseline. PCR and the neural network are outperformed by the historical mean return forecasting method. One possible explanation for this is the absence of size stratification rather than a failure of imputation, though this remains a hypothesis. We also document a negative momentum premium across all methods. One possible explanation is the momentum crash of 2008 falling into the training period, though this too remains a hypothesis rather than a confirmed finding.

Keywords: *missing data, imputation, cross-sectional asset pricing, return predictability, factor models, portfolio performance*

Table of Contents

List of Tables

List of Figures

Introduction

The presence of missing data is a pervasive issue in asset pricing research. The dataset created by Chen and Zimmermann (2022), containing more than 200 firm characteristics, has numerous missing observations for each variable since few firms observe all of the predictors simultaneously. However, excluding observations with missing data, which is the typical procedure used by many researchers, is not an unbiased method, as it tends to remove small and young firms from the sample.

The issue of how to deal with the missing data problem has increasingly been the focus of attention in recent years. For example, Chen & McCoy (2024) examine various imputation methods and find that mean imputation method performs very well and is difficult to beat when measuring portfolio performance, even when compared to some other imputation methods which perform better in recovering true characteristic values. In another study, Bryzgalova, Lerner, Lettau & Pelger (2025) develop a factor based imputation framework in which they separate missing data into three categories; those occurring at the beginning, middle, and end of the firms' observation window. This distinction makes a critical difference in dealing with the missing value problem. In yet another paper, Freyberger, Höppner, Neuhierl, and Weber (2025) incorporate imputation within GMM estimation of an asset pricing model and adjust the standard errors for imputation uncertainty.

Nevertheless, some questions still need answers. How do all those methods compare to each other in a common empirical context? Will higher accuracy of imputation in characteristic space lead to superior portfolio results? And is this answer contingent on the type of the downstream model used, linear or nonlinear?

This thesis answers these questions through the comparison of six imputation methods using one and the same data and evaluation setup. The methods go from a very basic baseline method of using the mean value of the variable across all stocks to more complex factor-based imputation techniques proposed by Bryzgalova et al. (2025) and a conditional mean imputation method inspired by FHNW (2025). All methods use the open-source data provided by Chen and Zimmermann (2022), limited to 50 signals by observability for the 2000-2021 sample period.

The evaluation consists of two aspects. The first aspect concerns the accuracy of the imputation which is assessed using a randomized masking process where 10% of observed values are randomly masked, and then the imputed values are compared to the actual values. This first dimension is purely statistical, measuring how well each method recovers true characteristic values in the characteristic space. The second aspect is financial: it evaluates whether imputation choices translate into better investment performance, assessed through univariate sorting, bivariate sorting based on both missingness and characteristics, beta sorting, and long-short portfolios.

The results of the research include the following. In terms of imputation accuracy, the conditional mean approach leads, while previous value imputation and the backward cross-sectional factor model are the second and the third best approaches respectively. All approaches outperform the naive mean imputation baseline when the cross-sectional structure of the data is taken into account properly. With respect to portfolio performance, results are more complex. The factor approaches XS and B-XS give the highest improvement over mean imputation in PCR portfolios, showing ΔSR of +0.498 and +0.489 respectively, even though these approaches are not the most accurate in the masking experiment. As far as neural networks are concerned, mean imputation gives the best baseline; more sophisticated imputation makes the results worse. We document a negative momentum premium observed across all methods. This is consistent with the 2008 financial crisis momentum crash falling within the training window.

The main conclusion is that there is no relationship between the success of the imputation methods and their effect on portfolio performance. The importance of the imputation methods is confirmed; however, which imputation technique is optimal will depend on the model used and the criterion of assessment.

The organization of the remainder of the dissertation is as follows. Chapter 2 explains the dataset used and outlines the patterns of missing data. Chapter 3 provides details about the six different imputation techniques used, the prediction models employed, and the criteria for evaluating portfolio performance. Chapter 4 reports the empirical results. Chapter 5 concludes.

Chapter 1: Literature Review

The literature on empirical asset pricing has undergone a fundamental transformation in the last several decades. The initial literature, stemming from Fama and French (1992) and extended in Fama and French (2015), typically employed a small number of firm characteristics to explain the cross section of asset returns, which were selected based on economic intuition and interpreted in a specific theoretical framework. In this context, these characteristics serve as proxies for the underlying state variables that determine the stochastic discount factor (SDF). However, in more recent years, a growing list of return predictors has been identified, which has led to a proliferation of potential predictors. This has led to the "dimensionality challenge," as described in Cochrane (2011), which refers to the problem of identifying which predictors contain incremental information about expected returns when the total number of characteristics becomes very large.

In response to the dimensionality challenge, the literature has increasingly turned towards a data mining approach, in which the objective is to distill relevant information from a broad range of characteristics. The adoption of machine learning techniques is a logical extension of this trend, as it is specifically designed to work in high-dimensional spaces. In particular, Gu, Kelly, and Xiu (2020) demonstrate the benefits of using a broader range of firm characteristics, which underlines the importance of being able to effectively summarize information in high-dimensional settings. Specifically, they show that machine learning methods improve out-of-sample predictive performance and that these methods are particularly effective at capturing nonlinear relationships in the data, rather than necessarily delivering higher economic value in terms of portfolio returns (Gu, Kelly, and Xiu, 2020).

However, this achievement also points to an essential limitation. To be specific, firm characteristics are frequently incomplete, and missing data are highly pervasive in large panel data sets. While early studies could largely ignore this issue because of the relatively small number of variables under consideration, this is clearly not feasible here, given the high-dimensional setting. In particular, discarding incomplete observations will result in a significant loss of information. This highlights a fundamental trade-off between working with a complete but limited dataset and exploiting a high-dimensional dataset at the cost of dealing with missing observations. Recent contributions, such as Freyberger, Höppner, Neuhierl, and Weber (2025) (hereafter FHNW) and Bryzgalova et al. (2025) (hereafter BLLP), propose methods to handle missing data in this context.

In response to the increasing availability of predictors, a wave of studies collected large panels of firm characteristics and applied sophisticated statistical methods to extract the signals for predicting excess returns. For example, Freyberger, Neuhierl, and Weber (2020) combine 62 predictors, while Kelly, Malamud, and Pedersen (2023) consider 138 predictors, and Han, He, Rapach, and Zhou (2022) extend the set to as many as 193 predictors. Overall, these studies show that expanding the set of predictors beyond Fama and French (2015) leads to economically

meaningful improvements in forecasting performance. When handling missing data, a common challenge in such high-dimensional panels, a common practice in the literature is to focus on the set of observations where all the characteristics are available, which is known as complete case analysis. The strategy is to remove all the observations with at least one missing value in the characteristics. For example, in the typical datasets employed in empirical asset pricing research, the set of complete case analysis may cover a limited proportion of the total available data, which is extremely limited (FHNW, 2025; Harvey, Liu, and Zhu, 2016).

In this context, recent contributions have explicitly focused on how to handle missing information in empirical asset pricing. The three most recent contributions, authored by Chen and McCoy (2024) (hereafter CM), BLLP, and FHNW (2025), are representative of the current state-of-the-art literature on handling missing data in high-dimensional asset pricing panels, as they approach this challenge from complementary perspectives. On one hand, CM assess the impact of imputation decisions using textbook solutions such as the Expectation-Maximization (EM) algorithm on portfolio outcomes and forecasting performance using different machine learning techniques. Conversely, BLLP propose a latent factor imputation framework that leverages cross-sectional and time-series dependencies inherent in firm characteristics and systematically characterizes the structure of missing data. In contrast, FHNW develop a Generalized Method of Moments (GMM) framework that integrates conditional mean imputation directly within the estimation procedure in order to account for the uncertainty introduced by imputed values and ensure valid statistical inference.

When taken as a whole, these contributions shift the attention from viewing missing data as a preprocessing issue to treating it as a methodological problem with quantifiable implications for model selection, inference, and portfolio construction.

1.1. The missing Data Problem

1.1.1. How Pervasive Is the Problem?

The complete case analysis, which only uses observations for which all characteristics are available at once, is the usual approach to dealing with missing data in large panels. However, since it discards all observations that contain at least one missing value, this approach leads to a very limited analysis in a high-dimensional setting. A key challenge is that expanding the set of predictors increases the prevalence of missing observations, which may restrict the use of certain empirical methods or lead to a loss of information. Moreover, this is not only a statistical problem. Given that various firm characteristics are not proxies for the same underlying phenomenon because they measure different dimensions of economic risk, this implies a financial and economic cost. Importantly, the structure of missingness itself can be informative, as it may reflect underlying economic mechanisms or data-generating processes rather than purely random omissions.

This problem is described in detail in the study by BLLP (2025): while missing data are often handled using simple approaches like historical means, medians, or zero imputation, these solution may be myopic as they ignore the underlying structure of the data. The study uses CRSP and Compustat data on more than 45 firm characteristics, including the aforementioned accounting characteristics, updated on a quarterly basis, covering more than 22,000 stocks from January 1967 to December 2020. While these characteristics are observed at a quarterly frequency, the key insight is that they exhibit strong persistence over time, which allows the model to exploit the time-series structure of the data. This persistence, rather than the sampling frequency itself, is what enables their imputation strategy to improve the imputation of missing values. In their data, over 40 percent of their quarter-level accounting data is missing before 1982, and even at the end of their sample, their measures of profitability, investment, and leverage remain missing for 8–14 percent of firms each month. More than 70 percent of firms and half of total market capitalization are missing if all 45 characteristics must be observed. These findings are also in line with those in the study by Koh and Reeb (2015), which finds R&D information is missing between 1980 and 2006 for 42 percent of their sample.

CM draw similar conclusions for their analysis of 159 predictors from the Chen and Zimmermann (2022) dataset, which covers a range of price-based, accounting-based, and analyst forecast variables. The result of combining 75 or more predictors with complete-case analysis is that all but 16 percent of the cross-section are eliminated, with the result of combining 125 predictors being essentially zero observations left. In contrast to BLLP, CM argue that the time-series dimension of the dataset is of limited use for imputation. This is because missingness is often highly persistent over time, leaving little historical variation to exploit, and because simple cross-sectional imputation methods are already able to capture most of the relevant information. In addition, they show that time-series-based approaches, such as autoregressive imputation, do not provide substantial improvements over these simpler methods.

FHNW (2025) verify this observation using the same data source as CM, namely Chen and Zimmermann (2022), by including 82 of the 159 characteristics from 1978 to 2021, covering approximately 2.4 million observations. The 82 characteristics are selected as a subset of the full dataset in order to retain a non-trivial complete case sample while preserving the most standard predictors used in the literature. Still, the complete case set makes up only 10 percent of the dataset. Meanwhile, almost half of all observations have at most five characteristics missing, meaning that these partially observed data points remain close to complete cases. Throwing these data points away goes against a fundamental statistical principle emphasized by FHNW and rooted in Zhang, Mykland, and Aït-Sahalia (2005), that one should not discard informative observations, as this leads to a loss of information and may distort the representation of the return distribution. Moreover, FHNW show that observations with missing characteristics tend to exhibit more dispersed realized returns, underscoring the implications of ignoring the missingness in terms of risk. Missing data tends not to be at random and rather reflect underlying economic structure, as

certain types of firms or characteristics are more likely to be unobserved, an aspect explicitly modeled by BLLP (2025).

1.1.2. The Structure of Missingness

A significant insight from these three papers is that the absence of firm characteristics contradicts the missing completely at random (MCAR) assumption. It is instead structured, persistent, and, in the case of BLLP, endogenous to firm fundamentals.

At a broad level, CM clearly state three stylized facts about missingness in cross-sectional predictor data. First, missingness is highly persistent: 81% of stocks that are currently missing book-to-market, a key valuation predictor in asset pricing, have been missing it in every prior month, indicating that missingness tends to persist over time rather than being transient. Second, only 40% of the total predictor variance can be explained by the first ten principal components due to the modest cross-sectional correlations between predictors; nearly all pairwise correlations fall between −0.25 and +0.25, in line with the previous findings of Green, Hand, and Zhang (2013) and Chen and Zimmermann (2022). This implies that there is little cross-sectional information available to impute missing predictors. Third, missingness clusters by firm type. For instance, when a stock lacks one accounting predictor, it usually lacks all accounting predictors at the same time, making it difficult to impute by using the correlation between characteristics that are close in terms of economic content.

A complementary analysis on the missingness in the Chen and Zimmermann dataset is provided by FHNW (2025). Instead of concentrating on the cross-sectional and time-series organization of missingness, they show that missingness varies across firms. First, they study missingness in terms of firm size: the fully observed subsample that complete-case analysis retains tends to include relatively larger and more established firms, as smaller firms are more likely to have missing characteristics. Additionally, they demonstrate that missingness is not an all-or-nothing phenomenon: in their 82-characteristics panel indicating that the majority of partially observed firms are extremely close to being complete. This feature is important because it implies that partially observed observations still contain substantial information and can be incorporated into the estimation procedure through appropriate weighting rather than being discarded entirely. This motivates the trade-off in their approach: their approach trades some flexibility for a unified estimation framework that jointly handles imputation and inference. By accounting for the additional uncertainty introduced by imputation and down-weighting noisier observations, their method improves efficiency while maintaining valid statistical inference. Formally, they characterize the missing data mechanism by allowing the probability of missingness to depend arbitrarily on always-observed characteristics such as firm size but not on the missing values themselves, a condition they treat as empirically plausible for most firm characteristics in their dataset and which appears reasonable in this context, as missingness is likely to depend on observable firm attributes such as size or reporting behavior rather than on the unobserved values themselves.

By categorizing missingness based on the timing of the missing information within a firm's history, BLLP aprovide more exhaustive characterization of the structure of missing observations. These groups are defined according to when missingness occurs in a firm's history: missing at the beginning (MAB), missing in the middle (MAM), and missing at the end (MAE). MAB corresponds to firms for which characteristics are missing at the beginning of the sample, typically due to insufficient historical data required to construct them. The most relevant case is MAM, where a characteristic is observed, becomes missing for a period, and then reappears. While MAB can largely be attributed to firms entering the sample, MAM is more likely to reflect firm-specific reporting behavior or economic conditions and therefore represents a more structural form of missingness. BLLP also show that missingness is persistent over time. They provide evidence for this using logistic regression analysis, which tests the likelihood that missing observations remain missing across periods. Importantly, this decomposition has implications for the choice of the imputation method: for example, for imputing MAM, one can leverage the B-XS imputation model, which combines cross-sectional and time-series information, given the persistence of the data and the availability of both past and future observations. This advantage is unavailable if data are missing at the beginning.

These different structures of missingness directly influence the selection of the imputation approach.

1.2. Approaches for Handling Missing Firm Characteristics

This section reviews the main approaches used in the literature to handle missing firm characteristics in empirical asset pricing. Rather than treating missing data as a simple preprocessing issue, recent contributions recognize that it can have important implications for both prediction and inference, and propose methods to address it accordingly. The literature can be broadly organized into three complementary approaches. The first approach focuses on prediction and provides justification, for using standard imputation methods. The second approach models firm characteristics using latent factor in order to recover missing observations by exploiting the common variation across characteristics. The third approach treats missing data as joint problem of imputation and prediction and develops estimation procedures that explicitly account for the uncertainty introduced by missingness.

These three approaches structure the remainder of this section. The prediction-based approach is presented in Section 1.2.1, with an emphasis on CM, who assess how common imputation techniques affect forecasting performance and portfolio results. The latent factor approach, which exploits the joint time-series and cross-sectional structure of firm characteristics, is presented in Section 1.2.2 and is developed by BLLP. The econometric approach proposed by FHNW is covered in Section 1.2.3; it integrates estimation and imputation within a GMM framework to ensure valid statistical inference.

1.2.1. Simple Imputation and Machine Learning Approaches

The initial strand of the literature approaches missing firm characteristics primarily from the perspective of prediction. In this setting, the main question is whether different imputation methods affect the accuracy of out-of-sample return predictions and the resulting portfolio performance. This viewpoint is particularly relevant to empirical asset pricing in high dimensions, where the objective is to construct portfolios based on predicted returns rather than to interpret individual coefficients.

In this literature, standard practice involves simple cross-sectional imputation methods such as mean or median substitution. As CM point out, many studies combine large sets of predictors imputed using these approaches, as in Gu, Kelly, and Xiu (2020) and Kozak, Nagel, and Santosh (2020). A more standard imputation method is the EM algorithm, proposed by Dempster, Laird, and Rubin (1977), which relies on a joint distribution of predictors to iteratively estimate missing values. CM implement this approach and compare it to simple cross-sectional imputation across multiple forecasting models and portfolio construction strategies. Within their empirical framework, they find that the gains from more sophisticated imputation methods are limited. In particular, improvements in out-of-sample forecasting performance are small, and differences in portfolio performance, as measured by Sharpe ratios, remain marginal.

These findings can be understood in light of the structural properties of firm characteristic panels discussed earlier, including the persistence of missingness over time, the low cross-sectional correlation between predictors, and the clustering of missingness by data source. In this context, the results of Gu, Kelly, and Xiu (2020) further suggest that predictive performance in high-dimensional settings depends primarily on the flexibility of the forecasting model rather than on the choice of imputation method.

Overall, this strand of the literature suggests that simple imputation techniques may be sufficient for prediction-oriented applications. However, this conclusion is specific to forecasting performance and does not extend to settings where statistical inference or estimation are the primary objectives. In such cases, the treatment of missing data can have first-order consequences for parameter estimation and inference.

1.2.2. Factor-Based Imputation Methods

The second strand of the literature considers missing firm characteristics not as an issue of forecasting but as a problem of exploiting the joint structure of firm characteristics to impute missing values. The core question is whether the common factor structure underlying firm characteristics can be used to recover the missing values.

This approach builds on the literature on large-dimensional factor models. For example, Bai (2003) establishes the inferential theory for factor models in high dimensions, showing that

principal component analysis can consistently estimate both factors and loadings as the cross-sectional and time-series dimensions increase. However, standard PCA requires fully observed panels, which is not the case in empirical asset pricing datasets. Xiong and Pelger (2023) extend this framework to incomplete panels, providing consistency results under general missing patterns, including cases where missingness depends on latent factors. Related contributions include Bai and Ng (2021), who study matrix completion and factor models with missing data, and Cahan, Bai, and Ng (2023), who develop factor-based imputation techniques for large panels.

Building on this literature, BLLP (2025) propose a factor-based imputation procedure. To estimate the factor structure, they construct a covariance matrix using all available pairs of observations, meaning that for each pair of characteristics, they rely on the subset of firms for which both are observed. This available-case approach allows them to recover the factor structure even when no single firm has all characteristics observed simultaneously. In the presence of missing data, factor estimates may become unstable for some firms. To address this issue, BLLP introduce regularization techniques such as ridge shrinkage, which reduce estimation variance at the cost of a small bias.

This purely cross-sectional approach is then extended through the B-XS model, which incorporates time-series information by combining current factor estimates with past observed characteristics and residual components in a cross-sectional regression estimated on partially observed data. In their sample, many firm characteristics exhibit substantial persistence over time, which motivates the use of time-series information in the imputation procedure.

The balance between cross-sectional and time-series information is determined in a data-driven way and depends on the properties of each characteristic. For instance, characteristics that are volatile and weakly autocorrelated, such as short-term momentum, rely more on the contemporaneous cross-sectional factor component, whereas more persistent accounting variables, such as total assets or leverage, rely more heavily on past observations.

To evaluate their method, BLLP conduct simulation exercises based on masking observed data, including a logit-based masking procedure designed to replicate realistic missingness patterns. In these experiments, the B-XS model achieves an out-of-sample explained variation R squared of approximately 0.58. This improvement in imputation accuracy also translates into asset pricing applications: in their masking exercises, portfolios constructed using B-XS-imputed characteristics exhibit correlations above 92 percent with portfolios based on fully observed data.

Overall, this approach highlights how exploiting the joint cross-sectional and time-series structure of firm characteristics can substantially improve the recovery of missing data, although it does not explicitly account for the impact of imputation on subsequent statistical inference.

1.2.3. Econometric Approaches to Missing Data

A third line of research addresses missing firm characteristics as a joint problem of estimation and imputation. Here, the main goal is to maintain the validity of statistical inference and parameter estimation in the presence of missing covariates, as opposed to purely imputation-oriented approaches. This viewpoint highlights how standard imputation techniques can bias both coefficient estimates and standard errors when used without properly accounting for the estimation step.

Methodologically, this approach focuses on how to handle partially observed covariates within the estimation procedure. Traditional references such as Little and Rubin (2002) show that valid inference requires explicit assumptions about the missing data mechanism as well as appropriate adjustments to estimation methods. If missing values are imputed and then treated as fully observed data, this can lead to biased estimates and incorrect inference due to the neglect of imputation uncertainty.

FHNW (2025) propose a framework that jointly handles imputation and estimation within asset pricing models. Rather than treating imputation as a separate preliminary step, they construct moment conditions that remain valid in the presence of missing covariates and estimate the model using a GMM approach. This joint treatment allows the use of all available return observations while preserving valid statistical inference.

A key feature of this framework is that missing characteristics are replaced by their conditional expectations given observed variables directly within the moment conditions, while explicitly accounting for the additional uncertainty introduced by this substitution. As a result, imputed values are not treated as actual observations, which might result as a source of bias when using imputed data in later statistical analysis in simpler approaches.

Crucially, this framework allows the researcher to retain partially observed firm-month observations without introducing selection bias or invalid inference. In contrast, complete-case analysis implicitly assumes that missingness is irrelevant, which can lead to non-representative samples and biased estimates when missingness is systematic. By incorporating all available observations and adjusting the estimation procedure accordingly, the GMM approach preserves the statistical properties of the estimator while improving efficiency.

This approach departs from the prediction-oriented literature. While simple imputation methods may be sufficient for forecasting purposes, they are generally inadequate for estimation or hypothesis testing. In such settings, missing data becomes a first-order econometric issue, as even small imputation errors can lead to biased parameter estimates or misleading inference (FHNW, 2025; Little and Rubin, 2002). This thesis implements only the imputation step of the FHNW framework, leaving the inference correction as a natural extension discussed in Section 5.3.

1.3. The implication of Imputation in Asset Pricing

Within empirical asset pricing, predictive models are evaluated both in-sample and out-of-sample. However, in high-dimensional settings, in-sample performance is often inflated due to overfitting, making out-of-sample evaluation and portfolio performance the primary criteria for assessing predictive models. Gu, Kelly, and Xiu (2020) show that using machine-learning techniques on large datasets containing many firm characteristics can substantially improve return predictions and portfolio performance.

This evaluation framework has important implications for the treatment of missing data. When the objective is to construct portfolios, the focus is not on the statistical accuracy of the imputed values themselves, but on whether different imputation strategies affect out-of-sample prediction and portfolio performance. Along these lines, CM show that, despite the challenges associated with alternative imputation methods, their impact on out-of-sample portfolio performance remains limited.

In contrast, BLLP emphasize the importance of handling missing data by exploiting the underlying structure of firm characteristics. Their factor-based approach improves the internal consistency of the dataset and preserves cross-sectional and time-series relationships when reconstructing missing values. This is particularly relevant in settings where the joint distribution of characteristics is central to estimation.

FHNW demonstrate that the treatment of missing data has first-order implications for statistical inference. While the effect of imputation methods on forecasting performance may be limited, failing to account for missing data within the estimation procedure can lead to biased parameter estimates and incorrect standard errors. Their GMM-based framework addresses this issue by incorporating missing data directly into the estimation step, thereby ensuring valid inference while retaining partially observed observations.

Taken together, these approaches reveal a fundamental tension between prediction and inference in the presence of missing data. On the one hand, prediction-oriented frameworks suggest that simple imputation methods may be sufficient, as flexible models can absorb imperfections in the data without materially affecting portfolio performance. On the other hand, inference-oriented approaches show that the same imputation choices can have substantial effects on parameter estimates and their statistical properties. This implies that the appropriate treatment of missing data ultimately depends on the objective of the analysis: forecasting performance or statistical inference.

Chapter 2: Data

2.1. Data Description

This study uses two main data sources. The first one is the Chen and Zimmermann (2022) open-source dataset, which combines information from CRSP and Compustat for firms listed on NYSE, NYSE MKT, and NASDAQ. The sample period goes from January 2000 to December 2021, covering around 22 years. The start date reflects data availability and computational constraints.

The second source of data is CRSP monthly stock return data. Monthly excess returns are used as the dependent variable in the forecasting analysis presented in Chapter 3. Following the standard approach in empirical asset pricing literature, the sample only includes common stocks with CRSP share codes 10, 11, or 12, and firms listed on the three main US exchanges. This restriction helps keep the focus on the type of firms usually studied in the literature and avoids missing data patterns coming from more unusual securities.

2.2. Firm Characteristics

The August 2023 release of the Chen and Zimmermann dataset contains 212 predictors, all drawn from peer-reviewed studies. Among all predictors, this study excludes the following three groups.

The first group consists of 33 discrete predictors. For these variables, a missing value does not represent genuinely unobserved information. It is instead assigned deliberately to encode a binary condition or an interaction effect. For example, '*AccrualsBM'* is equal to one only when a firm simultaneously ranks in the highest accrual quintile and the lowest book-to-market quintile, and is equal to zero otherwise. There is nothing to impute here because the missing entries are part of the variable's construction, not gaps in the data.

The second group covers 20 predictors with months where fewer than two valid stock observations exist, making cross-sectional estimation impossible. '*Activism1'*, which measures takeover vulnerability using institutional ownership data, is one such example, it appears very infrequently in the panel, leaving too few observations in many months for any reliable estimation.

The third group consists of predictors built on *OptionMetrics* data, which only becomes consistently available from the mid-1990s. Even within a sample starting in 2000, these signals have limited and uneven coverage across firms. '*CPVolSpread'*, which measures the difference between call and put implied volatility, is an example. Including such signals would introduce inconsistent coverage that could distort the missingness comparisons across predictor categories.

After removing these three groups, 159 continuous predictors remain eligible. From this pool, the 50 with the highest average cross-sectional observability are selected. In practice, applying this filter directly to all 212 predictors would yield the same 50, since discrete and options-based predictors are by construction less consistently observed and would not rank in the top 50 anyway. The two-step procedure is retained for transparency and to follow CM explicitly.

The 50 retained predictors span four categories, shown in Table 1. Price-based variables (24 signals) capture return-related information. Examples include *Mom12m* (12-month past return) and *MaxRet* (maximum daily return over the previous month). Accounting-based variables (16 signals) come from annual or quarterly financial statements and include measures like *Leverage* (total liabilities over market equity) and *RoE* (net income over book equity). Trading-based variables (8 signals) describe market microstructure and trading activity. Illiquidity (*Amihud's* measure of price impact per dollar traded) and *BidAskSpread* (the Corwin-Schulz effective spread) are representative examples. The distinction from price-based variables is that trading signals measure how stocks trade, not how they have performed. The remaining 2 signals: *FirmAge* and *hire* fall into another category.

Table 1: Composition of the Predictor Panel by Data Category

Category	**Number of Predictors**	**Median Missingness Rate (%)**
Price	24	7.7
Accounting	16	13.9
Trading	8	7.0
Other	2	6.9
Total	**50**	**9.7**

Price-based and trading variables are typically updated monthly and tend to be available for most firms, while accounting variables arrive at quarterly frequency and are more often missing, as measured by their missingness rates. particularly for smaller or younger firms. This heterogeneity across predictor types is part of what makes imputation a non-trivial problem here, a single method is unlikely to work equally well across all categories, which is one reason for comparing several approaches.

2.3. Missing Data Patterns

The missingness structure of the panel is examined along three dimensions: how prevalent missing data is overall, how it compounds when multiple predictors are combined, and how it is

distributed across the life cycle of a firm's presence in the sample. These findings are broadly consistent with BLLP and CM, who document similar patterns in related datasets.

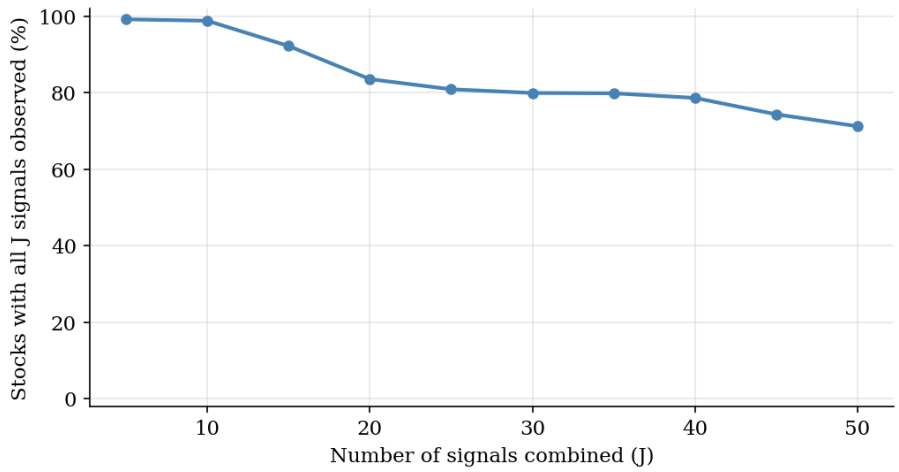
2.3.1. Overall observability

At the level of individual signals, the panel is fairly well-covered. The overall missing rate is 9.7 percent, so roughly nine in ten firm-predictor cells contain a valid observation in any given month. In other words, roughly 90 percent of firm-predictor cells contain a valid observation in any given month. In a given cross-section, 71.2 percent of firms have all 50 signals simultaneously observed. This is the complete-case rate, measured at the firm level within each month rather than across the full panel.

That number falls as more signals need to be jointly observed. Figure 1 shows how the complete-case rate changes as the required number of simultaneously observed signals increases, starting from the most observed ones and adding the less observed ones, in order of observability. It goes from close to 99 percent when only the five most-observed signals are required, down to 71 percent when all 50 must be present at the same time.

With the signals ranked by observability, the first ones correspond to price and trading signals, followed by accounting signals. The complete-case rate starts near 99 percent and declines gradually at first, then more steeply once accounting signals which tend to go missing together are added to the required set. These are available for most firms in most months, so the curve starts high and declines slowly. This is consistent with what CM document: missingness clusters within data categories, so adding signals from the same category compounds the joint missingness quickly.

Figure 1: Complete-Case Rate as a Function of Jointly Observed Predictors



2.3.2. Temporal Structure

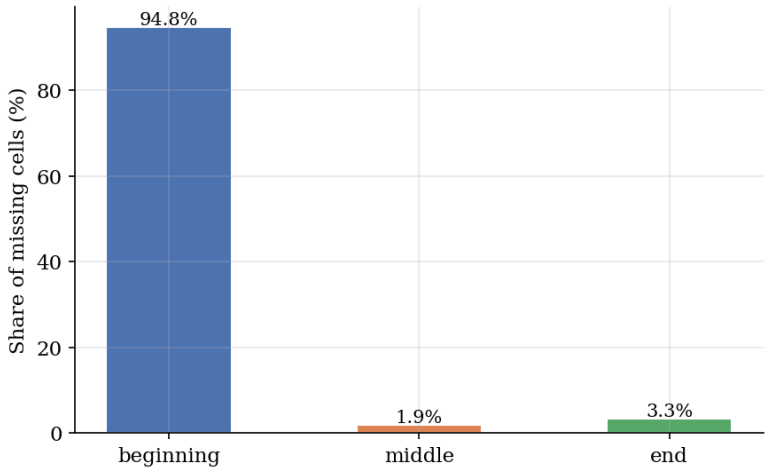
How missingness is distributed over time has direct consequences for which imputation strategies are actually usable. Three patterns are worth highlighting.

The first is persistence. When a firm-predictor pair is missing in a given month, it tends to have been missing in all prior months as well, especially for accounting and trading variables. This is not random dropout, it is structural absence. The practical implication is that time-series-based imputation methods have very little to work with for most missing observations, because there is no historical variation to exploit. CM make exactly this point in arguing that the time-series dimension adds little beyond what simpler cross-sectional methods already capture.

The second pattern is that most of the missingness is related to the beginning of the sample. To document this, we apply the MAB, MAM, and MAE classification introduced by BLLP. For a given firm and signal, each missing observation is classified based on where it falls relative to the signal's observability window for that firm. Missing values before the first observed value are classified as “beginning” missingness, and oppositely “end” missingness are the ones for which the last observed values are missing. Middle missingness arises when values are missing between two observed periods (i.e., observed, missed, observed).

Beginning and end missingness are more likely structural, linked to latent firm or market conditions. For instance, they tend to be tied to firm entry, exit, or persistent data gaps. Imputation models that we consider are not structural, therefore the last case; middle missingness; should be the best approximating a missing-at-random setting.

Figure 2: Distribution of Missing Observations by Position Type



As Figure 2 shows, MAB dominates the sample, accounting for 94.8 percent of all classified missing observations. MAM and MAE together represent only about 5 percent of the

total. This pattern makes sense given the sample construction: many of the signals used here require a minimum return history to compute, so firms entering the sample after January 2000 will be missing those signals at the start of their window. The heavy concentration of MAB also has a direct implication for the B-XS model, which relies on lagged values and therefore cannot improve on the pure cross-sectional XS approach for the large majority of missing observations in this dataset.

Reinforcing the idea that financial data is not missing at random, a third pattern is documented: missingness is not independent of firm size and other observable characteristics. Smaller and younger firms tend to have more missing data, which means complete-case analysis systematically excludes a particular type of firm and produces a sample that over-represents larger, more established companies. This is the endogeneity of the missingness problem discussed in BLLP, and it is one of the core reasons why simply dropping incomplete observations is not a neutral choice.

Table 2: Missingness Rates for the 50 Selected Predictors

Predictor	**Category**	**Missingness Rate (%)**
VolSD	Trading	18.4
MomOffSeason	Price	18.3
MomSeason	Price	18.3
roaq	Accounting	17.6
Beta	Price	16.3
PctTotAcc	Accounting	15.7
cfp	Accounting	15.7
PctAcc	Accounting	15.7
ShareIss1Y	Accounting	15.4
MRreversal	Price	15.4
NumEarnIncrease	Accounting	14.6
EBM	Accounting	13.9
BPEBM	Accounting	13.9

hire	Other	13.9
CashProd	Accounting	13.8
Leverage	Accounting	13.7
RoE	Accounting	13.6
BookLeverage	Accounting	13.6
CF	Accounting	13.6
AM	Accounting	13.6
SP	Accounting	13.6
OPLeverage	Accounting	13.6
zerotradeAlt12	Trading	11.3
IntMom	Price	11.2
CoskewACX	Price	11.2
PriceDelayTstat	Price	11.2
Coskewness	Price	11.1
Illiquidity	Trading	10.6
Mom12m	Price	10.5
MomSeasonShort	Price	10.4
PriceDelayRsq	Price	9.1
PriceDelaySlope	Price	9.1
VolMkt	Trading	8.8
AnnouncementReturn	Price	6.3
Mom12mOffSeason	Price	6.1
zerotrade	Trading	5.2

Mom6m	Price	5.2
IdioVolAHT	Price	4.8
betaVIX	Price	3.6
DolVol	Trading	2.2
High52	Price	1.1
zerotradeAlt1	Trading	1.1
IdioVol3F	Price	0.8
ReturnSkew3F	Price	0.8
ReturnSkew	Price	0.7
RealizedVol	Price	0.7
BidAskSpread	Trading	0.7
MaxRet	Price	0.1
IndMom	Price	0.0
FirmAge	Other	0.0

2.4. Preprocessing

Before any imputation is applied, the raw predictor values go through a three-step transformation. Each step is carried out separately within each month and each predictor, using only the cross-section of observed values for that month. Hence, no future information is used and no look-ahead bias is present. Missing values are left untouched.

Step 1: Winsorisation.

⇒ The observed values for each predictor in each month are clipped at the 1st and 99th percentiles of that month's cross-sectional distribution. Values outside those bounds are replaced by the corresponding threshold. This limits the influence of extreme observations, which would otherwise distort the covariance estimates that the EM and factor-based imputation methods rely on.

Step 2: Box-Cox transformation.

⇒ A generalised Box-Cox transformation is applied to each predictor within each month, using the extension of Hawkins and Weisberg (2017) to handle non-positive values. The shape parameter is estimated by maximum likelihood from the observed cross-section. The purpose is to make the marginal distribution of each predictor approximately normal within each month, which is a formal requirement of the Gaussian EM algorithm and also improves the behaviour of the factor-based methods. The sensitivity of results to this transformation is left as a robustness check for future work, as discussed in Section 5.3.

Step 3: Standardisation.

⇒ The transformed values are demeaned and scaled to unit standard deviation, using the mean and standard deviation of the observed values in that month's cross-section.

After Step 3, every predictor has a cross-sectional mean of exactly zero in every month by construction. This has a simple but important implication for the mean imputation baseline: replacing a missing value with zero is the same as replacing it with the cross-sectional (winsorized) mean. Mean imputation is therefore implemented as filling missing entries with zero, which is the standard approach in the machine learning asset pricing literature following Gu, Kelly, and Xiu (2020).

One detail worth emphasising is that none of these steps change the missingness pattern. A value that was missing before preprocessing is still missing afterwards. All transformation parameters are estimated from observed values only. The result is a standardised, approximately normal panel, thus preserving the original missing data structure and serving as the common input to all imputation methods.

Chapter 3: Methodology

3.1. Imputation Methods

This section describes the six imputation methods compared in the empirical analysis. Some imputation methods are purely cross-sectional and others use both dimensions of the panel. All methods take the preprocessed panel from Section 2.4 as input. Some work purely within the current month's cross-section. B-XS also uses values from the previous month, and conditional mean uses the other observed predictors for the same firm in the same month to condition the imputation. So not all methods are purely cross-sectional**.**

Table 3 provides a concise overview of all methods before each is described in detail.

Table 3: Summary of Imputation Methods

Method	**Information Used**	**Reference**	**Key Parameter**
Complete-case	None	—	—
Mean imputation	Cross-sectional mean	CM	—
Previous value	Last observed value	BLLP	—
EM algorithm	Full cross-sectional covariance matrix	CM	$\delta = 10^{-4}$
XS factor model	Low-rank approximation via latent factors (only cross-section)	BLLP	$K = 5$, $\gamma = 0.01/L$
B-XS factor model	Low-rank approximation via latent factors (cross-section and time series)	BLLP	$K = 5$, $\gamma = 0.01/L$
Conditional mean	OLS conditional on observed predictors within the month	FHNW	Min. obs. = 20

3.1.1. Complete-Case Analysis

Complete-case analysis is the null benchmark. The approach where missingness is simply ignored. No imputation is applied. Observations where all 50 signals are available are flagged and passed directly to the forecasting and portfolio analysis. This is the traditional approach in empirical asset pricing, and it serves as the baseline against which all imputation methods are compared.

3.1.2. Mean Imputation

Mean imputation is the simplest active imputation method and the main empirical benchmark, following CM. Since the preprocessing standardises each predictor to have zero mean within each month, mean imputation reduces to filling the missing values of the l signal at month t associated to firm i with zero:

$$\hat{X}_{i,l,t} = 0 \quad for\ all\ missing\ (i,l,t)$$

In other words, this is equivalent to filling with the cross-sectional mean or median. The two coincide here because the Box-Cox transformation in Section 2.4 makes each predictor approximately normally distributed within each month, and for a symmetric distribution mean and median are the same. Without that transformation, the two could diverge. CM show that this approach is surprisingly hard to beat in terms of portfolio performance, even by methods that are considerably more accurate in the masking experiment. The theoretical justification is that mean imputation corresponds to the special case of the EM algorithm in which the off-diagonal elements of the cross-sectional covariance matrix are set to zero, which is a reasonable approximation when cross-sectional correlations between predictors are small, as they are in this dataset.

3.1.3. Previous-Value Imputation

Previous-value (PV) imputation replaces a missing observation with the most recently observed value for that firm-predictor pair:

$$\hat{X}_{i,l,t} = X_{i,l,\,t^*} \quad where \ \ t^* = max\{s < t : \ X_{i,l,s}\ is\ observed\}$$

and $\hat{X} = 0$ when t^* does not exist. This method is particularly relevant for accounting-based variables, which are updated quarterly and where the most recent observation may be a reasonable approximation of the current value given the persistence documented in Section 2.3. Its main limitation is performance in the presence of block missingness: when a characteristic has been missing for many consecutive months, the last observed value is distant in time and likely uninformative. This limitation is most severe for MAE, where the gap between the last observed value and the current month can be very long. For MAM, where the interruption is usually short, the previous value tends to be a reasonable stand-in. Since 94.8 percent of missing observations in this dataset are of the beginning type, PV is in practice falling back to the cross-sectional mean for the large majority of cases.

3.1.4. Expectation-Maximisation (EM) Algorithm

The EM algorithm follows CM and is implemented as a cross-sectional Gaussian EM applied independently to each month t. The algorithm models the joint distribution of predictors

within each month as multivariate normal with mean zero; consistent with the standardisation applied during preprocessing; and unknown covariance matrix Σ_t.

The algorithm iterates two steps until convergence:

- **E-step:** For each firm i, the missing subvector is replaced by its conditional expectation given the observed subvector and the current estimate of Σ_t. Let $X_{obs,i}$ denote the observed predictors for firm i and $X_{miss,i}$ the missing ones. The conditional mean imputation is:

$$\hat{X}_{miss,i} = \Sigma_{mo,t}\Sigma_{oo,t}^{-1}X_{obs,i}$$

  where $\Sigma_{oo,t}$ is the submatrix of Σ_t corresponding to the observed predictors of firm i, and $\Sigma_{mo,t}$ is the submatrix of cross-covariances between the missing and observed predictors. Additionally, the conditional covariance of the missing block is:

$$Cov_{miss|obs} = \Sigma_{mm,t} - \Sigma_{mo,t}\Sigma_{oo,t}^{-1}\Sigma_{om,t}, t$$

- **M-step:** The covariance matrix is updated as:

$$\Sigma_t^{New} = \frac{1}{N}\sum_{i}^{N} \mathbb{E}_t\left[X_i X_{i,t}'\right]$$

  where the expectation uses the conditional mean from the E-step and adds back the conditional covariance for the missing block. Specifically, for the missing-missing block of each firm's outer product:

$$\mathbb{E}_t\left[X_{miss,i,t}X'_{miss,i,t}\right] = \hat{X}_{miss,i,t}\hat{X}'_{miss,i,t} + Cov_{miss|obs,t}$$

The algorithm[1] is initialised using the pairwise available-case covariance matrix, where each element (l, p) is estimated using only the firms for which both predictors l and p are observed simultaneously. Convergence is declared when the maximum absolute change in Σ_t across an iteration falls below 10^{-4}, matching the tolerance used by CM. A small ridge term εI is added to $\Sigma_{oo,t}$ before inversion for numerical stability; this is not a modelling choice but a practical safeguard against near-singular matrices in months with sparse coverage.

[1] For the full derivation see CM, Appendix A.

3.1.5. Cross-Sectional Factor Model (XS)

The XS model follows BLLP and is applied month-by-month. Rather than estimating a full $L \times L$ covariance matrix, it assumes that the L characteristics are driven by K latent cross-sectional factor:

$$C_{i,l,t} = F_{i,t}\Lambda'_{l,t} + e_{i,l,t}$$

where $F_{i,t}$ is a K-dimensional vector of factor scores for firm i and $\Lambda_{l,t}$ is a K-dimensional vector of loadings for characteristic l. The estimation proceeds in three steps. First, the characteristic covariance matrix is estimated using the available-case approach:

$$\widehat{\Sigma_{l,p}^{XS}} = \frac{1}{|Q_{l,p}|}\sum_{i \in Q_{l,p}} C_{i,l}C_{i,p}$$

where $Q_{l,p}$ is the set of firms for which both l and p are observed in month t. This pairwise estimator does not require any firm to have all characteristics observed simultaneously. Second, the loadings $\widehat{\Lambda}$ are extracted as the scaled eigenvectors of the K largest eigenvalues of $\hat{\Sigma}^{XS}$:

$$\widehat{\Lambda} = \widehat{\mathrm{V}}\widehat{\mathrm{D}}^{1/2}$$

where $\widehat{\mathrm{V}}$ contains the top K eigenvectors and $\widehat{\mathrm{D}}^{1/2}$ is the diagonal matrix of their square-rooted eigenvalues. Third, the factor scores for each firm are estimated via ridge regression using only that firm's observed characteristics:

$$\widehat{F_i^{\gamma}} = \left(\sum_l W_{i,l}\,\widehat{\Lambda_l}\widehat{\Lambda_l'} + \gamma I_K\right)^{-1}\sum_l W_{i,l}\,\widehat{\Lambda_l}\,C_{i,l}$$

where $W_{i,l} = 1$ if characteristic l is observed for firm i and 0 otherwise, and γ is the ridge regularisation parameter. Missing characteristics are then imputed as:

$$\widehat{C_{i,l}} = \widehat{F_i^{\gamma}}\widehat{\Lambda_l'}$$

The ridge parameter is set to $\gamma = \frac{0.01}{L}$, where L is the number of characteristics, following BLLP's finding that this value corresponds to the average noise level of the scaled covariance matrix $(\frac{1}{L})\hat{\Sigma}^{XS}$ and is close to optimal across different missingness patterns. The number of factors is set to $K = 5$. With $L = 50$ characteristics, standard asymptotic arguments suggest K should stay below roughly $\sqrt{L}$, which gives an upper bound of around 7. $K = 5$ is a conservative choice within that range. Setting K too large would also cause the XS model to gradually converge toward the

EM estimator, since a high-rank factor approximation starts to recover the full covariance matrix, a point CM make explicitly. A more formal selection based on the Bai and Ng (2002) information criterion is left as a robustness check in Section 5.3.

3.1.6. Backward Cross-Sectional Model (B-XS)

The B-XS model extends the XS model by incorporating time-series information, following BLLP, Definition 2. For each characteristic l and month t, the imputed value combines the contemporaneous factor component with information from the previous month:

$$\hat{C}_{i,l,t}^{B\text{-}XS} = \beta_{l,t,1}\hat{F}_{i,t}\widehat{\Lambda}'_{l,t} + \beta_{l,t,2}C_{i,l,t-1} + \beta_{l,t,3}\hat{e}_{i,l,t-1}$$

where $C_{i,l,t-1}$ is the lagged observed value of the characteristic, and

$$\hat{e}_{i,l,t-1} = C_{i,l,t-1} - \hat{F}_{i,t-1}\widehat{\Lambda}'_{l,t-1}$$

is the lagged residual from the XS model. The three coefficients $\beta_{l,t,1}, \beta_{l,t,2}, \beta_{l,t,3}$ are estimated by OLS in a cross-sectional regression at time t, using firms for which all three regressors and the target variable are simultaneously observed. The use of past observed values and past residuals separately allows the model to capture both the persistence in the common factor component and the persistence in the idiosyncratic component, as motivated by the AR(1) data-generating process discussed in BLLP.

When lagged values are not available which covers all beginning-type missing observations, representing 94.8 percent of missing values in this dataset. The method falls back to pure XS imputation. In practice this means B-XS and XS produce nearly identical results for most observations, and any advantage B-XS has is limited to the small fraction of middle-type missing values where a lagged observation actually exists.

3.1.7. Conditional Mean Imputation

Conditional mean imputation approximates the imputation component of the FHNW (2025) framework. In each month, firms are grouped by their missingness pattern, defined as the binary vector of length L indicating which predictors are observed. For each pattern and each missing predictor k, the conditional expectation is estimated by regressing $X_{i,k}$ on the firm's observed predictors within that month:

$$\hat{X}_{i,k} = X_{\text{obs},i}^{(l)}\hat{\beta}_{k,l}$$

where $\hat{\beta}_{k,l}$ is estimated by OLS with a small ridge term on the subset of firms in that month for which $X_{i,k}$ and all predictors in the conditioning set are observed. When the training sample for a given pattern is smaller than 20 firms, the fallback is the cross-sectional mean.

It is important to be precise about the scope of this implementation. FHNW (2025) do not propose conditional mean imputation as a standalone preprocessing step. Their contribution is to integrate this imputation directly into the GMM moment conditions of an asset pricing model, adjusting the estimation to account for the additional uncertainty that the imputed values introduce into the standard errors. The inference correction is not applied here, for a practical reason: implementing it requires embedding the imputation directly into the moment conditions of a specific asset pricing model and deriving adjusted standard errors for the GMM estimator. That is a substantial undertaking on its own and goes beyond the scope of this thesis. What is implemented is the imputation step only. The resulting panel is then used as input to the same forecasting and portfolio pipeline as the other five methods, so the comparison is on equal footing. This means the comparison in Section 4 measures the predictive value of conditional mean imputation as a preprocessing step, not the inferential benefits of the full FHNW framework. This distinction is discussed further in Section 5.3.

3.2. Forecasting Models

The imputed panels described in Section 3.1 serve as inputs to the machine learning forecasting models. Two complementary high-dimensional approaches are used: *principal component regression* , which provides a linear benchmark, and *neural network*, which captures nonlinearities. Both models share the same estimation design. In each month *t*, the model is trained on all available data up to *t* and the resulting forecasts are used to form portfolios for the following month, ensuring that the analysis is genuinely out-of-sample. Following CM, estimation is performed separately for three size groups based on NYSE market capitalisation breakpoints: *micro-cap stocks* below the 20th percentile, *small-cap stocks* between the 20th and 50th percentiles, and *large-cap stocks* above the 50th percentile. This stratification reflects the well-documented finding that predictability differs across the size distribution, and that pooling all stocks into a single model can obscure these differences.

3.2.1. Principal Component Regression

Principal component regression (PCR) reduces the dimensionality of the imputed predictor panel by extracting the principal components of the characteristic matrix within the training window and regressing monthly stock returns on the first K components using OLS. The predicted return for stock i in month $t + 1$ is:

$$\hat{r}_{i,t+1} = p'_{i,t}\,\widehat{\beta_t}$$

where $p_{i,t}$ is the vector of principal component scores for stock i and $\widehat{\beta}_t$ is estimated by OLS on the training sample. The number of components K is treated as a tuning parameter and selected by minimising the mean squared prediction error in a held-out validation window.

Because the principal components are constructed from the imputed panel, the choice of imputation method affects the principal components themselves and so the regression coefficients estimated from them. This sensitivity is one of the main objects of interest in the empirical analysis, as imputation methods that preserve the covariance structure of the characteristic panel more accurately should in principle produce more informative principal components.

3.2.2. Neural Networks

The neural network specification follows Gu, Kelly, and Xiu (2020) and uses a two-hidden-layer feed-forward architecture, corresponding to the NN2 specification in Gu, Kelly, and Xiu (2020). The hidden layers contain 16 and 8 neurons respectively, with a rectified linear unit (ReLU) activation function. The network maps the imputed predictor vector for each stock to a scalar predicted return, trained to minimise a squared prediction error loss with an L1 regularisation penalty on the weights. The Adam optimiser is used with early stopping based on validation loss to prevent overfitting. The final forecast averages three independently trained networks rather than the ten used in Gu, Kelly, and Xiu (2020). This reduces computation while keeping the main benefit of ensemble averaging.

The neural network is evaluated under two distinct data scenarios, designed to assess whether imputation genuinely recovers predictive information or artificially inflates performance.

⇒ **Scenario S1** uses the fully complete-case dataset, in which all firm-month observations with any missing predictor values are removed. This scenario represents the traditional approach and provides a clean upper bound on what is achievable without imputation but with a sufficiently flexible model.

⇒ **Scenario S2** uses the imputed panel, retaining all observations including those with more than 20 percent of predictor values originally missing before imputation.

A further robustness check would be to train on imputed data and evaluate predictions on complete-case observations only, and vice versa. This cross-evaluation would help separate whether any performance differences come from the training sample composition or from the quality of imputed inputs at prediction time. This is left for Section 5.3.

The Sharpe ratio produced under each scenario is then compared. If the imputed dataset S2 consistently delivers a higher Sharpe ratio than the clean complete-case dataset S1, this helps to understand whether the improvement can be attributed to genuine pricing information recovery.

This comparison therefore serves as a robustness check to understand whether the results of the forecasting analysis are economically meaningful.

Models are estimated using a rolling 120-month training window. The out-of-sample evaluation period covers January 2010 to December 2021, yielding 144 monthly return forecasts. The number of principal components K is selected by minimising mean squared prediction error on a 24-month held-out validation subsample at the end of each training window, with K ranging from 1 to 15. The neural network uses a three-network ensemble rather than the ten-network ensemble of Gu, Kelly, and Xiu (2020), adopted for computational tractability while preserving the core benefit of ensemble averaging. Given data availability constraints, a single model is estimated on the full cross-section rather than separately by size group as in CM; this is noted as a limitation in Section 5.3.

3.3. Portfolio Construction

Three portfolio construction approaches are used, ordered by increasing sophistication: univariate and bivariate characteristic sorts, beta sorting, and prediction-based long-short portfolios. Together, they allow the analysis to assess how imputation choices affect investment performance across different frameworks.

3.3.1. Univariate and Bivariate Sorting

Univariate sorting is the most direct method for evaluating whether imputation choices affect the return predictability of individual characteristics. In each month, stocks are assigned to one of four or six groups based on quantile breakpoints of the chosen characteristic, and the long-short portfolio goes long the top group and short the bottom group, held for one month with value-weighting following the convention in Fama and French (1993). This approach allows for a clean comparison of how imputation methods affect the characteristic-return relationship, since the portfolio signal is the characteristic value itself.

Bivariate sorting extends this by conditioning simultaneously on two dimensions, following the double-sort methodology of Daniel and Titman (1997) and its application in the missing data context of BLLP. Stocks are first sorted into groups based on their level of missingness; classified as None, Low, Medium, or High depending on the fraction of their predictor values that were originally missing, and then, within each missingness group, further sorted based on the value of a target characteristic. This produces a grid of portfolios, illustrated schematically in *Figure 3* below, where the rows represent missingness levels and the columns represent characteristic percentile groups.

Figure 3: Structure of the Bivariate Sort Portfolio Design

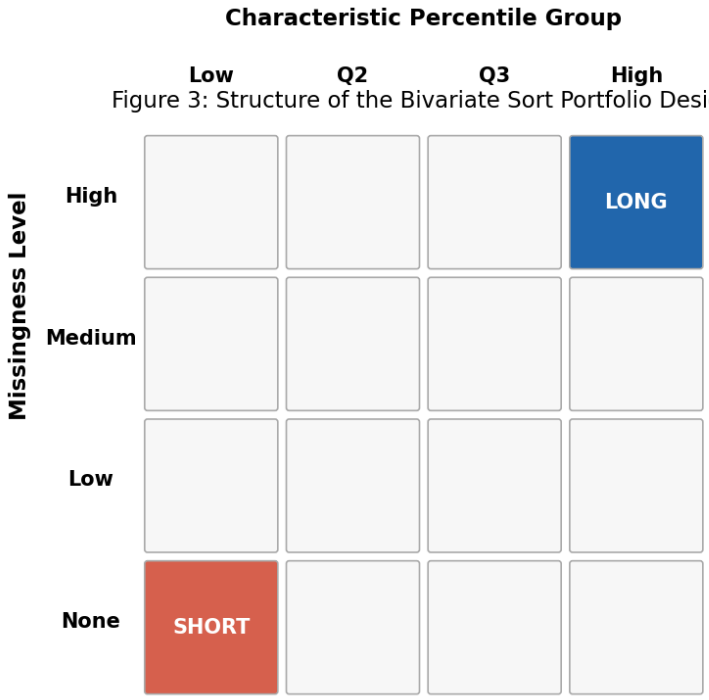


Figure 3: Structure of the Bivariate Sort Portfolio Design

This design is directly motivated by the central question of the thesis. If 1) imputation accurately recovers the missing values of the characteristics, and 2) the return spread is consistently estimated, then we expect the spread to be more or less stable regardless of which missingness row you are in. If imputation is distorting the characteristic values instead, the within-row spread will differ across missingness levels, and this pattern will vary depending on which method was used. The main spread portfolio reported is long the high-characteristic/high-missingness cell and short the low-characteristic/no-missingness cell, as shown in Figure 3. Other combinations, sorting on missingness alone, or on the interaction of both dimensions are possible and could provide further evidence, but are left as extensions.

3.3.2. Beta Sorting

Beta sorting constructs portfolios based on each stock's estimated exposure to a common risk factor rather than its characteristic value, following the approach of Black, Jensen, and Scholes (1972) and its more recent applications in FHNW. For each stock, beta is estimated using a rolling OLS regression of monthly excess returns on the Mom12m signal over a 60-month window. An alternative definition, more directly relevant to the thesis question, would be to regress returns on the characteristic itself and sort on the resulting slope coefficient. This would measure how sensitive returns are to each signal and could show more directly how imputation affects the characteristic-return relationship. The market-beta version is kept here as a simpler benchmark that is largely unaffected by the imputation choice, with the characteristic-beta approach left as a natural extension.

Stocks are then sorted into quintiles based on their estimated betas and the long-short value-weighted portfolio goes long the highest-beta quintile and short the lowest. The key feature of beta sorting in this context is that it uses only return data in its construction and does not directly depend on the imputed characteristics. It therefore provides a reference portfolio whose composition is largely invariant to the imputation method. Comparing the beta-sorted portfolio returns to the characteristic-sorted and prediction-based portfolio returns allows one to isolate how much of any observed performance difference across imputation methods is attributable to the quality of the imputed characteristics themselves, rather than to differences in systematic risk exposure.

Beta sorting is also used alongside PCR to examine whether the systematic risk profile of the PCR portfolios changes with the imputation method. FHNW show that the treatment of missing data can affect the estimated factor structure of asset pricing models, and comparing how strongly PCR predictions load on market beta across imputation methods provides direct evidence on this channel.

3.3.3. Prediction-Based Portfolios

Prediction-based portfolios are constructed directly from the return forecasts produced by the PCR and neural network models, following the portfolio construction framework of Gu, Kelly, and Xiu (2020) and CM. In each month, stocks are ranked by their predicted return and assigned to deciles. The long-short portfolio goes long the top decile and short the bottom decile, held for one month. Both equal-weighted and value-weighted versions are computed, as the two weighting schemes can give substantially different impressions of performance when predictability varies across the size distribution, a point documented in Fama and French (2008) and emphasised by CM in the context of missing data.

This is the primary portfolio construction approach in the empirical analysis, as it most directly addresses the question of whether imputation choices affect the economic value of machine learning return forecasts. The two-scenario comparison described in Section 3.2.2; complete-case S1 versus imputed panel S2; is applied here as well, so that for each imputation method the prediction-based Sharpe ratio from S2 can be benchmarked against the clean S1 result.

3.4. Evaluation Metrics

The empirical analysis evaluates the imputation methods along two distinct dimensions. The first is a direct measure of imputation accuracy, assessed through a masking experiment in the characteristic space. The second is an indirect measure based on forecasting and portfolio performance, which is the primary criterion of interest from an applied asset pricing perspective. These two dimensions do not necessarily move together, and the relationship between them is one of the central empirical questions of the thesis.

3.4.1. Imputation Accuracy

Imputation accuracy is evaluated using a random masking exercise following CM and BLLP. A random ten percent of the observed firm-predictor values are masked, that is, temporarily set to missing, and each imputation method is then applied to the resulting panel as if those values were genuinely unobserved. The masking is applied uniformly across all firm-predictor cells. An interesting extension would be to stratify the masked sample by firm size, since smaller firms tend to have both more missing data and lower observability, and it is possible that imputation accuracy differs systematically between large and small firms. The imputed values at the masked positions are compared to the true observed values using two standard metrics.

The root mean squared error (RMSE) for imputation method m is defined as:

$$\text{RMSE}_\text{m} = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{(i,l,t)\in\mathcal{M}} \left(C_{i,l,t} - \hat{C}_{i,l,t}^{(m)}\right)^2}$$

and the mean absolute error (MAE) is:

$$\text{MAE}_\text{m} = \frac{1}{|\mathcal{M}|} \sum_{(i,l,t)\in\mathcal{M}} \left|C_{i,l,t} - \hat{C}_{i,l,t}^{(m)}\right|$$

where $\mathcal{M}$ denotes the set of masked observations, $C_{i,l,t}$ is the true observed value, and $\hat{C}_{i,l,t}^{(m)}$ is the imputed value produced by method m. Since the characteristics are standardised to unit variance in the preprocessing step, the RMSE and MAE are directly comparable across predictors and across methods. As a reference, mean imputation, which replaces all missing values with zero, produces an RMSE of approximately 1.0 by construction, since the data have unit variance after standardisation. A method with RMSE below 1.0 therefore genuinely recovers information, while an RMSE at or above 1.0 indicates the imputation is no better than simply using the cross-sectional mean, as noted by CM in their masking analysis.

These metrics are reported both in aggregate across all predictors and stratified by the position type of the masked observation, following the beginning-middle-end decomposition of BLLP described in Section 2.3. This stratification is informative because the amount of information available for imputation differs substantially across position types, and imputation methods differ in their ability to exploit that information. In particular, the B-XS model is expected to perform substantially better than the pure XS model for MAM, where lagged values are available, while the two methods should produce similar results for MAB where no prior observations exist. It should be noted that in the present dataset, as documented in Section 2.3.2, 94.8 percent of missing observations are of the beginning type, which means the masking

experiment predominantly evaluates MAM where observed values are available to mask. The position-stratified results in Table 8 should therefore be interpreted with this in mind.

3.4.2. Forecasting Performance

Forecasting performance is assessed using the out-of-sample predictive R^2 of Goyal and Welch (2008), which measures the improvement in prediction accuracy relative to a historical mean return benchmark. For each stock i in month t, the benchmark prediction is the historical mean return $\bar{r}_t$ estimated over the training window. The out-of-sample R^2 is then:

$$R^2_{OOS} = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2}{\sum_{i,t}\left(r_{i,t+1} - \bar{r}_t\right)^2}$$

A positive R^2_{OS} indicates that the forecasting model outperforms the historical mean benchmark, while a negative value indicates underperformance. This metric is reported separately for each imputation method, forecasting model and by size group. It allows for a direct comparison with the results of CM and Gu, Kelly, and Xiu (2020), who report this metric as their primary measure of predictive performance.

3.4.3. Portfolio Performance

Portfolio performance is the primary evaluation criterion, as it translates forecasting accuracy into statistical significance. The main metric is the annualised Sharpe ratio of the long-short portfolio:

$$SR = \frac{\sqrt{12} \cdot \bar{r}_{LS}}{\hat{\sigma}_{LS}}$$

where $\bar{r}_{LS}$ is the sample mean of monthly long-short portfolio returns and $\hat{\sigma}_{LS}$ is their sample standard deviation, with the factor $\sqrt{12}$ converting to an annualised figure. The Sharpe ratio is reported for each combination of imputation method, forecasting model, and portfolio construction approach, under both equal-weighting and value-weighting.

Additional portfolio performance metrics include the annualised mean return and the annualised return standard deviation, reported separately so that differences in Sharpe ratios can be attributed to either the return or the volatility component. Following BLLP, the performance of univariate and bivariate characteristic-sorted portfolios is also reported in terms of the return spread between the top and bottom groups, which provides a direct measure of the economic magnitude of the imputation effect on characteristic-based strategies.

3.4.4. The Two-Scenario Comparison

As described in Section 3.2.2, all portfolio performance metrics are computed under both Scenario S1, the fully complete-case dataset, and Scenario S2, the imputed panel retaining all observations. The comparison of Sharpe ratios across the two scenarios is a key evaluation tool in this analysis. Formally, for each imputation method m and forecasting model, the incremental Sharpe ratio from imputation is defined as:

$$\Delta SR^{(m)} = SR_{S2}^{(m)} - SR_{S1}$$

A positive $\Delta SR^{(m)}$ indicates that imputation method m delivers higher risk-adjusted returns than the complete-case baseline. As discussed in Section 3.2.2, the interpretation of this difference depends on its magnitude and relates with the imputation accuracy results. A large positive $\Delta SR^{(m)}$ that is uniform across imputation methods of varying accuracy would suggest that the gain comes primarily from the larger sample size rather than from the quality of the imputed values. The intuition is that S2 retains more firms than S1 by construction including smaller and younger firms that complete-case analysis drops. If those firms happen to earn higher returns on average, any imputation method will mechanically improve the Sharpe ratio simply by including them, regardless of how accurately their characteristics were recovered. This is the selection bias argument of BLLP. A positive $\Delta SR^{(m)}$ that is concentrated among the most accurate imputation methods would instead support the interpretation that better imputation translates into better investment performance.

Formal tests for differences in Sharpe ratios, such as those in Barillas et al. (2020) or the Diebold-Mariano test adapted for portfolio returns, would put these comparisons on a more rigorous statistical footing. Given the scope of this thesis, the comparison is reported descriptively, with formal testing left for future work**.**

Chapter 4: Empirical Results

4.1. Missingness Patterns

This section briefly refreshes the key missingness patterns documented in Chapter 2 before presenting the imputation results.

Table 4: Summary Statistics for the Imputed Panel

Statistic	Value
Firm-month observations	372,359
Number of signals	50
Total cells	18,617,950
Observed cells	16,807,324
Missing rate	9.7%
Complete-case observations	265,211
Complete-case rate	71.2%

As documented in Section 2.3.1, the complete-case rate falls from 99 percent at J=5 to 71 percent at J=50.

The beginning-middle-end breakdown is illustrated in Table 5 and Figure 2. MAB makes up 94.8 percent of total missingness. This number will keep repeating in the results. For any approach that is based on lagged variables, there is almost nothing to use.

Table 5: Distribution of Missing Observations by Position Type

Position Type	Count	Share (%)
Beginning	1,715,862	94.8
Middle	34,676	1.9
End	60,088	3.3

Total	**1,810,626**	**100.0**

Figure 4 shows missingness by type. The missing percentage of accounting signals is 14.5 percent on average, almost double that of the price and trading signals. It is important to bear this in mind while analyzing the imputation results.

Figure 4: Average Missingness Rate by Predictor Category

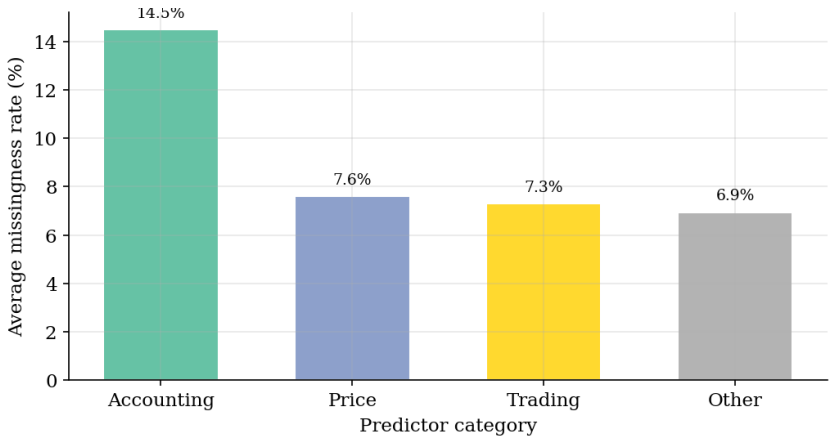


Table 6 gives the complete-case curve numerically for J = 5 to J = 50.

Table 6: Complete-Case Rate by Number of Jointly Observed Predictors

J (Predictors Required)	Complete-Case Rate (%)
5	99.2
10	98.8
15	92.2
20	83.5
25	80.9
30	79.9
35	79.8
40	78.6

45	74.3
50	71.2

4.2. Imputation Accuracy

The masking experiment hides 10 percent of observed values and checks how well each method gets them back. Table 7 has the full results.

Table 7: Imputation Accuracy by Method / Overall Masking Experiment Results

Method	MAE	RMSE
Mean	0.7047	0.9998
PV	0.2991	0.6659
EM	0.5747	0.8707
CMean	0.284	0.5541
B-XS	0.435	0.6993
XS	0.5578	0.8344

Note: RMSE is normalised relative to unit-variance standardised predictors. Mean imputation produces RMSE ≈ 1.0 by construction.

All six methods beat the naive mean baseline. CMean is the most accurate, with RMSE 45 percent below mean. PV and B-XS follow at around 30-33 percent below. EM and XS also improve on mean, though by less.

The masking experiment predominantly evaluates MAM, since those are the only observations with true values available to mask. MAM missingness is by nature a firm-level and time-series phenomenon: a characteristic was observed, disappeared, and reappeared for the same firm. Methods that exploit this structure, either by using lagged values directly (PV, B-XS) or by conditioning on the firm's other observed signals in the same month (CMean), are therefore better suited to this setting than purely cross-sectional methods like XS, EM, and mean imputation, which ignore the firm's history entirely. This explains why CMean, PV, and B-XS outperform XS, EM, and mean in the masking experiment. By conditioning on all observed signals for the same firm in the same month, CMean extracts more information than any method that works signal by signal without firm-level conditioning. PV works well for the simpler reason that characteristics don't

change much from month to month. B-XS showing improvement over XS is consistent with BLLP theoretical argument, even if the advantage is limited by the dominance of MAB.

Figure 5: Out-of-Sample Imputation Accuracy by Method

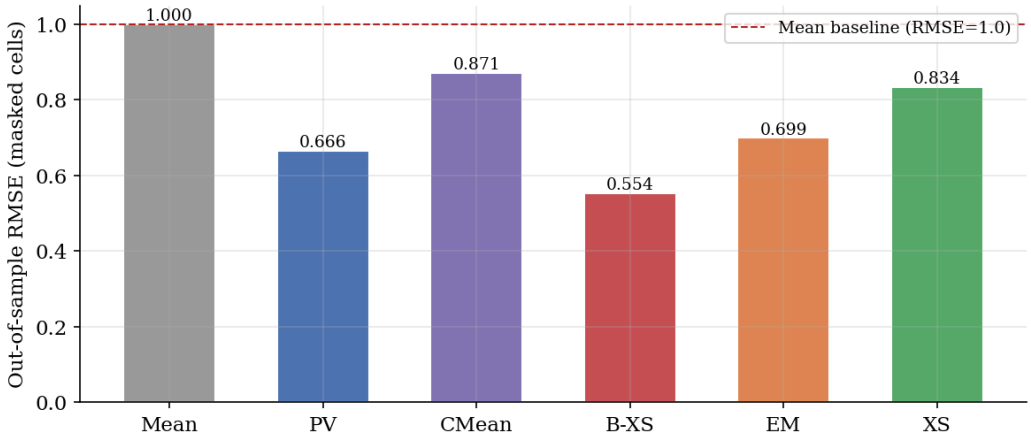


Table 8 breaks results down by position type. Since the masking experiment hides observed values, it mostly hits middle-type observations. Those are the only ones with something to mask.

Table 8: Imputation Accuracy by Method and Position Type

Method	Position	RMSE
Mean	Middle	0.9998
PV	Middle	0.6659
EM	Middle	0.6993
CMean	Middle	0.8707
B-XS	Middle	0.5541
XS	Middle	0.8344

Note: Beginning and end positions contain almost no masked observations since the masking experiment only masks observed values, which fall almost entirely in the middle category.

4.3. Characteristic-Based Portfolio Results

Accuracy of the estimation process refers to how well the procedures work to recover the true values in sample. What really counts, however, is if this level of precision results in a good

investment portfolio, that is in an out-of-sample analysis. This discussion focuses on that issue in terms of three techniques that rank stocks based on the actual estimates.

4.3.1. Univariate Sorts

Table 9 shows Sharpe ratios from long-short portfolios sorted on six signals, across all six methods.

Table 9: Univariate Sort Sharpe Ratios by Signal and Imputation Method

Signal	Category	Mean	PV	EM	XS	B-XS	CMean
Mom12m	Price	−0.157	−0.147	−0.137	−0.240	−0.240	−0.149
Illiquidity	Trading	+0.067	+0.059	+0.067	+0.032	+0.031	+0.074
BookLeverage	Accounting	+0.054	+0.085	−0.018	−0.015	−0.018	−0.039
RoE	Accounting	+0.043	+0.025	+0.312	+0.306	+0.304	+0.252
IdioVol3F	Price	+0.094	+0.094	+0.091	+0.090	+0.091	+0.092
DolVol	Trading	+0.036	+0.033	+0.041	+0.034	+0.036	+0.043

The methods produce broadly similar results across most signals. A few patterns stand out.

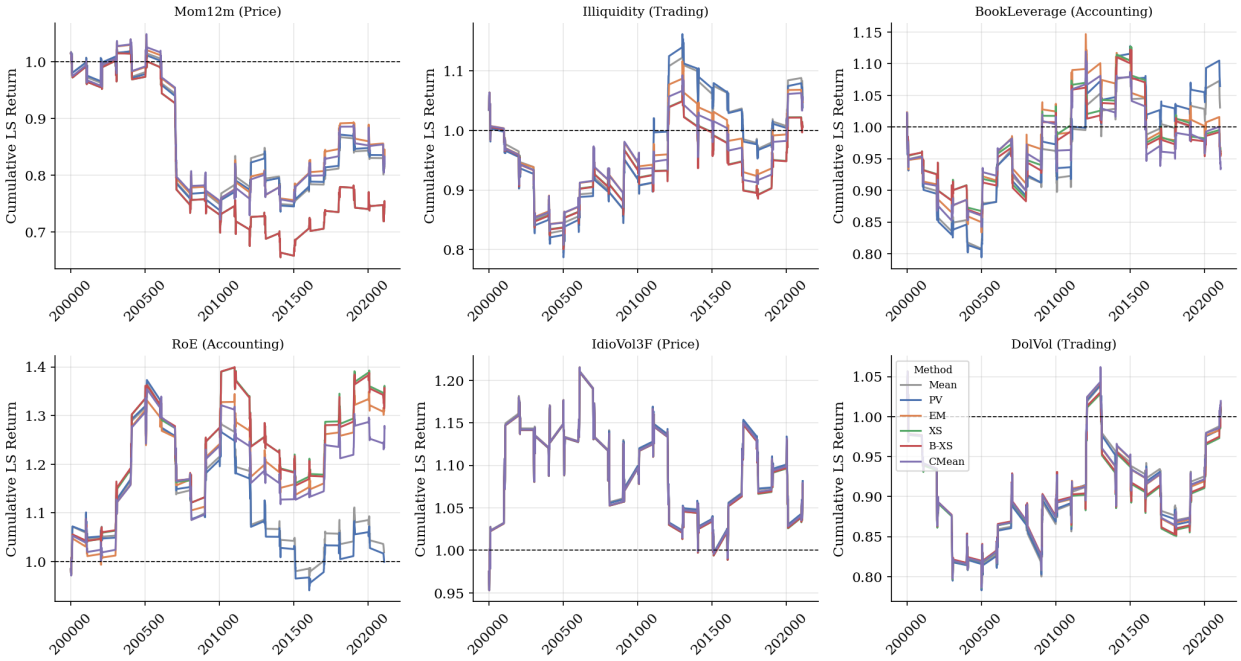
Mom12m, all methods give negative Sharpe ratios between -0.137 and -0.240. Initially this seems puzzling, but it may be due to the time span, which covers 2000 through 2009 and includes the 2008 financial crisis during which momentum famously crashed. It is also worth noting that XS and B-XS recover a stronger negative signal than the other methods, which suggests that momentum may be partly driven by a low-dimensional cross-sectional pattern rather than firm-specific or time-series variation.

For Illiquidity, the Sharpe ratio is larger for methods that take into account the firm-specific component. PV, CMean, and B-XS give Sharpe ratios between 0.059 and 0.074, whereas mean, EM, and XS give between 0.032 and 0.067. Since illiquidity is a persistent firm-level characteristic, methods that exploit firm history recover the signal somewhat more effectively.

For BookLeverage, mean and PV give positive Sharpe ratios while EM, XS, B-XS, and CMean give negative ones. This is probably better read as instability in the imputation strategies rather than a genuine economic difference: different methods are recovering different versions of the leverage signal.

For RoE, IdioVol3F, and DolVol, the Sharpe ratios are surprisingly stable across all six methods. Whatever the imputation technique used, the findings for these signals are reliable.

Figure 6: Cumulative Long-Short Returns: Univariate Characteristic Sorts by Imputation Method



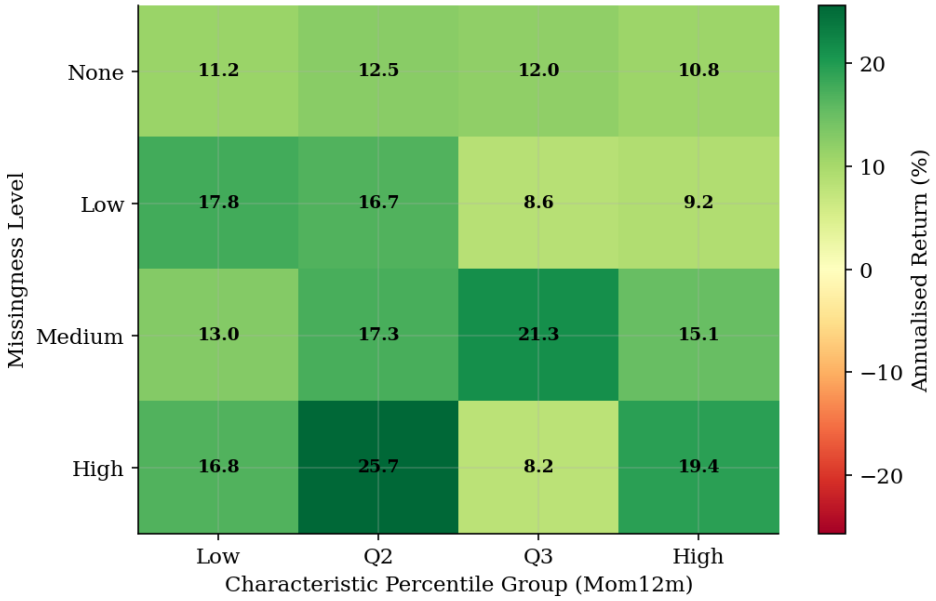
Long-short returns of each method are shown in Figure 6. Differences between the methods are seen in all the panels except for XS/B-XS vs. all others. In particular, the lines for Mom12m and BookLeverage run in opposite directions, providing one of the clearest examples of a change in signs.

4.3.2. Bivariate Sorts

The bivariate sort tests the central question more directly. If imputation works correctly, the return spread across characteristic groups should look broadly similar regardless of how much of a firm's data was missing. A high-momentum firm with lots of imputed data should not behave dramatically differently from a high-momentum firm with fully observed data. In particular, if the imputation is retrieving the missing information, we expect a monotonic effect in the spread when considering the degree of missingness. In other words, we expect the imputation to do roughly equal or better as missingness increases.

Stocks are sorted simultaneously on value and missingness level, producing a 4×4 grid. Rows are missingness groups (None, Low, Medium, High) and columns are characteristic percentile groups (Low, Q2, Q3, High). If imputation helps, you should observe a broadly consistent pattern across rows not a dramatic variation in the spread as missingness increases.

Figure 7: Bivariate Sort Returns: Missingness Level vs Characteristic Percentile (Mom12m)



For all methods, the patterns across rows are broadly consistent. The return gradient from low to high momentum is present in most missingness groups and does not shift dramatically as missingness increases.

One observation worth flagging: within some missingness rows, the return pattern across characteristic groups is not monotonic. Instead of a clean Low < Q2 < Q3 < High pattern, Q2 and Q3 sometimes show higher returns than the High group. This is puzzling and likely reflects a composition effect rather than a genuine momentum pattern. This is counterintuitive: more imputation leading to higher apparent returns. It likely reflects a composition effect: firms with more missing data may have specific characteristics that drive returns independently of momentum. Results for other signals beyond Mom12m would help clarify whether this pattern holds broadly.

Table 10: Bivariate Sort Annualised Returns (%) by Missingness Group and Mom12m Percentile

- **Mean imputation:**

	Low	**Q2**	**Q3**	**High**
None	11.2	12.5	12.0	10.8
Low	17.3	16.3	9.8	8.0
Medium	8.6	19.5	18.4	15.6

High	—	—	—	—

- **PV:**

	Low	**Q2**	**Q3**	**High**
None	11.2	12.5	12.0	10.8
Low	17.2	16.6	9.9	8.6
Medium	7.6	18.9	19.0	14.8
High	—	—	—	—

- **EM:**

	Low	**Q2**	**Q3**	**High**
None	11.2	12.5	12.0	10.8
Low	17.5	14.9	11.1	8.8
Medium	12.9	15.5	24.7	14.1
High	13.4	24.1	18.8	14.0

- **XS:**

	Low	**Q2**	**Q3**	**High**
None	11.2	12.5	12.0	10.8
Low	17.8	16.7	8.6	9.2
Medium	13.0	17.3	21.3	15.1
High	16.8	25.7	8.2	19.4

- **B-XS:**

	Low	Q2	Q3	High
None	11.2	12.5	12.0	10.8
Low	17.8	16.7	8.6	9.2
Medium	13.0	17.3	21.3	15.1
High	16.8	25.7	8.2	19.4

- **CMean:**

	Low	Q2	Q3	High
None	11.2	12.5	12.0	10.8
Low	18.9	13.9	9.2	10.2
Medium	12.4	15.1	24.7	15.0
High	14.2	21.5	21.9	12.7

Note: High missingness group is empty for Mean, PV, and CMean because no firms in this dataset have sufficiently high average missingness to populate the top quartile after the observability-based signal selection.

The return gradient from low to high momentum is present in most missingness groups and does not shift dramatically as missingness increases.

One observation worth flagging: the Q2 and Q3 columns for medium and high missingness groups show quite high annualised returns, sometimes approaching 25 percent, noticeably above the None missingness rows. It likely reflects a composition effect: firms with more missing data may have specific characteristics that drive returns independently of momentum. Results for other signals beyond Mom12m would help clarify whether this pattern holds broadly.

Three spreads are measured in each missingness row to characterize the shape of the return-characteristic relationship: (1) the classical long-short spread: long Q4, short Q1; (2) long Q3, short Q2, in order to check whether excess returns are present in the middle of the distribution; (3) long Q1 and Q4, short Q2 and Q3, which approximates the second derivative of returns with respect to the characteristic.

In the None missingness row, where there is no imputation, all three spreads are small and negative for all methods, and portfolios 1 and 2 are roughly the same. This means that, in the fully observed subsample, the momentum-return relationship is flat rather than monotonic and does not concentrate at the tails. When missingness increases, spreads become less stable and differ across methods, especially for XS and B-XS. Portfolio 3 is predominantly negative, meaning groups in the middle of the characteristic distribution earn somewhat more than the extremes, though this does not hold for all rows and methods. Given the absence of a momentum premium in this sample, conclusions about the shape of the relationship are hard to draw.

4.3.3. Beta Sorting

Table 11 shows results from beta sorting. Firms sorted by their estimated sensitivity of future returns to *Mom12m*, from a rolling 60-month regression.

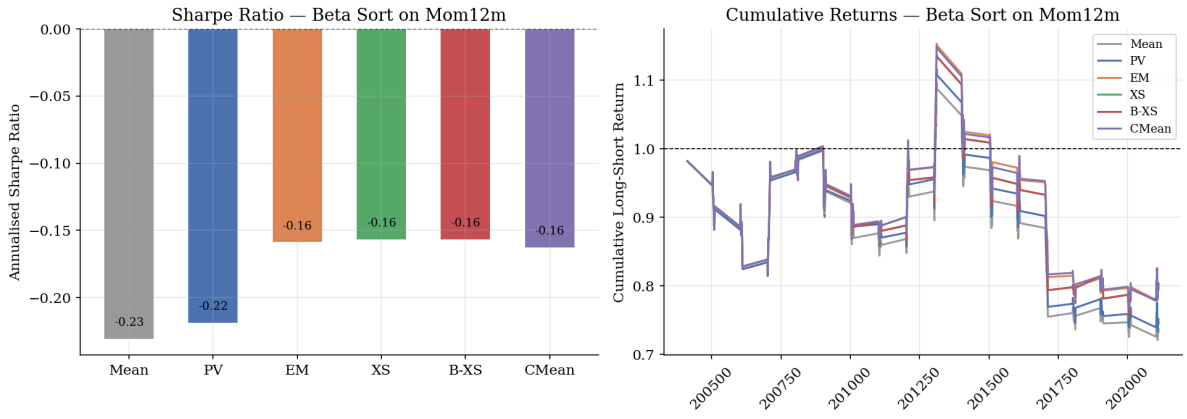
Table 11: Beta Sorting Results by Imputation Method (Signal: Mom12m)

Method	Annualised Return (%)	Sharpe Ratio
Mean	-1.593	-0.231
PV	-1.497	-0.219
EM	-1.094	-0.159
XS	-1.081	-0.157
B-XS	-1.081	-0.157
CMean	-1.116	-0.163

All methods give negative Sharpe ratios, consistent with the negative momentum premium documented in the univariate sorts. The variation across methods is modest roughly -0.16 to -0.23. EM, XS, B-XS, and CMean cluster around -0.16, while mean and PV are more negative at -0.23 and -0.22. This mirrors the RoE pattern from the univariate sorts, methods exploiting cross-sectional information perform somewhat better.

Beta sorting imposes both a monotonic and a linear relationship between the signal and returns, which characteristic sorting does not require. The more negative results here compared to the direct characteristic sort may partly reflect the cost of imposing this structure when the true relationship is nonlinear.

Figure 8: Beta Sorting Results by Imputation Method



4.4. Forecasting Results

4.4.1. Out-of-Sample R^2

Table 12: Out-of-Sample Predictive R^2 (%) by Imputation Method

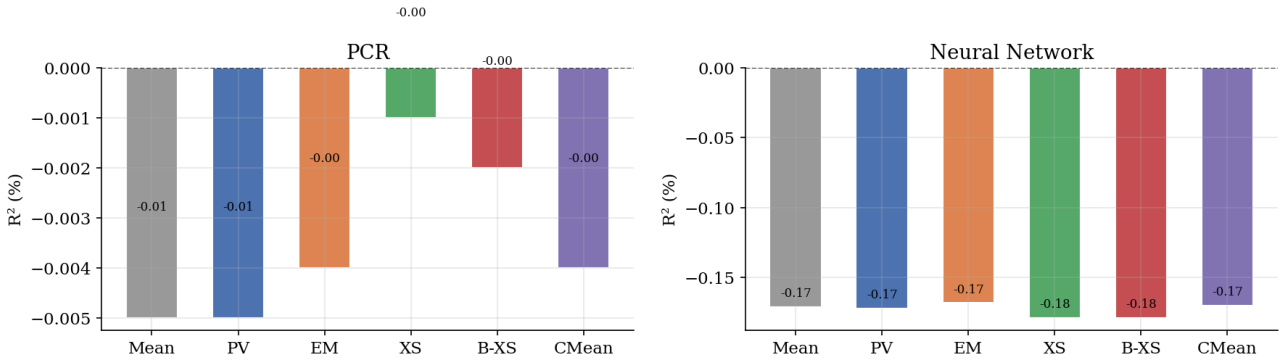
Method	PCR R^2_OS	NN R^2_OS
Mean	-0.005	-0.171
PV	-0.005	-0.172
EM	-0.004	-0.168
XS	-0.001	-0.179
B-XS	-0.002	-0.179
CMean	-0.004	-0.17

All values are negative for both models and all methods. The differences across methods are minimal. Neither PCR nor NN beats the historical mean as a return forecast, and the imputation choice makes almost no difference.

The models are estimated on the full cross-section without separating firms by size. CM show that predictability is concentrated in small stocks pooling all firms dilutes that signal significantly. The negative R^2_OS most likely reflects this limitation rather than a failure of imputation.

NN R^2_OS is more negative than PCR (-0.17 versus -0.005). This is unexpected because Gu, Kelly, and Xiu (2020) find NN outperforms PCR. One possible explanation is overfitting without size stratification. The 2008 momentum crash in the training window may also disproportionately affect the NN given its greater flexibility.

Figure 9: Out-of-Sample Predictive R^2 by Imputation Method



4.4.2. Portfolio Performance

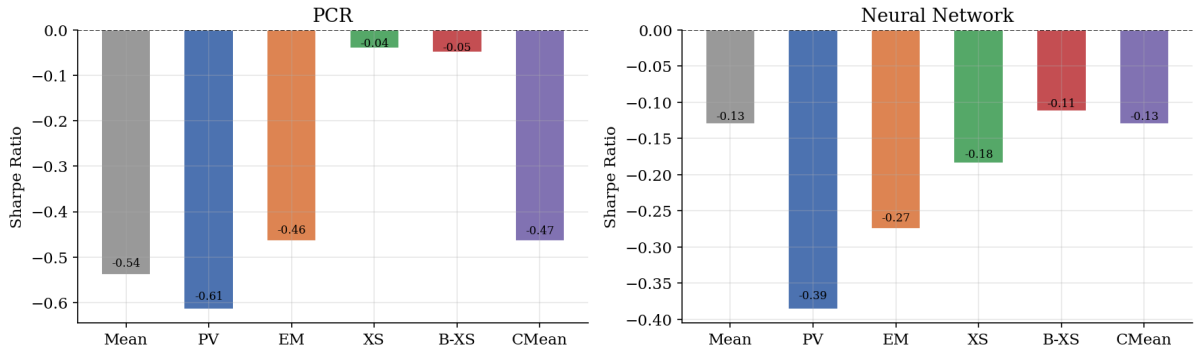
Table 13 gives annualised Sharpe ratios from prediction-based long-short portfolios.

Table 13: Annualised Sharpe Ratio by Imputation Method

Method	**PCR**	**NN**
Mean	-0.5389	-0.1298
PV	-0.6143	-0.3858
EM	-0.4638	-0.2749
XS	-0.0409	-0.1845
B-XS	-0.0499	-0.1121
CMean	-0.465	-0.1299

All Sharpe ratios are negative. The most striking result is XS and B-XS for PCR. Sharpe ratios of -0.041 and -0.050, dramatically better than all other methods which range from -0.46 to -0.61. For NN, B-XS gives the best result at -0.112, close to mean at -0.130. PV gives the worst result for both models despite being among the more accurate imputation methods.

Figure 10: Annualised Sharpe Ratio by Imputation Method



4.4.3. The S1/S2 Comparison

Table 14: Incremental Sharpe Ratio vs Mean Imputation (ΔSR)

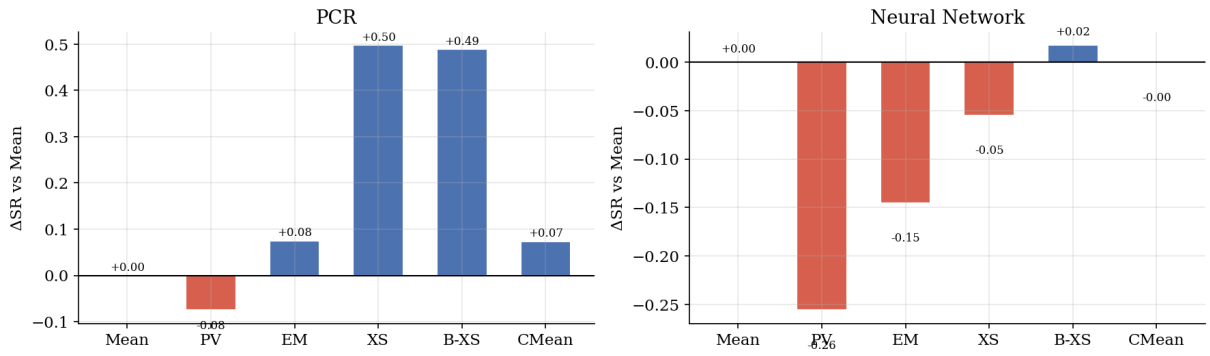
Method	**PCR ΔSR**	**NN ΔSR**
Mean	0	0
PV	-0.0754	-0.256
EM	0.0751	-0.1451
XS	0.498	-0.0547
B-XS	0.489	0.0177
CMean	0.0739	-0.0001

XS produces ΔSR = +0.498 for PCR, whereas B-XS produces +0.489. Both EM and CMean improve slightly on mean (+0.075 and +0.074 respectively). PV is the only method that makes things worse (-0.075). For NN, B-XS is the only method with a positive ΔSR (+0.018), though the improvement is marginal. Mean and CMean produce similar results (ΔSR ≈ 0). Every other method makes NN worse than mean.

These results need to be read carefully. All Sharpe ratios are negative, meaning all long-short portfolios lose money regardless of the imputation method. The large ΔSR for XS and B-XS does not reflect genuine outperformance. It simply means their PCR Sharpe ratios (-0.041 and -0.050) are unusually close to zero compared to the other methods, which range from -0.46 to -0.61. In other words, XS and B-XS lose less money than the others, but they are still losing money. The reverse strategy would have been more profitable.

One possible explanation for the less negative PCR Sharpe ratios of XS and B-XS is overfitting. Both methods perform well in the masking experiment for MAM, which is precisely the setting the masking exercise evaluates. However, good in-sample recovery does not guarantee good out-of-sample portfolio performance. XS and B-XS may be fitting middle-type observations too closely in sample, introducing noise into the PCR principal components out of sample. PV, which never tries to exploit complex structure, may avoid this problem entirely which could partly explain why it holds up relatively better in portfolio performance despite weaker masking results.

Figure 11: Incremental Sharpe Ratio vs Mean Imputation Baseline (ΔSR)



4.5. Summary of Main Findings

Table 15 brings together the key metrics across all six methods and all evaluation dimensions.

Table 15: Summary of Results Across All Evaluation Dimensions

Method	RMSE↓	Mom12m SR	Beta SR	PCR SR	PCR ΔSR	NN SR	NN ΔSR
Mean	1.000	−0.157	−0.231	−0.539	0.000	−0.130	0.000
PV	0.666	−0.147	−0.219	−0.614	−0.075	−0.386	−0.256
EM	0.871	−0.137	−0.159	−0.464	+0.075	−0.275	−0.145
XS	0.834	−0.240	−0.157	−0.041	+0.498	−0.185	−0.055
B-XS	0.699	−0.240	−0.157	−0.050	+0.489	−0.112	+0.018
CMean	0.554	−0.149	−0.163	−0.465	+0.074	−0.130	−0.000

Four findings stand out.

⇒ Observation 1: CMean shows the highest in-sample accuracy (RMSE = 0.554) while the out-of-sample portfolio performance is similar to that of mean imputation. On the other hand, PV method is accurate in sample while providing the lowest Sharpe ratios out of sample. While being precise in estimating missing values, these methods do not provide good portfolio results.

⇒ Observation 2: XS and B-XS methods are not the most accurate in the masking experiment, but the Sharpe ratios based on PCR for them are the least negative (-0.041 and -0.050 respectively). Nevertheless, as mentioned in section 4.4.3, all Sharpe ratios are negative. It should be noted that high ΔSR for XS and B-XS methods should be interpreted with caution.They just lost less money than the others, rather than provided gains. Possible reasons include their better factor structure resulting in better stability of the out-of-sample PCR forecast or overfitting on middle-type missing values in sample.

⇒ Observation 3: For both PCR and NN, R^2_OS is negative for all methods, which means that none of them is superior to the historical mean out-of-sample forecast. This observation probably stems from the lack of stratification by size and not imputation failure. Another strange finding is that NN models are inferior to PCR ones although Gu, Kelly, and Xiu (2020) find otherwise.

⇒ Observation 4: One method is not optimal across all criteria. According to statistical accuracy, CMean is the winner in-sample, but XS and B-XS yield the least negative Sharpe ratio when applied to out-of-sample PCR portfolios. When it comes to NN, the most stable choices are mean and CMean. It is up to the user to decide which one to choose.

Chapter 5: Conclusion

5.1. Main Findings

This study compared six imputation methods for missing data in cross-sectional asset pricing and tested whether the choice matters for portfolio performance. The answer is yes but the relationship between accuracy and portfolio performance is not what you would expect.

On imputation accuracy, CMean comes out on top with RMSE 45 percent below the mean baseline. PV and B-XS follow at around 30-33 percent below. EM and XS also improve on mean, though by less. All methods beat the naive baseline, which is consistent with what the literature would predict.

B-XS and XS give nearly identical imputation accuracy results. Since 94.8 percent of missing observations are at the beginning of a firm's history where no lags exist, B-XS falls back to XS most of the time. For the small fraction of middle-type missing observations, B-XS does show some advantage: RMSE 0.699 versus 0.834 for XS.

The univariate sorts show broadly similar results across methods for most signals, with the most notable variation appearing for RoE where methods exploiting cross-sectional information produce higher Sharpe ratios. We document a negative momentum premium across all methods. One possible explanation is the momentum crash of 2008 falling into the training window, though this remains a hypothesis.

On forecasting, both PCR and NN produce negative R^2_OS. The imputation choice makes almost no difference there. This reflects the absence of size stratification rather than a fundamental limitation of imputation. The NN underperforming PCR is also unexpected and warrants further investigation.

The most striking result is the PCR ΔSR: XS gives +0.498 and B-XS gives +0.489, far above all other methods. However, as discussed in Section 4.4.3, all Sharpe ratios remain negative. XS and B-XS lose less money than the others, but they are still losing money. One possible explanation is that their factor structure produces more stable out-of-sample PCR forecasts, though overfitting to MAM in sample cannot be ruled out.

5.2. Implications

The clearest takeaway is that imputation accuracy and portfolio performance pull in different directions. CMean and PV are the most accurate methods but not the best for portfolios. XS and B-XS score lower on accuracy but produce the least negative PCR Sharpe ratios. Though

all remain negative, so this should be interpreted carefully. Using the masking experiment alone to pick an imputation method would lead you to the wrong choice for PCR portfolio construction.

For practitioners building linear return prediction models, factor-based imputation appears to be the better choice even when it scores lower on accuracy metrics. For neural networks, simpler imputation works better, mean or CMean are the most suitable options.

The bivariate sort is a useful diagnostic tool. It tests whether the characteristic-return relationship is preserved across different levels of missingness. The results here show broadly consistent patterns across all methods, which suggests imputation is not dramatically distorting the momentum signal in the cross-section.

5.3. Limitations and Future Work

The study has several shortcomings that have not been taken care of:

⇒ Stratification by size is obviously one of them. This entire analysis is based on pooled cross-section estimates. In contrast, CM estimate separate models for small-cap, mid-cap, and large-cap firms. That way, they achieve a substantially higher forecasting accuracy, solving the negative R^2_OS problem.

⇒ Factor selection criteria: The choice of $K = 5$ for $L = 50$ has been made on the basis of some rules of thumb. A better approach, perhaps involving Bai-Ng factor selection criterion, would be a good addition to what we currently have.

⇒ Box-Cox: As a preprocessing technique, it transforms our signal distribution into normality prior to imputation. However, whether this transformation has any effect; particularly on EM as this method depends on normality assumption, would certainly be worth investigating.

⇒ Complete FHNW procedure: What is done here is just the imputation part of FHNW (2025), not their inference adjustment. The real contribution is adjusting the standard errors for the uncertainty associated with imputation, and that needs to be done by incorporating the imputation in some particular asset pricing model.

⇒ Tests of significance: all ΔSR differences presented here are purely descriptive. For proper tests of statistical significance: Barillas et al. (2020) or the Diebold-Mariano type test for portfolio returns can tell whether the ΔSR differences are statistically significant. Considering the low ΔSR values, there is a good chance none of them will pass.

⇒ Neural network: the ensemble uses three networks instead of ten, meaning three layers of neurons and links are trained independently and averaged. This is partly due to computational limitations but perhaps a bigger set of networks would yield better results. Another test to run: Train on imputed data and evaluate the performance on non-missing data only; vice versa.

⇒ Size-dependent masking: masking procedure is neutral to firm size. Small firms generally have more missing data and could be easier or harder to impute, we don't know yet. Testing the masking procedure stratified by firm size will reveal more clearly where each imputation strategy works best.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1), 135–171.

Bai, J., and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.

Bai, J., and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536), 1746–1763.

Barillas, F., Kan, R., Robotti, C., and Shanken, J. (2020). Model comparison with Sharpe ratios. *Journal of Financial and Quantitative Analysis*, 55(6), 1840–1874.

Black, F., Jensen, M., and Scholes, M. (1972). The capital asset pricing model: Some empirical tests. In M. Jensen (Ed.), *Studies in the Theory of Capital Markets*. Praeger.

Bryzgalova, S., Lerner, S., Lettau, M., and Pelger, M. (2025). Missing financial data. *Review of Financial Studies*, forthcoming.

Cahan, E., Bai, J., and Ng, S. (2023). Factor-based imputation of missing values and covariances in panel data of large dimensions. *Journal of Econometrics*, 233(1), 113–132.

Chen, A., and McCoy, M. (2024). Missing data in asset pricing panels. *Journal of Financial Economics*, forthcoming.

Chen, A., and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2), 207–264.

Cochrane, J. (2011). Presidential address: Discount rates. *Journal of Finance*, 66(4), 1047–1108.

Daniel, K., and Titman, S. (1997). Evidence on the characteristics of cross-sectional variation in stock returns. *Journal of Finance*, 52(1), 1–33.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38.

Fama, E., and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2), 427–465.

Fama, E., and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

Fama, E., and French, K. (2008). Dissecting anomalies. *Journal of Finance*, 63(4), 1653–1678.

Fama, E., and French, K. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

Freyberger, J., Höppner, B., Neuhierl, A., and Weber, M. (2025). Missing data in asset pricing. Working paper.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5), 2326–2377.

Goyal, A., and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.

Green, J., Hand, J., and Zhang, F. (2013). The supraview of return predictive signals. *Review of Accounting Studies*, 18(3), 692–730.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Han, Y., He, A., Rapach, D., and Zhou, G. (2022). Expected stock returns and firm characteristics: E-LASSO in high-dimensional factor models. Working paper.

Harvey, C., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.

Hawkins, D., and Weisberg, S. (2017). Combining the Box-Cox power and generalized log transformations to accommodate nonpositive responses in linear and mixed-effects linear models. *South African Statistical Journal*, 51(2), 317–328.

Kelly, B., Malamud, S., and Pedersen, L. (2023). Principal portfolios. *Journal of Finance*, 78(1), 347–387.

Koh, P., and Reeb, D. (2015). Missing R&D. *Journal of Accounting and Economics*, 60(1), 73–94.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2), 271–292.

Little, R., and Rubin, D. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

Xiong, R., and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1), 271–301.

Zhang, L., Mykland, P., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472), 1394–1411.