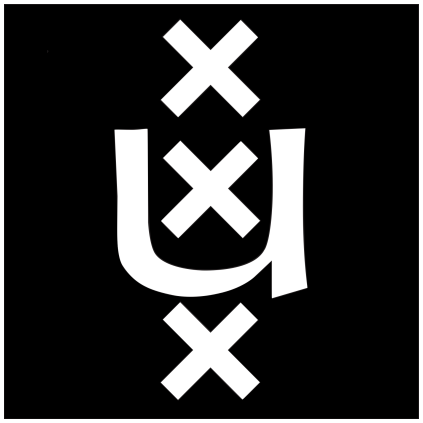


University of Amsterdam
Amsterdam Business School

MSc Finance, Quantitative Finance Track
Master Thesis

Missing Data Everywhere

Medina R. O., Eduardo

June, 2023

Thesis Supervisor: Dr. Simon Rottke

Statement of Originality: This document is written by Student Eduardo Medina R. O. who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. The Faculty of Economics and Business is responsible solely for the supervision of completion of the work, not for the contents.

Abstract

This paper reports on the missingness of observations in data for more than 120 factors found in the finance literature. No stock has all factors observed at any point in time, a problem that is highlighted by the multidimensional nature of asset pricing data. Due to the complex nature of missing patterns and their importance for inference on anomalies, I perform a comparative analysis of imputation methods for missing data in the cross-section of characteristics. Through an EM algorithm, I perform imputations for 120 signals. Imputations have a marginal effect when considering 120 signals alone, resulting in a .5% drop in the average cumulative return of anomalies with a .1 drop in Sharpe ratio. However, performance disparity becomes economically significant when considering a larger set of signals. Average signal cumulative return drops by more than 40% when considering 204 signals in the sample.

Contents

Chapter 1

Introduction

Recent literature in finance has brought attention to the pressing issue of missing data in asset pricing panels. The handling of missing data is carefully discussed in the statistics, data science, and psychology fields. Nonetheless, the issue is hardly addressed in the finance field despite being especially relevant in asset pricing, given the high dimensionality of its datasets.

A multitude of factors which explain the cross-section of returns have been published in the literature. Harvey et al. (2016) identified 316 anomalies as potential predictors of returns. This makes the task of developing relevant pricing models challenging and exacerbates the impact of missing data (Chochrane, 2011; Hou, Xue, & Zhang, 2015; Mclean & Pontiff, 2016). However, it is common practice to drop firms with missing observations or use ad-hoc imputation methods such as mean and median estimates. The result is a likely bias on estimations and a large loss of information on the cross-section of returns. Approximately 70% of all firms listed in the US have missing data, accounting for half of its total market cap (Bryzgalova, Lerner, Lettau, & Pelger, 2022), which highlights the need to proactively address the missing data problem.

One approach to reducing this bias is through improved asset pricing models. Hou, Xue, and Zhang (2015) (HXZ) as well as Feng et al. (2020) have

developed novel techniques to capture the true predictability of anomalies, significantly narrowing down the scope of the factor zoo to those with relevant predictive power. However, these approaches have not addressed the issue of missingness.

Missing variables impact asset pricing research in various ways. Bryzgalova, Lerner, Lettau, and Pelger (2022) (BLLP) shed light on the complexity of missing data. There is cross-sectional correlation across 45 firm characteristics, and the same correlation is observed in their missing patterns. Firm characteristics play a crucial role in portfolio construction. Yet, it is common for scholars and practitioners to exclude companies with missing information from their stock pools, which is impractical. For example, applying this approach to the 125 most observed predictors in the Chen and Zimmerman (2022) (CZ) dataset would result in the exclusion of over 80% of stocks (Chen & McCoy, 2023) (CM). This reduces coverage, and it eliminates important parts of the return distribution, thereby affecting Sharpe ratios and cumulative returns. BLLP further reports that anomalies exhibit increased excess returns after imputation using novel methods. However, Beckmeyer and Wiedemann (2023) (BW) show lower anomaly returns after imputations through attention-based machine learning, while others report better predictability when applying conditional mean imputations based on gerenalized method of moments (GMM) (Freyberger, Höppner, Neuhierl, & Weber, 2022) (FHNW). This shows that literature on imputations of missing data in finance is still in its early stages, and there is a lack of consensus on the best models and on the predictive power of anomalies with a complete dataset.

Hence, this paper puts forward three relevant points: (1) Missing data patterns in asset pricing datasets are complex and predominant, (2) Imputed missing observations change the performance of anomalies, and (3) performance and effects of imputation models change across datasets.

Further, contributions to the literature are threefold. First, this writ-

ing reports a comprehensive study on missingness, using a relevant dataset published by CZ. Second, a comparative study of different imputation methods, some of which were not tested before nor applied to such an extensive dataset, is performed. Third, the performance of anomalies after imputations in different subsets of data is examined, which display novel effects of imputations on the cross-section of returns.

Results reveals a high complexity in missing patterns. When considering all 204 signals, there is no point in the sample where all factors are observed simultaneously. CM's EM algorithm demonstrates superior imputations compared to other models. Furthermore, the average anomaly excess return decreases after missing observations are imputed. In a subset of 120 predictors, the effect is marginal, but when considering the 204 signals, the alpha of the average anomaly return in the Fama-French five factor model (FF5-UMD) (Fama & French, 2015a), is reduced by half (Fama & French, 2015a).

The remainder of this paper is organized as follows. The Literature Review (2) covers the scarce but growing body of research on missing data in finance. It explores alternative approaches and recent developments in the field. Data (3) provides a detailed description of the data used, focusing on a subset of the dataset by CZ and its missing patterns. The Methodology (4), introduces the EM imputation model, and formal methods used for testing its performance. Subsequently, Results (5) compares the performance of the average anomaly return with missing observations (Vanilla) to performance with the imputed dataset. The implications of my findings are reported in the Discussion (6). Finally, the Conclusion (7) summarizes the paper.

Chapter 2

Literature Review

2.1 Too Many Factors

Since the pioneering work of Sharpe (1964) and Lintner (1965) on the Capital Asset Pricing Model (CAPM), numerous asset pricing factors have emerged. Due to their simplicity and effectiveness, linear factor models explaining asset returns have gained large popularity in the finance field. However, the relevance of many of these factors is debatable. Harvey et al. (2016) draw attention to the excessive number of factors, with over 300 published in the literature, many of which are redundant. This abundance of factors, often referred to as the "factor zoo," poses a significant challenge in recent finance literature (Chochrane, 2011).

The challenge of the factor zoo can be attributed to several potential causes. First, the competitive research environment often leads to overfitting. Researchers face pressure to produce significant results in order to garner citations, grants, and publications. As a result, papers with significant findings are more likely to be published in top journals, while those with insignificant results are overlooked (Rosenthal, 1979; Fanelli, 2013). Additionally, Harvey (2017) argues that researchers may engage in p-hacking, manipulating the data and analysis process to achieve significant results.

However, even under perfect research practices, erroneous inferences can still occur when using incorrect models. Biased results often arise due to incorrect identification, particularly when factors are weakly correlated with assets (Gospodinov, Kan, & Robotti, 2014; Kan, Robotti, & Shanken, 2013). Consequently, many factors that are considered significant may actually be spurious, crowding out the effects of truly relevant factors (Bryzgalova, 2017). In response to these challenges, new modeling techniques have been developed. Sparse models, for instance, focus on a limited number of factors that explain a significant portion of the variation in asset returns. For instance, HXZ propose a q-factor model, while Fama and French (2015b) propose a 5-factor that incorporates profitability and investment factors in addition to traditional factors.

Contrastingly, dense models take a different approach by initially considering the entire universe of potential predictors and subsequently reducing the number of relevant factors using regularization techniques (Giannone, Lenza, & Primiceri, 2021). For instance, stochastic discount factor (SDF) models measure the joint explanatory power of a broad set of cross-sectional factors (Kozak, Nagel, & Santosh, 2020). Feng, Giglio and Xiu (2020) propose a novel model that assesses the marginal contribution of each factor to explaining the cross-section of returns.

Bayesian techniques have also increased in popularity, due to their precision and efficiency. Bryzgalova et al. (2023) propose pricing models through a Bayesian framework, while Barillas and Shanken (2018) propose tests for comparing competing asset pricing models using similar techniques. Gu et al. (2020) apply machine learning (ML) such as decision trees and neural networks to improve return predictions. However, even with good and improved models, the quality of the input data remains crucial. Ultimately, if the data used in the models are of poor quality or contain biases, the output will be similarly flawed.

Missing data is an additional source of biased estimates in asset pricing

(Beaver, McNichols, & Price, 2007). The multidimensional nature of panels in finance exacerbates the problem of missing data due to the complexity and interconnectedness of the variables involved. Datasets typically consist of multiple securities observed over time, and each asset is associated with various predictors. When missing data is present in one predictor, it can propagate across other variables within the panel due to the interdependence and correlation between covariates (FHNW; BLLP). More variables and dimensions in panels increase the likelihood of encountering missing data, which makes the task of handling missingness more complex.

Hence, Imputation methods need to consider the dependencies among variables and capture the joint distribution of the variables in order to generate plausible imputations. Failure to properly account for this multidimensional structure can lead to biased imputations and potentially distort the estimated relationships between predictors and the target variable.

2.2 On Missing Data

The issue of missing data has been extensively discussed in various empirical sciences. Econometrics, data science, and statistics, among other fields, have a vast body of literature that delves into missing data and imputation methods, which goes beyond the scope of this paper. However, in the context of asset pricing, I focus specifically on the literature that addresses such. It is worth noting that (Little & Rubin, 2019) offer a comprehensive overview of these topics, particularly in the realm of statistics (BLLP).

Understanding the missing mechanism, which refers to *how* and *why* data is missing, is crucial in the handling of missing observations. Given its complex structure, efficient modeling requires well-defined assumptions that revolve around the probability distribution of the missing pattern. Imputation models based on EM algorithms typically assume that data is "missing at random" (MAR), meaning that missing values are independent of the observed

data, and have the same probability of being missing across all observations (Little, 2021).

Missing completely at random" (MCAR) is a stronger assumption, where the missingness is not only independent of the observed data but is also independent of the unobserved data. In other words, the missingness occurs completely at random, without any systematic pattern. However, empirical findings generally do not support the assumption of missing data being MCAR (Xiong & Pelger, 2023).

Conversely, if the probability of missing observations is dependent on the outcome variable, it is considered "missing not at random" (MNAR) (Freyberger et al., 2022). In this case, the distribution of the missing mechanism is explicitly defined through a selection model. However, applying this approach to panel datasets with larger dimensions, as is often the case in finance, becomes infeasible (Harvey et al., 2015.).

In asset pricing, models that do not explicitly model the missing pattern distribution dominate. The GMM framework relies on moment restrictions and accounts for imputation uncertainty (Hansen, Heaton, & Yaron, 1996). Inverse probability weighting (IPW) is another approach worth noting for deriving reliable estimations on missing data in finite samples.

It is important to note that the asset pricing field still lacks a consensus on the appropriate modeling of missing data, and the choice of imputation method may depend on the specific research context and data characteristics. In fact, most papers fail to explicitly state how they deal with missing data, and a common approach is dropping incomplete observations (Lewellen, 2015; Freyberger, Neuhierl, & Weber, 2020; Kelly, Pruitt, & Su, 2019).

Notably, Beaver et al. (2007) highlight the impact of omitting observations on returns of strategies based on firm characteristics, which often involve dropping firms with missing observations. For instance, portfolios based on earnings and cash flows realize increased returns when delisted firm-years are included. As aforementioned, this problem is exacerbated by

the multidimensional nature of recent asset pricing research. As the number of characteristics and signals increases, so does the number of firms that have missing observations. Simply ignoring those has an even larger impact on anomaly performance.

2.3 Contemporaneous Literature

Kozal et al. (2020) and Gu et al. (2020) use imputation models using cross-sectional means. However, imputations are not the primary focus of their paper, and there is a lack of performance measurements provided. More recent studies on missing data include Freyberger et al. (2022), Bryzgalova et al. (2022), Beckermeyer and Wiedermann (2023), and Chen and McCoy (2023). Each paper proposes different approaches and assumptions for imputations. Interestingly, all studies, except for Chen and McCoy (2023) report economically different performance of signals after imputations.

In the paper by FHNW, they propose moment conditions to derive estimators within the GMM framework, considering missing observations. This approach is similar to the one employed by Lynch and Warter (2013). Their model incorporates both cross-sectional and temporal information of firm characteristics to perform imputations. They assume a linear regression model in which firm returns depend on a complete set of observations as well as an incomplete set. The paper explores various estimation methods, with a computationally efficient approach being the imputation of observations for parameter estimation and subsequent downweighting of imputed values. Notably, they demonstrate lower mean squared errors (MSE) compared to standard imputation methods. In addition, the paper examines the performance of anomalies in long-short portfolios, and finds higher excess returns for these anomalies (Freyberger et al., 2022). However, it is important to note that the performance of this model relies on the availability of observed data to estimate covariates.

BLLP take a different approach by employing latent factor models. They build upon Xion and Pelger (2023), who provide an estimator for latent factors that accommodates general missing patterns. This approach is also utilized by Jin et al. (2021), Cahan et al. (2023), Bai and Ng (2021). The authors estimate a weighted covariance matrix based based solely on observed values to capture cross-sectional information. By utilizing principal component analysis (PCA), they derive a latent 6-factor model. Imputations are conducted using characteristic loadings and factors, which are estimated through regressions on vectors of firm information, including characteristics and past characteristics to account for time series information. As in FHNW, they do not model the missing mechanism explicitly, and conduct a comparative study on imputation methods. Further, their results indicate positive excess returns in strategies based on anomalies after imputation.

BW take a contrasting approach to the previous studies by focusing more on machine learning (ML) applications. They adopt an attention-based ML model borrowed from natural language processing, and apply it to various masking patterns. However, the theoretical properties of the model and assumptions on patterns of missing data are not discussed. While they do report improved performance compared to other imputation methods, they do not provide a comparison with simple mean imputations, as highlighted by Chen and McCoy (2023). In contrast to BLLP and FHNW, BW find lower excess returns from strategies constructed on imputed characteristics.

Finally, CM implemented an EM algorithm for imputation purposes. Their approach assumes that cross-sectional signals follow a joint multivariate normal distribution. The expected values of the missing predictors are estimated using maximum-likelihood estimation. They compared the out-of-sample results of their method with unconditional mean imputation and found a slight improvement in Sharpe ratios and excess returns after the imputations. This is primarily due to the low correlation between the signals, indicating that the return predictors are independent. Consequently,

the performance of the EM algorithm and other methods is only marginally better than that of standard imputation methods.

2.4 Variations in Data and Methods

It is important to highlight the diversity of datasets used in recent literature on missing data, as noted by Chen and McCoy (2023). Each publication relies on vastly different datasets. BLLP use 45 firm characteristics by Kozat et al. (2020), while BW work with 151 characteristics from the open source dataset by Jensen et al. (2021), keeping stocks with at least 20% of characteristics observed. FHNW and CM use the dataset from Chen and Zimmermann (2022), although with differences in signal selection and time periods. FHNW use 82 predictors from 1978 to 2021, while CM employ 120 predictors spanning from 1985 to 2022. Notably, these differences in data can have a significant impact on imputation results.

All of the aforementioned papers note the complexity and magnitude of missing data patterns. They get even more complex when considering that empirical analyses are conducted using different data. Rules and thresholds for factor construction may differ across datasets, consequently changing the pattern of missingness. Characteristics present in one dataset may change or become missing when changing between datasets. Hence, it is important to note that performance of imputation methods may differ depending on the data itself. Additionally, it is unclear whether the relationships between predictors remain constant across different data. For instance, BLLP note the cross-sectional dependence of firm characteristics, whereas the main assumption of CM is that predictors are independent. This can have a significant impact on the implications and effectiveness of imputation methods.

Chapter 3

Data

I use the open-source dataset by CZ, consisting of 204 predictors that have been published in academic journals, mechanically missing predictors are dropped as well as redundant ones, following FHNW and CM. For instance, predictors such as long-term reversal require at least 5 years of data, naturally leading to a larger proportion of missing observations compared to other predictors. The same applies to binary and double-sorted predictors, where a predictor that is not equal to one is set to missing.

CM and FHNW utilize subsets of the predictors available in the CZ data set. CM use the 120 predictors with the most observations in 1985, and extend their analysis until 2022. Conversely, FHNW keep only 82 characteristics from 1978 to 2021. This analysis shows that the FHNW data set is contained within CM's, signals in the FHNW dataset are also present in CM. However, even when using CM's data set, it means dropping almost half of the 204 predictors from the original dataset. Although well motivated, dropped predictors have a considerable impact on the occurrence of missing values in the sample. The magnitude of missing data is highly correlated with the specific data set being used. Table 3.1 provides a summary of stylized facts regarding the three data sets: CZ, CM, and FHNW

The table shows two important considerations. Firstly, it highlights the

Table 3.1: Contasts in Data Sets

Sample Subsets	CZ	CM	FHNW
Avg. Missing %	56.2	36.5	38.2
	(25.6)	(17.9)	(15.0)
Coverage	204	120	80
Range	1925-2023	1985-2022	1978-2021

considerable proportion of missing values observed in the samples, which is a common characteristic across all papers in contemporary literature. Secondly, the average missing percentage varies depending on the specific sample used, as shown by the varying missing percentages across data sets. Additionally, through the process of dropping predictors, the average missing share decreases from 56% to 38%. This demonstrates that the data itself and the number of signals employed are factors that contribute to the presence of missing values and can potentially impact the missing mechanism.

Nonetheless, the CM subset has broader coverage and is more representative of the current zoo of predictors. Hence, all tests, imputations, and portfolios are performed based on their dataset. For robustness tests, involving a larger set of signals, I utilize the complete CZ data set.

I follow CM in changes from the original data. I exclude discrete predictors, event study-based predictors, and double sorts. As noted previously, stocks that do not belong to either the long or short portfolio for a particular predictor have machanically high missing values. This can artificially inflate the number of missing observations in the sample (CM, 2022). Moreover, highly specialized predictors and those observed fewer than two times in the sample are dropped. This screening process ensures that the missing mechanism is not erroneously affected by the behavior of those predictors. Lastly, the final data set consists of the 120 most frequently observed predictors in 1985, utilizing data from CRSP, Compustat, and IBES.

3.1 Signals

In the same vein as CM, I report the 120 most observed signals in 1985, sorted on the share of stocks with observed predictors in Table A.1 in the appendix.

Except for short term reversal (*streversal*), all predictors have missing observations. CRSP signals based on data of the previous month have shares of observations exceeding 90% (CM, 2022). Signals requiring longer periods of CRSP data and those based on accounting information have significantly lower observation rates. For instance, real dirty surplus (rds), investment growth (*invgrowth*), and cash (*cash*) are missing for more than 55% of stocks. Below 50% observation rates are predictors that often have missing data and are unable to be constructed, e.g. analyst revision (*analystrevision*), as well as those that require even longer spans of observable data, for instance long term reversal. The least observed predictor included in the sample is payout yield (*payoutyield*). As expected, these observation shares align consistently with CM's.

3.2 Missing Patterns

Missing values persist over time and a significant proportion of firm-month observations have missing observations. Figures 3.1 and 3.2 illustrate the pattern of missing data for different signals, including streversal, classic momentum (*mom6m*), asset growth (*assetgrowth*), *rds*, *cash*, and *payoutyield*. Signals within the group have varying degrees of missing observations, and rely on both accounting data and past CRSP data. This allows for examination of any distinct patterns in missingness based on the predictor type and determine if those patterns differ across data types required for the predictor. As mentioned earlier, *streversal* has 100% observed values, being ranked first. Subsequently, the other signals have 85.3%, 66.9%, 63.4%, 50.3%, and 33.4% average observation rates respectively.

Figure 3.1: Number of Firms Over Time

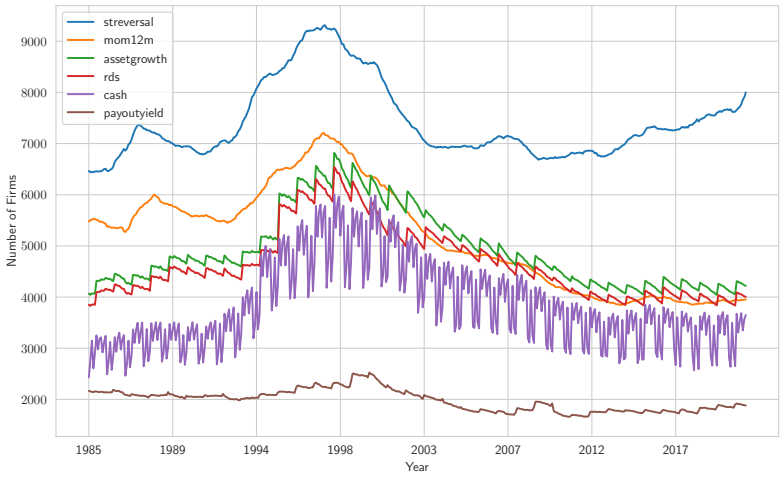


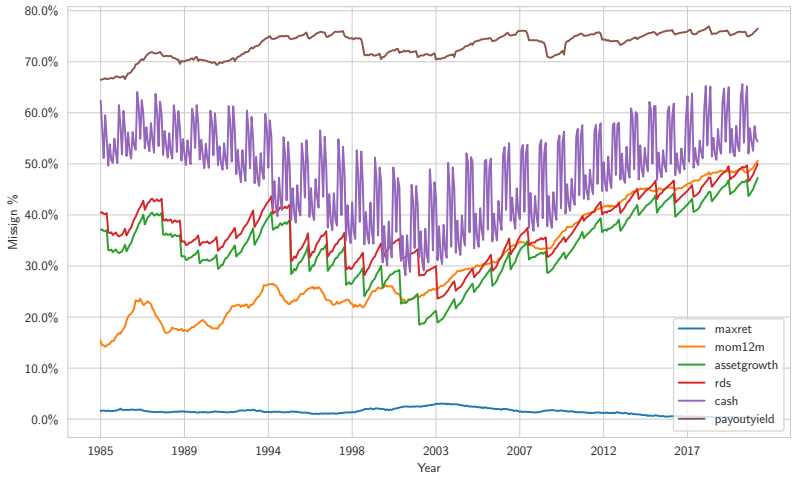
Figure 3.1 plots the number of firms per characteristic. Note that *streversal* also represents the total number of firms since it has no missing observations, and the remaining lines show the number of stocks that have the respective factor observed. The number of firms peaked around 1997 consisting of more than 9000 unique stocks due to the inclusion of Nasdaq (BLLP). The number of firms has declined over the last decades, although the trend reversed around 2012. It is worth noting that signals based on accounting data have a different pattern compared to those based on CRSP observations, as seen by the teeth-like pattern for *cash*, *rds*, and *assetgrowth*. Moreover, consistent with their proportions in table A,1, signals are observed in a varying number of unique firms. For instance, *mom12m* is observed in 7000 firms in 1998, whereas cash only has 5500 firms with observations. *Payoutyield* is roughly observed in less than 2000 stocks for the entire sample.

Interestingly, starting from 2017, the number of firms displayed an increasing trend that was not mirrored by the signal lines. This suggests that despite improved data measures and greater coverage of firms, the addition

of new public firms to the US stock exchanges outpaces the availability of observations for their respective signals.

Figure 3.2 shows the proportion of missing observations per signal. *Streversal* is replaced by the second-ranked signal, *maxret*, for easier interpretation of the graph. *Maxret* has a constant missing percentage, However, the remaining factors display a clear increase in the proportion of missing observations after 2003. This finding contrasts with the results of BLLP, where the missing percentage share of their firm characteristics decreased over time. These discrepancies highlight the inconsistency of key missingness trends across data sets. Further, across most signals, missing values consist of a significant share of total observations.

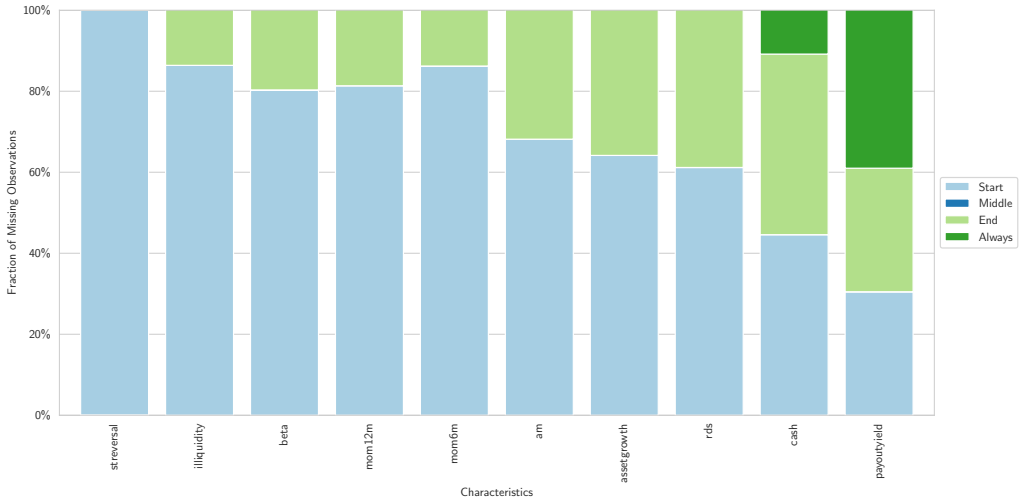
Figure 3.2: Missing Proportions Over Time



For the same set of predictors, the following Figure 3.3 depicts the missing share of signals in more detail. The percentage of missing observations is split into three categories. Across firms in the sample, signal observations that have missing values in previous months are missing at the "Start". Conversely, if a signal is observed and subsequently has missing values it is

missing at the "End". Observations that do not fall into either of these cases are labeled as "Middle". Lastly, if an observation is never observed, it is marked as "Always."

Figure 3.3: Types of Missing Variables



Missing data is distributed in different ways across signals. For the majority of signals displayed, observations are missing at the start of the sample. However, *rds* shows a considerable share of missing observations present in the end of the sample, whereas the majority of the *payoutyield* signal is never observed. *Cash* is missing in a similar way, although concentrating at the start and end of the sample. Moreover, two of the signals with always missing data are based on accounting data. The plot shows that the missing patterns can vary depending on the type of signal and data.

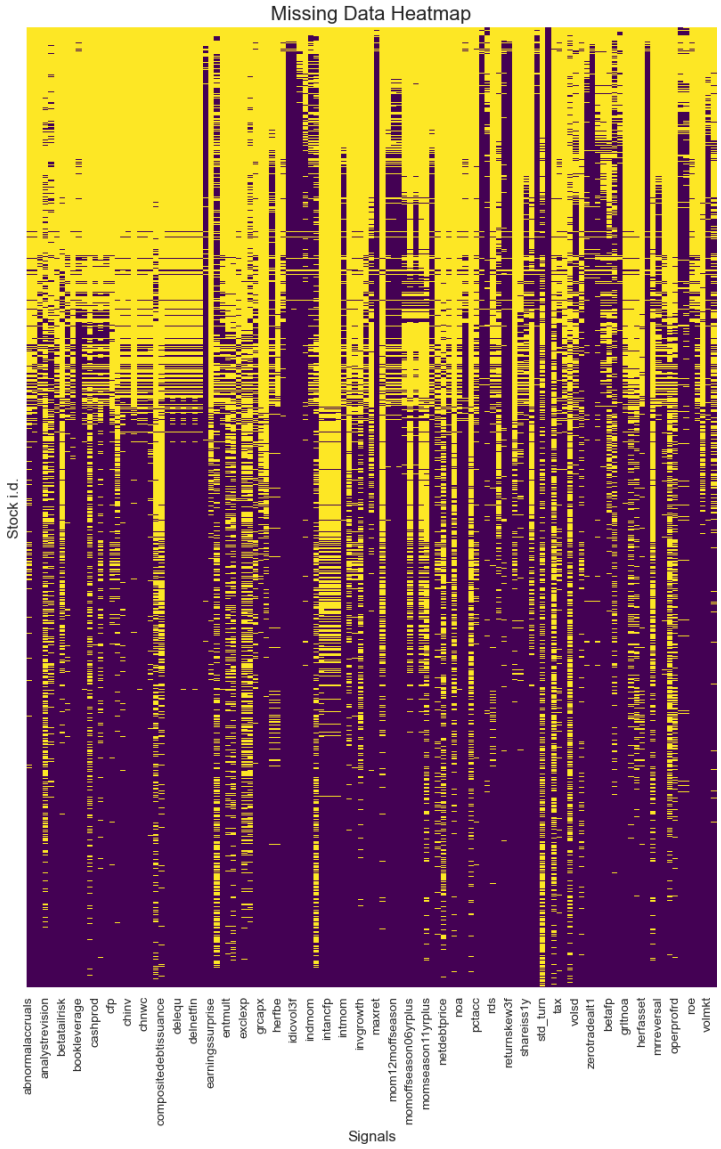
Overall, missing data is prevalent over time and across signals, constituting a significant proportion of total observations. This underscores the need for careful imputation. However, the missing pattern in the data requires thorough analysis. Consistent with the literature, I observe a complex and heterogeneous pattern of missing observations across most signals and firms

in the sample.

Figure 3.4 provides further insights into these patterns, depicting the missing and observed signals for each firm in January 1990, sorted by the percentage of missing values. The heatmap color-codes the observations, with purple representing observed signals and yellow representing missing ones. While missing observations often come in blocks, the missing pattern is mostly sporadic and unpredictable, and the majority of firms have more than one signal missing at the same time.

The issue of missingness is more apparent when multiple predictors need to be observed simultaneously. For almost every stock, there is at least one missing signal. Therefore, when dealing with a set of return predictors, which is often the case in finance, it becomes crucial to impute missing values in order to meaningfully utilize the predictors. Further, it becomes clear that missing patterns are not MCAR, as proposed by BLLP, due the apparent dependency of missing observations across signals. However, this dependency is, on average, close to zero, as it is shown in the following section.

Figure 3.4: Missing Data Across Firms and Signals in Jan-1990



Note: In the Y-axis, stock i.d. represents the unique firm identifier present for each stock in the sample. The actual identifiers are omitted. In the X-axis are the signals for which each stock has either a observed value, in purple, or a missing value, in yellow. Note that only some signals have labels, as having 120 labels would cluster the graph.

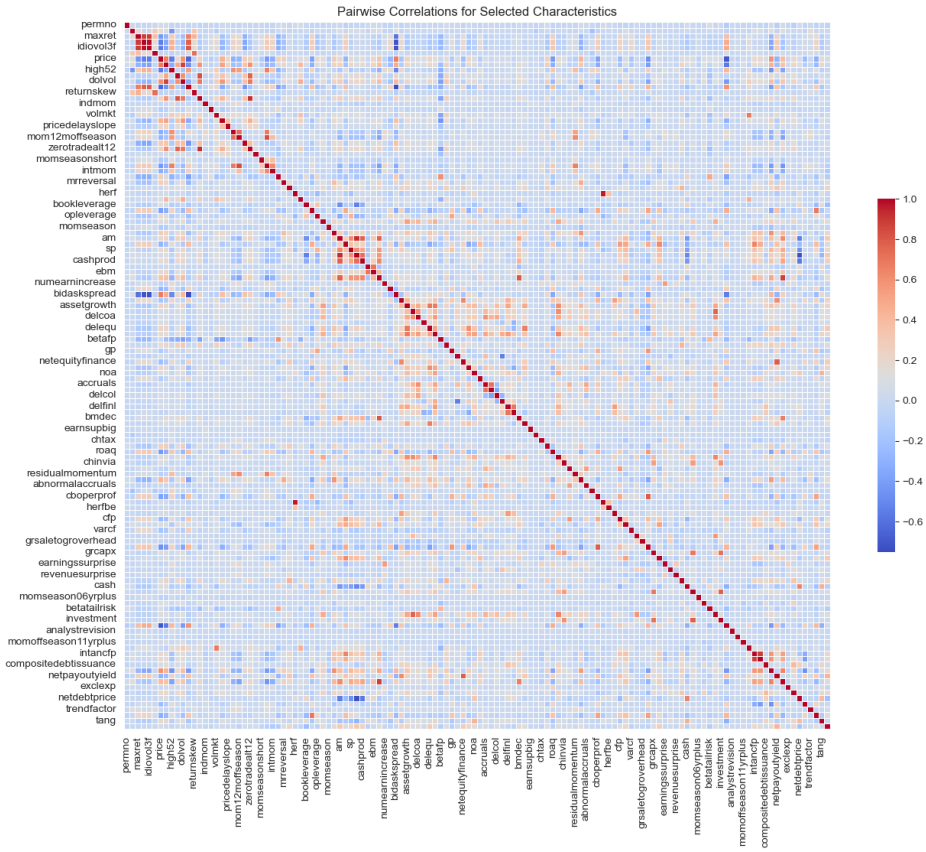
3.3 Signal Dependency

Here I report the dependency and correlations between signals in detail. These factors are relevant for any imputation model, as they offer further insights into the appropriate assumptions of the missing pattern.

The findings in this section are intriguing. Signals display low negative serial correlations across signals, and average pairwise correlations cluster around zero. This is a stark contrast to BLLP, who have characteristics with autocorrelations close to one. This may be driven by the differences in signal construction or coverage. BLLP have only a small subset of signals they miss out on a large set of information provided by other predictors, and that can be a factor influencing the correlations in the data set. Furthermore, BLLP rely on Kozak et al.'s (2020) dataset, which likely has different rules for signal construction compared to CZ'S.

Nonetheless these differences are intriguing, statistical properties like such should be consistent across datasets. A more detailed examination of characteristic construction in BLLP is necessary and additional tests are needed to understand the drivers of said changes. I leave this task for future research.

Figure 3.5: Pairwise Correlation Matrix



Note: The heatmap shows signal correlations in the sample. Negative correlations are depicted in light blue, with negative correlations having a darker shade of blue. Strongly positive correlations are in bright red, as described by the legend in the right.

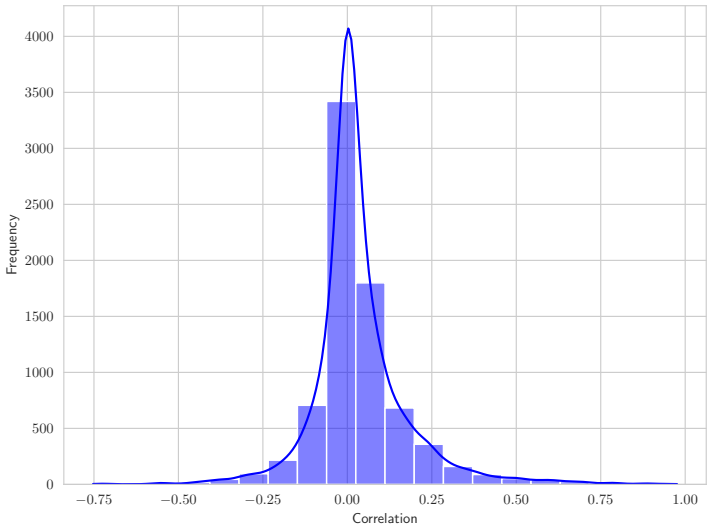
Correlations between all signals in the sample are displayed by the heatmap in Figure 3.5. There is cross-sectional dependency between factors, though to a small extent. Strong correlation clusters in red or dark blue are few. For instance, *am* and *ebm* are highly positively correlated, represented in bright red. However, the majority of signals has a correlation close to zero, characterized by the predominant light blue shades in the headmap.

Again, this runs counter to the impact the cross-section has on missing observations, argued by BLLP. Instead, our findings are consistent with CM,

who find a pairwise correlation between predictors clustering around zero. CM argue that near-zero correlations are a result of the high dimensionality of data, leading to a more flat distribution of returns. CM's dataset has more than double the predictors than BLLP's which may also contribute to the decrease in correlations. This is intuitive once we consider the variety of signals in the dataset, many of which are based on diverse factors and data.

Figure 3.6 plots average pairwise correlations quantitatively in a histogram. As suggested by the heatmap, average correlations between signals are clustered around zero. However, it is important to note that the correlations have a slight positive skew. This suggests a degree of cross-sectional dependence, albeit marginal, allowing for the assumption of a MAR pattern in the data for the implementation of the EM algorithm (CM, 2023)

Figure 3.6: Pairwise correlation Distribution



Note: On the Y-axis frequency represents the number of correlation pairs. Along the X-axis are correlation coefficients, ranging from -0.25 to 1. The bars can be interpreted as follows: take the highest bar, it shows that around 3400 correlation pairs have a correlation of roughly zero.

In summary, three important conclusions are drawn from the careful study of the data. First, across time and firms, missing observations are persistent and impactful. Second, missing patterns are complex, and vary across signals. Third, missingness depends on data, in this high-dimensional setting, dependency across signals is weak.

Chapter 4

Methodology

A relevant imputation model should be able to handle the complex patterns of missingness observed in the data. I follow CM's model and implement an EM algorithm. In what follows, I formalize their model they along with its necessary assumptions as well as the measures for imputation performance.

4.1 Expectation Maximization

EM estimates parameters of probabilistic models with unobserved variables. It is an iterative algorithm that alternates through the expectation step (E-step) and the maximization step (M-step) (Nielsen, 2000). The former consists of computing the expected values of missing variables with parameters derived from the observed data. In the second, the model updates the parameters to maximize the log-likelihood of the observed data. This optimization problem is solved to find new parameter values and perform the E-step again. The process repeats itself until convergence. In other words, for each month of observations, the model uses the observed values in the cross-section of signals as parameters and gets estimates for missing values. Similar to FHNW and BLLP, the model relies on the observed cross section at time t to estimate missing values.

MAR holds in the observed missing mechanism, and due to the weak signal dependence among the predictors, one can assume *ignorability* Rubin (1976). Meaning, the process that causes missing data does not need to be modeled. Hence, following CM, we assume that in a set month t, stock i has a vector of predictors X, which follow

$$X_{i,t} \sim f(x_{i,t}|\theta_t) \tag{4.1}$$

Where x_{it} is the observed vector and θ_{it} is a vector of parameters. Given the parameters that maximize the likelihood function based on observed data, missing values are imputed with respect to their conditional expected values

$$\hat{x}_{i,t}^{\text{miss}} = E\left(X_{i,t}^{\text{miss}} \middle| X_{i,t}^{\text{obs}} = x_{i,t}^{\text{obs}}, \hat{\theta}_t\right) \tag{4.2}$$

In which $\hat{x}_{\text{it}}$ is a vector of estimated missing observations after maximization for stock i at time t. In other words, this vector represents the imputed values across all signals that are missing for firm i at time t, that were estimated through the EM algorithm on parameters based on cross-sectional information.

The maximization problem is complex. To improve efficiency, predictors are winsorized at 1% and standardized to have a mean of zero through Hawkins and Weisberg's (2017) methods. We argue, then, that the data follows a multivariate normal distribution (MVN)

$$f(x_{i,t}|\theta_t) \sim \text{MVN}(0, \Sigma_t) \tag{4.3}$$

One may argue that this assumption is somewhat restrictive. However, it has been shown to work effectively even when the underlying observations do not follow a normal distribution, and it has also been utilized by BLLP and others in the field.

Lastly, it is worth mentioning that implementing the model itself can be challenging, and it becomes computationally expensive in larger dataset. It

would require several hours to impute just one month, and days for a year. As a result, I opted to utilize the final imputed dataset provided in Chen's website. For a more comprehensive discussion of the model, its assumptions, and its implementation, I refer to (Chen & McCoy, 2023).

4.2 Imputation Performance

As stated above, the high-dimensionality of returns is computationally demanding. I did not have the hardware resources necessary to perform tests in the entire dataset. With over 3,000,000 rows and 120 columns, the process of imputing one month of data alone would take several hours.

To circumvent the problem I built a Python script that randomly selected a set of random signals. This set was masked with respect to three distinct patterns: MAR, MCAR, and BLOCK, where the latter assumes that missing values occur consecutively in a block. Missing patterns in blocks were often observed in certain signals in the data. Subsequently, masked values are imputed following various imputation methods.

For robustness, I applied the framework of Monte Carlo simulations and repeated this process multiple times for each imputation method. Finally, I calculated the average MAE for each method across all simulations and presented them in the table below

After data transformations, the signal means should be equal to zero, making for an easy interpretation of the MAE. In this setting, it measures the absolute difference between the imputed values and the true values in the masked subset of data. The performance of various imputation methods are evaluated, including the EM algorithm, ad-hoc methods common in asset pricing, e.g. mean and median, as well as less usual ones such as the previous value (PV) are reported.

It is important to note that although some methods may appear to have a zero score in the table, they are not truly zero but instead have extremely

Table 4.1: Imputation Errors

	Imputaion MSE		
	MCAR	MAR	BLOCK
EM	0.00	0.00	0.00
KNN	0.16	0.02	0.00
MICE	0.16	0.01	3.47
Mean	0.17	0.17	0.00
Median	0.16	0.17	0.00
PV	0.61	0.06	1.98

small MAE values. Their scores are rounded to zero for better notation.

The performance of different imputation methods was varied. PV performed the worst under MCAR but surprisingly well in MAR. Mean and median imputation methods had similar performances under both MCAR and MAR patterns. Additionally, they performed exceptionally well under the BLOCK pattern. Likelihood-based models generally outperformed ad-hoc methods, except for MICE under BLOCK which had the highest average MAE of 3.47. K nearest neighbors with the K=10 parameter, and MICE had performances comparable to EM under MAR. However, the EM algorithm consistently outperformed other models in all subsets, with MAE values close to zero under every masking pattern. Overall, despite being superior, the improvement from ad-hoc methods is marginal, consistent in CM.

Overall, the findings are consistent with results from other contemporaneous papers on imputation models. Common imputation methods yield less precise estimates than more complex ones, tailored to handle intricate patterns. While the improvement against ad-hoc methods is marginal, the superiority of the EM algorithm over other methods is clear, consistent with CM's findings.

4.3 Other Signals and Portfolio Construction

A small note is worth reading my subsequent analysis of anomaly returns. As stated before, I argue that missing patterns are highly dependent on data, such that the performance of imputation methods should vary across different samples.

Consequently, imputations are compared in two instances. First, I compare the imputed signals of the 120 predictors to the non-imputed signals of the same 120 predictors. This allows me to assess the impact of imputation on the performance of a subset of signals. Second, I use the entire set of 204 predictors in the same time window and compare the average returns of the original signals (all 204) to the case where 120 of them have imputed observations. This analysis takes into account the added cross-sectional variability, and sheds light on how the model performs out of sample.

Moreover, note that each long-short signal portfolio on the signal follows rebalancing and weighting criteria used on their respective original paper (Chen & Zimmermann, 2022). Due to the wide variety of signals, applying one method to all signal portfolios can often lead to erroneous results. Furthermore, I calculate the average characteristic return, which represents the return of a portfolio holding an equal-weighted position in every signal.

This paper, then, proceeds to discuss the results and implications of using imputed data for the set of predictors, taking into consideration the findings from the aforementioned analysis on data and imputation methods.

Chapter 5

Results

Here, I perform imputations on the selected signals comparing Sharpe ratios, alphas, and cumulative returns of the average anomalies. My results are in line with CM and BM. After imputations the excess returns of anomalies decrease, but to a small extent. However, when comparing the average returns of the entire set of 204 predictors, the difference becomes economically significant.

5.1 Cumulative returns

Figure 5.1 presents the cumulative return of the average return of the 120 signals using the original data (vanilla) and data with the imputed missing observations obtained through the EM algorithm (imputed). The cumulative returns are essentially the same until 1995. Subsequently, vanilla cumulative returns are slightly higher than the imputed. After the year 2000, the difference in performance grows larger. Surprisingly, both perform well during the financial crisis period, with the imputed return still showcasing weaker performance. The higher volatility at the end of the sample is also expected due to the COVID pandemic. Throughout the sample period, vanilla signals have consistently performed better, especially in the latter half. These

Figure 5.1: Avg. Cumulative Returns of 120 Signals Compared

Note: Figure plots the cumulative returns of the vanilla (blue) 120 signals and imputed (orange) 120 signals spanning from 1985 until 2022. The Y-axis represents the dollar value of investment, and the X-axis dates. Returns consist of an equal weighted position on all 120 signals, hence average signal return. The graph shows the cumulative return to investing 1 dollar on the respective signals during the sample

findings suggest that imputing missing values using the EM algorithm has a noticeable effect on the cumulative return of the average characteristic, leading to a slightly lower performance compared to using the original data.

In Table 5.1, the cumulative return and Sharpe Ratio (SR) of the average signal returns are reported, considering both the subset of 120 signals and the entire dataset of 204 signals. The results show that differences in performance become more pronounced when considering the full set of predictors. Additionally, performance increases significantly when utilizing more predictors.

Table 5.1: Sharpe Ratios and Returns Compared

	Sharpe Ratio	Cumulative Return (%)
	120 Signals	
Vanilla	0.36	0.22
Imputed	0.26	0.17
	204 Signals	
Vanilla	1.97	2.30
Imputed	1.69	1.11

5.2 Performance in Asset Pricing Models

The subset of 120 signals shows marginal performance changes between the vanilla and imputation cases. Average cumulative return and SR of the vanilla signals are .22% and .36 respectively. The imputed case has a marginal decrease in performance, dropping by roughly 10% in both metrics. As stated before, this is in line with CM. But they find small improvements in return predictability after imputations. Hence, despite the magnitude of the effect being consistent with their results, its direction is not.

In contrast, SRs and returns for 204 predictors are considerably higher. Vanilla average signals yield returns of 2.3%, and cumulative returns are equal to 1.11% for the imputed case. Moreover, SRs for both cases are higher than 1, showing strong improvement in performance through the addition of signals, with it being likely driven by higher diversification. Noticeably, differences in returns are greater, with returns falling by more than 50% in the case for 204 predictors. Hence, it is likely that imputations are not as precise when using 204 predictors. Further, additional predictors can change the missing patterns in the data and add new covariates with imputed data.

These results indicate that the imputed dataset leads to significantly lower cumulative returns and lower SR compared to the vanilla case. This suggests that the imputation of missing values can have a notable impact on the

performance of the average signal. This contrasts the published results by CM. The imputations through the EM model lead to considerably lower anomaly returns when considering a larger set of predictors.

Subsequently, average anomaly returns are explored in more detail by running several regressions on the average signal return. This provides a more comprehensive measure of performance compared to SRs and cumulative returns. I employ four models: the Capital Asset Pricing Model (CAPM), the Fama and French three-factor model (FF3), their extension of the model with the additional profitability (RMW) and investment (CMA) factors (FF5), and the modified FF5 model with the inclusion of the momentum factor by (Carhart, 1997) (FF5-UMD)

The regression table below presents the alpha coefficients multiplied by 100 for each of the mentioned models. The CAPM and the subsequent models assume that all returns should be explained by the components in the model. Therefore, a significant coefficient for alpha represents the average signal return that remains unexplained by the factors in the model. This is interpreted as the average outperformance solely attributed to betting on the signals.

Table 5.2: Alphas for 120 Signals

Coefficient	Model			
	CAPM	FF3	FF5	FF5-UMD
Panel A: 120 Signals Vanilla				
α	0.11	0.12	0.13	0.11
t(α)	(2.176)	(2.406)	(2.357)	(2.131)
Adj. R^2	0.025	0.055	0.058	0.064
Panel A: 120 Signals Imputed				
	CAPM	FF3	FF5	FF5-UMD
α	0.10	0.12	0.13	0.12
t(α)	(1.946)	(2.343)	(2.479)	(2.266)
Adj. R^2	0.029	0.088	0.093	0.098

Note: Panel (A) reports alpha coefficients for the non-imputed signals. Panel (B) shows results for the imputed signals using EM. Average signal return can be interpreted as the return of an equal weighted position on the 120 signals in the CM subset. Alphas are multiplied by 100 and their respective t-stats are in parenthesis. CAPM is a simple regression of the average signal return on the excess return of the market from 1985 until beginning 2022. FF3 adds two additional factors to the regression, HML and SMB. FF5 includes RMW and CMA to the previous three factors. FF5 has the addition of the Carhart momentum factor, UMD. Since, analysis is focused in comparing performance between imputed and vanilla cases, factor coefficients are not reported. Nonetheless, their behaviour is consistent with previous results in the literature. All asset pricing factors were retrieved from French's website.

Table 5.2 presents the alphas and their t-scores (in parenthesis) for the respective models on the 120 signals subset. In the restrictive dataset of 120 predictors, my findings align with CM's paper, imputed data has a marginal impact on return predictability. Panel (a) reports the alphas of the vanilla dataset. Betting on anomalies outperforms other factors in the models, depicted by the statistically significant alphas. Further, alphas in all model specifications are significant at the 5% confidence level. They are, addition-

ally, robust to the addition of other factors. Despite falling from .13% to .11% with the addition of factors, anomalies display positive and significant alphas in every morel. Thus, without imputations, anomalies consistently generate abnormal returns in the sample.

Panel b displays the coefficients of the same 120 signals with imputed missing values. With the exception of the CAPM model, the t-stats remain relatively unchanged and the alphas are virtually the same. Imputations seem to only marginally impact performance within the 120 predictors. Notably, the Adj. R-Squared of the imputed data regression models improves, compared to the vanilla data. Across all specifications, models with imputed data explain a larger share of variation in returns. For instance, while the FF5-UMD initially has an adj. R2 of 6.4%, the score increases to 9.8% with imputed data. showing that by reliably estimating missing data one can better explain the variation in returns.

Table 5.3: Alphas for 204 Signals

Coefficient	Model			
	CAPM	FF3	FF5	FF5-UMD
Panel A: 204 Signals Vanilla				
α	0.59	0.57	0.49	0.44
t(α)	(18.47)	(19.36)	(17.74)	(20.836)
Adj. R^2	0.34	0.46	0.55	0.74
Panel A: 204 Signals Imputed				
	CAPM	FF3	FF5	FF5-UMD
α	0.30	0.30	0.29	0.25
t(α)	(12.792)	(12.759)	(11.730)	(12.338)
Adj. R^2	0.041	0.037	0.093	0.38

Note: This table follows table 5.2 in every aspect. Except that for average signal returns, an equal-weighted position on all 204 signals is taken. Similarly, no imputations are made in vanilla, and in imputed 120 of the 204 signals are imputed using EM.

Table 5.3 presents the alphas for the entire set of 204 signals. For both vanilla and imputed cases, alphas are considerably larger. In panel (a), the alphas range from 0.59% to 0.44%, while in panel (b), they range from 0.30% to 0.25%. Furthermore, the alphas remain highly significant in both panels, and the average anomaly return aligns in magnitude with the findings in CZ.

As noted earlier, the larger set of signals demonstrates a more pronounced performance difference between vanilla and imputed data. All models using the imputed data display significantly lower alphas. In the case of FF5-UMD, for example, the vanilla signals have an alpha of 0.44% compared to 0.25% in the imputed data, resulting in a nearly 50% reduction in excess returns. Moreover, Adj. R-Squared decreases when missing variables are imputed. While subtle changes in performance and explanatory power are

expected when using different data subsets, the magnitude of these changes is surprising. Several factors that may contribute to this change are thoroughly discussed in the subsequent analysis.

Chapter 6

Discussion and Implications

The disparity between the alphas of imputed and vanilla strategies is striking but expected. As emphasized throughout this paper, the presence of missing observations across predictors, their persistence over time, and their substantial proportion in the total dataset can introduce biases in the parameters estimated on anomaly returns. The literature has provided support for this idea, showcasing different anomaly performances after imputations, as observed in this study. Interestingly, there is conflicting evidence regarding the magnitude and direction of the performance change resulting from imputations. While the specific reasons for these disparities in performance changes are not yet fully understood, they may arise from various factors such as the specific imputation method employed, the missing data patterns, the nature of the predictors, and the underlying data generating process.

I argue here that inconsistent results are expected once we carefully consider the implications of imputing data. It is important to note that the cross-section of returns is in itself a complex set of data and covariates. Understanding its mechanism and even changing them is, therefore, a challenging task. For instance, consider that one decided to drop the stocks which have unobserved characteristics. Naturally, these stocks contain important information on the distribution of returns, therefore by dropping them one

loses that part of the distribution. By imputing missing variables, one aims to retrieve the lost information on returns. However, this assumption holds true only if imputations were free from estimation errors. In reality, the imputed observations cannot perfectly capture the behavior of returns. In fact, the type of information that is imputed is highly dependent on its model. For instance, all five contemporaneous papers that tackle missing data in asset pricing have contrasting model designs, assumptions, and data. Their relevance depends heavily on these parameters, and therefore, it becomes challenging to assess the external validity of any contemporaneous imputation model.

Furthermore, the problem is exacerbated when considering the impact that imputations have on the cross section. In table 6.1 I report the average signal return against the (FF5-UMD) in more detail, now considering all six factor components. In the model, each coefficient represents the contribution or loading of the corresponding factor to the average signal returns, and alpha represents the average signal return that is not explained by the six factors. In other words, it is the average signal return when all factor exposures are equal to zero. For instance, the SMB factor in 120-Vanilla measures the impact of the small minus big factor on the average signal's return. The factor is significant at the 5% level, meaning that on average, holding smaller stocks (shorting larger stocks) contributes to 7.06% higher signal returns.

Table 6.1: FF5-UMD Model for 120 and 204 Signals

Coefficient	Subsets			
	120-Vanilla	120-Imputed	204-Vanilla	204-Imputed
α	0.11	0.12	0.44	0.25
	(2.131)	(2.266)	(17.742)	(12.338)
MktRF	2.60	2.09	-5.73	-0.96
	(2.09)	(1.632)	(-11.388)	(-2.022)
SMB	7.06	7.04	-0.42	0.22
	(3.660)	(3.670)	(-0.558)	0.306
HML	-1.08	-3.56	8.03	3.59
	(-0.441)	(-1.468)	(8.422)	(3.980)
RMW	2.59	0.24	5.97	1.61
	(1.044)	(0.098)	(6.154)	(1.756)
CMA	-6.52	-6.56	12.32	-2.67
	(-1.765)	(-1.785)	(8.534)	(-1.951)
UMD	2.00	1.86	9.04	7.19
	(1.655)	(1.544)	(18.121)	(16.083)
Adj. R^2	0.025	0.098	0.745	0.381

Note: The table reports the FF5-UMD in four specifications. From left to right, the first two use the 120 signals for average signal returns, while the last two use 204, following the same criteria and intuition from the previous regression tables. Factor loadings are now reported to understand their behaviours when there are imputations. Following other tables, all coefficients are multiplied by 100, and t-scores are given in parenthesis.

When comparing the 120 Vanilla and Imputed cases, as aforementioned, alphas stay constant in magnitude and significance. Imputations, however, impact the factor loadings of the average signal return. Take the robust minus weak (RMW), and high minus low coefficients (HML). Despite not being significant, both showcase considerable changes in their magnitudes and t-values. Where the coefficient of the former decreases from -1.08% to -

3.56% and the latter from 2.59% to 0.24% between vanilla and imputed data. Hence, imputations added or inlfuenced cross-sectional information in the distribution of returns which decreased the impact of HML and RMW factors on the average anomaly return. Alternatively, one can consider imputations as an added covariate in the cross-section, and its correlations with other factors impacts the loadings on the average signal return. When considering the cross-section of only 120 signals, imputations increase the explanatory power of the model, providing an improved distribution of the cross-section of returns, as showcased by the increased Adj. R-Squared measures also highlighted in the previous table.

The impact of imputations on factor loadings becomes more apparent when considering the models for the 204 signals. However, before analyzing the last two regressions in more detail, it is important to note that the 120 imputed predictors were based on an EM model for 120 signals, imputations will be less precise when using a set of 204 signals, since estimation was done for 120 only. Nonetheless, the end goal of imputations is to reasonably estimate missing information. A relevant imputation model should provide good enough estimates that are robust to a larger set of predictors or different data altogether, even if estimation was done using a smaller set of covariates.

However, I find the opposite when comparing the regressions for the entire set of 204 signals. Coefficients and alphas change considerably more in both magnitude and significance. From the vanilla to the imputed case, the market coefficient increases from -5.73% to -0.96%. Despite it still being strongly significant, its t-value falls considerably. Moreover, both SMB and CMA coefficients have their signs flipped between the vanilla and imputed case, and the latter with a considerable decrease in magnitude, from 12.32% to -2.67%. SMB stays insignificant, whereas CMA goes from strongly significant with a t-value of 8.434 in the vanilla dataset, to marginally significant, showcased by the decrease to -1.951 in its t-value. HML also decreases considerably, by roughly 4%, with UMD following the same pattern.

It is challenging to determine if those changes are due to a decreased bias in estimation or the product of added noise in the sample. The large decrease in Adj. R-Squared between the vanilla and imputed models on 204 predictors leads me to believe it is the latter. Imputations, in that case, likely added additional noise to the data, and the imputed model only explains 31% of return variation, compared to 74.5% in the vanilla model. Imputations do not add to the explanation of the cross-section in the second set of models. Hence, this supports the aforementioned argument on the role of data on imputation models.

The main issue pertains to the data used to train the model and its performance measures. Contemporaneous writings, despite performing tests on different missing patterns as well as, in-sample, and out-sample datasets, they overlook test on the robustness and consistency of their models. Cross-sectional information should remain relatively consistent across datasets. If small firms consistently outperform larger ones based on a particular signal or factor, they should exhibit outperformance in every dataset containing that signal. Therefore, a reliable imputation model should yield consistent performance across different datasets, as the underlying information being retrieved does not change, assuming that it is able to correctly estimate the missing cross sectional information.

Moreover, loading differences may be driven by confounding factors between missing variables and signals. Confounding factors are variables or factors that would be related to both the missing mechanism and signal returns. These factors can introduce bias and influence the imputation results. For example, if the missingness of certain variables is related to firm size, and size is also associated with anomaly returns, then imputing missing values without accounting for this confounding factor could lead to biased results.

My findings suggest that the EM algorithm fails to provide reliable estimations for missing values in a large set of signals. Furthermore, to gain a better understanding of imputation models, a more comprehensive set of tests

is necessary, and a more detailed analysis of the relationship between missing patterns and signal returns is due. It is crucial to explore how imputations impact the cross-section of data. As reported in this study, imputations have significant impact on the six factor loadings in the FF5-UMD model.

I show that current testing is too restrictive, and does not guarantee a robust performance of models. Nevertheless, my analysis comes with a series of limitations

My conclusions are limited primarily by a lack of generalizability. The tests, datasets, and models I have utilized are too limited in number to draw definitive conclusions regarding the impact of imputations on anomaly returns. For instance, the impact of imputations may differ if using alternative asset pricing models such as a stochastic discount factor (SDF) model or ones using alternative factors (Huang et al., 2022).

Moreover, while I argue that lower performance can be attributed to estimation noise, one may argue that different results would be found if using an alternative model or dataset. Indeed, the reliability and effectiveness of an imputation model may vary depending on the characteristics of the data. Hence, despite hinting at the effect of imputations on the cross-section of returns, and on the difference of their performance across datasets, a more comprehensive analysis is warranted.

I leave for further research a more robust analysis of contemporary models using diverse signals and datasets. Additionally, a detailed examination of the impact of adding additional covariates through imputations on the remaining cross-section of returns would be valuable.

Chapter 7

Conclusion

Missing observations have a significant impact on asset pricing panels, both quantitatively and economically, posing challenges for researchers in understanding risk and return drivers. The problem is further exacerbated by the multitude of factors currently employed, especially when analyzing groups of predictors. Existing practices for handling missing values, such as dropping observations or imputing with industry means or signal medians, result in a substantial loss of valuable cross-sectional and time series information.

The literature on asset pricing has largely overlooked this issue, with only a handful of contemporary papers explicitly addressing missing values and proposing imputation models. However, these models and tests are not without their limitations. I argue that the performance of an imputation model is highly dependent on the type of data used for estimation. This is demonstrated by the effects of imputations on anomaly performance across datasets, as well as their impact on factor loadings. While the EM algorithm proves reliable for a 120-signal panel, its precision diminishes when applied to a larger set of 204 signals. Additionally, in both 120 and 204 signal cases, imputations result in lower average anomaly returns. However the estimates yield substantial differences in anomaly performance between the two datasets.

My findings highlight the unreliability of current tests and focus on the limitations of imputing variables. Imputations involve estimating cross-sectional information that significantly influences returns and their behavior, as evidenced by changes in factor loadings in signal returns following imputations.

Given that the ultimate objective of imputation studies is to provide a dependable and efficient method for recovering missing observations, implementation of complex imputation models should be done with caution. An inefficient imputation model can introduce biases comparable to our existing approaches for handling missing values. Taking this into account, simple methods of substituting missing values for means and medians become, ironically, valid alternatives. Future researchers should carefully evaluate the impact of imputing variables using sophisticated models for imputations.

Appendix A

Appendix

Table A.1: Signal List

Signal Name	Author	% Obs.	Rank
streversal	Jegadeesh 1989	100.0	1
maxret	Bali, Cakici, and Whitelaw 2010	98.3	2
idiorisk	Ang et al. 2006	97.7	3
idiovol3f	Ang et al. 2006	97.6	4
returnskew3f	Bali, Engle and Murray 2015	97.6	5
price	Blume and Husic 1972	97.5	6
size	Banz 1981	97.5	7
high52	George and Hwang 2004	97.5	8
zerotradealt1	Liu 2006	97.2	9
dolvol	Brennan, Chordia, Subra 1998	96.2	10
idiovolaht	Ali, Hwang, and Trombley 2003	95.0	11
returnskew	Bali, Engle and Murray 2015	93.8	12
zerotrade	Liu 2006	93.3	13
indmom	Grinblatt and Moskowitz 1999	92.7	14
coskewness	Harvey and Siddique 2000	90.9	15
volmkt	Haugen and Baker 1996	90.7	16
pricedelayrsq	Hou and Moskowitz 2005	90.6	17
pricedelayslope	Hou and Moskowitz 2005	90.6	18
mom6m	Jegadeesh and Titman 1993	89.4	19
mom12moffseason	Heston and Sadka 2008	88.7	20
illiquidity	Amihud 2002	88.6	21
zerotradealt12	Liu 2006	88.1	22
pricedelaytstat	Hou and Moskowitz 2005	87.2	23
momseasonshort	Heston and Sadka 2008	85.6	24
mom12m	Jegadeesh and Titman 1993	85.3	25
intmom	Novy-Marx 2012	84.8	26
beta	Fama and MacBeth 1973	81.4	27
mrreversal	De Bondt and Thaler 1985	79.6	28
shareiss1y	Pontiff and Woodgate 2008	79.5	29
herf	Hou and Robinson 2006	79.1	30
volsd	Chordia, Subra, Anshuman 2001	77.0	31
bookleverage	Fama and French 1992	76.2	32
roe	Haugen and Baker 1996	76.2	33
opleverage	Novy-Marx 2010	75.9	34
hire	Bazdresch, Belo and Lin 2014	75.6	35
momseason	Heston and Sadka 2008	72.8	36
momoffseason	Heston and Sadka 2008	72.7	37
am	Fama and French 1992	72.4	38
cf	Lakonishok, Shleifer, Vishny 1994	72.4	39
sp	Barbee, Mukherji and Raines 1996	72.1	40

Signal Name	Author	% Obs.	Rank
leverage	Bhandari 1988	72.0	41
cashprod	Chandrashekar and Rao 2009	71.8	42
bpebm	Penman, Richardson and Tuna 2007	71.7	43
ebm	Penman, Richardson and Tuna 2007	71.7	44
bm	Rosenberg, Reid, and Lanstein 1985	71.2	45
numearnincrease	Loh and Warachka 2012	71.2	46
tax	Lev and Nissim 2004	69.0	47
bidaskspread	Amihud and Mendelsohn 1986	67.8	48
xfin	Bradshaw, Richardson, Sloan 2006	67.3	49
assetgrowth	Cooper, Gulen and Schill 2008	66.9	50
chinv	Thomas and Zhang 2002	66.9	51
delcoa	Richardson et al. 2005	66.9	52
dellti	Richardson et al. 2005	66.9	53
delequ	Richardson et al. 2005	66.8	54
dnoa	Hirshleifer, Hou, Teoh, Zhang 2004	66.8	55
betafp	Frazzini and Pedersen 2014	66.7	56
volumetrend	Haugen and Baker 1996	66.7	57
gp	Novy-Marx 2013	66.6	58
grltnoa	Fairfield, Whisenant and Yohn 2003	66.6	59
netequityfinance	Bradshaw, Richardson, Sloan 2006	66.6	60
totalaccruals	Richardson et al. 2005	66.5	61
noa	Hirshleifer et al. 2004	66.5	62
equityduration	Dechow, Sloan and Soliman 2004	66.4	63
accruals	Sloan 1996	66.4	64
chnwc	Soliman 2008	66.4	65
delcol	Richardson et al. 2005	66.4	66
chnncoa	Soliman 2008	66.2	67
delfinl	Richardson et al. 2005	66.2	68
delnetfin	Richardson et al. 2005	66.2	69
bmdec	Fama and French 1992	66.2	70
cheq	Lockwood and Prombutr 2010	65.0	71
earnsupbig	Hou 2007	64.4	72
indretbig	Hou 2007	64.4	73
chtax	Thomas and Zhang 2011	63.6	74
rds	Landsman et al. 2011	63.4	75
roaq	Balakrishnan, Bartov and Faurel 2010	63.1	76
investppeinv	Lyandres, Sun and Zhang 2008	62.7	77
chinvia	Abarbanell and Bushee 1998	62.4	78
pctacc	Hafzalla, Lundholm, Van Winkle 2011	62.0	79
residualmomentum	Blitz, Huij and Martens 2011	61.9	80

Signal Name	Author	% Obs.	Rank
lrreversal	De Bondt and Thaler 1985	61.9	81
abnormalaccruals	Xie 2001	61.8	82
announcementreturn	Chan, Jegadeesh and Lakonishok 1996	61.2	83
cboperprof	Ball et al. 2016	61.0	84
herfasset	Hou and Robinson 2006	60.5	85
herfbe	Hou and Robinson 2006	60.5	86
netdebtfinance	Bradshaw, Richardson, Sloan 2006	58.8	87
cfp	Desai, Rajgopal, Venkatachalam 2004	58.6	88
entmult	Loughran and Wellman 2011	58.2	89
varcf	Haugen and Baker 1996	56.9	90
grsaletogrinv	Abarbanell and Bushee 1998	55.8	91
grsaletogroverhead	Abarbanell and Bushee 1998	55.2	92
operprofrd	Ball et al. 2016	55.1	93
grcapx	Anderson and Garcia-Feijoo 2006	54.9	94
ep	Basu 1977	54.5	95
earningssurprise	Foster, Olsen and Shevlin 1984	54.2	96
chassetturnover	Soliman 2008	52.7	97
revenuesurprise	Jegadeesh and Livnat 2006	51.5	98
shareiss5y	Daniel and Titman 2006	51.4	99
cash	Palazzo 2012	50.3	100
grcapx3y	Anderson and Garcia-Feijoo 2006	50.1	101
momseason06yrplus	Heston and Sadka 2008	49.0	102
momoffseason06yrplus	Heston and Sadka 2008	48.7	103
betatailrisk	Kelly and Jiang 2014	47.6	104
invgrowth	Belo and Lin 2012	45.9	105
investment	Titman, Wei and Xie 2004	43.9	106
feps	Cen, Wei, and Zhang 2006	43.7	107
analystrevision	Hawkins, Chamberlin, Daniel 1984	43.1	108
momseason11yrplus	Heston and Sadka 2008	42.0	109
momoffseason11yrplus	Heston and Sadka 2008	41.9	110
std_turn	Chordia, Subra, Anshuman 2001	41.7	111
intancfp	Daniel and Titman 2006	41.2	112
intanep	Daniel and Titman 2006	41.2	113
compositedebtissuance	Lyandres, Sun and Zhang 2008	41.1	114
intansp	Daniel and Titman 2006	41.1	115
netpayoutyield	Boudoukh et al. 2007	40.7	116
intanbm	Daniel and Titman 2006	40.4	117
exclexp	Doyle, Lundholm and Soliman 2003	39.6	118
meanrankrevgrowth	Lakonishok, Shleifer, Vishny 1994	39.5	119
netdebtprice	Penman, Richardson and Tuna 2007	38.3	120

References

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

Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance*, *73*(2), 715-754. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12607` doi: https://doi.org/10.1111/jofi.12607

Beaver, W., McNichols, M., & Price, R. (2007). Delisting returns and their effect on accounting-based market anomalies. *Journal of Accounting and Economics*, *43*(2-3), 341–368.

Beckmeyer, H., & Wiedemann, T. (2023). Recovering missing firm characteristics with attention-based machine learning. *Available at SSRN 4003455*.

Bryzgalova, S. (2017). Spurious factors in linear asset-pricing models.

Bryzgalova, S., Huand, J., & Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance*, *78*(1), 487-557. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13197`

Bryzgalova, S., Lerner, S., Lettau, M., & Pelger, M. (2022). Missing financial data. *Available at SSRN 4106794*.

Cahan, E., Bai, J., & Ng, S. (2023). Factor-based imputation of missing

values and covariances in panel data of large dimensions. *Journal of Econometrics*, *233*(1), 113–131.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57-82. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1997.tb03808.x doi: https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chen, A. Y., & McCoy, J. (2023). *Missing values and the dimensionality of expected returns.*

Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, *27*(2), 207–264.

Chochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, *66*(4), 1047-1108. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x doi: https://doi.org/10.1111/j.1540-6261.2011.01671.x

Fama, E. F., & French, K. R. (2015a). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X14002323 doi: https://doi.org/10.1016/j.jfineco.2014.10.010

Fama, E. F., & French, K. R. (2015b). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X14002323 doi: https://doi.org/10.1016/j.jfineco.2014.10.010

Fanelli, D. (2013). Positive results receive more citations, but only in some disciplines. *Scientometrics*, *94*(2), 701–709.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Freyberger, J., Höppner, B., Neuhierl, A., & Weber, M. (2022). *Missing data in asset pricing panels* (Tech. Rep.). National Bureau of Economic Research.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteris-

tics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.

Gospodinov, N., Kan, R., & Robotti, C. (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *The Review of Financial Studies*, *27*(7), 2139–2170.

Gu, S., Kelly, B., & Xiu, D. (2020, 02). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223-2273. Retrieved from `https://doi.org/10.1093/rfs/hhaa009` doi: 10.1093/rfs/hhaa009

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, *14*(3), 262–280.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*(4), 1399–1440.

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

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650–705.

Jensen, T. I., Kelly, B. T., & Pedersen, L. H. (2021). *Is there a replication crisis in finance?* (Tech. Rep.). National Bureau of Economic Research.

Jin, S., Miao, K., & Su, L. (2021). On factor models with random missing: Em estimation, inference, and cross validation. *Journal of Econometrics*, *222*(1), 745–777.

Kan, R., Robotti, C., & Shanken, J. (2013). Pricing model performance and the two-pass cross-sectional regression methodology. *The Journal of Finance*, *68*(6), 2617-2649. Retrieved from

https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12035 doi: https://doi.org/10.1111/jofi.12035

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

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

Lewellen, J. (2015). *The cross-section of expected stock returns. critical finance review 4 (1): 1–44.*

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The journal of finance*, *20*(4), 587–615.

Little, R. J. (2021). Missing data assumptions. *Annual Review of Statistics and Its Application*, *8*, 89–107.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Lynch, A. W., & Wachter, J. A. (2013). Using samples of unequal length in generalized method of moments estimation. *Journal of Financial and Quantitative Analysis*, *48*(1), 277–307.

Mclean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5-32. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12365 doi: https://doi.org/10.1111/jofi.12365

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

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