Data Uncertainty in Financial Information*

Jiantao Huang† Serhiy Kozak‡

August 2025

Abstract

We develop a Bayesian tensor model that addresses three fundamental data challenges in empirical asset pricing: missing observations, infrequent measurements, and inherent noise in financial information. The model combines two ingredients: multiple imputations of firm characteristics and tensor representations jointly modeling time-series and cross-sectional dependencies in the high-dimensional panel. Our method outperforms traditional approaches, delivering higher prediction accuracy and valid confidence intervals for firm characteristic data. Imputing missing and infrequently updated data, although having negligible impacts on factor tangency portfolios, reduces the number of statistically significant alphas in characteristic-managed portfolios. Imputation of all characteristics, including the observed ones, can be used as synthetic data to perform pseudo out-of-sample tests for cross-sectional asset pricing models.

Keywords: Asset Pricing, Data Uncertainty, Tensor Models, Multiple Imputations.

JEL Classification Codes: G11, G12, C11, C33, C58.

* Any errors or omissions are the responsibility of the authors. For helpful comments, discussions, and suggestions, we thank Christian Julliard, Stefan Nagel, Guofu Zhou, and seminar and conference participants at the University of Maryland and CFE-CMStatistics 2024.

†Faculty of Business and Economics, The University of Hong Kong, huangjt@hku.hk.

‡R.H. Smith School of Business, University of Maryland, sekozak@umd.edu.

1 Introduction

Empirical asset pricing focuses on the estimation uncertainty of model parameters given the observed data. However, the potential uncertainty originating from the data is broadly overlooked. It is well-known that missing financial information, such as missing data on firm characteristics, is prevalent in real datasets. Moreover, data uncertainty arises from the usual practice of extrapolating firm fundamentals observed at annual or quarterly frequencies using their most recent values in monthly datasets. In extreme circumstances, even observed data may exhibit measurement uncertainty.[1] Ignoring data uncertainty, as commonly seen in existing approaches, can have sizeable and unpredictable effects on estimation and statistical inference that rely on such data. Our paper proposes a Bayesian tensor model to formally address these data challenges, with a focus on the high-dimensional panel data of firm characteristics.

Our model is built on two fundamental notions that are essential for satisfactorily addressing data uncertainty in financial information. First and foremost, we use *multiple imputations*. The basic idea is that imputing one value (single imputation) for a data point is generally incorrect and leads to erroneous inferences based on these imputed data. Instead, we create multiple imputations to reflect the uncertainty of reproducing these data points. To accomplish this, we use a Bayesian framework, which delivers a posterior distribution of firm characteristics instead of estimating only the "best" value, as has been common in earlier literature. That is, we treat firm characteristics as probabilistic quantities rather than fixed values, acknowledging that even observed data may contain measurement errors.

Second, we model characteristics as a three-dimensional *tensor* (characteristics across stocks and time) and find its *tensor rank decomposition*—a generalization of the singular value decomposition to high-dimensional panel data. Intuitively, when predicting the characteristics of a given firm, our model combines information from both contemporaneous

[1]There are several sources of measurement issues within the observed financial data. First, some firm characteristics—e.g., volatility, skewness, beta, and earnings surprises—are estimated, which inherently introduce estimation errors. Second, many accounting variables, including R&D expenses and intangible assets, are difficult to measure accurately. Third, certain firm characteristics proposed in economic theories are often latent. For example, the "expected profitability" in Hou, Xue, and Zhang (2015) is unobservable, so observable accounting metrics are often used as noisy proxies for this characteristic. Lastly, accounting manipulation, such as earnings management, adds further noise to the financial data.

and historical data of that firm and closely related companies. The tensor decomposition drastically reduces the number of parameter estimates, thus improving the out-of-sample imputation performance. This significant reduction in dimensionality allows us to handle pervasive missingness. Within the tensor model, all the required ingredients—the time series and cross-sectional dimensions of stock characteristics—are integrated into a unified parametric framework, which allows us to conduct valid statistical inference. In contrast, the inference is generally problematic in disjoint models of time series and the cross-section of equity characteristics.

We apply our framework to study the Global Factor Data constructed by Jensen, Kelly, and Pedersen (2023), which contain an extensive set of firm characteristics. We retain the continuous variables, ending up with 151 financial variables, and construct the US, European, and Asian datasets.

Two features of the Global Factor Data pose challenges to the traditional tensor estimation methods, such as the alternating least squares method (Carroll and Chang, 1970; Harshman, 1970). First, firm-month panel data is highly unbalanced. The traditional tensor model often performs poorly in unbalanced panels due to the unrealistic assumption that the nature of a company remains unchanged in the entire history. Second, data missingness is a paramount issue in firm characteristics, whereas the alternating least squares method requires a complete dataset. Data missingness becomes even more pronounced when some characteristics are observed infrequently. Consequently, recent papers applying the tensor decomposition (e.g., Lettau (2024)) often rely on complete and balanced datasets.

We address these issues in two steps. In the first step, we deviate from the static tensor model by adding a separate lag dimension to preserve the time-series information in firm characteristics. Importantly, we use only a short window (two years) of historical information about any given firm in estimating model parameters, therefore avoiding the assumption of stationarity for individual firms over time. In our model, the nature of a company is constant within a short rolling window yet is allowed to be time-varying across rolling windows.

We handle the second issue—incomplete dataset—by adopting a Bayesian framework, in which all model parameters are recovered by using only the observed information. Our framework is legitimate as long as the data is missing at random. Under this assumption,

the likelihood of missing an observation depends only on the observed information, implying that we can use the observed firm characteristics to impute the missing and infrequently updated data. The missing-at-random assumption is more likely to hold if we condition on a broader set of information that could relate to data missingness, which is exactly delivered by our tensor model—an extensive set of firm characteristics across different periods.

We find that missing financial information substantially differs from the available data. We use the tensor model to forecast the values for the missing elements and compare them with the observed quantities. As we show, the firms with missing characteristics are predicted to be less financially healthy in that they are less profitable, possess less cash, and have lower intrinsic value-to-market ratios. Moreover, these firms are more illiquid, pay fewer dividends, and invest less in R&D. These observations also imply that the missing-completely-at-random assumption—which posits that the causes of missing an observation are entirely unrelated to the data-generating process—is strongly rejected in the data. Therefore, our paper adopts the more realistic missing-at-random assumption.

We conduct a series of simulation studies to explore the performance of the Bayesian tensor model. We randomly simulate a sample of 20% observations and denote it as out-of-sample. We estimate model parameters using the remaining 80% of data points and evaluate the model performance based on the 20% simulated hold-out sample.

In simulations, our method provides accurate predictions for out-of-sample observations, with an R^2 above 65% in the US dataset. As a comparison, the principal component analysis (PCA) yields an R^2 of around 50%, whereas the R^2 is zero if we conduct predictions using cross-sectional means. We observe similar outperformance of tensor models over PCA in European and Asian datasets. The advantages of our tensor model are twofold. First, the tensor model captures the lagged information, while PCA extracts only cross-sectional dependencies. Moreover, the data-compression rate—defined as the ratio of the number of parameters to the number of observations—for the tensor decomposition is much lower than that of PCA. Hence, our approach is less susceptible to overfitting the data. Additionally, although PCA does not yield accurate confidence intervals for the firm characteristic data in the simulated hold-out sample, our tensor model produces precise and reliable confidence intervals, which are essential for effectively managing data uncertainty.

Finally, we examine the implications of accounting for data uncertainty in financial information within the context of cross-sectional asset pricing.

We begin by constructing univariate characteristic-managed portfolios based on both the observed data and the imputed data using our tensor model. We document that recovering the missing information significantly impacts the Sharpe ratios of some managed portfolios, with mixed effects on performance post-imputation. For the infrequently updated characteristics, using the higher-frequency information to infer their values can have notable effects on their performance. In the example of "Change in operating cash flow to assets" in the US dataset, filling only missing values leads to a slight change in the Sharpe ratio from 0.5 to 0.54. However, after further imputing its quarterly unchanged values by monthly model predictions, we observe a substantial increase in the Sharpe ratio to 0.8 per year. This example highlights the importance of accounting for both missing and infrequently observed data.

Next, we decompose managed portfolio returns into idiosyncratic and systematic components using the IPCA model (Kelly, Pruitt, and Su, 2019). Using the estimated systematic latent factors, we construct the factor tangency portfolios. We show that imputing missing or even updating stale information has minimal effects on the mean-variance efficiency of factor tangency portfolios. This observation echoes the finding in Chen and McCoy (2024) that simple cross-sectional mean imputation performs comparably to more sophisticated methods in predicting stock returns. Nevertheless, accounting for missing and stale information leads to a reduction in the number of statistically significant unconditional alphas of managed portfolios. At the 5% significance level, we observe 13, 12, and 27 significant pricing errors in the US, European, and Asian datasets, respectively, when missing data are filled in with zeros. These counts of significant alphas decrease to 8, 11, and 19 after accounting for data uncertainty in both missing and infrequently updated observations.

Multiple imputation is important beyond inference. The imputed datasets can be used as synthetic data to conduct pseudo out-of-sample tests. Different from the cross-validation approach, which must balance estimation uncertainty stemming from limited in-sample data against evaluation uncertainty from a short out-of-sample period, our pseudo out-of-sample tests rely on the entire time-series sample. This approach is especially valuable for refining model complexity when the available time-series data are limited.

In the first experiment, we leverage the synthetic datasets—100 simulated imputations of all data from the Bayesian tensor model—to determine the number of latent factors in IPCA models. While the large systematic latent factors maintain robust performance across pseudo out-of-sample tests, smaller latent factors tend to exhibit unstable performance when some randomness is introduced into the pseudo characteristic data. According to the pseudo out-of-sample exercise, we need 10 to 12 latent factors in the US data, whereas the IPCA model is denser in the European and Asian data, requiring 15 to 20 latent factors.

In the second experiment, we explore the pseudo out-of-sample performances of IPCA models. In addition to the factor tangency portfolios, we construct parametric arbitrage portfolios conditional on firm characteristics. Next, given the "best" in-sample estimates of model parameters, we evaluate the performance of both arbitrage and factor tangency portfolios across the 100 pseudo out-of-sample datasets. Although only minor drops are detected in the Sharpe ratios of the factor tangency portfolios, the performance of arbitrage portfolios declines significantly by more than 50% in these pseudo out-of-sample tests. In particular, the annualized Sharpe ratios of these parametric arbitrages range between 0.29 and 0.77, aligning more closely with the absence of near-arbitrage opportunities in the arbitrage pricing theory (Ross, 1976; Huberman, 1982; Chamberlain and Rothschild, 1983; Ingersoll, 1984). This finding highlights the importance of considering data uncertainty in evaluating the statistical arbitrage, which extends beyond the traditional estimation uncertainty that prevents arbitrageurs from exploiting all pricing errors as emphasized by the past literature (e.g., Da, Nagel, and Xiu (2024)).

Related Literature. The paper builds on three strands of existing literature. First, our paper is motivated by the recent growing literature on studying missing values in financial data (e.g., Freyberger, Höppner, Neuhierl, and Weber (2025); Bryzgalova, Lerner, Lettau, and Pelger (2025); Xiong and Pelger (2022); Chen and McCoy (2024); Beckmeyer and Wiedemann (2022)). Similar to the existing papers, we exploit both time-series and cross-sectional dependencies in the data imputation process. Unlike previous approaches, we model these two ingredients in a tensor model that significantly reduces the number of parameter estimates and enhances the prediction accuracy for missing financial data. Moreover, due to its unified parametric feature, we ensure valid statistical inference. In contrast, splitting

the cross-sectional and time-series steps makes statistical inferences challenging, and it is unclear how to construct proper confidence intervals. Most notably, we treat firm characteristics as probabilistic quantities and employ multiple imputations (Rubin, 1976, 1996; Little and Rubin, 2019) to capture the uncertainty in financial data. In our paper, missing observations, infrequently updated accounting information, and even observed data are all considered noisy measures of the true signals and contribute to data uncertainty.

Moreover, our paper is related to the recent application of high-dimensional tensor models in economics and finance (e.g., Wang, Zheng, and Li (2024); Cen and Lam (2025); Huang, Lu, Zheng, and Li (2025); Chen, Han, and Yu (2024); Ghysels, Babii, and Pan (2024); Lettau (2024)). We refine these traditional tensor models in two dimensions. First, we depart from the static tensor model by adding a lag dimension to capture the historical information in firm characteristics, which is essential in accommodating the unbalanced panel data. Second, we adopt a Bayesian framework to allow for missing observations in the tensor model, whereas the traditional methods of tensor decomposition, such as the alternating least squares approach, often require a complete dataset.

Finally, our paper studies a novel source of uncertainty in empirical asset pricing—data uncertainty in financial information. Previous literature often emphasizes parameter and model uncertainty, conditional on the observed dataset, or uncertainty about the distributional assumption of these data, in evaluating asset pricing models (e.g., Pástor (2000); Pástor and Stambaugh (2000); Avramov (2002, 2004); Kan and Zhou (2007); Garlappi, Uppal, and Wang (2007); Tu and Zhou (2004, 2010); Anderson and Cheng (2016); Barillas and Shanken (2018); Kozak, Nagel, and Santosh (2020); Chib, Zeng, and Zhao (2020); Bryzgalova, Huang, and Julliard (2023); Huang and Shi (2025)). We argue that the financial dataset is subject to data uncertainty due to the missing, infrequently observed, and noisily measured observations. Empirically, imputing the missing and stale financial information can significantly affect the performance of characteristic-managed portfolios. Moreover, multiple imputations of all characteristic data can be used in the pseudo out-of-sample tests for cross-sectional asset pricing models. These tests suggest that accounting for data uncertainty significantly reduces the Sharpe ratios of statistical arbitrage, which represents a critical economic mechanism different from the previous literature.

2 Model

This section lays out the econometric framework to handle data uncertainty in financial information. Throughout this paper, we use y_{tijk} to denote the value of the j-th characteristic of firm i at month $t-k$. In particular, the 3-D tensor $\mathcal{Y}_t \in \mathbb{R}^{N_t \times J \times K}$ represents all information at time t, where N_t is the number of firms at time t, J is the number of firm characteristics, and K is the number of time lags. Similarly, $\mathcal{Y} = \{\mathcal{Y}_t\}_{t=1}^{T}$ denotes the full sample from period 1 to T.

In empirical works, the data missingness is omnipresent. To handle missing values of firm characteristics, we create an indicator function, I_{tijk}, which equals 1 if we observe y_{tijk} and 0 if it is missing. Moreover, we differentiate the observed and missing data, i.e., $\mathcal{Y} = (\mathcal{Y}^{obs}, \mathcal{Y}^{mis})$. We observe $\mathcal{Y}^{obs}$ and I_{tijk} and aim to infer the distribution of $\mathcal{Y}^{mis}$.

We study a *tensor* model, which summarizes the cross-sectional and time-series dependencies of firm characteristics in a unified parametric framework. Simply put, the tensor model extends the two-dimensional (2-D) latent factor models. We use Figure 1a to illustrate the singular value decomposition (SVD) for the 2-D data. Suppose the matrix stores the cross-section of firm characteristics at a specific time. The vertical axis denotes different firms, whereas the horizontal axis represents various firm characteristics. The PCA/SVD extracts the cross-sectional dependencies. One limitation of this approach is that it ignores valuable time-series information.

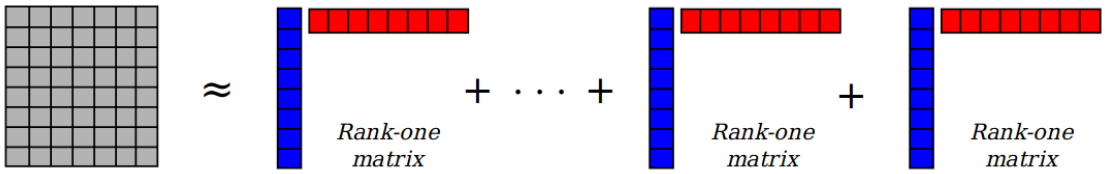
High-dimensional tensor models fix this issue. Specifically, we decompose y_{tijk} into common and idiosyncratic components:

$$y_{tijk} = x_{tijk} + \epsilon_{tijk}, \quad \epsilon_{tijk} \sim \mathcal{F}(0, \tau_{\epsilon,j}^{-1}),$$

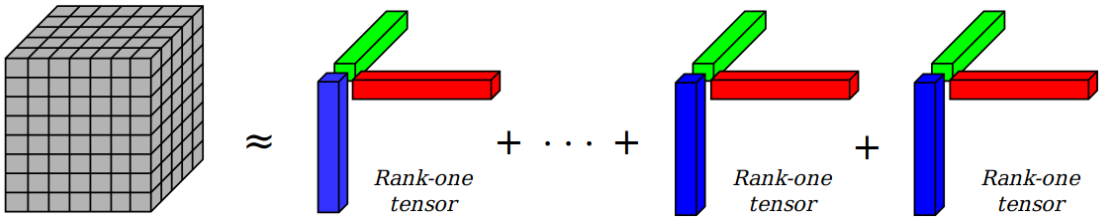
where ϵ_{tijk} is the prediction error following a zero-mean distribution $\mathcal{F}$, and x_{tijk} is the common component that has a tensor decomposition as follows:

$$x_{tijk} = \sum_{r=1}^{R} s_{tir} c_{jr} l_{kr}, \tag{1}$$

where s_{tir}, c_{jr}, and l_{kr} represent the firm-specific, characteristic-specific, and lag-specific



(a) Singular value decomposition



(b) CP decomposition

Figure 1: Matrix vs Tensor decomposition

information. Suppose we collect all x_{tijk} and ϵ_{tijk} using 3-D tensors, i.e., $\mathcal{X}_t, \mathcal{E}_t \in \mathbb{R}^{N_t \times J \times K}$. The tensor decomposition can be succinctly written as

$$\mathcal{X}_t = \sum_{r=1}^{R} \boldsymbol{s}_{tr} \circ \boldsymbol{c}_r \circ \boldsymbol{l}_r, \quad \mathcal{Y}_t = \mathcal{X}_t + \mathcal{E}_t, \tag{2}$$

$$\boldsymbol{c}_r = (c_{1r}, \ldots, c_{Jr})^\top, \quad \boldsymbol{l}_r = (l_{1r}, \ldots, l_{Kr})^\top, \quad \boldsymbol{s}_{tr} = (s_{t1r}, \ldots, s_{t,N_t,r})^\top,$$

and $\circ$ is the outer product. This is an instance of the CANDECOMP/PARAFAC (CP) decomposition, initially proposed by Carroll and Chang (1970) and Harshman (1970).[2] Figure 1b gives a graphical explanation of the CP decomposition. Unlike the SVD in Figure 1a, the CP decomposition allows us to extract both time-series (from $\boldsymbol{l}_r$) and cross-sectional (from $\boldsymbol{c}_r$) information from the 3-D data. Similar to SVD, we normalize the loadings $\boldsymbol{c}_r$ and $\boldsymbol{l}_r$ such that they have unit lengths. Therefore, the vector norm of $\boldsymbol{s}_r$ quantifies the importance of the r-th common component, which is similar to the r-th singular value in the 2-D case.

In contrast to the standard tensor decomposition, we pool the time-series and cross-sectional observations but add a separate lag dimension to preserve the time-series information in firm characteristics. We choose this model design since both the number of individual stocks and the nature of firms are time-varying; hence, we cannot impose constancy in this

[2]Interested readers can read Kolda and Bader (2009) for a detailed introduction to tensor decomposition.

dimension. Within our model, s_{tir} summarizes the most recent history—from period $t-K$ to t—of firm i in the r-th dimension. As a comparison, we often use the full sample to decompose the static tensor in the traditional tensor decomposition. This method is problematic in reality because of the unbalanced panel data and time-varying nature of firms; that is, it imposes that the nature of firm i is summarized by time-invariant s_{ir}. Instead, our method uses only the recent time-series observations, that is, in a short window of K months, to create the embeddings of firm i and assumes that these embeddings are constant in a short window. Meanwhile, the characteristic- and lag-specific loadings ($\boldsymbol{c}_r$ and $\boldsymbol{l}_r$) are estimated using the full-sample information, thus enhancing estimation efficiency.

One advantage of the tensor decomposition is that it significantly reduces the amount of model parameters. To illustrate this idea, we compare it with PCA applied to the cross-section of N firms and J characteristics. The numbers of free parameters in the tensor decomposition and PCA are about $(N + J + K)R$ and $(N + R)R$, without considering the restrictions on the vector norm.[3] What matters in the estimation is the ratio of the number of parameters over the number of elements in tensors/matrices. In particular, the data-compression ratio for the tensor decomposition is $\frac{(N+J+K)R}{N\times J\times K}$, whereas it equals $\frac{(N+R)R}{N\times J}$ for PCA. Suppose we study 1,500 firms, 36 firm characteristics, and 36 lags. If $R = 20$, the compression rates of the CP decomposition and PCA are 1.6% and 56.3%, respectively. Therefore, the tensor model is often able to recover more latent components (equivalently, a larger R) than PCA, without overfitting the data.

The estimation of tensor models is challenging within our context for two reasons. First, there is no closed-form solution to tensor decomposition, so we generally rely on some numerical algorithms (e.g., alternating least squares method) to compute the tensor decomposition. Moreover, traditional estimation approaches require a complete dataset, which is clearly not the case in studies of individual stocks.

We propose a Bayesian framework to estimate the tensor model in equation (1). Our method does not intend to "choose" the optimal tensor rank. As Håstad (1990) indicates, the rank determination of tensors is NP-hard; that is, tensor decomposition in equation (1)

[3] A more rigorous comparison is to include the coefficient restrictions. For example, we generally require the factor loadings in PCA to be orthonormal. However, the number of restrictions will not greatly affect the data-compression ratio.

is generally not unique (see, e.g., Kolda and Bader (2009)). Instead, we assign a prior to $\boldsymbol{s}_{tr}$ and let the data determine the relevance of the r-th component. We start with a relatively conservative choice of R and use our Bayesian algorithm to shrink unessential dimensions (that is, unimportant $\boldsymbol{s}_{tr}$) toward zeros. Furthermore, our Bayesian model estimates the full posterior distribution of model parameters and delivers the entire distribution of both missing and observed firm characteristics. Hence, we properly handle data uncertainty in financial information.

In the following three subsections, we first specify our assumption about the missing data mechanism in subsection 2.1. We next explain data uncertainty and multiple imputation in subsection 2.2. Finally, in section 2.3, we present our Bayesian tensor model.

2.1 Missing Data Mechanisms

We now describe several assumptions about missing data mechanisms. Note that we use $\mathcal{Y}$ and $\mathcal{I}$ to denote the sample data and response indicators, and $\mathcal{Y} = (\mathcal{Y}^{obs}, \mathcal{Y}^{mis})$. Let ψ contain the parameters governing the missingness mechanism. Following the definition in the statistical literature, the data are said to be *missing-completely-at-random (MCAR)* if

$$p(\mathcal{I} \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis}, \psi) = p(\mathcal{I} \mid \psi). \tag{3}$$

In other words, the causes of missing an observation are entirely unrelated to the data-generating process. This assumption is rather restrictive and unrealistic in financial data. For instance, smaller, younger, and less profitable firms are more likely to omit reporting their financial indicators.

Our paper instead studies a milder assumption, called *missing-at-random (MAR)*:

$$p(\mathcal{I} \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis}, \psi) = p(\mathcal{I} \mid \mathcal{Y}^{obs}, \psi). \tag{4}$$

Under this assumption, the probability of missing an observation depends only on the observed information, which implies that we can use the observed firm characteristics to impute the missing ones. As Rubin (1976) argues, MAR is the weakest assumption under which we can ignore the missingness mechanism and still maintain correct statistical inference.

Some literature studies the *missing-not-at-random (MNAR)*[4] assumption, such as selection models (e.g., Heckman (1979)) and pattern-mixture models (e.g., Little (1993)). Under this condition, we cannot simplify $p(\mathcal{I} \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis}, \psi)$, so the likelihood of missing data depends on unobserved information $\mathcal{Y}^{mis}$. Specific assumptions need to be made to handle the MNAR patterns of missingness.

Our paper assumes MAR, under which the distribution of missing values can be estimated using only the observed information. In the tensor model, we also assume that the common component x_{tijk} is sufficient to capture the missingness mechanism while the error term ϵ_{tijk} is irrelevant.[5] Empirically, we can make this assumption more likely to hold by conditioning on more firm characteristics that could relate to data missingness. That is, conditional on a sufficiently large set of (both cross-sectional and time-series) information, the probability of data missingness is approximately unrelated to the unobserved information.

2.2 Data Uncertainty and Multiple Imputations

Imputing one value (single imputation) for the missing element is generally incorrect because this undermines data uncertainty in the missing component. To handle data uncertainty, we adopt multiple imputations, first proposed by Donald Rubin. Specifically, he proposed a simple and intuitive solution—create multiple imputations to reflect uncertainty of filling in missing data.

Multiple imputations are the key to properly accounting for estimation uncertainty. Suppose Q is a quantity of scientific interest, for example, the Sharpe ratio of the tangency portfolio conditional on firm characteristics. We attempt to infer the distribution of Q using the firm characteristics $\mathcal{Y}$ as conditional information. However, only part of $\mathcal{Y}$ is observable, so the conditional distribution of Q is

$$p(Q \mid \mathcal{Y}^{obs}) = \int p(Q \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis}) p(\mathcal{Y}^{mis} \mid \mathcal{Y}^{obs}) d\mathcal{Y}^{mis}, \tag{5}$$

where $p(Q \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis})$ is the distribution of Q conditional on the hypothetical complete data, and $p(\mathcal{Y}^{mis} \mid \mathcal{Y}^{obs})$ is the posterior distribution of the missing data given the observed

[4] Testing MAR versus MNAR is impossible since the required information for hypothesis testing is missing.

[5] In fact, other methods assuming the latent factor models (for example, PCA) make the same assumption.

one. If we use a single imputation, i.e., setting $\mathcal{Y}^{mis} = \hat{\mathcal{Y}}^{mis}$, we will underestimate the uncertainty of Q. A proper way is to generate $\mathcal{Y}^{mis}$ from $p(\mathcal{Y}^{mis} \mid \mathcal{Y}^{obs})$ and then draw Q from $p(Q \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis})$ for each realization of $\mathcal{Y}^{mis}$. We call the process in equation (5) *integrating out data uncertainty*.

2.3 Bayesian Inference

We estimate the tensor model with a Bayesian paradigm. We assume that prediction errors are independent and identically distributed (IID) and follow normal distributions:

$$y_{tijk} \mid x_{tijk} \overset{\text{iid}}{\sim} \mathcal{N}(x_{tijk}, \tau_{\epsilon,j}^{-1}), \tag{6}$$

where $\tau_{\epsilon,j}$ denotes the noise precision of the j-th firm characteristic. Under this assumption, we allow for heterogeneous predictability in different characteristics. This distributional assumption further implies that the likelihood function of the data $\mathcal{Y}$ is:

$$p(\mathcal{Y} \mid \mathcal{X}, \tau_{\epsilon,j}) \propto \prod_{j=1}^{J} \tau_{\epsilon,j}^{\frac{N_{obs,j}}{2}} \exp\{-\frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \sum_{j=1}^{J} \sum_{k=1}^{K} I_{tijk} (y_{tijk} - x_{tijk})^2 \tau_{\epsilon,j}\}, \tag{7}$$

where $N_{obs,j} = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \sum_{k=1}^{K} I_{tijk}$, and $N_{obs} = \sum_{j=1}^{J} N_{obs,j}$.

We go on to assign prior distributions to model parameters. Since we normalize $\boldsymbol{c}_r$ and $\boldsymbol{l}_r$, it is reasonable to assign the following priors to them:

$$c_{jr} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad l_{kr} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{8}$$

The prior variances of $\boldsymbol{c}_r$ and $\boldsymbol{l}_r$ are homogeneous across r since they have the same L_2 norm.

The (cross-sectional) variance of $\boldsymbol{s}_{tr}$ quantifies the importance of the r-th component. We assign the normal-gamma prior to $\boldsymbol{s}_{tr}$ and its precision parameter:

$$s_{tir} \sim \mathcal{N}(\mathbf{0}, \tau_{sr}^{-1}), \quad \tau_{sr} \sim \Gamma(a_0, b_0), \quad r = 1, \ldots, R. \tag{9}$$

When τ_{sr}^{-1} is small, we shrink the r-th component towards zero. Ex-ante, there is no theoretical guidance on the tensor rank. Therefore, our strategy is to start with a reasonably

large truncation level of R and allow the data to shrink irrelevant dimensions toward zero through the posterior distribution of τ_{sr}.

Finally, we put a gamma prior on the noise precision parameter $\{\tau_{\epsilon,j}\}_{j=1}^{J}$:

$$\tau_{\epsilon,j} \sim \Gamma(a_0, b_0), \quad j = 1, \ldots, J. \tag{10}$$

We use Gibbs sampling to update model parameters: $\theta := (s_{tir}, c_{jr}, l_{kr}, \tau_{sr}, \tau_{\epsilon,j})$. In the empirical analysis, we choose small values of a_0 and b_0 (equal to 0.01) to reflect that we do not have strong prior information regarding τ_{sr} and $\tau_{\epsilon,j}$. Our algorithm is simple and fast since all posterior distributions are either normal or gamma distributions, as explained in Proposition 1.

Proposition 1. *Under the prior distributions in equations (8)–(10) and the data likelihood in equation (7), the posterior distributions of model parameters* $\theta := (s_{tir}, c_{jr}, l_{kr}, \tau_{sr}, \tau_{\epsilon,j})$ *are the normal-gamma distributions in equations (A2)–(A6).*

We present a detailed version of the Gibbs sampler in Proposition A1, and the proofs are in Appendix B.[6] Our Bayesian estimator converges fast, generally within 100 steps. Suppose we obtain M posterior draws, $\{(s_{tir}^{(m)}, c_{jr}^{(m)}, l_{kr}^{(m)}, \tau_{sr}^{(m)}, \tau_{\epsilon,j}^{(m)})\}_{m=1}^{M}$. For the m-th draw, we construct the common component $x_{tijk}^{(m)}$, defined as:

$$x_{tijk}^{(m)} = \sum_{r=1}^{R} s_{tir}^{(m)} c_{jr}^{(m)} l_{kr}^{(m)}. \tag{11}$$

In the multiple imputations, we draw the missing y_{tijk} from the full posterior distribution:

$$y_{tijk}^{(m)} = x_{tijk}^{(m)} + \epsilon_{tijk}^{(m)}, \quad \epsilon_{tijk}^{(m)} \sim \mathcal{F}^{(m)}\left(0, \frac{1}{\tau_{\epsilon,j}^{(m)}}\right), \; m = 1, \ldots, M, \tag{12}$$

where $\mathcal{F}^{(m)}$ is the distribution of prediction errors. Without taking a stance on the error term's distribution $\mathcal{F}^{(m)}$, we design a Bayesian Bootstrap algorithm, which is described in Definition 1, to generate multiple imputations of firm characteristics.

[6] Our Bayesian framework also makes the ignorability assumption, as we outline in Appendix A.

Definition 1 (Bayesian Bootstrap). *Our Bayesian Bootstrap consists of two steps:*
Step 1. For each draw of $x_{tijk}^{(m)}$*, compute* $\hat{\epsilon}_{tijk}^{(m)} = y_{tijk} - x_{tijk}^{(m)}$ *for only the observed entries.*
Step 2. For each time t *and characteristic* j*, draw* $\epsilon_{tijk}^{(m)}$ *with replacement from* $\{\hat{\epsilon}_{tijk}^{(m)}\}$ *and define* $y_{tijk}^{(m)} = x_{tijk}^{(m)} + \epsilon_{tijk}^{(m)}$.

The Bayesian-Bootstrap algorithm provides robust confidence intervals for missing financial data. As Rubin and Schenker (1986) argue, it is preferred to let the observed data infer the shape of the distribution of $\mathcal{Y}^{mis}$ when the assumption of normality is questionable. Our framework can be used not only for missing values but also to infer the stale values that are infrequently updated, such as quarterly or annual accounting variables. Furthermore, we can use multiple imputations of all data points in the model estimation and evaluation if we believe that even the observed data are noisily measured. In other words, our framework treats the firm characteristics as probabilistic quantities rather than fixed values.

Remark. *Note that* $y_{t+l,i,j,l} = y_{tij0}$*, so we can use either* $y_{t+l,i,j,l}$ *or* y_{tij0} *to impute the missing entries of characteristic* j *of firm* i *at time* t*. However,* $y_{t+l,i,j,l}$ *is constructed using the rolling window between time* $t + l - K$ *and* $t + l$ $(l < K)$*, which implies that we use future information to impute the missing data at time* t*. To avoid the forward-looking bias in* $y_{t+l,i,j,l}$*, we use* y_{tij0} *in data imputation throughout all empirical analyses.*

3 Data

We use the Global Factor Data constructed by Jensen, Kelly, and Pedersen (2023), which contains 153 firm characteristics in the original dataset. We remove two dummy variables ("f_score" and "ni_inc8q") because our approach is designed to model continuous variables. Using the exact definition in Jensen, Kelly, and Pedersen (2023), we further group the remaining 151 characteristics into 13 cluster themes: Accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, size, short-term reversal, and value.[7]

The Global Factor Data is comprehensive not only because it provides an extensive set of firm characteristics but also because it covers international datasets of more than 90

[7]The data and related documentation can be found at https://jkpfactors.com.

countries. We study both the US and international markets. The sample of the US data spans from January 1973 to December 2021. We further construct the European and Asian datasets by following the approach in the Ken French library.[8] The European and Asian data samples range from July 1986 to December 2021. We keep the firms with market values larger than the 20th percentile of NYSE stocks. Moreover, the returns in the Global Factor Data are excess returns denominated in US dollars.

Following past literature, we normalize firm characteristics cross-sectionally in each period (*per month*). That is, we transform the raw characteristics into rank quantiles and scale them to be located in the interval $[-0.5, 0.5]$. This preprocessing makes data stationary across time and closer to our distributional assumption in equation (6). We use these rank-transformed characteristics to build a three-dimensional tensor $\mathcal{Y}_t \in \mathbb{R}^{N_t \times J \times K}$, which represents all observed information at time t, where N_t is the number of firms at time t, J is the number of firm characteristics, and K is the number of time lags. To build the lags, at each time t, we focus on all stocks observable at this time, creating full time-series observations for only these stocks over the last K periods. Any missing observations are marked as such in our tensor. Lastly, we stack all $\mathcal{Y}_t$ tensors across time, $\mathcal{Y} = \{\mathcal{Y}_t\}_{t=1}^{T}$, which denotes the full observed sample.

Missing data is ubiquitous in firm characteristics. Table 1 reports the percentages of firm-month observations missing a certain amount of data. Complete data are clearly elusive: Even in the US data, less than 6% of firm-month observations have complete data in all 151 variables, and only 44% have less than 10 missing signals. Therefore, the missing data are extremely problematic in multivariate applications.

Figure 2 plots the time series of missing rates in each of the 13 cluster themes. We observe that data missingness tends to be more severe in earlier subsamples, which poses a huge challenge for asset pricing applications, such as backtesting. Moreover, there is significant heterogeneity across different clusters. For example, the clusters "accruals" and "seasonality" have higher missing rates than "short-term reversal" and "momentum" clusters. While the US sample has a fairly low average missing rate of 16% over the full sample, missing rates

[8]The Asian dataset contains five markets: Hong Kong, Singapore, Australia, New Zealand, and Japan. The European dataset consists of 16 countries: the United Kingdom, France, Germany, Switzerland, Netherlands, Sweden, Italy, Spain, Denmark, Belgium, Norway, Finland, Ireland, Austria, Portugal, and Greece.

Table 1: Percentages of firms missing certain amount of signals

	US	EUR	ASIA
Complete	5.39%	2.00%	0.11%
< 10 missing	44.38%	19.89%	6.67%
< 20 missing	62.44%	36.88%	33.79%
< 30 missing	72.39%	46.82%	43.94%
< 40 missing	78.00%	53.44%	53.66%
< 50 missing	83.19%	61.44%	60.90%

This table reports the percentages of firms missing a certain amount of signals in the US, European, and Asian datasets. The US data sample spans from January 1973 to December 2021, whereas the European and Asian data samples range from July 1986 to December 2021.

of European and Asian datasets are about 31% and 32%, respectively.

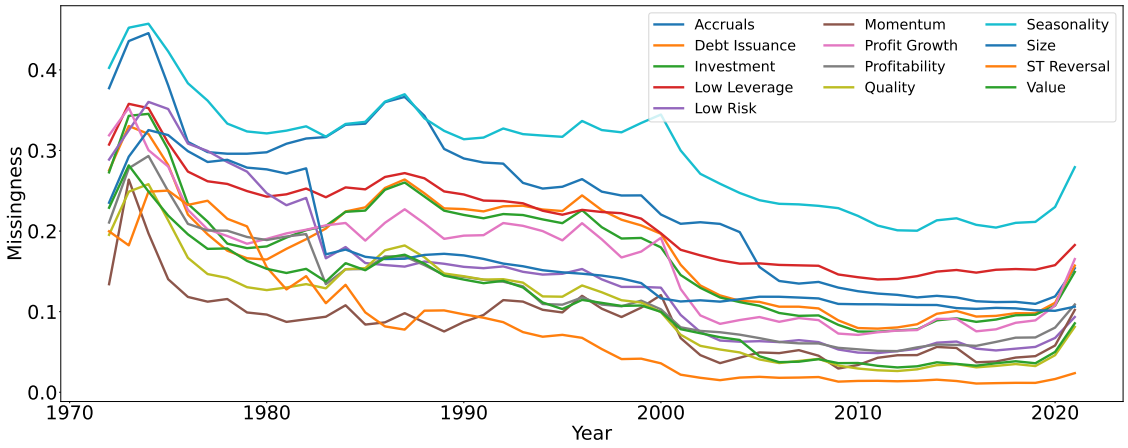
Missing financial information might differ substantially from the observed data. For instance, if less profitable companies are more likely not to report their accounting details, we should anticipate that the average profitability indicators of the missing entries will be lower than those recorded in the dataset. To explore this, we employ tensor models with $K = 24$ monthly lags and tensor rank $R = 200$ to study the patterns of missing characteristics.

Figure 3 presents the average imputed firm characteristics for the US, European, and Asian datasets using our tensor model. Note that the average values of the observed characteristics are normalized to have zero cross-sectional means. In Figure 3, we display the top 15 variables whose model-predicted values of the missing entries deviate the most from their cross-sectional means. Two key observations stand out.

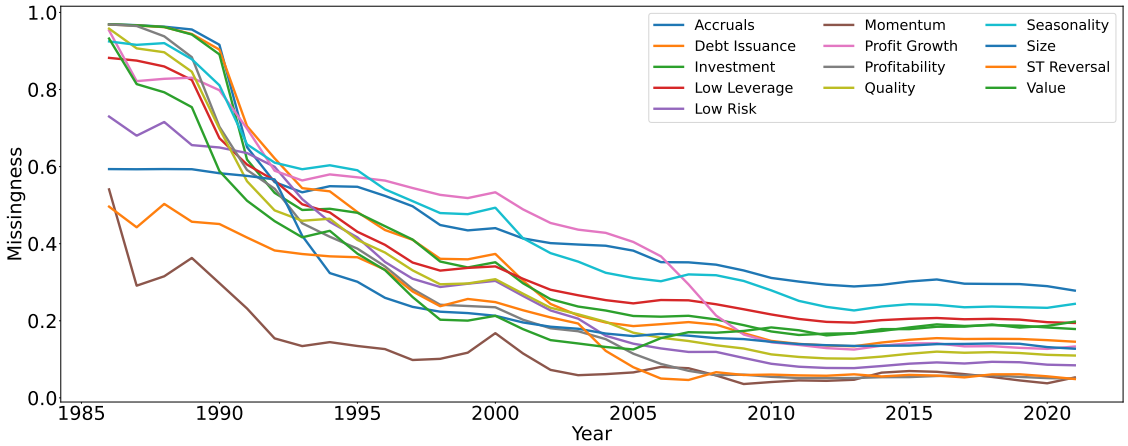
First, the data strongly rejects the missing-completely-at-random assumption. If this assumption were valid, we would expect to see roughly the same averages (around zero) for both missing and observed values. However, this is not the case in any of the three plots shown in Figure 3. Even when accounting for data uncertainty, as demonstrated in the black intervals denoting the 95% confidence bands across 100 multiple imputations, the conclusion remains unchanged.[9]

Moreover, we identify universal patterns in missing firm characteristics across all three

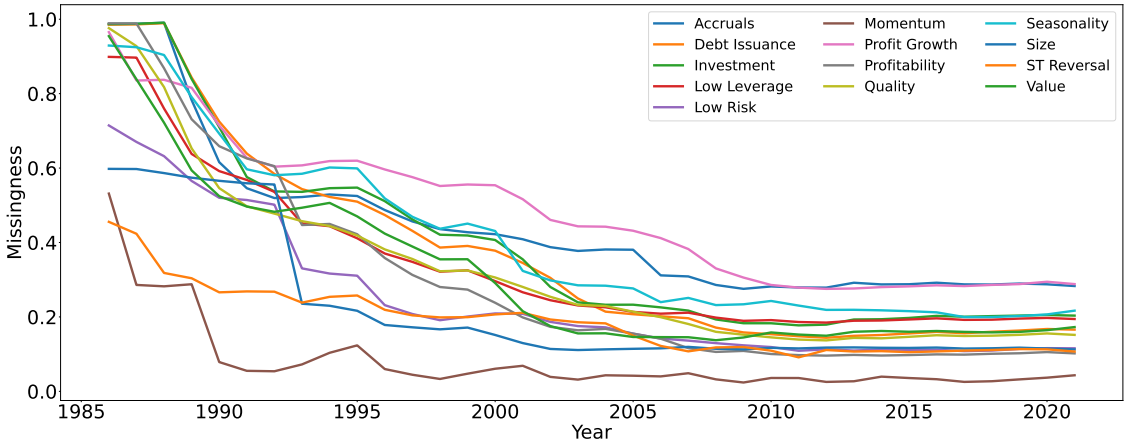
[9]To clarify, the black intervals in Figure 3 only account for the data uncertainty regarding the mean predictions for missing values. They do not consider the sampling uncertainty within each imputation. Given that the mean predictions are derived from at least tens of thousands of data points, the estimation uncertainty in these mean predictions is minimal.



(a) US Data



(b) European Data



(c) Asian Data

Figure 2: Historical missing rates of firm characteristics

These figures plot the time series of missing rates of firm characteristics in the US, European, and Asian datasets. The 151 firm characteristics are grouped into 13 cluster themes. The US data sample spans from January 1973 to December 2021, whereas the European and Asian data samples range from July 1986 to December 2021.

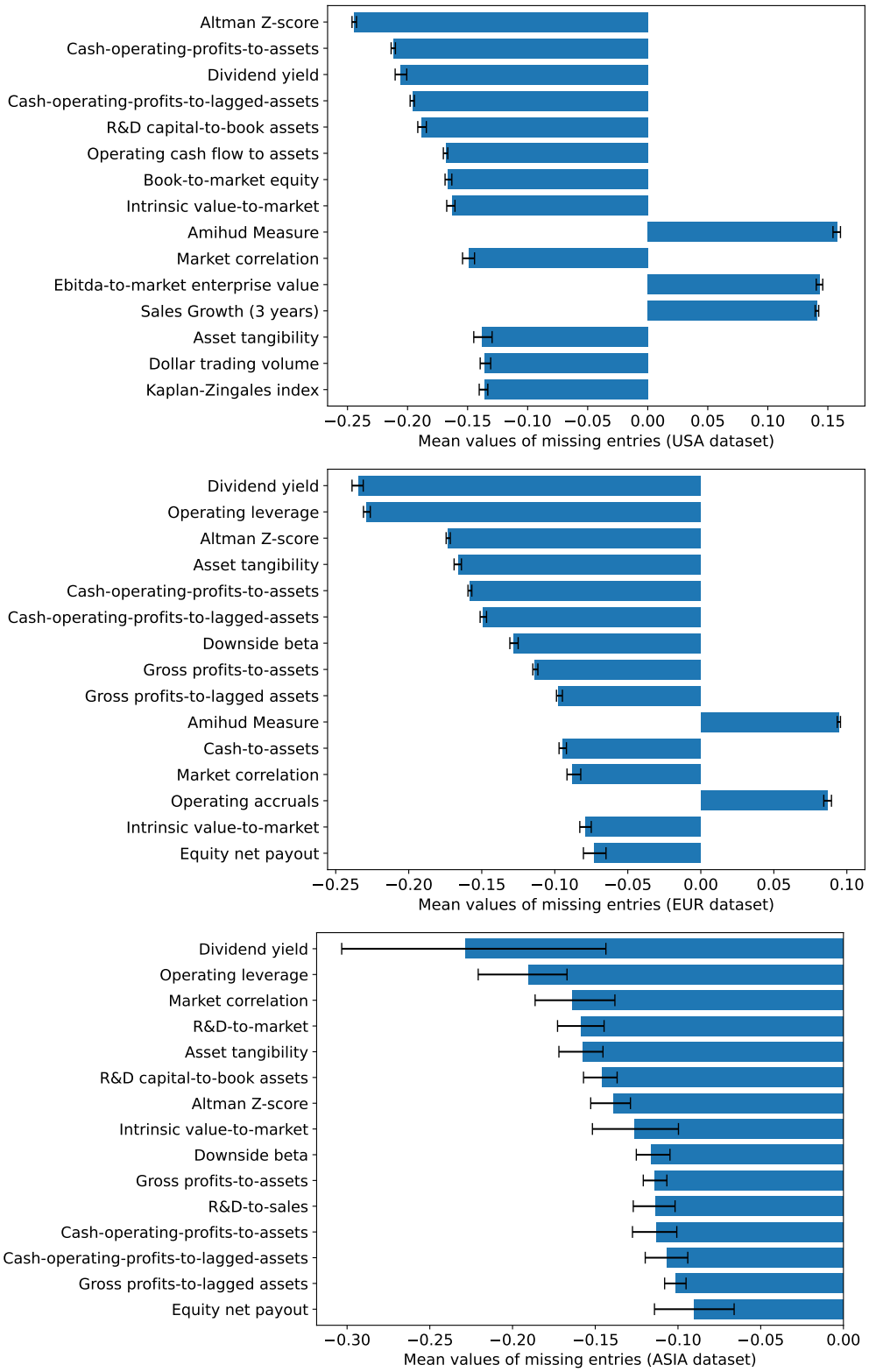


Figure 3: Model-implied mean values of missing firm characteristics

This figure plots the mean values of missing firm characteristics in the US, European, and Asian datasets. Blue bars (black intervals) denote the averages (95% confidence bands) across 100 multiple imputations. We consider tensor models with $K = 24$ lags and tensor rank $R = 200$. We report the top 15 characteristics whose mean values of the missing entries deviate the most from zero (the mean values of observed entries). The US data sample starts from January 1973 to December 2021, whereas the European and Asian data samples range from July 1986 to December 2021.

datasets. Specifically, firms with missing characteristics tend to be less financially healthy (e.g., lower Altman Z-scores), less profitable (e.g., lower cash-operating-profits-to-(lagged)-assets and gross-profit-to-(lagged)-assets), pay fewer dividends (e.g., lower dividend yields and equity net payouts), and invest less in R&D (e.g., lower R&D capital-to-book assets). These firms are also more illiquid (e.g., higher Amihud measure and lower dollar trading volume), have lower intrinsic value-to-market ratios, are less correlated with the market portfolio, possess less cash (e.g., lower cash-flows-to-assets and cash-to-assets), and exhibit lower operating leverages. Overall, missing firm characteristics, on average, differ significantly from the observed information across many dimensions.

4 Simulations

In this section, we conduct a series of simulation studies to examine the performance of our Bayesian tensor model. For each dataset, we first remove the stale observations. For example, some accounting-based variables are updated at the quarterly frequency, so their values are precisely the same within one quarter. By removing the repeated observations, we avoid mechanically fitting the model with unchanged historical values in the simulations.

The next step is to randomly simulate a sample of 20% missing entries ($\mathcal{Y}^{mis}$) in the dataset without stale observations, which we refer to as the out-of-sample. We estimate model parameters using only the observed information ($\mathcal{Y}^{obs}$), i.e., the sample of the remaining 80% observations, denoted as in-sample. The main objective is to examine whether the tensor model can reconstruct the posterior distribution of data points in the hold-out sample.

We evaluate the model performance based on two metrics. First, we consider the prediction accuracy, which is measured as the predictive R^2:

$$\text{In-sample based on } \mathcal{Y}_t^{obs}: \quad R^2_{tj,in} = 1 - \frac{\sum_{i=1}^{N_t^{obs}} (y_{tij0}^{obs} - \hat{y}_{tij0}^{obs})^2}{\sum_{i=1}^{N_t^{obs}} [y_{tij0}^{obs}]^2}, \tag{13}$$

$$\text{Out-of-sample based on } \mathcal{Y}_t^{mis}: \quad R^2_{tj,oos} = 1 - \frac{\sum_{i=1}^{N_t^{mis}} (y_{tij0}^{mis} - \hat{y}_{tij0}^{mis})^2}{\sum_{i=1}^{N_t^{mis}} [y_{tij0}^{mis}]^2}. \tag{14}$$

Remark. *As we center each characteristic cross-sectionally per period, the unconditional*

mean is zero. Hence, equations (13) and (14) compare the model predictions with cross-sectional means that are often used in the finance literature. For example, Kozak, Nagel, and Santosh (2020) replace the missing firm characteristics with zeros. If we observe positive R^2, our model performs better than imputing the data with cross-sectional means.

The second metric is the coverage of confidence intervals. We construct the 80%, 90%, 95%, and 98% confidence intervals based on Definition 1. Theoretically, we expect 90% confidence intervals to cover 90% of hypothetical "true" values on average. The same logic holds for other confidence levels. A method can adequately handle data uncertainty only if it provides appropriate confidence intervals of data points in the hold-out sample.

4.1 Prediction Accuracy

This subsection studies the prediction performance of our Bayesian tensor decomposition. As a benchmark, we consider the (cross-sectional) PCA. That is, at each time t, we compute the pairwise correlation between the j_1-th and j_2-th firm characteristics as follows:

$$C_{t,j_1,j_2} = \sum_{i=1}^{N_t} (y_{t,i,j_1} - \bar{y}_{t,i,j_1}) \cdot (y_{t,i,j_2} - \bar{y}_{t,i,j_2}) \cdot I_{t,i,j_1} \cdot I_{t,i,j_2}. \tag{15}$$

That is, we include the firm-month observations to estimate the pairwise correlations whenever we observe both y_{t,i,j_1} and y_{t,i,j_2} for firm i. Next, we use the eigendecomposition of the correlation matrix in equation (15) period-by-period. Suppose we consider a R-factor model, and U collects the first R eigenvectors of the correlation matrix. We predict Y_t ($N_t \times J$ matrix) as follows:

$$\hat{Y}_t = \tilde{Y}_t \cdot U \cdot U^\top,$$

where $\tilde{Y}_t$ is the $N_t \times J$ matrix where missing values are replaced with zeros.

Our tensor model further captures time-series dependencies. Using the Gibbs sampler in Proposition 1, we estimate model parameters using $\mathcal{Y}^{obs}$ and obtain M posterior draws of $\boldsymbol{s}_{tr}$, $\boldsymbol{c}_r$, and $\boldsymbol{l}_r$. The predicted values of firm characteristics are constructed as follows:

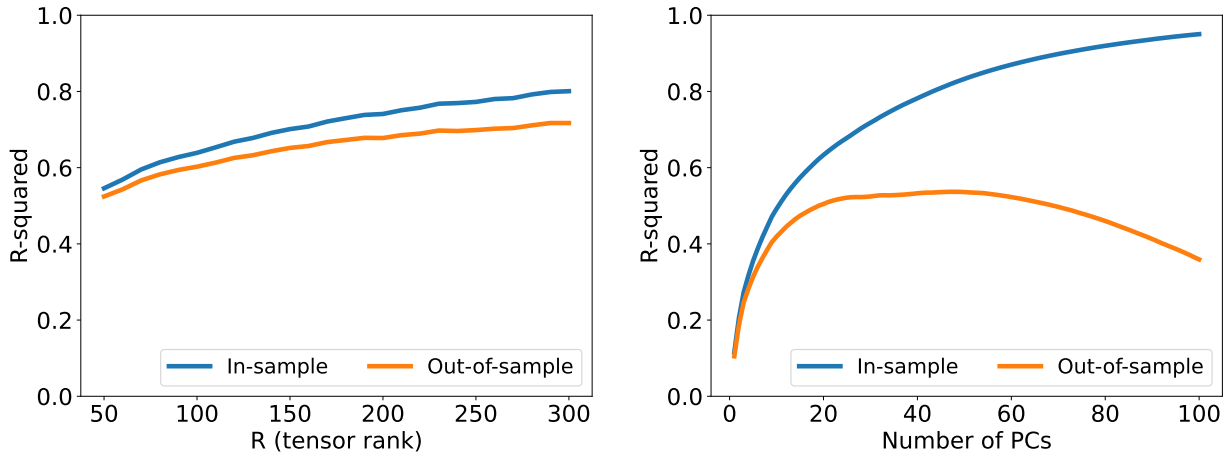
$$\hat{\mathcal{Y}}_t^{(m)} = \sum_{r=1}^{R} \boldsymbol{s}_{tr}^{(m)} \circ \boldsymbol{c}_r^{(m)} \circ \boldsymbol{l}_r^{(m)}, \quad \text{and } \hat{\mathcal{Y}}_t = \frac{1}{M} \sum_{m=1}^{M} \hat{\mathcal{Y}}_t^{(m)}.$$

We use $\hat{\mathcal{Y}}_t$ to predict both $\mathcal{Y}^{obs}$ and $\mathcal{Y}^{mis}$ and measure the in- and out-of-sample fit.

Our tensor model has two crucial tuning parameters—the number of lags K and the tensor rank R. As we indicate in Section 2, the data-compression rate for the tensor decomposition is roughly $\frac{(N+J+K)R}{N\times J\times K}$. Therefore, increasing K enlarges the denominator and prevents the model from overfitting the data. Intuitively, we require the model to explain the cross-section of firm characteristics in many different periods, which helps discipline the model. A downside is that a large value of K imposes that the nature of a firm is time-invariant in a long rolling window, which is unrealistic and can potentially lead to inferior performance. In our simulations, we consider $K = 24$, equivalent to using the past two-year historical information for imputations. The second critical dimension is the tensor rank. A larger tensor rank improves the in-sample fit, with a higher likelihood of overfitting data. We consider a wide range of values for the tensor rank in simulations.

Figure 4 plots the in- and out-of-sample prediction accuracy in the US characteristic dataset. Panel (a) shows the performance of our tensor model. We need a large tensor rank, e.g., $R \geq 200$, to achieve good prediction accuracy in the hold-out sample. In this case, the out-of-sample R-squared is larger than 65%. As the tensor rank becomes larger than 200, the orange curve tends to be flat. Moreover, the in-sample R-squared is rather close to the out-of-sample value, which implies that we can estimate a tensor model of a large rank without overfitting the data. As a comparison, we present the performance of the 2-D PCA in Panel (b). As suggested by previous literature, PCA delivers a decent prediction accuracy of about 50%. There is also no sparsity in the PCA models: We need 20–40 PCs to maximize the out-of-sample prediction accuracy. Nevertheless, the PCA model underperforms our tensor model and can easily overfit the data.

Why do the tensor models outperform the 2-D PCA? First, PCA uses only the cross-sectional dependencies to model the data. Instead, the 3-D tensor model additionally incorporates the lagged information, which empirically enables more accurate predictions. Second, the data compression rate for the tensor decomposition is much lower than PCA, because adding the lagged information imposes a specific structure on the panel data of firm characteristics. Consequently, our tensor model can handle a high-dimensional space of firm characteristics across many periods without overfitting the data. This salient feature of



(a) Bayesian Tensor ($K = 24$) **(b)** PCA

Figure 4: Prediction accuracy of missing values: tensor decomposition vs PCA, US dataset

This figure plots the in- and out-of-sample prediction accuracy in the US characteristic dataset. The data sample spans from January 1973 to December 2021. In the simulations, we first remove the repeated observations and next randomly simulate a sample of 20% missing entries (out-of-sample). We estimate model parameters using only the observed information, that is, the sample of the remaining 80% observations (in-sample). The R-squared is computed using equations (13)–(14). Panel (a) considers the tensor model, with $K = 24$ lags and tensor rank R between 50 and 300. Panel (b) shows the results based on the PCA approach, with the number of PCs between 1 and 100.

tensor models is essential for the missing-at-random assumption, which is more likely to be satisfied when conditioning on a larger information set.

Figure 5 plots the prediction accuracy in the US dataset for 13 cluster themes. Our forecast is based on a tensor model with $K = 24$ lags and tensor rank $R = 200$ in this figure. Some characteristics are more predictable than others. For instance, short-term reversals and seasonality variables are challenging to impute, with out-of-sample R-squared of about 5% and 35%, respectively. To properly handle the data missingness of these variables, not only do we need to impute their missing values, but we should also acknowledge the significant data uncertainty in the imputation process. On the contrary, our tensor model shows good imputation performance for the value, size, and quality characteristics.

We repeat the simulation studies in European and Asian datasets. Results are reported in Figures A1 and A2 of the Appendix. We get slightly lower out-of-sample prediction accuracies in international markets, probably due to the higher missing rates and shorter time-series samples. In these plots, we still detect very heterogeneous imputation performances across different clusters: Short-term reversals are the most difficult to forecast, while we have the

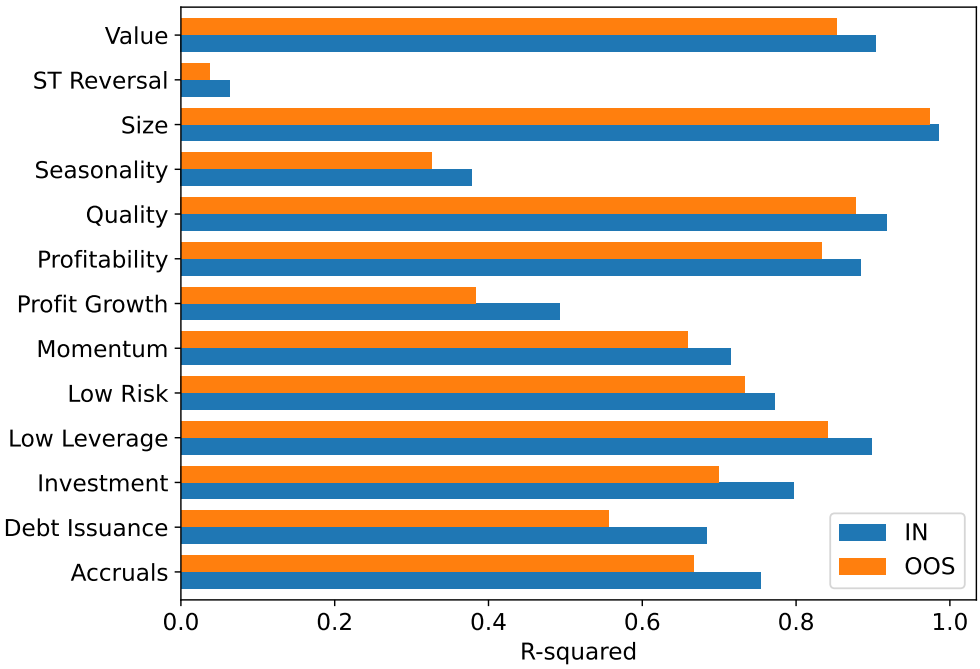


Figure 5: Prediction accuracy by 13 Themes: US data, $R = 200$

This figure plots the in- and out-of-sample prediction accuracy in the US characteristic dataset. The 151 firm characteristics are grouped into 13 cluster themes. The data sample starts from January 1973 to December 2021. In the simulations, we first remove the repeated observations and then randomly simulate a sample of 20% missing entries (out-of-sample). We estimate model parameters using only the observed information, that is, the sample of the remaining 80% observations (in-sample). The R-squared is computed using equations (13)–(14). We consider a tensor model with $K = 24$ lags and tensor rank $R = 200$.

highest R-squared in value, size, and quality clusters, entirely consistent with the simulation results in the US data.

4.2 Coverage of Confidence Intervals

In this subsection, we study the posterior coverages of confidence intervals provided by our estimates. Panel A of Table 2 presents the coverages of confidence intervals for the tensor model. These confidence intervals are based on 200 Bootstrap steps explained in Definition 1. We confirm that the tensor models generate appropriate confidence intervals. For instance, in the model with a tensor rank of 200, the 80%, 90%, 95%, and 98% confidence intervals contain 80.8%, 89.1%, 93.4%, and 95.8% of the true values in the simulated hold-out sample. However, as Panel B suggests, the Bootstrap based on PCA models fails to provide appropriate confidence bands. For example, when the number of PCs is 40, the 90%

Table 2: Coverages of confidence intervals in simulations

CI :=	80%	90%	95%	98%
Panel A. Bayesian tensor				
$R = 100$	0.797	0.888	0.934	0.960
$R = 150$	0.800	0.887	0.932	0.959
$R = 200$	0.808	0.891	0.934	0.959
$R = 250$	0.813	0.893	0.934	0.958
Panel B. Bootstrap based on PCA				
Number of PCs = 20	0.725	0.837	0.900	0.939
Number of PCs = 30	0.667	0.790	0.864	0.913
Number of PCs = 40	0.617	0.744	0.827	0.883
Number of PCs = 50	0.572	0.699	0.783	0.842

This table reports the coverages of confidence intervals in Monte Carlo simulations, where the missing rate is 20%. In Panel A, we consider tensor models with $K = 24$ lags and tensor rank R of 100, 150, 200, and 250. The confidence intervals in tensor models are constructed based on Definition 1. In Panel B, we study the PCA model with the number of PCs between 20 and 50. We construct the 80%, 90%, 95%, and 98% confidence intervals for the missing firm characteristics. All confidence intervals are based on 200 Bootstrapped samples.

confidence interval from PCA covers the hypothetical true values only in 74% of simulations.

Why is PCA unable to provide proper confidence intervals in the hold-out sample? Theoretically, we do not observe the actual distribution of prediction errors. Hence, our Bayesian-Bootstrap procedure draws the prediction errors for $\mathcal{Y}^{mis}$ from the in-sample distribution conditional on $\mathcal{Y}^{obs}$ and the in-sample parameter estimates. This requires the in-sample distribution of prediction errors to be close to the out-of-sample one. However, according to Figure 4, PCA tends to overfit the data, leading to much tighter distributions of in-sample prediction errors than the out-of-sample ones; hence, its confidence intervals based on in-sample estimates are distorted. This is confirmed by Table 2: The coverages of PCA confidence intervals decline as the number of PCs increases because the overfitting becomes more severe in a larger PCA model.

In summary, our tensor model accurately predicts the characteristic data in the hold-out sample and provides a proper distribution of these data points. Moreover, our framework can be used to generate probabilistic distributions of infrequently observed information and even all the data points, which is essential in handling data uncertainty in financial information.

5 Asset Pricing Implications of Data Uncertainty

In this section, we explore the asset pricing implications of data uncertainty. We employ tensor models with $K = 24$ monthly lags and tensor rank $R = 200$ because, as we show in previous simulations, the prediction accuracy remains almost constant when the tensor rank is greater than 200. We start the empirical analysis by exploring how accounting for data uncertainty in financial information impacts the performance of characteristic-managed portfolios, focusing on total returns in subsection 5.1, and separating mispricing from the systematic risk in subsection 5.2. Finally, in subsection 5.3, we demonstrate how the multiple imputations of firm characteristics can be used as synthetic data to perform pseudo out-of-sample testing for asset pricing models.

5.1 Univariate Characteristic-Managed Portfolios

First, we investigate how accounting for data uncertainty in firm characteristics affects the performance of univariate characteristic-managed portfolios. In European and Asian datasets, some firm characteristics are entirely missing at the beginning of the samples. To handle this issue, we require a firm characteristic to have at least ten cross-sectional observations per month; as a result, 118 European (and 110 Asian) characteristics are ultimately included in the asset pricing analysis. Details can be found in Table A1 of the Appendix.

Specifically, we construct the managed portfolios based on firm characteristics, as follows:

$$g_{j,t+1} = \frac{\boldsymbol{z}_{tj}^{\top} \boldsymbol{R}_{t+1}}{N_t}, \tag{16}$$

where $\boldsymbol{R}_{t+1}$ is an $N_t \times 1$ vector of individual stock returns at time $t+1$, and $\boldsymbol{z}_{tj}$ is the vector of rank-transformed firm characteristic j at time t. After imputing missing or infrequently updated values of characteristic j using the tensor model, we rescale their values to be located in the interval $[-0.5, 0.5]$. Due to the demeaned rank transformation, each characteristic-managed portfolio is a zero-investment long-short strategy.[10]

Using our Bayesian tensor models with $K = 24$ monthly lags and a tensor rank of 200,

[10] Past literature (e.g., Kozak, Nagel, and Santosh (2020), Gu, Kelly, and Xiu (2020)) often fills in missing Y_{tj} with cross-sectional means or zeros, so missing values therein do not affect the portfolio construction.

we obtain 100 posterior draws of firm characteristics, denoted as $\{\mathcal{Y}_t^{mis,(m)}\}_{m=1}^{100}$. These draws are then combined with the observed information $\mathcal{Y}_t^{obs}$ to create managed portfolios following equation (16). We explore two scenarios: 1) imputing only the missing values, and 2) imputing both missing values and infrequently observed information. In the latter case, we consider stale observations as noisy measures of the true signals and utilize the tensor model to study data uncertainty in these infrequently updated observations.

To handle parameter and data uncertainty simultaneously, we employ a hierarchical Bootstrap approach. For each imputed dataset, $\mathcal{Y}_t^{(m)}$, we construct the rank-transformed characteristics $\boldsymbol{Z}_t^{(m)}$ and characteristic-managed portfolios $\boldsymbol{G}_{t+1}^{(m)} = \frac{(\boldsymbol{Z}_t^{(m)})^\top \boldsymbol{R}_{t+1}}{N_t}$. Subsequently, we use the standard Bootstrap approach to generate the distributions of mean returns, denoted as $\{\bar{\boldsymbol{G}}^{(m,b)}\}_{b=1}^{B}$, for the m-th imputed dataset, where B is the number of Bootstrap samples and chosen to be 1000. In this step, we assume that asset returns are serially independent. Finally, we estimate the confidence intervals of $\bar{\boldsymbol{G}}$ using $\{\bar{\boldsymbol{G}}^{(m,b)}\}_{b=1,m=1}^{1000,100}$.

Table 3 reports the *annualized* Sharpe ratios of these univariate characteristic-managed portfolios. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics (that is, zeros due to the normalization). The two columns "Tensor models: only missing" and "Tensor models: missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: missing&stale" and "cs-mean". We rank the characteristics based on the absolute changes in Sharpe ratios in *descending* order. We focus on the top 15 variables with the largest absolute changes in Table 3, with more extensive results reported in Tables A2–A4 in Appendix. Several empirical findings are noteworthy.

First, accounting for data uncertainty in firm characteristics significantly impacts the Sharpe ratios of some managed portfolios, which is a consistent observation across all three datasets. For example, in the European data, the Sharpe ratio of the managed portfolio sorted by "rd5_at" (R&D capital-to-book assets) declines from 0.507 per annum (significantly positive) to only about 0.14 after the imputation of missing values. Similar declines in Sharpe ratios of characteristic-managed portfolios are observed in the US (e.g., "dsale_dinv") and Asian (e.g., "fcf_me") datasets. Nevertheless, imputing missing values does not mechan-

Table 3: Annualised Sharpe ratios of characteristic-managed portfolios

Label	Name	cs-mean	Tensor models: only missing	Tensor models: missing&stale	Δ
	Panel A. US data				
ocf_at_chg1	Change in operating cash flow to assets	0.497**	0.541**	0.802**	0.304**
seas_16_20na	Years 16-20 lagged returns	-0.161	0.141	0.141	0.303*
dsale_dinv	Change sales minus change Inventory	0.896**	0.705**	0.617**	-0.279**
seas_11_15na	Years 11-15 lagged returns, nonannual	-0.207	0.064	0.064	0.272**
ni_ar1	Earnings persistence	0.019	-0.081	-0.212	-0.231
seas_6_10na	Years 6-10 lagged returns	-0.521**	-0.301**	-0.301**	0.220**
dsale_dsga	Change sales minus change SG&A	-0.201	-0.128	-0.011	0.190**
sti_gr1a	Change in short-term investments	-0.055	-0.047	0.131	0.186*
cop_at	Cash-operating-profits-to-assets	1.412**	1.236**	1.229**	-0.184**
dbnetis_at	Net debt issuance	-0.723**	-0.732**	-0.872**	-0.149
debt_gr3	Growth in book debt (3 years)	-0.799**	-0.952**	-0.941**	-0.142*
taccruals_at	Total accruals	-0.303**	-0.251*	-0.163	0.139**
sale_emp_gr1	Labor force efficiency	-0.071	-0.051	0.064	0.135
lti_gr1a	Change in long-term investments	-0.150	-0.174	-0.285	-0.135
inv_gr1a	Inventory change	-0.783**	-0.759**	-0.649**	0.134*
	Panel B. European data				
rd5_at	R&D capital-to-book assets	0.507**	0.143	0.137	-0.370**
cowc_gr1a	Change in current operating working capital	-1.065**	-0.882**	-0.733**	0.332*
lti_gr1a	Change in long-term investments	-0.311*	-0.191	-0.022	0.289
sti_gr1a	Change in short-term investments	-0.136	-0.085	0.144	0.280
cop_at	Cash-operating-profits-to-assets	1.123**	0.862**	0.858**	-0.265**
niq_su	Standardized earnings surprise	0.585**	0.925**	0.846**	0.261
dgp_dsale	Change gross margin minus change sales	0.243	0.534**	0.498**	0.254
ncol_gr1a	Change in noncurrent operating liabilities	-0.393**	-0.320*	-0.139	0.254
eqnpo_me	Net payout yield	0.368**	0.617**	0.615**	0.247**
rd_me	R&D-to-market	0.489**	0.245	0.243	-0.246*
nfna_gr1a	Change in net financial assets	1.116**	1.047**	0.879**	-0.237*
fnl_gr1a	Change in financial liabilities	-1.094**	-1.050**	-0.861**	0.233*
dsale_dsga	Change sales minus change SG&A	-0.010	0.177	0.216	0.226
aliq_at	Liquidity of book assets	-0.415**	-0.233	-0.191	0.224**
sale_emp_gr1	Labor force efficiency	0.077	0.154	0.295	0.218
	Panel C. Asian data				
eqpo_me	Payout yield	0.649**	0.967**	0.957**	0.308**
eqnpo_me	Net payout yield	0.674**	0.926**	0.925**	0.251*
dgp_dsale	Change gross margin minus change sales	-0.236	-0.094	0.008	0.243
eqnetis_at	Net equity issuance	-0.168	-0.286	-0.408**	-0.240
ncol_gr1a	Change in noncurrent operating liabilities	0.190	0.160	-0.047	-0.238
dsale_dsga	Change sales minus change SG&A	-0.020	0.121	0.207	0.226
oaccruals_ni	Percent operating accruals	-0.008	-0.071	-0.220	-0.212*
cowc_gr1a	Change in current operating working capital	-0.501**	-0.374**	-0.294	0.207
lti_gr1a	Change in long-term investments	-0.084	-0.120	-0.278	-0.194
ni_ar1	Earnings persistence	-0.078	-0.024	0.099	0.177
fcf_me	Free cash flow-to-price	0.766**	0.599**	0.600**	-0.167
ocf_at_chg1	Change in operating cash flow to assets	0.057	0.033	0.214	0.157
ppeinv_gr1a	Change PPE and Inventory	-0.019	-0.021	-0.172	-0.153
cash_at	Cash-to-assets	-0.030	-0.132	-0.164	-0.134*
chcsho_12m	Net stock issues	-0.143	-0.141	-0.268	-0.126

This table reports the annualized Sharpe ratios of characteristic-managed portfolios. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**).

ically decrease Sharpe ratios. In the Asian dataset, the portfolios sorted by variables related to payout yields earn higher Sharpe ratios in imputed datasets compared to those based on cross-sectional mean imputation. This highlights the variability in the consequence of imputing missing values on portfolio performance across different datasets and characteristics.

Table 3 also underscores the importance of imputing infrequently observed accounting information. For instance, consider the quarterly updated variable "Change in operating cash flow to assets" in the US dataset. Imputing missing values in this characteristic results in a slight change in the Sharpe ratio from 0.497 to 0.541. However, further replacing the stale quarterly observations with monthly model predictions leads to a substantial increase in the Sharpe ratio to 0.802 per year. This demonstrates the importance of accounting for data uncertainty in both missing and infrequently observed data.

5.2 Multivariate Asset Pricing Models

The portfolio analysis in subsection 5.1 does not account for the systematic risk carried by characteristic-managed portfolios. In this subsection, we study a multivariate asset pricing model—the IPCA model developed by Kelly, Pruitt, and Su (2019)—to decompose managed portfolio returns into mispricing (alpha) and systematic (beta) components. In particular, we aim to explore how accounting for data uncertainty reshapes the risk-return tradeoffs.

The functional form of the IPCA model is as follows:

$$\boldsymbol{R}_{t+1} = \boldsymbol{\beta}_t \boldsymbol{f}_{t+1} + \boldsymbol{e}_{t+1}, \quad \boldsymbol{\beta}_t = \boldsymbol{Z}_t \boldsymbol{\Gamma}_{\boldsymbol{\beta}}, \tag{17}$$

where $\boldsymbol{Z}_t$ is an $N_t \times (J+1)$ matrix of instruments (consisting of a constant term plus rank-transformed firm characteristics), $\boldsymbol{f}_{t+1}$ is a $P \times 1$ vector of common latent factors, the $(J+1) \times P$ matrix $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$ defines the mapping between the instruments and time-varying beta loadings $\boldsymbol{\beta}_t$, and $\boldsymbol{e}_{t+1}$ represents idiosyncratic shocks. We follow the exact alternating least squares method in Kelly, Pruitt, and Su (2019) to estimate $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{f}_{t+1}$ iteratively, that is,

$$\hat{\boldsymbol{f}}_{t+1} = \left(\boldsymbol{\Gamma}_\beta^\top \boldsymbol{Z}_t^\top \boldsymbol{Z}_t \boldsymbol{\Gamma}_\beta\right)^{-1} \boldsymbol{\Gamma}_\beta^\top \boldsymbol{Z}_t^\top \boldsymbol{R}_{t+1}, \quad \text{and} \tag{18}$$

$$\text{vec}(\mathbf{\Gamma}_\beta^\top) = \left[\sum_{t=1}^{T}(\boldsymbol{Z}_t^\top \boldsymbol{Z}_t) \otimes (\boldsymbol{f}_{t+1}\boldsymbol{f}_{t+1}^\top)\right]^{-1}\left[\sum_{t=1}^{T}(\boldsymbol{Z}_t^\top \boldsymbol{R}_{t+1}) \otimes \boldsymbol{f}_{t+1}\right]. \tag{19}$$

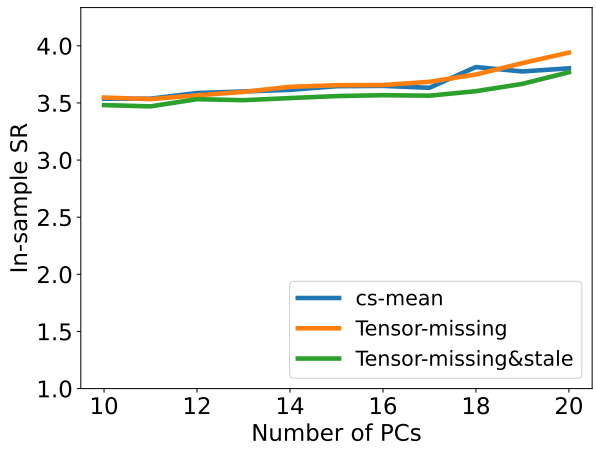
With the estimates of $\mathbf{\Gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{f}_{t+1}$, we define the unconditional alphas of characteristic-managed portfolios as the difference between their realized and predicted mean returns:

$$\boldsymbol{\alpha}_{\text{char}} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\boldsymbol{Z}_t^\top \boldsymbol{R}_{t+1}}{N_t} - \frac{\boldsymbol{Z}_t^\top \boldsymbol{Z}_t}{N_t}\mathbf{\Gamma}_{\boldsymbol{\beta}}\boldsymbol{f}_{t+1}\right]. \tag{20}$$

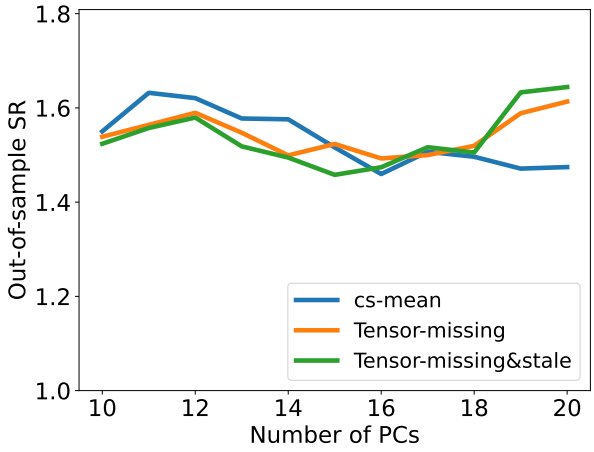
We first investigate whether the estimation of the priced systematic components in stock returns—that is, the factor tangency portfolios composed of systematic factors $\boldsymbol{f}_{t+1}$—is affected by data uncertainty. Specially, for each imputed dataset, we consider a wide range of P values, extract the P latent factors, and report the in- and out-of-sample Sharpe ratios derived from these latent factor estimates. For the in-sample exercises, the US data sample spans from January 1973 to December 2021, while the European and Asian data samples cover the period from July 1986 to December 2021. In the out-of-sample exercises, we estimate all model parameters, including $\mathbf{\Gamma}_{\boldsymbol{\beta}}$ and portfolio weights on $\boldsymbol{f}_{t+1}$, given by $\mathbb{E}[\boldsymbol{f}_t]^\top \text{cov}(\boldsymbol{f}_t)^{-1}$, using the pre-2007 samples and assess the out-of-sample performance in the subsample from January 2007 to December 2021.

Figure 6 demonstrates that ten latent factors are nearly sufficient to capture all the priced systematic information in the US cross-section of stock returns. In comparison, fifteen latent factors are almost sufficient for the European and Asian cross-sections, regardless of whether the performance is evaluated in-sample or out-of-sample. Moreover, imputing missing and infrequently observed information has almost negligible effects on the mean-variance efficiency of the systematic components. This aligns with the observation in Chen and McCoy (2024) that simply imputing with cross-sectional means performs well compared to more sophisticated imputation methods in the context of stock return predictions. Our findings, based on extensive international datasets, reinforce this argument within the context of systematic latent factor models.

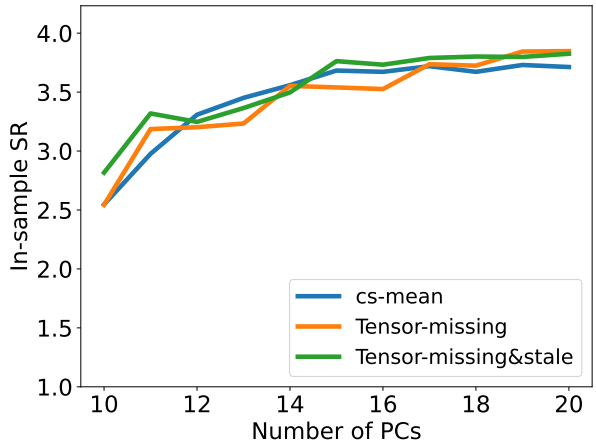
Motivated by the evidence in Figure 6, we use 15-factor models to estimate the alphas of characteristic-managed portfolios as outlined in equation (20). Table 4 specifically examines the characteristics that exhibit changes in their statistical significance (at the 5% level),



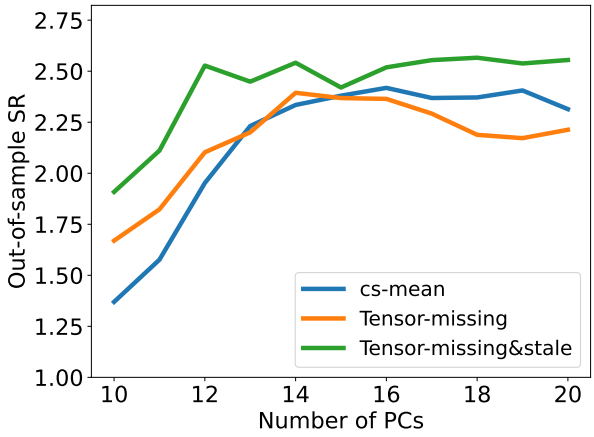
(a) In-sample SR, US data



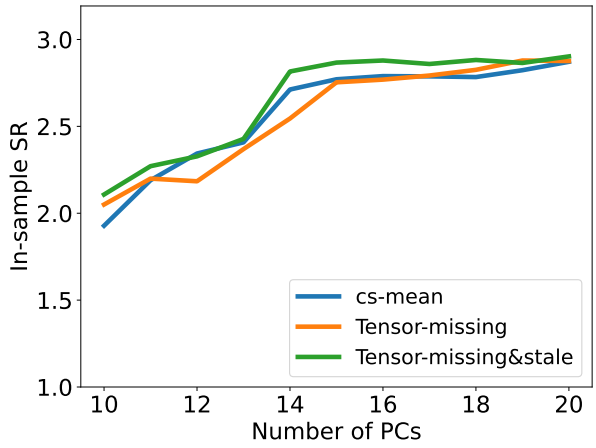
(b) Out-of-sample SR, US data



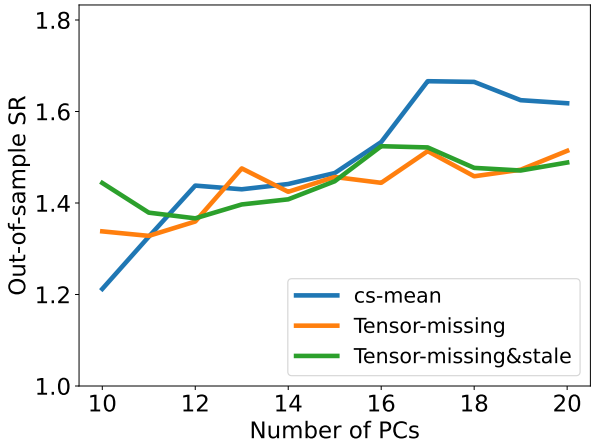
(c) In-sample SR, European data



(d) Out-of-sample SR, European data



(e) In-sample SR, Asian data



(f) Out-of-sample SR, Asian data

Figure 6: In- and out-of-sample Sharpe ratios of IPCA factor tangency portfolios

This figure plots the in- and out-of-sample Sharpe ratios of the factor IPCA tangency portfolios. The number of latent factors ranges from 10 to 20. The label "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means. The labels "Tensor-missing" and "Tensor-missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. In the in-sample exercises, the US data sample starts from January 1973 to December 2021, whereas the European and Asian data samples range from July 1986 to December 2021. In the out-of-sample exercises, we estimate all model parameters ($\mathbf{\Gamma}_{\boldsymbol{\beta}}$ and portfolio weights on $\boldsymbol{f}_{t+1}$) using the pre-2007 samples and evaluate the out-of-sample performance in the subsample between January 2007 and December 2021.

following the imputation of missing and stale data. Table A5 in the Appendix presents the estimates for the remaining characteristics whose IPCA alphas are consistently significant regardless of data imputation. For comparability, we report the Sharpe ratios of the alphas.

First, accounting for data uncertainty in both missing and stale observations reduces the number of significant alphas of managed portfolios. At the 5% significance level, the counts of significant alphas—with missing values imputed as zeros—are 13, 12, and 27 in the US, European, and Asian datasets, respectively. However, when employing the tensor model to impute both missing and infrequently observed data, there are only 8, 11, and 19 significant alphas in these datasets.

Furthermore, data imputation does not always decrease the magnitudes of the managed-portfolio alphas. For example, in the European cross-section, the alpha of the equity-net-payout managed portfolio is 0.18 and insignificant when using cross-sectional mean imputation. However, with the missing entries replaced by our tensor model predictions, its alpha significantly increases to 0.407.

In a nutshell, leveraging the IPCA model of Kelly, Pruitt, and Su (2019), we find that accounting for data uncertainty in financial information does not affect the performance of the systematic component of asset returns. In contrast, we highlight the importance of taking data uncertainty into consideration when estimating the alphas of managed portfolios.

5.3 Pseudo Out-of-Sample Tests via Multiple Imputations

Empirical asset pricing studies often conduct out-of-sample analyses by dividing the full sample into several subsamples, such as the analysis illustrated in Figure 6. The primary goal of out-of-sample testing is to assess a model's performance on data that was not used during model training. In this subsection, we demonstrate how our tensor model can generate simulated panels of firm characteristics that can be used in the pseudo out-of-sample tests. Specifically, we estimate all model parameters using a single sample and then examine how the asset pricing model behaves across multiple imputations of characteristic data.

First, we use the pseudo out-of-sample tests to determine the number of latent factors in the IPCA model. Intuitively, the large and dominant latent factors should maintain consistent performance even when some randomness is introduced into the firm characteristic

Table 4: IPCA alphas of characteristic-managed portfolios

Label	Name	cs-mean	Tensor models: only missing	Tensor models: missing&stale	Δ
	Panel A. US data				
corr_1260d	Market correlation	-0.231	-0.268*	-0.326**	-0.095
inv_gr1a	Inventory change	-0.367**	-0.361**	-0.278*	0.089
eqnpo_12m	Equity net payout	0.296**	0.255*	0.240	-0.057
niq_be	Quarterly return on equity	0.329**	0.346**	0.281*	-0.049
seas_6_10na	Years 6-10 lagged returns	-0.305**	-0.257*	-0.275*	0.030
cowc_gr1a	Change in current operating working capital	-0.292**	-0.335**	-0.266*	0.026
ret_12_7	Momentum t-12 to t-7	-0.306**	-0.337**	-0.284*	0.022
	Panel B. European data				
nncoa_gr1a	Change in net noncurrent operating assets	-0.169	-0.317*	-0.408**	-0.239
eqnpo_12m	Equity net payout	0.180	0.355*	0.407**	0.227**
ncoa_gr1a	Change in noncurrent operating assets	-0.251	-0.403**	-0.422**	-0.171
noa_gr1a	Change in net operating assets	-0.290*	-0.420**	-0.454**	-0.164
debt_me	Debt-to-market	-0.287*	-0.349*	-0.427**	-0.140
eq_dur	Equity duration	-0.386**	-0.225	-0.178	0.208
ival_me	Intrinsic value-to-market	0.393**	0.223	0.185	-0.208
be_me	Book-to-market equity	0.401**	0.274	0.214	-0.187
taccruals_at	Total accruals	0.379**	0.355*	0.253	-0.126
taccruals_ni	Percent total accruals	0.364**	0.361*	0.319*	-0.046
ni_ivol	Earnings volatility	0.356**	0.335*	0.317	-0.039
	Panel C. Asian data				
netis_at	Net total issuance	-0.297*	-0.382**	-0.406**	-0.109
ocf_at	Operating cash flow to assets	0.287*	0.291	0.376**	0.090
iskew_capm_21d	Idiosyncratic skewness, CAPM	0.322*	0.392**	0.399**	0.076
noa_at	Net operating assets	-0.337*	-0.409**	-0.371**	-0.034
seas_1_1na	Year 1-lagged return	0.471**	0.374*	0.199	-0.273
rd_sale	R&D-to-sales	0.459**	0.199	0.272	-0.187
chcsho_12m	Net stock issues	-0.379**	-0.325*	-0.194	0.186
div12m_me	Dividend yield	0.407**	0.355*	0.252	-0.154
cowc_gr1a	Change in current operating working capital	-0.370**	-0.255	-0.235	0.135
inv_gr1a	Inventory change	-0.379**	-0.297	-0.253	0.125
mispricing_perf	Mispricing factor: Performance	0.421**	0.453**	0.320*	-0.101
eqnpo_12m	Equity net payout	0.375**	0.332*	0.281	-0.094
fcf_me	Free cash flow-to-price	0.374**	0.327*	0.297	-0.077
kz_index	Kaplan-Zingales index	-0.344**	-0.260	-0.277	0.066
at_gr1	Asset Growth	-0.393**	-0.342*	-0.363*	0.029
cop_atl1	Cash-operating-profits-to-lagged-assets	0.377**	0.328*	0.356*	-0.022

This table reports the IPCA alphas of characteristic-managed portfolios in 15-factor models. The alphas are defined as the difference between their realized mean returns and model-predicted mean returns, as in equation (20). For comparability, we report the Sharpe ratios of the alphas. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: only missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: only missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**).

data, whereas this added noise will have a greater effect on the small, low-variance factors that are not priced. The results are reported in Figure 7. For each imputed dataset ("CS-mean", "Tensor model: imputing missing", and "Tensor model: imputing missing&stale"), we obtain the in-sample estimates of $\mathbf{\Gamma}_{\boldsymbol{\beta}}$ and next construct the pseudo out-of-sample factor tangency portfolios in each of the 100 multiple imputations of firm characteristics. In each plot, the x-axis shows the number of latent factors, whereas the y-axis shows the out-of-sample annualized Sharpe ratio. We report both the average performance and the one-standard-deviation error bars, denoted by the shaded areas, following the one-standard-error rule in typical cross-validation exercises.[11] According to Figure 7, the IPCA model is the most sparse in the US dataset, requiring 10 to 12 latent factors under the one-standard-error rule. On the contrary, the models in the European and Asian data tend to be denser. For example, Panel (b) suggests a 15-factor model in the "CS-mean" imputation but 17- and 20-factor models in the two tensor-model imputations.

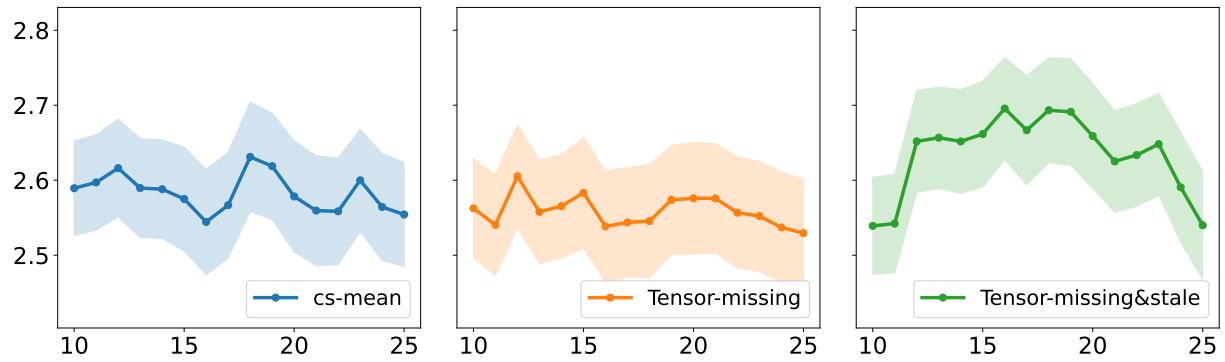
Although the pseudo out-of-sample exercise shares similar spirits to the cross-validation, it differs in that it utilizes the entire time-series information in both the model estimation and evaluation, thereby enhancing statistical efficiency. In comparison, methods such as K-fold cross-validation, split-window, and rolling-window analyses must manage the trade-off between estimation uncertainty from a limited in-sample period and evaluation uncertainty from a short out-of-sample period. As a result, the pseudo out-of-sample method serves as a useful complementary tool for refining model complexity.

Next, we present a "modified" version of the IPCA model. Compared with equation (17), conditional alphas of the individual stock returns, denoted as $\boldsymbol{\alpha}_t$, are incorporated in the return data-generating process:

$$\boldsymbol{R}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t \boldsymbol{f}_{t+1} + \boldsymbol{e}_{t+1}, \text{ and } \boldsymbol{\beta}_t = \boldsymbol{Z}_t \mathbf{\Gamma}_{\boldsymbol{\beta}}. \tag{21}$$

Kelly, Pruitt, and Su (2019) assume that conditional alphas are linear in the instruments, i.e., $\boldsymbol{\alpha}_t = \boldsymbol{Z}_t \mathbf{\Gamma}_{\boldsymbol{\alpha}}$. However, there are some pitfalls of this parametric assumption, as outlined

[11] In cross-validation exercises, the one-standard-error rule requires us to find the model with the highest out-of-sample Sharpe ratio and next select the simplest model whose average performance falls within one standard deviation of the best performance.



(a) US data



(b) European data



(c) Asian data

Figure 7: Pseudo out-of-sample Sharpe ratios of IPCA factor tangency portfolios, 100 multiple imputations of all data

This figure plots the pseudo out-of-sample performance of the IPCA factor tangency portfolios in 100 multiple imputations of all data. The label "CS-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The labels "Tensor model: imputing missing" and "Tensor model: imputing missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. For each imputed dataset, we obtain the in-sample estimates of $\mathbf{\Gamma}_{\boldsymbol{\alpha}}$ and $\mathbf{\Gamma}_{\boldsymbol{\beta}}$ and next construct the pseudo out-of-sample factor tangency portfolios in each of the 100 multiple imputations of firm characteristics. In each plot, the x-axis shows the number of latent factors, whereas the y-axis shows the out-of-sample Sharpe ratio per annum. We report the average performance, with the shaded areas denoting the one-standard deviation error.

in the following remark.

Remark. *Under the assumption of* $\boldsymbol{\alpha}_t = \boldsymbol{Z}_t\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$*, Kelly, Pruitt, and Su (2019) identify the latent factors as* $\hat{\boldsymbol{f}}_{t+1} = \left(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\right)^{-1}\boldsymbol{\beta}_t^\top\left(\boldsymbol{R}_{t+1} - \boldsymbol{\alpha}_t\right) = \left(\boldsymbol{\Gamma}_\beta^\top\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\boldsymbol{\Gamma}_\beta\right)^{-1}\boldsymbol{\Gamma}_\beta^\top\boldsymbol{Z}_t^\top\left(\boldsymbol{R}_{t+1} - \boldsymbol{Z}_t\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}\right)$*, which implies that the estimated latent factors are nontradable unless* $\boldsymbol{\alpha}_t^\top\boldsymbol{\beta}_t = \mathbf{0}$*. This condition is almost always violated when* $\boldsymbol{\alpha}_t = \boldsymbol{Z}_t\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$*. However, if the factors are nontradable,* $\boldsymbol{\alpha}_t$ *are not necessarily stock alphas. To see this, note that alphas are defined as the residual terms unexplained by systematic exposures* $\boldsymbol{\beta}_t$*, that is,* $\mathbb{E}_t\left[\boldsymbol{R}_{t+1}\right] - \boldsymbol{\beta}_t\boldsymbol{\lambda}_{ft}$*, where* $\boldsymbol{\lambda}_{ft}$ *are risk premia of* $\boldsymbol{f}_{t+1}$*. Equation (21) implies that* $\boldsymbol{\alpha}_t$ *are equal to* $\mathbb{E}_t\left[\boldsymbol{R}_{t+1}\right] - \boldsymbol{\beta}_t\boldsymbol{\lambda}_{ft}$ *if and only if* $\boldsymbol{\lambda}_{ft} = \mathbb{E}_t\left[\boldsymbol{f}_{t+1}\right]$*, which will hold naturally for tradable* $\boldsymbol{f}_{t+1}$*.*

Motivated by the above remark, we assume the following parametric form of $\boldsymbol{\alpha}_t$:

$$\boldsymbol{\alpha}_t = \left[\boldsymbol{I} - \boldsymbol{\beta}_t\left(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\right)^{-1}\boldsymbol{\beta}_t^\top\right]\boldsymbol{Z}_t\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}. \tag{22}$$

The parametric form of alphas in equation (22) satisfies several favorable properties: 1) it is naturally orthogonal to $\boldsymbol{\beta}_t$; 2) the estimates of $\boldsymbol{f}_{t+1}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$ are exactly the same as those given in equations (18)–(19) and are unrelated to alphas, representing a minimal deviation from Kelly, Pruitt, and Su (2019); 3) it upholds the intuition that alphas are residuals unexplained by systematic risk. We estimate $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$ via the least squares method described in Appendix C. Finally, we define the parametric arbitrage portfolio as

$$R_{\alpha,t+1} = \frac{\boldsymbol{\alpha}_t^\top\boldsymbol{R}_{t+1}}{N_t} = \frac{\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}^\top\boldsymbol{Z}_t^\top\left[\boldsymbol{I} - \boldsymbol{\beta}_t\left(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\right)^{-1}\boldsymbol{\beta}_t^\top\right]\boldsymbol{R}_{t+1}}{N_t}.^{12}$$

We now study the pseudo out-of-sample performance of the model in equation (21). The existing literature often emphasizes estimation uncertainty of model parameters—e.g., $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$—conditional on the observed data. Our paper focuses on an entirely different source of uncertainty—data uncertainty in financial information. We aim to assess whether the model performance, based on these optimal in-sample estimates of $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$, remains robust to permutations of firm characteristics.

[12]Note that the Sharpe ratio of $R_{\alpha,t+1}$ provides only a lower bound of the true degree of mispricing. To maximize the performance, one should construct $\boldsymbol{\alpha}_t^\top\mathrm{cov}_t(\boldsymbol{e}_{t+1})^{-1}\boldsymbol{R}_{t+1}$. Nevertheless, this portfolio requires us to measure the high-dimensional conditional covariance of idiosyncratic shocks, which is empirically challenging. Moreover, the true stock alphas may deviate from the parametric assumption in equation (22).

Table 5: Data uncertainty in the arbitrage and factor tangency portfolios

Number of latent factors = 15		US		EUR		ASIA	
		SR_α	SR_β	SR_α	SR_β	SR_α	SR_β
Panel A. In-sample performance							
CS-mean	imputing missing	1.49	3.64	1.50	3.68	1.46	2.77
Tensor model	imputing missing	1.53	3.65	1.56	3.54	1.69	2.75
	imputing missing&stale	1.49	3.56	1.39	3.76	1.58	2.86
Panel B. Pseudo out-of-sample performance in 100 multiple imputations of all data							
C-S mean	imputing missing	0.66	2.74	0.32	2.84	0.29	2.27
		[0.50, 0.85]	[2.60, 2.86]	[0.16, 0.50]	[2.66, 3.06]	[0.13, 0.40]	[2.16, 2.43]
Tensor model	imputing missing	0.67	2.75	0.77	2.74	0.48	2.28
		[0.50, 0.85]	[2.60, 2.91]	[0.58, 1.00]	[2.57, 2.94]	[0.31, 0.62]	[2.12, 2.45]
	imputing missing&stale	0.51	2.76	0.41	2.79	0.59	2.28
		[0.33, 0.68]	[2.62, 2.89]	[0.19, 0.60]	[2.57, 2.98]	[0.34, 0.79]	[2.13, 2.46]

This table reports the Sharpe ratios of the arbitrage (SR_α) and factor tangency (SR_β) portfolios in IPCA models with 15 latent factors. The row "CS-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The rows "Tensor model: imputing missing" and "Tensor model: imputing missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. Panel A shows the in-sample performance, whereas Panel B displays the pseudo out-of-sample performance in 100 multiple imputations. That is, after we obtain the in-sample estimates of $\mathbf{\Gamma_\alpha}$ and $\mathbf{\Gamma_\beta}$, we construct the pseudo out-of-sample arbitrage and beta portfolios in each of the 100 multiple imputations of firm characteristics. The numbers within square brackets in Panels B denote the 95% confidence intervals across 100 multiple imputations (pseudo out-of-sample datasets).

Table 5 reports the results based on 15-factor IPCA models. We show the in-sample Sharpe ratios in Panel A, whereas Panel B reports the median pseudo out-of-sample Sharpe ratios of the arbitrage and factor tangency portfolios in 100 random samples of all firm characteristics, with the 95% confidence intervals in the square brackets. That is, we treat all observed data as probabilistic quantities and use multiple imputations to account for data uncertainty in these data.

We observe that the performance of factor tangency portfolios only declines slightly in the pseudo out-of-sample data. Additionally, the imputation methods do not significantly affect the Sharpe ratios of the systematic components, which echoes the findings in Figure 6. However, the Sharpe ratios of the arbitrage portfolios decline by more than 50%, ranging between 0.29 and 0.77 in three datasets. The observation of relatively low Sharpe ratios of the parametric arbitrage portfolios is aligned with the absence of near-arbitrage opportunities in the arbitrage pricing theory (Ross, 1976; Huberman, 1982; Chamberlain and Rothschild, 1983; Ingersoll, 1984). Unlike the current literature emphasizing estimation errors that hinder arbitrageurs from exploiting all pricing errors (e.g., Da, Nagel, and Xiu (2024)), our paper underscores the important role of data uncertainty in assessing statistical arbitrage.

6 Conclusion

Our paper introduces a new Bayesian framework to address data uncertainty in the high-dimensional panel of firm characteristics. The model incorporates two key components. First, our model employs multiple imputations of firm characteristics to properly handle uncertainty about financial data. This involves treating firm characteristics as probabilistic variables rather than fixed values, acknowledging that even observed data contains measurement errors. Second, our model is built on the high-dimensional tensor representation, which jointly captures time-series and cross-sectional dependencies in a unified parametric framework. Our approach overcomes the limitations of existing approaches that model these two dimensions separately. Unlike the traditional tensor models, we refine the tensor decomposition to accommodate highly unbalanced panel data with pervasive missingness. Hence, our method can be applied to general panel datasets with these data features and is not restricted to data on firm characteristics.

Our empirical analysis, focusing on the characteristic data in Jensen, Kelly, and Pedersen (2023), demonstrates several key advantages of our approach. The tensor model produces higher prediction accuracy for missing elements than traditional methods like cross-sectional PCA. Moreover, our Bayesian framework can provide valid confidence intervals for the missing firm characteristics, which is essential in addressing potential data uncertainty.

We explore the asset pricing implications of accounting for data uncertainty. We find that recovering the missing and infrequently observed information has substantial impacts on the performance of univariate characteristic-managed portfolios. In the multivariate asset pricing analysis, although the performance of the factor tangency portfolios does not improve, the number of significant alphas in these managed portfolios tends to decline after considering data uncertainty. We also demonstrate how the multiple imputations of all characteristic data can be used as synthetic data to conduct pseudo out-of-sample tests for asset pricing models. In particular, the Sharpe ratios of the statistical arbitrage, although being high in-sample, become significantly lower in the pseudo out-of-sample and within the reasonable bounds—more aligned with the absence of near-arbitrage opportunities—after accounting for data uncertainty in financial information.

Looking forward, our framework opens new avenues for research in empirical asset pricing. By providing the profession with the distribution of firm characteristics, we enable researchers to conduct more powerful tests, make valid statistical inferences, and develop more robust investment strategies that properly account for data uncertainty. Our results suggest that careful treatment of data uncertainty in financial information is not merely a technical consideration but can fundamentally enhance our understanding of the risk-return tradeoff in financial markets.

References

Anderson, E. W., and A.-R. Cheng (2016): "Robust Bayesian portfolio choices," *Review of Financial Studies*, 29(5), 1330–1375.

Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64(3), 423–458.

——— (2004): "Stock return predictability and asset pricing models," *Review of Financial Studies*, 17(3), 699–738.

Barillas, F., and J. Shanken (2018): "Comparing asset pricing models," *Journal of Finance*, 73(2), 715–754.

Beckmeyer, H., and T. Wiedemann (2022): "Recovering missing firm characteristics with attention-based machine learning," *Available at SSRN 4003455*.

Bryzgalova, S., J. Huang, and C. Julliard (2023): "Bayesian solutions for the factor zoo: We just ran two quadrillion models," *Journal of Finance*, 78(1), 487–557.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2025): "Missing financial data," *Review of Financial Studies*, 38(3), 803–882.

Carroll, J. D., and J.-J. Chang (1970): "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition," *Psychometrika*, 35(3), 283–319.

Cen, Z., and C. Lam (2025): "Tensor time series imputation through tensor factor modelling," *Journal of Econometrics*, 249, 105974.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica: Journal of the Econometric Society*, 51(5), 1281–1304.

Chen, A. Y., and J. McCoy (2024): "Missing values handling for machine learning portfolios," *Journal of Financial Economics*, 155, 103815.

Chen, B., Y. Han, and Q. Yu (2024): "Estimation and Inference for CP Tensor Factor Models," *arXiv preprint arXiv:2406.17278*.

Chib, S., X. Zeng, and L. Zhao (2020): "On comparing asset pricing models," *Journal of Finance*, 75(1), 551–577.

Da, R., S. Nagel, and D. Xiu (2024): "The statistical limit of arbitrage," *Available at SSRN 4986807*.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2025): "Missing data in asset pricing panels," *Review of Financial Studies*, 38(3), 760–802.

Garlappi, L., R. Uppal, and T. Wang (2007): "Portfolio selection with parameter and model uncertainty: A multi-prior approach," *Review of Financial Studies*, 20(1), 41–81.

Ghysels, E., A. Babii, and J. Pan (2024): "Tensor PCA for Factor Models," *Available at SSRN 4791809*.

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

Harshman, R. A. (1970): "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multimodal factor analysis," .

Håstad, J. (1990): "Tensor rank is NP-complete," *Journal of Algorithms*, 11(4), 644–654.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica: Journal of the Econometric Society*, pp. 153–161.

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

Huang, F., K. Lu, Y. Zheng, and G. Li (2025): "Supervised factor modeling for high-dimensional linear time series," *Journal of Econometrics*, 249, 105995.

Huang, J., and R. Shi (2025): "Model uncertainty in the cross-section of stock returns," *Journal of Econometrics*, p. 106066.

Huberman, G. (1982): "A simple approach to arbitrage pricing theory," *Journal of Economic Theory*, 28(1), 183–191.

Ingersoll, J. E. (1984): "Some results in the theory of arbitrage pricing," *Journal of Finance*, 39(4), 1021–1039.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2023): "Is there a replication crisis in finance?," *Journal of Finance*, 78(5), 2465–2518.

Kan, R., and G. Zhou (2007): "Optimal portfolio choice with parameter uncertainty," *Journal of Financial and Quantitative Analysis*, 42(3), 621–656.

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

Kolda, T. G., and B. W. Bader (2009): "Tensor decompositions and applications," *SIAM review*, 51(3), 455–500.

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

Lettau, M. (2024): "High-dimensional factor models and the factor zoo," Discussion paper, National Bureau of Economic Research.

Little, R. J. (1993): "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, 88(421), 125–134.

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

Pástor, L. (2000): "Portfolio selection and asset pricing models," *Journal of Finance*, 55(1), 179–223.

Pástor, L., and R. F. Stambaugh (2000): "Comparing asset pricing models: an investment perspective," *Journal of Financial Economics*, 56(3), 335–381.

Ross, S. A. (1976): "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13(3), 341–360.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63(3), 581–592.

——— (1996): "Multiple imputation after 18+ years," *Journal of the American statistical Association*, 91(434), 473–489.

Rubin, D. B., and N. Schenker (1986): "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse," *Journal of the American statistical Association*, 81(394), 366–374.

Tu, J., and G. Zhou (2004): "Data-generating process uncertainty: What difference does it make in portfolio decisions?," *Journal of Financial Economics*, 72(2), 385–421.

——— (2010): "Incorporating economic objectives into Bayesian priors: Portfolio choice under parameter uncertainty," *Journal of Financial and Quantitative Analysis*, 45(4), 959–986.

Wang, D., Y. Zheng, and G. Li (2024): "High-dimensional low-rank tensor autoregressive time series modeling," *Journal of Econometrics*, 238(1), 105544.

Xiong, R., and M. Pelger (2022): "Large dimensional latent factor modeling with missing observations and applications to causal inference," *Journal of Econometrics*.

Appendices

Appendix A Ignorability

Let's use the notation θ to denote the parameters that govern the data model, $p(\mathcal{Y} \mid \theta)$. The posterior distribution of all parameters (θ, ψ) is

$$p(\theta, \psi \mid \mathcal{Y}^{obs}, \mathcal{I}) \propto p(\theta, \psi) \cdot p(\mathcal{Y}^{obs}, \mathcal{I} \mid \theta, \psi). \tag{A1}$$

We further impose two assumptions: (a) the parameters θ and ψ are a priori independent: $p(\theta, \psi) = p(\theta)p(\psi)$; and (b) MAR, which implies

$$\begin{aligned} p(\mathcal{Y}^{obs}, \mathcal{I} \mid \theta, \psi) &\propto \int p(\mathcal{Y}^{obs}, \mathcal{Y}^{mis} \mid \theta) p(\mathcal{I} \mid \mathcal{Y}^{obs}, \mathcal{Y}^{mis}, \psi) d\mathcal{Y}^{mis} \\ &= \int p(\mathcal{Y}^{obs}, \mathcal{Y}^{mis} \mid \theta) p(\mathcal{I} \mid \mathcal{Y}^{obs}, \psi) d\mathcal{Y}^{mis} \quad \text{(MAR)} \\ &= p(\mathcal{I} \mid \mathcal{Y}^{obs}, \psi) \int p(\mathcal{Y}^{obs}, \mathcal{Y}^{mis} \mid \theta) d\mathcal{Y}^{mis} \\ &= p(\mathcal{I} \mid \mathcal{Y}^{obs}, \psi) p(\mathcal{Y}^{obs} \mid \theta) \implies \end{aligned}$$

$$p(\theta, \psi \mid \mathcal{Y}^{obs}, \mathcal{I}) \propto [p(\theta)p(\mathcal{Y}^{obs} \mid \theta)] \times [p(\psi)p(\mathcal{I} \mid \mathcal{Y}^{obs}, \psi)].$$

In reality, we are more interested in model parameters θ that govern the data-generating process of $\mathcal{Y}$ than the ψ-parameters that almost have no intrinsic scientific value. Under the two assumption specified above, we can ignore ψ and focus on the first component $p(\theta)p(\mathcal{Y}^{obs} \mid \theta)$. On the contrary, if we consider MNAR, we have to explicitly model the joint distribution of $\mathcal{Y}$ and $\mathcal{I}$, which is challenging and sensitive to the model assumption.

Appendix B Proof of Proposition 1

This Appendix presents a detailed version of Proposition 1 in the main text and further derives the posterior distributions of model parameters in this proposition.

Proposition A1. *Under the prior distributions in equations (8), (9), and (10) and the*

data likelihood in equation (7), we can derive the posterior distributions of model parameters $\theta := (s_{tir}, c_{jr}, l_{kr}, \tau_{sr}, \tau_{\epsilon,j})$*:*

$$\boldsymbol{c}_{j:} \mid - \sim \mathcal{N}\left(\left(\boldsymbol{I}_R + \tau_{\epsilon,j}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{X}}_c\right)^{-1}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{y}}_{cj}\tau_{\epsilon,j}, \quad \left(\boldsymbol{I}_R + \tau_{\epsilon,j}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{X}}_c\right)^{-1}\right), \tag{A2}$$

$$\boldsymbol{l}_{k:} \mid - \sim \mathcal{N}\left(\left(\boldsymbol{I}_R + \tilde{\boldsymbol{X}}_l^\top \tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{X}}_l\right)^{-1}\tilde{\boldsymbol{X}}_l^\top \tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{y}}_{lk}, \quad \left(\boldsymbol{I}_R + \tilde{\boldsymbol{X}}_l^\top \tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{X}}_l\right)^{-1}\right), \tag{A3}$$

$$\boldsymbol{s}_{\tilde{i}:} \mid - \sim \mathcal{N}\left(\left(\boldsymbol{\Sigma}_s^{-1} + \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1}\tilde{\boldsymbol{X}}_s\right)^{-1}\tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1}\tilde{\boldsymbol{y}}_{s\tilde{i}}, \quad \left(\boldsymbol{\Sigma}_s^{-1} + \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1}\tilde{\boldsymbol{X}}_s\right)^{-1}\right), \tag{A4}$$

$$\tau_{sr} \mid - \sim \Gamma\left(a_0 + \frac{\sum_{t=1}^T N_t}{2}, b_0 + \frac{1}{2}\sum_{t=1}^T\sum_{i=1}^{N_t} s_{tir}^2\right), \tag{A5}$$

$$\tau_{\epsilon,j} \mid - \sim \Gamma\left(a_0 + \frac{N_{obs,j}}{2}, b_0 + \frac{1}{2}\sum_{t=1}^T\sum_{i=1}^{N_t}\sum_{k=1}^K I_{tijk}(y_{tijk} - x_{tijk})^2\right), \tag{A6}$$

where $p(\cdot \mid -)$ *denotes the posterior distribution of the parameter conditional on all other parameters and the observed data* $\mathcal{Y}^{obs}$*, and*

$$\underbrace{\tilde{\boldsymbol{I}}_{cj}}_{NK\times 1} = vec(\mathcal{I}_{:j:}), \ \underbrace{\tilde{\boldsymbol{I}}_{cj\times R}}_{NK\times R} = \left[vec(\mathcal{I}_{:j:}) \dots vec(\mathcal{I}_{:j:})\right],$$

$$\underbrace{\tilde{\boldsymbol{I}}_{lk}}_{NJ\times 1} = vec(\mathcal{I}_{::k}), \ \underbrace{\tilde{\boldsymbol{I}}_{lk\times R}}_{NJ\times R} = \left[vec(\mathcal{I}_{::k}) \dots vec(\mathcal{I}_{::k})\right],$$

$$\underbrace{\tilde{\boldsymbol{I}}_{s\tilde{i}}}_{JK\times 1} = vec(\mathcal{I}_{\tilde{i}::}), \ \underbrace{\tilde{\boldsymbol{I}}_{s\tilde{i}\times R}}_{JK\times R} = \left[vec(\mathcal{I}_{\tilde{i}::}) \dots vec(\mathcal{I}_{\tilde{i}::})\right],$$

$$\underbrace{\tilde{\boldsymbol{X}}_c}_{NK\times R} = \left[\boldsymbol{S}\odot\boldsymbol{L}\right] * \tilde{\boldsymbol{I}}_{cj\times R}, \ \underbrace{\tilde{\boldsymbol{y}}_{cj}}_{NK\times 1} = vec(\mathcal{Y}_{:j:}) * \tilde{\boldsymbol{I}}_{cj}, \ \underbrace{\tilde{\boldsymbol{\epsilon}}_{cj}}_{NK\times 1} = vec(\mathcal{E}_{:j:}) * \tilde{\boldsymbol{I}}_{cj},$$

$$\underbrace{\tilde{\boldsymbol{X}}_l}_{NJ\times R} = \left[\boldsymbol{S}\odot\boldsymbol{C}\right] * \tilde{\boldsymbol{I}}_{lk\times R}, \ \underbrace{\tilde{\boldsymbol{y}}_{lk}}_{NJ\times 1} = vec(\mathcal{Y}_{::k}) * \tilde{\boldsymbol{I}}_{lk}, \ \underbrace{\tilde{\boldsymbol{\epsilon}}_{lk}}_{NJ\times 1} = vec(\mathcal{E}_{::k}) * \tilde{\boldsymbol{I}}_{lk},$$

$$\underbrace{\tilde{\boldsymbol{X}}_s}_{JK\times R} = \left[\boldsymbol{C}\odot\boldsymbol{L}\right] * \tilde{\boldsymbol{I}}_{s\tilde{i}\times R}, \ \underbrace{\tilde{\boldsymbol{y}}_{s\tilde{i}}}_{JK\times 1} = vec(\mathcal{Y}_{\tilde{i}::}) * \tilde{\boldsymbol{I}}_{s\tilde{i}}, \ \underbrace{\tilde{\boldsymbol{\epsilon}}_{s\tilde{i}}}_{JK\times 1} = vec(\mathcal{E}_{\tilde{i}::}) * \tilde{\boldsymbol{I}}_{s\tilde{i}}.$$

We first introduce the notation used in this Appendix. We use boldface lowercase letters, e.g., $\boldsymbol{a}$, to denote vectors. Matrices are represented by boldface capital letters, e.g., $\boldsymbol{A}$.

The higher-order tensors (with orders larger than two) are denoted by boldface Euler script letters, e.g., $\mathcal{A}$. We often slice matrices or tensors, fixing one or two dimensions. For matrices, we extract the rows or columns. In particular, the i-th row and j-th column of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{a}_{i:}$ and $\boldsymbol{a}_{:j}$, respectively. Alternatively, we often denote the columns more compactly, e.g., $\boldsymbol{a}_j$. The slicing of three-dimensional (3-D) tensors is similar, and the resultant arrays are *fibers*. Specifically, the horizontal, lateral, and frontal slides of a 3-D tensor $\mathcal{A}$ are denoted by $\mathcal{A}_{i::}$, $\mathcal{A}_{:j:}$, and $\mathcal{A}_{::k}$.

Several matrix operations are important in this section, so we briefly define them here. First, we vectorize an $I \times J$ matrix $\boldsymbol{A}$ by rows as follow:

$$\underbrace{\text{vec}(\boldsymbol{A})}_{IJ\times 1} = \big(a_{11}, a_{12}, \ldots, a_{1J} \ldots\ldots a_{I1}, a_{I2}, \ldots, a_{IJ}\big)^\top.$$

Second, the *Kronecker product* of matrices $\boldsymbol{A} \in \mathbb{R}^{I\times J}$ and $\boldsymbol{B} \in \mathbb{R}^{K\times L}$ is

$$\underbrace{\boldsymbol{A} \otimes \boldsymbol{B}}_{IK\times JL} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \ldots & a_{1J}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \ldots & a_{2J}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{I1}\boldsymbol{B} & a_{I2}\boldsymbol{B} & \ldots & a_{IJ}\boldsymbol{B} \end{pmatrix}.$$

Third, the *Khatri-Rao product* of two matrices $\boldsymbol{A} \in \mathbb{R}^{I\times K}$ and $\boldsymbol{B} \in \mathbb{R}^{J\times K}$ is

$$\underbrace{\boldsymbol{A} \odot \boldsymbol{B}}_{IJ\times K} = \begin{pmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 & \ldots & \boldsymbol{a}_K \otimes \boldsymbol{b}_K \end{pmatrix}.$$

The final useful operation is *Hadamard product*, which is the elementwise matrix product. Given two matrices of the same size $I \times J$, the Hadamard product is

$$\underbrace{\boldsymbol{A} * \boldsymbol{B}}_{I\times J} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \ldots & a_{1J}b_{1J} \\ a_{21}b_{21} & a_{22}b_{22} & \ldots & a_{2J}b_{2J} \\ \vdots & \vdots & & \vdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \ldots & a_{IJ}b_{IJ} \end{pmatrix}.$$

For derivations, we stack the latent factors $\boldsymbol{S}$, $\boldsymbol{C}$, and $\boldsymbol{L}$ as follow:

$$\underbrace{\boldsymbol{S}}_{N\times R} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_R) = \begin{pmatrix} \boldsymbol{s}_{1:}^\top \\ \boldsymbol{s}_{2:}^\top \\ \vdots \\ \boldsymbol{s}_{N:}^\top \end{pmatrix}, \underbrace{\boldsymbol{C}}_{J\times R} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_R) = \begin{pmatrix} \boldsymbol{c}_{1:}^\top \\ \boldsymbol{c}_{2:}^\top \\ \vdots \\ \boldsymbol{c}_{J:}^\top \end{pmatrix}, \underbrace{\boldsymbol{L}}_{K\times R} = (\boldsymbol{l}_1, \ldots, \boldsymbol{l}_R) = \begin{pmatrix} \boldsymbol{l}_{1:}^\top \\ \boldsymbol{l}_{2:}^\top \\ \vdots \\ \boldsymbol{l}_{K:}^\top \end{pmatrix}.$$

B.1 Update c_{jr}

We update the parameters in $\boldsymbol{C}$ row-by-row; that is, we update $\boldsymbol{c}_{j:}$ for each characteristic j. First, we rewrite the tensor decomposition:

$$x_{tijk} = \sum_{r=1}^{R} s_{tir} c_{jr} l_{kr} = (s_{ti1} l_{k1}, s_{ti2} l_{k2}, \ldots, s_{tiR} l_{kR}) \boldsymbol{c}_{j:}.$$

We next stack the observations in matrices:

$$\begin{aligned} &\underbrace{\tilde{\boldsymbol{I}}_{cj}}_{NK\times 1} = \text{vec}(\mathcal{I}_{:j:}), \ \underbrace{\tilde{\boldsymbol{I}}_{cj\times R}}_{NK\times R} = \big[\text{vec}(\mathcal{I}_{:j:}) \ldots \text{vec}(\mathcal{I}_{:j:})\big], \\ &\underbrace{\tilde{\boldsymbol{X}}_c}_{NK\times R} = \big[\boldsymbol{S} \odot \boldsymbol{L}\big] * \tilde{\boldsymbol{I}}_{cj\times R}, \ \underbrace{\tilde{\boldsymbol{y}}_{cj}}_{NK\times 1} = \text{vec}(\mathcal{Y}_{:j:}) * \tilde{\boldsymbol{I}}_{cj}, \ \underbrace{\tilde{\boldsymbol{\epsilon}}_{cj}}_{NK\times 1} = \text{vec}(\mathcal{E}_{:j:}) * \tilde{\boldsymbol{I}}_{cj}. \end{aligned} \tag{A7}$$

Equation (A7) sheds lights on how the Bayesian model is valid in incomplete datasets. The indicator matrices, $\tilde{\boldsymbol{I}}_{cj}$ and $\tilde{\boldsymbol{I}}_{cj\times R}$, select only the rows of $\tilde{\boldsymbol{X}}_c$ and $\tilde{\boldsymbol{y}}_{cj}$ that have the observed information. Under this data structure, we derive the distribution of the observed data:

$$\tilde{\boldsymbol{y}}_{cj} = \tilde{\boldsymbol{X}}_c \boldsymbol{c}_{j:} + \tilde{\boldsymbol{\epsilon}}_{cj} \implies \tilde{\boldsymbol{y}}_{cj} \mid \tilde{\boldsymbol{X}}_c, \tau_{\epsilon,j} \sim \mathcal{N}(\tilde{\boldsymbol{X}}_c \boldsymbol{c}_{j:}, \frac{1}{\tau_{\epsilon,j}} \boldsymbol{I}_{NK}).^{13} \tag{A8}$$

Hence, the posterior distribution of $\boldsymbol{c}_{j:}$ is

$$\begin{aligned} p(\boldsymbol{c}_{j:} \mid -) &\propto \exp\{-\frac{\boldsymbol{c}_{j:}^\top \boldsymbol{c}_{j:}}{2}\} \exp\{-\frac{\tau_{\epsilon,j}}{2} (\tilde{\boldsymbol{y}}_{cj} - \tilde{\boldsymbol{X}}_c \boldsymbol{c}_{j:})^\top (\tilde{\boldsymbol{y}}_{cj} - \tilde{\boldsymbol{X}}_c \boldsymbol{c}_{j:})\} \\ &\propto \exp\{-\frac{1}{2}\big[\boldsymbol{c}_{j:}^\top \big(\boldsymbol{I}_R + \tau_{\epsilon,j} \tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{X}}_c\big) \boldsymbol{c}_{j:} - 2\boldsymbol{c}_{j:}^\top \tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{y}}_{cj} \tau_{\epsilon,j}\big]\} \end{aligned}$$

[13] A more rigorous statement should include only the rows of $\tilde{\boldsymbol{y}}_{cj}$ that are not missing. However, we assign zeros to those rows, so they do not enter the likelihood function of the observed data.

$$\implies \boldsymbol{c}_{j:} \mid - \sim \mathcal{N}\left(\left(\boldsymbol{I}_R + \tau_{\epsilon,j}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{X}}_c\right)^{-1}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{y}}_{cj}\tau_{\epsilon,j},\quad \left(\boldsymbol{I}_R + \tau_{\epsilon,j}\tilde{\boldsymbol{X}}_c^\top \tilde{\boldsymbol{X}}_c\right)^{-1}\right).$$

Next, we apply the similar tricks to derive the posterior distributions of l_{kr} and s_{tir}.

B.2 Update l_{kr}

We update the parameters in $\boldsymbol{L}$ row-by-row; that is, we update $\boldsymbol{l}_{k:}$ for each lag k. First, we rewrite the tensor decomposition:

$$x_{tijk} = \sum_{r=1}^{R} s_{tir}c_{jr}l_{kr} = (s_{ti1}c_{j1}, s_{ti2}c_{j2}, \ldots, s_{tiR}c_{jR})\boldsymbol{l}_{k:}.$$

We next stack the observations in matrices:

$$\begin{aligned}
&\underbrace{\tilde{\boldsymbol{I}}_{lk}}_{NJ\times 1} = \mathrm{vec}(\mathcal{I}_{::k}),\ \underbrace{\tilde{\boldsymbol{I}}_{lk\times R}}_{NJ\times R} = \big[\mathrm{vec}(\mathcal{I}_{::k})\ldots\mathrm{vec}(\mathcal{I}_{::k})\big],\\
&\underbrace{\tilde{\boldsymbol{X}}_l}_{NJ\times R} = \big[\boldsymbol{S}\odot\boldsymbol{C}\big] * \tilde{\boldsymbol{I}}_{lk\times R},\ \underbrace{\tilde{\boldsymbol{y}}_{lk}}_{NJ\times 1} = \mathrm{vec}(\mathcal{Y}_{::k}) * \tilde{\boldsymbol{I}}_{lk},\ \underbrace{\tilde{\boldsymbol{\epsilon}}_{lk}}_{NJ\times 1} = \mathrm{vec}(\mathcal{E}_{::k}) * \tilde{\boldsymbol{I}}_{lk}.
\end{aligned} \tag{A9}$$

$$\tilde{\boldsymbol{y}}_{lk} = \tilde{\boldsymbol{X}}_l\boldsymbol{l}_{k:} + \tilde{\boldsymbol{\epsilon}}_{lk} \implies \tilde{\boldsymbol{y}}_{lk} \mid \tilde{\boldsymbol{X}}_l, \{\tau_{\epsilon,j}\}_{j=1}^J \sim \mathcal{N}(\tilde{\boldsymbol{X}}_l\boldsymbol{l}_{k:}, \tilde{\boldsymbol{\Sigma}}_{lk}), \tag{A10}$$

where $\tilde{\boldsymbol{\Sigma}}_{lk} = \mathrm{diag}\{\frac{1}{\tau_{\epsilon,1}},\ldots,\frac{1}{\tau_{\epsilon,J}}\ldots\ldots\frac{1}{\tau_{\epsilon,1}},\ldots,\frac{1}{\tau_{\epsilon,J}}\}$. Hence, the posterior distribution of $\boldsymbol{l}_{k:}$ is

$$\begin{aligned}
p(\boldsymbol{l}_{k:} \mid -) &\propto \exp\{-\frac{\boldsymbol{l}_{k:}^\top\boldsymbol{l}_{k:}}{2}\}\exp\{-\frac{1}{2}(\tilde{\boldsymbol{y}}_{lk} - \tilde{\boldsymbol{X}}_l\boldsymbol{l}_{k:})^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}(\tilde{\boldsymbol{y}}_{lk} - \tilde{\boldsymbol{X}}_l\boldsymbol{l}_{k:})\}\\
&\propto \exp\{-\frac{1}{2}\big[\boldsymbol{l}_{k:}^\top\big(\boldsymbol{I}_R + \tilde{\boldsymbol{X}}_l^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{X}}_l\big)\boldsymbol{l}_{k:} - 2\boldsymbol{l}_{k:}^\top\tilde{\boldsymbol{X}}_l^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{y}}_{lk}\big]\}
\end{aligned}$$

$$\implies \boldsymbol{l}_{k:} \mid - \sim \mathcal{N}\left(\left(\boldsymbol{I}_R + \tilde{\boldsymbol{X}}_l^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{X}}_l\right)^{-1}\tilde{\boldsymbol{X}}_l^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{y}}_{lk},\quad \left(\boldsymbol{I}_R + \tilde{\boldsymbol{X}}_l^\top\tilde{\boldsymbol{\Sigma}}_{lk}^{-1}\tilde{\boldsymbol{X}}_l\right)^{-1}\right).$$

B.3 Update s_{tir}

We update the parameters in $\boldsymbol{S}$ row-by-row; that is, we update $\boldsymbol{s}_{i:}$ for each firm-month observation i. First, we rewrite the tensor decomposition:

$$x_{tijk} = \sum_{r=1}^{R} s_{tir} c_{jr} l_{kr} = (c_{j1}l_{k1}, c_{j2}l_{k2}, \ldots, c_{jR}l_{kR})\boldsymbol{s}_{\tilde{i},:}, \quad \text{where } \tilde{i} = i + \sum_{\tau=1}^{t-1} N_\tau.$$

We next stack the observations in matrices:

$$\begin{aligned}
&\underbrace{\tilde{\boldsymbol{I}}_{s\tilde{i}}}_{JK\times 1} = \mathrm{vec}(\mathcal{I}_{\tilde{i}::}),\ \underbrace{\tilde{\boldsymbol{I}}_{s\tilde{i}\times R}}_{JK\times R} = \big[\mathrm{vec}(\mathcal{I}_{\tilde{i}::}) \ldots \mathrm{vec}(\mathcal{I}_{\tilde{i}::})\big], \\
&\underbrace{\tilde{\boldsymbol{X}}_s}_{JK\times R} = \big[\boldsymbol{C} \odot \boldsymbol{L}\big] * \tilde{\boldsymbol{I}}_{s\tilde{i}\times R},\ \underbrace{\tilde{\boldsymbol{y}}_{s\tilde{i}}}_{JK\times 1} = \mathrm{vec}(\mathcal{Y}_{\tilde{i}::}) * \tilde{\boldsymbol{I}}_{s\tilde{i}},\ \underbrace{\tilde{\boldsymbol{\epsilon}}_{s\tilde{i}}}_{JK\times 1} = \mathrm{vec}(\mathcal{E}_{\tilde{i}::}) * \tilde{\boldsymbol{I}}_{s\tilde{i}}.
\end{aligned} \tag{A11}$$

$$\tilde{\boldsymbol{y}}_{s\tilde{i}} = \tilde{\boldsymbol{X}}_s \boldsymbol{s}_{\tilde{i}:} + \tilde{\boldsymbol{\epsilon}}_{s\tilde{i}} \implies \tilde{\boldsymbol{y}}_{s\tilde{i}} \mid \tilde{\boldsymbol{X}}_s, \{\tau_{\epsilon,j}\}_{j=1}^{J} \sim \mathcal{N}(\tilde{\boldsymbol{X}}_s \boldsymbol{s}_{\tilde{i}:}, \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}), \tag{A12}$$

where $\tilde{\boldsymbol{\Sigma}}_{s\tilde{i}} = \mathrm{diag}\{\frac{1}{\tau_{\epsilon,1}}, \ldots, \frac{1}{\tau_{\epsilon,1}} \ldots\ldots \frac{1}{\tau_{\epsilon,J}}, \ldots, \frac{1}{\tau_{\epsilon,J}}\}$. Hence, the posterior distribution of $\boldsymbol{s}_{\tilde{i}:}$ is

$$\begin{aligned}
p(\boldsymbol{s}_{\tilde{i}:} \mid -) &\propto \exp\{-\frac{\boldsymbol{s}_{\tilde{i}:}^\top \boldsymbol{\Sigma}_s^{-1} \boldsymbol{s}_{\tilde{i}:}}{2}\} \exp\{-\frac{1}{2}(\tilde{\boldsymbol{y}}_{s\tilde{i}} - \tilde{\boldsymbol{X}}_s \boldsymbol{s}_{\tilde{i}:})^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} (\tilde{\boldsymbol{y}}_{s\tilde{i}} - \tilde{\boldsymbol{X}}_s \boldsymbol{s}_{\tilde{i}:})\} \\
&\propto \exp\{-\frac{1}{2}\big[\boldsymbol{s}_{\tilde{i}:}^\top \big(\boldsymbol{\Sigma}_s^{-1} + \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} \tilde{\boldsymbol{X}}_s\big) \boldsymbol{s}_{\tilde{i}:} - 2\boldsymbol{s}_{\tilde{i}:}^\top \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} \tilde{\boldsymbol{y}}_{s\tilde{i}}\big]\}
\end{aligned}$$

$$\implies \boldsymbol{s}_{\tilde{i}:} \mid - \sim \mathcal{N}\left(\big(\boldsymbol{\Sigma}_s^{-1} + \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} \tilde{\boldsymbol{X}}_s\big)^{-1} \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} \tilde{\boldsymbol{y}}_{s\tilde{i}},\ \ \big(\boldsymbol{\Sigma}_s^{-1} + \tilde{\boldsymbol{X}}_s^\top \tilde{\boldsymbol{\Sigma}}_{s\tilde{i}}^{-1} \tilde{\boldsymbol{X}}_s\big)^{-1}\right),$$

where $\boldsymbol{\Sigma}_s$ is the prior covariance matrix of $\boldsymbol{s}_{\tilde{i}:}$, i.e., $\boldsymbol{\Sigma}_s = \mathrm{diag}\big\{\tau_{s1}^{-1}, \tau_{s2}^{-1}, \ldots, \tau_{sR}^{-1}\big\}$.

B.4 Update τ_{sr}

$$\begin{aligned}
p(\tau_{sr} \mid -) &\propto \tau_{sr}^{a_0 - 1} \exp\{-b_0 \tau_{sr}\} \tau_{sr}^{\frac{\sum_{t=1}^{T} N_t}{2}} \exp\{-\frac{\tau_{sr}}{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} s_{tir}^2\} \\
&\propto \tau_{sr}^{\frac{\sum_{t=1}^{T} N_t}{2} + a_0 - 1} \exp\{-\tau_{sr}\big[b_0 + \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} s_{tir}^2\big]\},
\end{aligned}$$

$$\tau_{sr} \mid - \sim \Gamma(a_0 + \frac{\sum_{t=1}^{T} N_t}{2}, b_0 + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t} s_{tir}^2).$$

B.5 Update $\tau_{\epsilon,j}$

$$\begin{aligned} p(\tau_{\epsilon,j} \mid -) &\propto \tau_{\epsilon,j}^{a_0-1} \exp\{-b_0\tau_{\epsilon,j}\}\tau_{\epsilon,j}^{\frac{N_{obs,j}}{2}} \exp\{-\frac{\tau_{\epsilon,j}}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{k=1}^{K} I_{tijk}(y_{tijk} - x_{tijk})^2\} \\ &\propto \tau_{\epsilon,j}^{\frac{N_{obs,j}}{2}+a_0-1} \exp\{-\big[b_0 + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{k=1}^{K} I_{tijk}(y_{tijk} - x_{tijk})^2\big]\tau_{\epsilon,j}\}, \end{aligned}$$

$$\tau_{\epsilon,j} \mid - \sim \Gamma(a_0 + \frac{N_{obs,j}}{2}, b_0 + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{k=1}^{K} I_{tijk}(y_{tijk} - x_{tijk})^2).$$

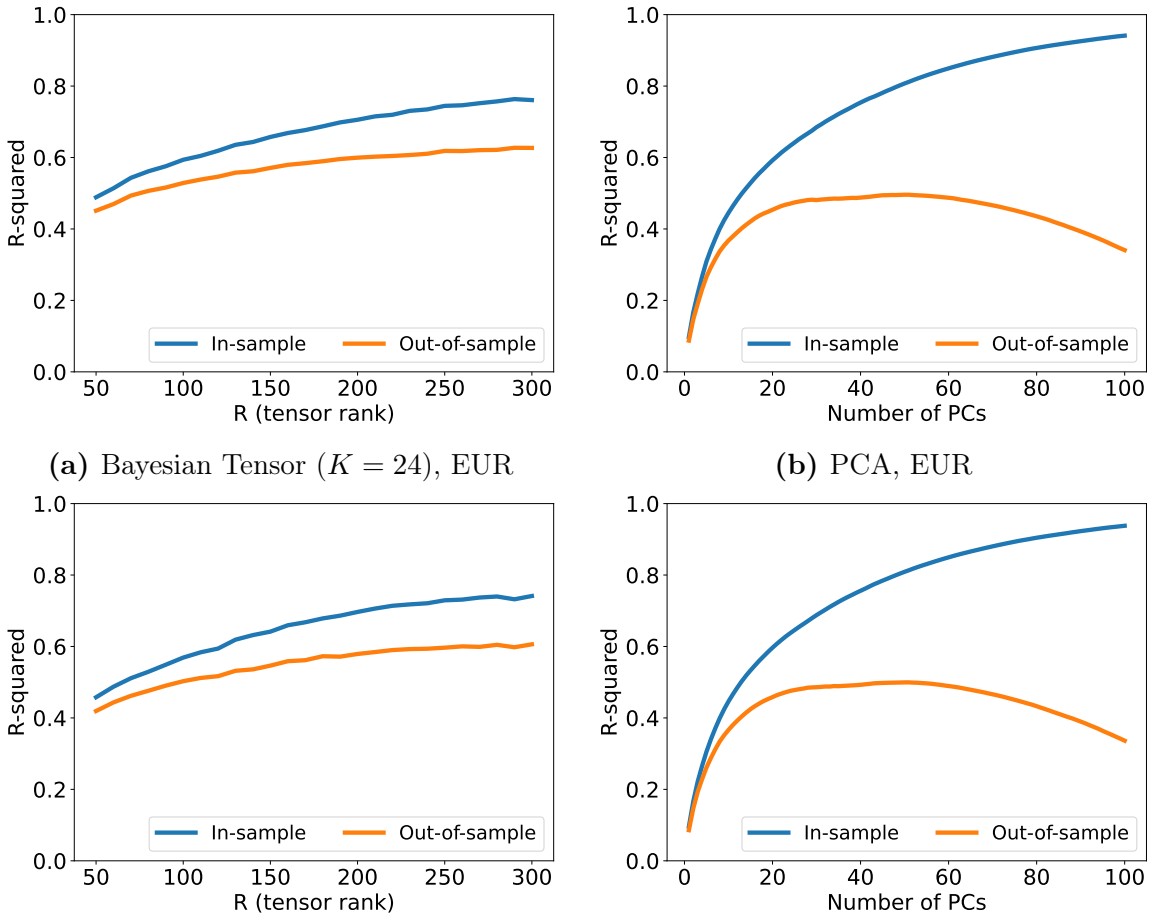
Appendix C Estimation of Γ_α in the IPCA Model

Conditional on $\boldsymbol{f}_{t+1}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\beta}}$, we estimate $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$ via the least squares method. Since the matrix $\big[\boldsymbol{I} - \boldsymbol{\beta}_t\big(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\big)^{-1}\boldsymbol{\beta}_t^\top\big]\boldsymbol{Z}_t$ is rank deficient, with a column rank of $(J-P)$, we can identify only $(J-P)$ parameters in $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}$. We assume that $\boldsymbol{\Gamma}_{\boldsymbol{\alpha}} = \boldsymbol{U}\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}}$, where $\boldsymbol{U}$ is a $J \times (J-P)$ matrix with a full column rank, and estimate the $(J-P)\times 1$ vector $\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}}$:

$$\widehat{\tilde{\boldsymbol{\Gamma}}}_\alpha = \Big(\sum_{t=1}^{T}\tilde{\boldsymbol{Z}}_t^\top\tilde{\boldsymbol{Z}}_t\Big)^{-1}\cdot\Big(\sum_{t=1}^{T}\tilde{\boldsymbol{Z}}_t^\top(\boldsymbol{R}_{t+1} - \boldsymbol{\beta}_t\boldsymbol{f}_{t+1})\Big) = \Big(\sum_{t=1}^{T}\tilde{\boldsymbol{Z}}_t^\top\tilde{\boldsymbol{Z}}_t\Big)^{-1}\cdot\Big(\sum_{t=1}^{T}\tilde{\boldsymbol{Z}}_t^\top\boldsymbol{R}_{t+1}\Big),$$

where $\tilde{\boldsymbol{Z}}_t = \big[\boldsymbol{I} - \boldsymbol{\beta}_t\big(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\big)^{-1}\boldsymbol{\beta}_t^\top\big]\boldsymbol{Z}_t\boldsymbol{U}$, and $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}} = \boldsymbol{U}\cdot\widehat{\tilde{\boldsymbol{\Gamma}}}_\alpha$. Theoretically, we can choose $\boldsymbol{U}$ as any matrix with a full column rank. In our application, $\boldsymbol{U}$ collects the $(J-P)$ eigenvectors of $\sum_{t=1}^{T}\boldsymbol{Z}_t^\top\big[\boldsymbol{I} - \boldsymbol{\beta}_t\big(\boldsymbol{\beta}_t^\top\boldsymbol{\beta}_t\big)^{-1}\boldsymbol{\beta}_t^\top\big]\boldsymbol{Z}_t$ that correspond to the non-zero eigenvalues. Note that $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\alpha}}$ $(= \boldsymbol{U}\cdot\widehat{\tilde{\boldsymbol{\Gamma}}}_\alpha)$ is invariant to a nonsingular rotation of $\boldsymbol{U}$.

Appendix D Additional Figures and Tables



(a) Bayesian Tensor ($K = 24$), EUR

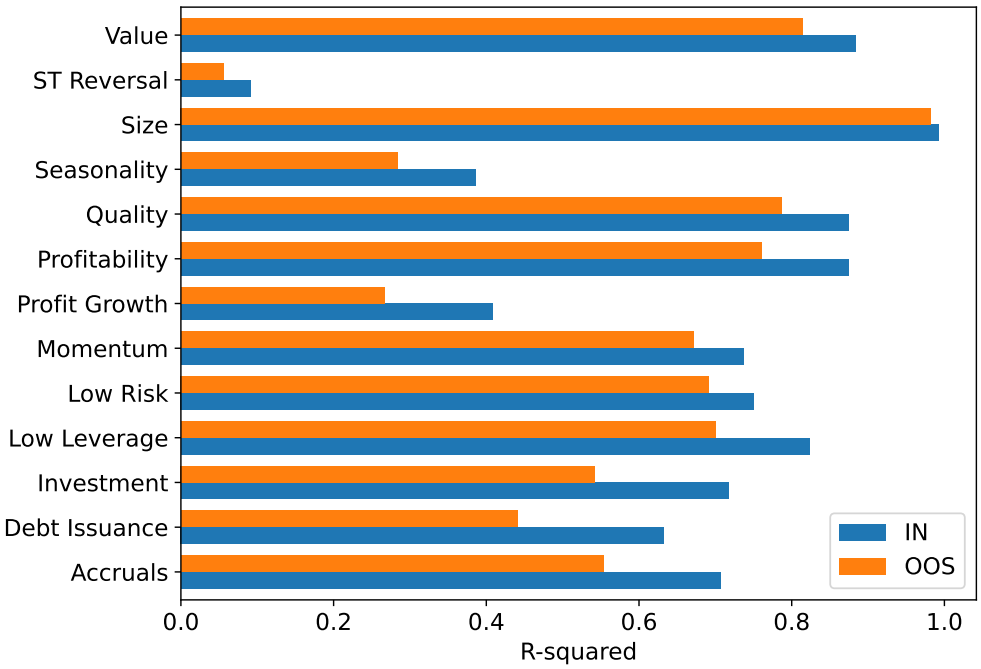
(b) PCA, EUR

(c) Bayesian Tensor ($K = 24$), ASIA

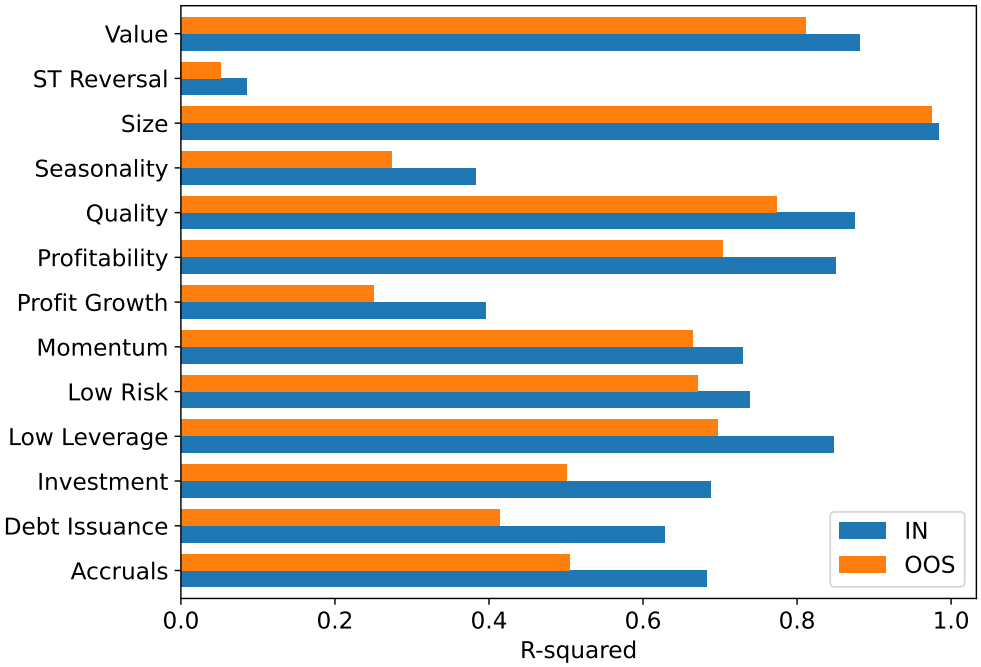
(d) PCA, ASIA

Figure A1: Prediction accuracy of missing values: tensor decomposition vs PCA, European and Asian datasets

This figure plots the in- and out-of-sample prediction accuracy in the European and Asian datasets characteristic dataset. The data sample spans from July 1986 to December 2021. In the simulations, we first remove the repeated observations and then randomly simulate a sample of 20% missing entries (out-of-sample). We estimate model parameters using only the observed information, that is, the sample of the remaining 80% observations (in-sample). The R-squared is computed using equations (13)–(14). Panels (a) and (c) consider the tensor model, with $K = 24$ lags and tensor rank R between 50 and 300. Panels (b) and (d) show the results based on the PCA approach, with the number of PCs between 1 and 100.



(a) European data



(b) Asian data

Figure A2: Prediction accuracy by 13 Themes: European and Asian data, $R = 200$

This figure plots the in- and out-of-sample prediction accuracy in the European and Asian datasets characteristic dataset. The 151 firm characteristics are grouped into 13 cluster themes. The data sample spans from July 1986 to December 2021. In the simulations, we first remove the repeated observations and then randomly simulate a sample of 20% missing entries (out-of-sample). We estimate model parameters using only the observed information, that is, the sample of the remaining 80% observations (in-sample). The R-squared is computed using equations (13)–(14). We consider a tensor model with $K = 24$ lags and tensor rank $R = 200$.

Table A1: Firm characteristics used in asset pricing applications

Label	US	EUR	ASIA	Label	US	EUR	ASIA	Label	US	EUR	ASIA
cowc_gr1a	Y	Y	Y	earnings_variability	Y	Y	Y	niq_at	Y	N	N
oaccruals_at	Y	Y	Y	ivol_capm_21d	Y	Y	Y	op_at	Y	Y	Y
oaccruals_ni	Y	Y	Y	ivol_capm_252d	Y	Y	Y	op_atl1	Y	Y	Y
seas_16_20na	Y	N	N	ivol_ff3_21d	Y	N	Y	opex_at	Y	Y	Y
taccruals_at	Y	Y	Y	ivol_hxz4_21d	Y	N	N	qmj	Y	Y	Y
taccruals_ni	Y	Y	Y	ocfq_saleq_std	Y	N	N	qmj_growth	Y	Y	Y
capex_abn	Y	N	N	rmax1_21d	Y	Y	Y	qmj_prof	Y	Y	Y
debt_gr3	Y	Y	Y	rmax5_21d	Y	Y	Y	qmj_safety	Y	Y	Y
fnl_gr1a	Y	Y	Y	rvol_21d	Y	Y	Y	sale_bev	Y	Y	Y
ncol_gr1a	Y	Y	Y	seas_6_10na	Y	N	N	corr_1260d	Y	N	N
nfna_gr1a	Y	Y	Y	zero_trades_21d	Y	N	N	coskew_21d	Y	Y	Y
ni_ar1	Y	Y	Y	zero_trades_126d	Y	N	N	dbnetis_at	Y	Y	Y
noa_at	Y	Y	Y	zero_trades_252d	Y	N	N	kz_index	Y	Y	Y
aliq_at	Y	Y	Y	prc_highprc_252d	Y	Y	Y	lti_gr1a	Y	Y	Y
at_gr1	Y	Y	Y	resff3_6_1	Y	N	N	pi_nix	Y	Y	Y
be_gr1a	Y	Y	Y	resff3_12_1	Y	N	N	seas_2_5an	Y	N	N
capx_gr1	Y	Y	N	ret_3_1	Y	Y	Y	seas_6_10an	Y	N	N
capx_gr2	Y	Y	N	ret_6_1	Y	Y	Y	seas_11_15an	Y	N	N
capx_gr3	Y	Y	N	ret_9_1	Y	Y	Y	seas_11_15na	Y	N	N
coa_gr1a	Y	Y	Y	ret_12_1	Y	Y	Y	seas_16_20an	Y	N	N
col_gr1a	Y	Y	Y	seas_1_1na	Y	Y	Y	sti_gr1a	Y	Y	Y
emp_gr1	Y	Y	N	dsale_dinv	Y	Y	Y	ami_126d	Y	N	N
inv_gr1	Y	Y	Y	dsale_drec	Y	Y	Y	dolvol_126d	Y	N	N
inv_gr1a	Y	Y	Y	dsale_dsga	Y	Y	Y	market_equity	Y	Y	Y
lnoa_gr1a	Y	Y	N	niq_at_chg1	Y	N	N	prc	Y	Y	Y
mispricing_mgmt	Y	Y	Y	niq_be_chg1	Y	N	N	rd_me	Y	Y	Y
ncoa_gr1a	Y	Y	Y	niq_su	Y	Y	N	iskew_capm_21d	Y	Y	Y
nncoa_gr1a	Y	Y	Y	ocf_at_chg1	Y	Y	Y	iskew_ff3_21d	Y	N	Y
noa_gr1a	Y	Y	Y	ret_12_7	Y	Y	Y	iskew_hxz4_21d	Y	N	N
ppeinv_gr1a	Y	Y	Y	sale_emp_gr1	Y	Y	N	ret_1_0	Y	Y	Y
ret_60_12	Y	N	N	saleq_su	Y	Y	N	rmax5_rvol_21d	Y	Y	Y
sale_gr1	Y	Y	Y	seas_1_1an	Y	Y	Y	rskew_21d	Y	Y	Y
sale_gr3	Y	Y	Y	tax_gr1a	Y	Y	Y	at_me	Y	Y	Y
saleq_gr1	Y	Y	N	dolvol_var_126d	Y	N	N	be_me	Y	Y	Y
seas_2_5na	Y	N	N	ebit_bev	Y	Y	Y	bev_mev	Y	Y	Y
age	Y	Y	Y	ebit_sale	Y	Y	Y	chcsho_12m	Y	Y	Y
aliq_mat	Y	Y	Y	ni_be	Y	Y	Y	debt_me	Y	Y	Y
at_be	Y	Y	Y	niq_be	Y	N	N	div12m_me	Y	Y	Y
bidaskhl_21d	Y	N	N	o_score	Y	Y	Y	ebitda_mev	Y	Y	Y
cash_at	Y	Y	Y	ocf_at	Y	Y	Y	eq_dur	Y	Y	Y
netdebt_me	Y	Y	Y	ope_be	Y	Y	Y	eqnetis_at	Y	Y	Y
ni_ivol	Y	Y	Y	ope_bel1	Y	Y	Y	eqnpo_12m	Y	Y	Y
rd_sale	Y	Y	Y	turnover_var_126d	Y	N	N	eqnpo_me	Y	Y	Y
rd5_at	Y	Y	N	at_turnover	Y	Y	Y	eqpo_me	Y	Y	Y
tangibility	Y	Y	Y	cop_at	Y	Y	Y	fcf_me	Y	Y	Y
z_score	Y	Y	Y	cop_atl1	Y	Y	Y	ival_me	Y	Y	Y
beta_60m	Y	N	N	dgp_dsale	Y	Y	Y	netis_at	Y	Y	Y
beta_dimson_21d	Y	Y	Y	gp_at	Y	Y	Y	ni_me	Y	Y	Y
betabab_1260d	Y	N	N	gp_atl1	Y	Y	Y	ocf_me	Y	Y	Y
betadown_252d	Y	Y	Y	mispricing_perf	Y	Y	Y	sale_me	Y	Y	Y

This table presents the firm characteristics that are used in asset pricing applications in the US, European, and Asian datasets. We remove "turnover_126d" since it is almost perfectly correlated with the other variable "zero_trades_126d". To be included in the final sample, a firm characteristic should have at least ten observations per month. The definitions of these 150 firm characteristics can be found in Jensen, Kelly, and Pedersen (2023). Y: included; N: excluded.

Table A2: Annualised Sharpe ratios of characteristic-managed portfolios, US data

Label	Name	cs-mean	Tensor models: only missing	Tensor models: missing&stale	Δ
saleq_su	Standardized Revenue surprise	0.073	0.043	-0.057	-0.130*
cop_atl1	Cash-operating-profits-to-lagged-assets	1.016**	0.867**	0.889**	-0.128**
niq_su	Standardized earnings surprise	0.503**	0.506**	0.376**	-0.127
oaccruals_at	Operating accruals	-0.988**	-0.934**	-0.864**	0.124**
seas_2_5an	Years 2-5 lagged returns, annual	0.588**	0.469**	0.469**	-0.119**
qmj	Quality minus Junk: Composite	0.449**	0.568**	0.568**	0.119**
capex_abn	Abnormal corporate investment	-0.730**	-0.807**	-0.842**	-0.112
inv_gr1	Inventory growth	-0.831**	-0.796**	-0.728**	0.103
fnl_gr1a	Change in financial liabilities	-1.124**	-1.153**	-1.227**	-0.103
oaccruals_ni	Percent operating accruals	-1.084**	-1.053**	-0.984**	0.100*
capx_gr3	CAPEX growth (3 years)	-0.508**	-0.585**	-0.606**	-0.098**
rd_me	R&D-to-market	0.393**	0.431**	0.489**	0.095
taccruals_ni	Percent total accruals	-0.323**	-0.276*	-0.228	0.095
niq_be	Quarterly return on equity	0.589**	0.605**	0.495**	-0.094**
ni_ivol	Earnings volatility	-0.061	-0.148	-0.152	-0.091**
netis_at	Net total issuance	-0.646**	-0.697**	-0.733**	-0.087*
be_gr1a	Change in common equity	-0.334**	-0.325**	-0.250*	0.084**
ppeinv_gr1a	Change PPE and Inventory	-0.812**	-0.721**	-0.729**	0.083
ret_60_12	Long-term reversal	-0.264*	-0.346**	-0.346**	-0.082*
niq_at_chg1	Change in quarterly return on assets	0.400**	0.452**	0.482**	0.082
ope_bel1	Operating profits-to-lagged book equity	0.313**	0.359**	0.394**	0.081**
rd_sale	R&D-to-sales	-0.002	-0.044	-0.082	-0.079
corr_1260d	Market correlation	-0.053	0.027	0.027	0.079**
sale_bev	Assets turnover	0.752**	0.807**	0.830**	0.078**
capx_gr1	CAPEX growth (1 year)	-0.685**	-0.668**	-0.611**	0.073
sale_gr3	Sales Growth (3 years)	-0.258*	-0.335**	-0.330**	-0.072**
emp_gr1	Hiring rate	-0.406**	-0.454**	-0.477**	-0.071
cowc_gr1a	Change in current operating working capital	-0.946**	-0.953**	-0.877**	0.069
capx_gr2	CAPEX growth (2 years)	-0.626**	-0.653**	-0.694**	-0.067
nncoa_gr1a	Change in net noncurrent operating assets	-0.973**	-0.909**	-0.906**	0.067*
ebit_bev	Return on net operating assets	0.547**	0.592**	0.611**	0.064**
op_atl1	Operating profits-to-lagged book assets	0.327**	0.341**	0.391**	0.063**
gp_atl1	Gross profits-to-lagged assets	0.249**	0.261*	0.312**	0.062**
seas_16_20an	Years 16-20 lagged returns, annual	0.733**	0.675**	0.675**	-0.059
saleq_gr1	Sales growth (1 quarter)	-0.159	-0.156	-0.218	-0.059

This table reports the annualized Sharpe ratios of characteristic-managed portfolios in the US dataset. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: only missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: only missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**). Unlike Table 3, we report in this table the characteristics whose absolute changes in the column Δ are ranked 16th to 50th.

Table A3: Annualised Sharpe ratios of characteristic-managed portfolios, European data

Label	Name	Tensor models: cs-mean	only missing	missing&stale	Δ
rd_sale	R&D-to-sales	0.141	-0.088	-0.075	-0.216
kz_index	Kaplan-Zingales index	-0.121	-0.343*	-0.335*	-0.214**
o_score	Ohlson O-score	-0.655**	-0.840**	-0.868**	-0.213**
dbnetis_at	Net debt issuance	-1.034**	-1.012**	-0.829**	0.205
noa_gr1a	Change in net operating assets	-0.996**	-0.874**	-0.795**	0.201*
col_gr1a	Change in current operating liabilities	-0.145	-0.080	0.028	0.173
dsale_drec	Change sales minus change receivables	0.169	0.038	-0.002	-0.170
ncoa_gr1a	Change in noncurrent operating assets	-0.765**	-0.735**	-0.599**	0.166
ocf_at_chg1	Change in operating cash flow to assets	0.772**	0.798**	0.934**	0.162
oaccruals_ni	Percent operating accruals	-0.308*	-0.201	-0.148	0.160
ni_ivol	Earnings volatility	-0.085	-0.256	-0.244	-0.160*
coa_gr1a	Change in current operating assets	-0.520**	-0.416**	-0.361**	0.159
capx_gr2	CAPEX growth (2 years)	-0.439**	-0.540**	-0.597**	-0.158
ope_bel1	Operating profits-to-lagged book equity	0.518**	0.600**	0.663**	0.145**
taccruals_ni	Percent total accruals	0.209	0.344*	0.349*	0.139
nncoa_gr1a	Change in net noncurrent operating assets	-0.762**	-0.719**	-0.627**	0.135
inv_gr1	Inventory growth	-0.717**	-0.671**	-0.585**	0.132
capx_gr1	CAPEX growth (1 year)	-0.462**	-0.530**	-0.592**	-0.130
inv_gr1a	Inventory change	-0.631**	-0.600**	-0.504**	0.127
ocf_at	Operating cash flow to assets	0.779**	0.808**	0.901**	0.123**
at_gr1	Asset Growth	-0.400**	-0.410**	-0.280	0.120*
saleq_gr1	Sales growth (1 quarter)	0.033	-0.029	-0.087	-0.120
capx_gr3	CAPEX growth (3 years)	-0.385**	-0.520**	-0.502**	-0.117
pi_nix	Taxable income-to-book income	-0.028	0.017	-0.139	-0.111
tangibility	Asset tangibility	0.294	0.393**	0.403**	0.109
oaccruals_at	Operating accruals	-0.580**	-0.418**	-0.471**	0.108
op_atl1	Operating profits-to-lagged book assets	0.409**	0.417**	0.516**	0.107**
sale_bev	Assets turnover	0.534**	0.617**	0.641**	0.107**
cash_at	Cash-to-assets	0.094	-0.020	-0.011	-0.106
netis_at	Net total issuance	-1.037**	-1.192**	-1.141**	-0.104
cop_atl1	Cash-operating-profits-to-lagged-assets	0.760**	0.585**	0.661**	-0.099
at_turnover	Capital turnover	0.380**	0.427**	0.474**	0.095**
taccruals_at	Total accruals	0.338*	0.389**	0.244	-0.093
qmj	Quality minus Junk: Composite	0.530**	0.623**	0.622**	0.092
aliq_mat	Liquidity of market assets	0.368**	0.459**	0.459**	0.091**

This table reports the annualized Sharpe ratios of characteristic-managed portfolios in the European dataset. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: only missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: only missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**). Unlike Table 3, we report in this table the characteristics whose absolute changes in the column Δ are ranked 16th to 50th.

Table A4: Annualised Sharpe ratios of characteristic-managed portfolios, Asian data

Label	Name	cs-mean	Tensor models: only missing	Tensor models: missing&stale	Δ
rd_sale	R&D-to-sales	0.280*	0.138	0.166	-0.114
ni_ivol	Earnings volatility	0.042	0.023	-0.062	-0.105
pi_nix	Taxable income-to-book income	-0.239	-0.204	-0.140	0.098
fnl_gr1a	Change in financial liabilities	-0.091	-0.100	-0.183	-0.092
aliq_at	Liquidity of book assets	-0.156	-0.218	-0.245	-0.089
dsale_drec	Change sales minus change receivables	0.451**	0.508**	0.363*	-0.088
op_atl1	Operating profits-to-lagged book assets	0.352**	0.282	0.269	-0.083
ocf_me	Operating cash flow-to-market	0.493**	0.573**	0.574**	0.081**
at_be	Book leverage	-0.203	-0.169	-0.122	0.081
cop_at	Cash-operating-profits-to-assets	0.510**	0.422**	0.433**	-0.078
dbnetis_at	Net debt issuance	-0.152	-0.163	-0.228	-0.076
at_turnover	Capital turnover	0.012	0.011	0.087	0.075
op_at	Operating profits-to-book assets	0.408**	0.372**	0.337*	-0.072
ncoa_gr1a	Change in noncurrent operating assets	-0.091	-0.123	-0.161	-0.070
inv_gr1a	Inventory change	-0.195	-0.177	-0.127	0.068
sale_bev	Assets turnover	0.000	-0.002	0.064	0.064
o_score	Ohlson O-score	-0.341**	-0.287*	-0.280	0.061
sale_gr1	Sales Growth (1 year)	0.081	0.076	0.021	-0.060
inv_gr1	Inventory growth	-0.152	-0.113	-0.095	0.058
nncoa_gr1a	Change in net noncurrent operating assets	-0.096	-0.134	-0.147	-0.051
taccruals_ni	Percent total accruals	-0.068	-0.097	-0.118	-0.050
netis_at	Net total issuance	-0.372**	-0.355**	-0.422**	-0.050
sale_gr3	Sales Growth (3 years)	0.099	0.091	0.049	-0.050
debt_gr3	Growth in book debt (3 years)	-0.125	-0.127	-0.174	-0.049
tax_gr1a	Tax expense surprise	0.058	0.058	0.106	0.049
nfna_gr1a	Change in net financial assets	0.080	0.075	0.128	0.047
at_gr1	Asset Growth	-0.063	-0.077	-0.108	-0.045
div12m_me	Dividend yield	0.850**	0.892**	0.892**	0.042
z_score	Altman Z-score	-0.069	-0.077	-0.109	-0.040
coa_gr1a	Change in current operating assets	-0.190	-0.144	-0.151	0.039
tangibility	Asset tangibility	0.039	-0.028	-0.000	-0.039
cop_atl1	Cash-operating-profits-to-lagged-assets	0.429**	0.366**	0.391**	-0.038
aliq_mat	Liquidity of market assets	0.434**	0.472**	0.472**	0.038
rskew_21d	Total skewness	-0.049	-0.086	-0.086	-0.037*
ocf_at	Operating cash flow to assets	0.464**	0.447**	0.501**	0.037

This table reports the annualized Sharpe ratios of characteristic-managed portfolios in the Asian dataset. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: only missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: only missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**). Unlike Table 3, we report in this table the characteristics whose absolute changes in the column Δ are ranked 16th to 50th.

Table A5: IPCA alphas of characteristic-managed portfolios, additional results

Label	Name	cs-mean	Tensor models: only missing	Tensor models: missing&stale	Δ
	Panel A. US data				
qmj	Quality minus Junk: Composite	0.283**	0.475**	0.476**	0.193*
qmj_growth	Quality minus Junk: Growth	0.374**	0.482**	0.520**	0.146
chcsho_12m	Net stock issues	-0.390**	-0.363**	-0.309**	0.081
iskew_hxz4_21d	Idiosyncratic skewness, HXZ4	0.355**	0.312**	0.332**	-0.024
ret_60_12	Long-term reversal	0.347**	0.321**	0.331**	-0.016
seas_2_5na	Years 2-5 lagged returns	0.327**	0.326**	0.315**	-0.011
resff3_12_1	Residual momentum t-12 to t-1	-0.303**	-0.322**	-0.305**	-0.002
	Panel B. European data				
mispricing_mgmt	Mispricing factor: Management	0.343**	0.482**	0.530**	0.187
rskew_21d	Total skewness	0.556**	0.454**	0.418**	-0.138
nfna_gr1a	Change in net financial assets	0.481**	0.482**	0.414**	-0.067
ivol_capm_252d	Idiosyncratic volatility, CAPM (252 days)	0.434**	0.378*	0.399**	-0.035
fnl_gr1a	Change in financial liabilities	-0.433**	-0.456**	-0.407**	0.026
iskew_capm_21d	Idiosyncratic skewness, CAPM	0.504**	0.527**	0.528**	0.024
	Panel C. Asian data				
fnl_gr1a	Change in financial liabilities	-0.585**	-0.503**	-0.443**	0.142
nfna_gr1a	Change in net financial assets	0.552**	0.480**	0.432**	-0.121
netdebt_me	Net debt-to-price	-0.448**	-0.352*	-0.362**	0.086
dbnetis_at	Net debt issuance	-0.528**	-0.461**	-0.448**	0.080
ncoa_gr1a	Change in noncurrent operating assets	-0.529**	-0.538**	-0.456**	0.074
mispricing_mgmt	Mispricing factor: Management	0.582**	0.582**	0.516**	-0.066
z_score	Altman Z-score	0.497**	0.377**	0.436**	-0.061
noa_gr1a	Change in net operating assets	-0.476**	-0.457**	-0.425**	0.051
cop_at	Cash-operating-profits-to-assets	0.435**	0.368**	0.386**	-0.049
debt_me	Debt-to-market	-0.483**	-0.394**	-0.434**	0.049
cash_at	Cash-to-assets	0.419**	0.407**	0.466**	0.047
iskew_ff3_21d	Idiosyncratic skewness, FF3	0.338**	0.434**	0.378**	0.040
qmj_safety	Quality minus Junk: Safety	0.536**	0.475**	0.498**	-0.038
oaccruals_at	Operating accruals	-0.367**	-0.365**	-0.393**	-0.026
nncoa_gr1a	Change in net noncurrent operating assets	-0.440**	-0.465**	-0.432**	0.008

This table reports the IPCA alphas of characteristic-managed portfolios in 15-factor models. The alphas are defined as the difference between their realized mean returns and model-predicted mean returns, as in equation (20). For comparability, we report the Sharpe ratios of the alphas. The column "cs-mean" corresponds to the setting in which missing values are imputed by cross-sectional means of characteristics. The two columns "Tensor models: only missing" and "Tensor models: only missing&stale" display the results obtained by employing our tensor model to impute missing or infrequently updated information. The final column (Δ) reports the difference between the columns "Tensor models: only missing&stale" and "cs-mean". We estimate the confidence intervals of the Sharpe ratios and their differences (in column Δ) using a hierarchical Bootstrap approach. If the 90% (95%) confidence interval of the estimate does not contain zero, the estimate will be highlighted by * (**). Unlike Table 4, we present the estimates for the remaining characteristics whose IPCA alphas are always significant at the 5% significance level with and without data imputation.