Uncertainty everywhere: Integrating conceptual uncertainty in the stochastic discount factor*

Luis Gruber, Gregor Kastner, Stefan Voigt, and Patrick Weiss†

December 11, 2025

Abstract

Model and parameter uncertainty are well-established concerns in estimating the stochastic discount factor. We show that conceptual uncertainty, stemming from defensible yet arbitrary choices in factor construction, like portfolio breakpoints or weighting schemes, is equally consequential. We develop a path-augmented Bayesian model averaging framework that systematically integrates trillions of plausible risk factor permutations into the stochastic discount factor. Accounting for conceptual uncertainty reduces pricing errors and yields 50% higher out-of-sample Sharpe ratios relative to established benchmark models. Our findings underscore the critical role of data preprocessing in asset pricing and demonstrate that explicitly modeling such choices enhances empirical inference.

Keywords: Factor models, stochastic discount factor, model averaging, non-standard errors
JEL Classification: G11, G12, C11, C52, C58

*We thank participants at the Copenhagen Business School (CBS) BIGFI/Finance seminar, University of Konstanz, Karlsruhe Institute for Technology (KIT), WU Vienna University of Economics and Business, NUS Quantitative Finance Conference 2025, Collegio Carlo Alberto, and the FinEML Conference 2025 in Rotterdam for helpful comments and suggestions. Support from Inquire Europe is gratefully acknowledged. Stefan Voigt is grateful to Long-Term Investors@UniTo and the Danish Finance Institute for financial support.

†Luis Gruber is at University of Klagenfurt, e-mail: luis.gruber@aau.at. Gregor Kaster is at University of Klagenfurt, e-mail: gregor.kastner@aau.at. Stefan Voigt is at University of Copenhagen and the Danish Finance Institute, e-mail: stefan.voigt@econ.ku.dk. Patrick Weiss is at Reykjavik University, e-mail: patrickw@ru.is.

1 Introduction

In empirical finance, researchers have considerable discretion in translating conceptual research questions into empirical designs. Seemingly innocuous choices, such as the treatment of missing data or the weighting scheme for portfolios, can meaningfully affect key quantities such as a risk premium estimate. Such shifts influence perceived risk and, in turn, shape capital allocation decisions. Often underappreciated, such leeway in empirical design choices introduces conceptual uncertainty. Menkveld et al. (2024) refers to this variation as "non-standard errors" with magnitudes on par with the classical sampling error.

Estimating the stochastic discount factor (SDF) m_t is prone to substantial conceptual uncertainty. Under no-arbitrage conditions, the conditional expected excess return of an asset r_t equals its negative conditional covariance with the SDF. A canonical proxy for the unobservable marginal utility of the representative investor, the defining component of the SDF, assumes that m_t is linear in K tradeable risk factor returns f_t, i.e., $m_t = a - b'f_t$. Substituting and rearranging yields the beta pricing equation[1]

$$\mathbb{E}(r_t) = -\mathbb{C}ov(m_t, r_t) = \beta'\lambda, \tag{1}$$

where β is a $K \times 1$ vector of factor loadings and λ is a $K \times 1$ vector of factor risk premia. A central challenge in estimating Equation (1) is to choose the relevant risk factors f_t. The *correct* collection of risk factors remains subject to research controversy. Harvey et al. (2015), for instance, documents hundreds of proposed risk factors. Yet, even *if* the relevant sources of systematic risk were known, conceptual uncertainty persists: risk factor return time series f_t are typically constructed from firm characteristics. Empirically, the relation between r_t and f_t is often investigated via portfolio sorting, a non-parametric approach to capture the risk premium associated with a stock characteristic (Cattaneo

[1] All expectations and covariances are conditional on information available at time $t-1$. We omit the conditioning notation for ease of exposition. In our empirical application, we explicitly consider conditional SDFs.

et al., 2020). Portfolio sorting involves grouping assets based on a firm characteristic (e.g., size or value) and tracking the average returns of these groups to approximate the relation between expected returns and factor exposures β in Equation (1).

How to construct, say, a size factor, remains a subjective decision. While theory may suggest *which* characteristics should matter, the empirical implementation relies on specific choices regarding the construction of risk factor proxies. Walter et al. (2024) document that such design choices for portfolio sorts can meaningfully affect both the time series of risk factors and their estimated risk premia, leading to considerable conceptual uncertainty at the risk-factor level.

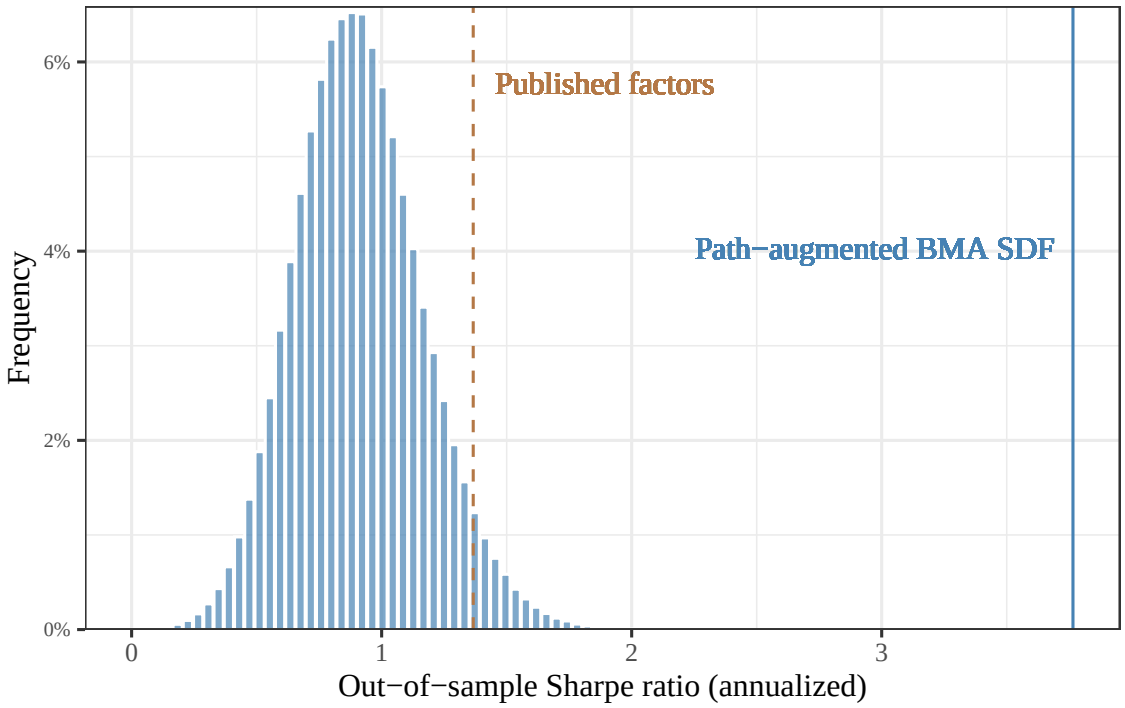
As a motivating example, suppose the SDF spanned by 51 risk factors.[2] Figure 1 illustrates the fundamental challenge: the implied efficient tangency portfolio based on the published risk factors yields an out-of-sample Sharpe ratio of 1.5, yet randomly sampling alternative construction paths for each risk factor generates Sharpe ratios between near zero to two. Solely relying on the single preprocessing path that produces the published risk factors ignores conceptual uncertainty entirely and yields a severely limited picture of the risk-return trade-off.

Given such pervasive conceptual uncertainty, how should we formulate optimal decisions? In this paper, we propose a novel solution to handle conceptual uncertainty as a new dimension of uncertainty: the *path-augmented BMA SDF*. Rather than selecting a single arbitrary preprocessing path, our approach integrates over all plausible specifications simultaneously using Bayesian model averaging. The blue solid line in Figure 1 shows the substantially increased out-of-sample Sharpe ratio, documenting the importance of systematically incorporating conceptual uncertainty in the path-augmented BMA SDF.

Our framework relies on economically motivated priors for three key dimensions: model parameters, risk factor inclusion, and preprocessing choices. The derived posterior distributions allow us to systematically integrate over both model uncertainty (which risk factors to include) and conceptual uncertainty (how to construct the included risk

[2]We present the underlying database of risk factor returns and their construction details in Section 3.

Figure 1: Sharpe ratios of one million efficient tangency portfolios. In this figure, we show a distribution of annualized out-of-sample Sharpe ratios. We compute efficient tangency weights for a fixed asset universe of 51 risk factors, randomly varying only the data preprocessing choices discussed in Section 3. The weights are optimized based on sample moments from 1973 - 2006, and we evaluate the resulting Sharpe ratio for the period 2006 - 2022. The performance of the published risk factors from Open Source Asset Pricing (Chen and Zimmermann, 2022) is indicated with a brown dashed line, and the performance of the proposed path-augmented BMA SDF is indicated with a blue solid line.



factors), yielding an SDF that reflects all sources of uncertainty. We assess risk factor relevance using well-established conditional linear factor models that nest common SDF specifications (see, e.g., Avramov et al., 2023). The Bayesian approach offers important inferential advantages. As in Jensen et al. (2023), we impose a conservative prior where all alphas are expected to be zero. Risk factor inclusion probabilities emerge from joint inference across all risk factors simultaneously, integrate out the effect of preprocessing choices, and naturally address the multiple-testing concerns of frequentist approaches.

Our model space contains trillions of possible SDF specifications. In particular, we apply our path-augmented BMA SDF to 51 widely-used risk factors and consider eleven common

data preprocessing choices that yield up to 3,840 plausible specifications per risk factor.[3] The resulting comprehensive risk factor database is part of the open-source Tidy Finance project (Scheuch et al., 2023) and publicly available at app-download-center.cloud.sdu.dk. Our database reveals substantial variation in estimated risk premia and in factors' time-series correlations across specifications. While our risk factor library highlights that data preprocessing matters, we confirm the robustness of many risk factors to conceptual uncertainty (similar to, e.g., Jensen et al., 2023). By exploring an arguably small set of such preprocessing choices, we can rigorously assess the relevance of conceptual uncertainty in estimating the SDF and derive economically interpretable insights (Hellum et al., 2025).

Our empirical results yield three main insights. First, our posterior model probabilities support a stochastic discount factor dense in risk factors. Our path-augmented BMA SDF selects, on average, 31 out of 50 risk factors. Virtually all models include long-term reversal, trading volume, and size as relevant factors. As a consequence of our economically motivated priors, risk factors are only deemed relevant if they contribute to pricing accuracy across multiple data preprocessing choices. Interestingly, a BMA SDF based solely on published risk factors in the spirit of Avramov et al. (2023) selects a very similar number of risk factors (on average, 30 out of 50). Thus, conceptual uncertainty itself does not increase the model complexity in terms of the number of selected risk factors.

Second, the path-augmented BMA SDF consistently outperforms standard factor models regarding out-of-sample pricing accuracy. We document superior performance in (i) pricing a cross-section of asset returns for a period not used for estimation and (ii) pricing a cross-section of novel test assets in a period not used for estimation. Overall, we consistently observe lower pricing errors relative to established benchmark models such as the Fama and French (1993) three-factor model and a conditional BMA SDF based on published risk factors in the spirit of Avramov et al. (2023).

[3]Overall, our model space comprises approximately 10^{174} potential models, far more than atoms in the observable universe. We developed suitable MCMC techniques to explore this vast model space efficiently.

Finally, we show that conceptual uncertainty represents a substantial and previously unquantified source of investment risk. To show this, we decompose the posterior predictive variance of the asset returns implied by our model into three components: the average model-specific risk, model disagreement (variation in expected returns across models), and conceptual uncertainty (variation in expected returns across data preprocessing choices within a model). Empirically, conceptual uncertainty contributes almost as much to total variance as model disagreement, which has been extensively studied: in the presence of model disagreement, optimal capital allocation changes and asset pricing inference shifts (Avramov, 2002; Avramov and Chordia, 2006). Consequently, investors who ignore conceptual uncertainty as a source of investment risk must accept substantially smaller risk-adjusted returns, as illustrated in Figure 1.

Our paper is related to several strands of the literature. First, we contribute to the growing literature on model and estimation uncertainty in empirical (Bayesian) asset pricing. Barillas and Shanken (2018), Chib et al. (2020), and Bryzgalova et al. (2023) developed statistical methodologies for identifying promising factor pricing models in a static framework, leading to an unconditional SDF. In contrast, and in line with Avramov et al. (2023), we explicitly focus on a conditional SDF. Our results confirm the importance of time variation in factor loadings and mispricing to enhance pricing performance. We use their conditional BMA SDF, which does not incorporate conceptual uncertainty, as one of our primary benchmark models.

The literature on the factor zoo (see, e.g., Cochrane, 2011; Hou et al., 2020) is substantial. While Harvey (2017) warns of false discoveries in factor research, McLean and Pontiff (2016) show many factors fail to replicate out-of-sample. Our paper contributes to the quest for identifying relevant risk factors by highlighting the importance of conceptual uncertainty when constructing risk factor returns. The question of which risk factors to include is incomplete without simultaneously considering how to construct them. To the best of our knowledge, we are the first paper to explicitly model the relevance of conceptual uncertainty for estimating the SDF and provide guidance on optimal portfolio

choice with a multitude of plausible data preprocessing choices.

In financial economics, the importance of data preprocessing choices has been recognized, for instance, by Mitton (2022) studying corporate finance regressions.[4] Historically, Stambaugh (1982) already tested the CAPM using a range of different proxies for the market portfolio to partially address the critique of Roll (1977). Other disciplines also realized that data set construction can introduce noise and potentially bias. For example, Steegen et al. (2016) and Simonsohn et al. (2020) propose *multiverse* analyses that report results across alternatively processed data sets. In contrast, we integrate over all uncertainty rather than merely documenting it. In the context of financial decision-making, however, conceptual uncertainty is ultimately a source of investment risk that needs to be integrated into the investment decision, rather than only quantifying its magnitude.

Our paper is also related to the literature on machine-learning methods in finance (see, e.g., Gu et al., 2020). Ultimately, considering different risk factor specifications can be viewed as a highly structured, non-linear approach to mapping expected returns onto potentially relevant characteristics. What distinguishes our approach from classical machine learning is that instead of relying on necessary cross-validation or penalization procedures to avoid overfitting, we rely on the guidance imposed by the academic profession to select appropriate choices. The result is a set of interpretable models that can be used to derive economically meaningful conclusions. In line with this interpretation, our findings relate to recent work on model complexity in asset pricing. While Kelly et al. (2024) argue that allowing for higher model complexity can improve predictive performance, Cartea et al. (2025) show that the opposite holds if risk factor returns are measured with error. Our approach navigates a middle ground. We allow for model complexity by integrating over a vast set of plausible models induced by conceptual uncertainty. By averaging across plausible data preprocessing choices, we reduce measurement error while retaining interpretability.

[4]Several papers take an asset-pricing perspective: Soebhag et al. (2024) studies portfolio sorts, Chen et al. (2024) highlights conceptual uncertainty for machine learning, and Kessler et al. (2020) and Hasler (2025) analyse conceptual uncertainty for the value premium.

2 Bayesian conditional linear factor models

We consider a conditional stochastic discount factor in the spirit of Avramov et al. (2023), which is linear in a set of risk factors with potentially time-varying mispricing and factor loadings. We first present the asset pricing model and corresponding prior elicitation for a fixed set of risk factors and data preprocessing choices. Subsequently, we detail how to integrate over both model uncertainty and conceptual uncertainty by means of Bayesian model averaging.

2.1 Asset pricing model

We consider a general class of conditional asset pricing models where the $N \times 1$ vector of expected excess returns r_t for the N test assets in period $t = 1, \dots, T$ is a linear function of K tradeable risk factor returns f_t[5]

$$r_t = \alpha(z_{t-1}) + \beta(z_{t-1}) f_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{2}$$

where z_{t-1} is an $L \times 1$ vector of macroeconomic predictors (where L is the number of predictors), and the $N \times 1$ vector ε_t is assumed Normally distributed with mean zero and $N \times N$ variance-covariance matrix Σ. The $N \times 1$ vector $\alpha(z_{t-1}) = \alpha_0 + \alpha_1 z_{t-1}$ reflects mispricing and is determined by a fixed component through the $N \times 1$ vector α_0 and a time-varying component through the $N \times M$ matrix α_1.[6] Similarly, the $N \times K$ matrix $\beta(z_{t-1})$ of factor loadings can be time-varying as a function of macroeconomic predictors such that $\beta(z_{t-1}) = \beta_0 + \beta_1(I_K \otimes z_{t-1})$ where $I_K \in \mathbb{R}^{K \times K}$ is the identity matrix and $\otimes$ denotes the Kronecker product. Stacking the observations r_t, z_{t-1}, and f_t, for all T periods, such that $R \in \mathbb{R}^{T \times N}$, $F \in \mathbb{R}^{T \times K}$, $Z \in \mathbb{R}^{T \times L}$, and collecting the regression coefficients in the $(K+1)(L+1) \times N$ matrix $\theta = [\alpha_0, \alpha_1, \beta_0, \beta_1]'$ yields the canonical

[5] In all of our specifications, we consider the Capital Asset Pricing Model as a plausible benchmark, so the market return $f_{t,m}$ is always included into the set of tradeable risk factor returns.

[6] Among others, Ferson and Harvey (1999) highlight the relevance of time-varying mispricing in asset pricing tests, Jagannathan and Wang (1996) link factor loading dynamics to the business cycle.

conditional asset pricing model in matrix form

$$R = W\theta + E, \tag{3}$$

with $W = [\iota_T, Z, F, \Psi]$, where $\Psi = (\psi_1, \ldots, \psi_T) \in \mathbb{R}^{T\times(KM)}$ is defined through $\psi_t = (I_K \otimes z_{t-1})f_t$ and ι_T is a T-vector of ones. Ordinary-least square estimation of the model in Equation (3) yields the well-known estimator $\theta^{\text{OLS}} = (W'W)^{-1}W'R$.

In line with a large body of work in financial economics, we motivate economically meaningful priors on the parameter space (see, e.g., Pastor and Stambaugh, 1999; Kozak et al., 2020). To estimate the model parameters $\{\theta, \Sigma\}$, we impose prior beliefs which are weighted towards the static *Capital Asset Pricing Model*. The prior is rooted in the belief that (i) there is no mispricing in the market, (ii) factor loadings are constant, and (iii) only the market risk factor is relevant for pricing. Thus, instead of choosing the estimator that delivers the lowest possible squared in-sample pricing errors, θ^{OLS}, our Bayesian approach weights the prior conviction and the evidence from the data to arrive at a posterior estimate of θ^{post}, balancing both sources of information. The prior implies $\alpha_0^{\text{prior}} = 0_{N\times 1}$, $\alpha_1^{\text{prior}} = 0_{N\times L}$, and $\beta_1^{\text{prior}} = 0_{N\times K(L+1)}$ where $0_{i\times j}$ is an $i \times j$ matrix of zeroes. For β_0^{prior}, the factor loading on the market risk factor f_m is $\beta_m^{\text{prior}} = (f_m'f_m)^{-1} f_m'R$, the CAPM beta coefficient stemming from a regression of test asset returns on the market risk factor excess returns without an intercept term. For all other risk factors loadings, $\beta_0^{\text{prior}} = 0$. Taken together, the prior beliefs imply

$$\theta^{\text{prior}} = \left[0_{N\times(L+1)}, \beta_0^{\text{prior}}, 0_{N\times K(L+1)}\right].$$

We choose our empirical Bayes prior of conditional conjugate form such that θ is multivariate normal distributed around θ^{prior} with covariance matrix proportional to the covariance matrix $(W'W)^{-1}$ of the OLS estimator θ^{OLS}, scaled by an hypothetical sample size $T_0 > 0$ such that

$$\pi\left(\text{vec}\left(\theta\right)|\Sigma\right) \propto \mathcal{N}\left(\text{vec}\left(\theta^{\text{prior}}\right), \frac{T}{T_0}\Sigma \otimes (W'W)^{-1}\right). \tag{4}$$

The corresponding inverse Wishart prior for the covariance matrix Σ is specified as $\Sigma \sim \mathcal{IW}(S^{\text{prior}}, \nu^{\text{prior}})$, with scale matrix $S^{\text{prior}} = \frac{T_0}{T}\left(R - W\theta^{\text{prior}}\right)'\left(R - W\theta^{\text{prior}}\right)$ and degrees of freedom $\nu^{\text{prior}} = T_0 - (K+1)(L+1)$. Combining the likelihood of the data $D = \{R, F, Z\}$ with the prior distribution leads to the canonical Gaussian conditional posterior distribution $\pi\left(\text{vec}(\theta)|\Sigma, \mathcal{D}\right)$ with conditional posterior mean

$$\theta^{\text{post}} = \frac{T_0}{T+T_0}\theta^{\text{prior}} + \left(1 - \frac{T_0}{T+T_0}\right)\theta^{\text{OLS}} \tag{5}$$

and conditional posterior covariance matrix $\frac{T}{T+T_0}\Sigma \otimes (W'W)^{-1}$. The resulting conditional posterior mean of mispricing, $\mathbb{E}\left(\alpha^{\text{post}}(z_{t-1})|\mathcal{D}, \Sigma\right) = \frac{T}{T+T_0}\alpha(z_{t-1})^{\text{OLS}} < \alpha(z_{t-1})^{\text{OLS}}$ is effectively shrunk towards zero, with the degree of shrinkage controlled by the hypothetical sample size T_0. Depending on the strength of the prior, large levels of mispricing $\alpha^{\text{OLS}}(z_{t-1})$ are substantially reduced in the posterior assessment, in spirit similar to the framework of Jensen et al. (2023). Intuitively, a large T_0 increases the weight on the prior beliefs and renders mispricing but also time-variation in the factor loadings less plausible. More complex models are thus penalized a priori, effectively guarding against overfitting.

To control the size of the shrinkage effect, we rely on economic reasoning to calibrate T_0. In the spirit of Barillas and Shanken (2018) and Chib et al. (2020), our choice of T_0 implicitly puts a prior on the extent of expected mispricing. We set T_0 such that the maximum value for the admissible Sharpe ratio spanned by the SDF of a given factor model relative to the Sharpe ratio of the market factor f_m is $\tau \geq 1$.[7] The connection between T_0 and the admissible Sharpe ratio is a direct consequence of the shrinkage of mispricing in the posterior distribution: a very high hypothetical sample size T_0 implies strong shrinkage of mispricing towards zero and, consequently, a low admissible Sharpe ratio. Conversely, a small T_0 allows for larger levels of mispricing and, thus, a higher admissible Sharpe ratio. To pin down the exact choice of T_0, we follow Avramov et al. (2023) and show in the Appendix that the conditional variance of total mispricing $\alpha(z_{t-1})$

[7] In all empirical applications, we use $\tau = 2$, i.e., we consider Sharpe ratios twice as high as the Sharpe ratio of the market plausible. Our main results are robust to varying τ between 1.5 and 2.5.

is given by $\mathbb{V}\text{ar}\left(\alpha|\Sigma,\mathcal{D}\right)=\left(1+L\right)\frac{1+Sh(F)^2}{T_0}\Sigma$ where $Sh(F)^2=\mu_F'\Sigma_F^{-1}\mu_F$ with $\mu_F=\mathbb{E}(F)$ and $\Sigma_F=\mathbb{V}\text{ar}(F)$ the $K\times K$ variance-covariance matrix of F.

Then, note that the Gaussian conditional posterior distribution directly implies that $\mathbb{E}\left(\alpha'\Sigma^{-1}\alpha|\Sigma,\mathcal{D}\right)=\left(N+K\right)\left(1+L\right)\frac{1+Sh(F)^2}{T_0}$. In a regression context, Gibbons et al. (1989) also show that $\mathbb{E}\left(\alpha'\Sigma^{-1}\alpha\right)=Sh(F,R)^2-Sh(F)^2$, corresponding to the improvement of the performance of the efficient tangency portfolio spannend by all test assets and risk factor returns relative to the risk factor returns only. We set the upper bound on reasonable improvements of the squared Sharpe ratio of the test asset returns R over the market factor returns f_m such that $\mathbb{E}\left(\alpha'\Sigma^{-1}\alpha|\Sigma,\mathcal{D}\right)=\left(\tau^2-1\right)Sh(f_m)^2$ where $\tau\geq 1$. Rearranging terms leads to the corresponding hypothetical sample size

$$T_0=\frac{\left(N+K\right)\left(1+L\right)\left(1+Sh(F)^2\right)}{\left(\tau^2-1\right)Sh(f_m)^2}. \tag{6}$$

Intuitively, a large τ increases the admissible Sharpe ratio and reduces T_0 relative to T. The resulting posterior conditional mean of the misspricing coefficient $\alpha|\sigma^2,\mathcal{D}$ thus *decreases* with *increasing* ex-ante Sharpe ratios of the SDF spanned by the risk factor returns F. As a result, risk factors with higher ex-ante Sharpe ratios lead to stronger shrinkage of mispricing towards zero, effectively "leveling the field" and guarding against overfitting. In the context of conceptual uncertainty, such an economical prior rooted in admissiable Sharpe ratios penalizes data preprocessing steps with the aim of identifying large in-sample Sharpe ratios.

2.2 Bayesian model comparison and averaging

For a given set of risk factors F, the posterior distribution of the parameters in the asset pricing model of Equation (2) in combination with the data-driven choice of the hypothetical sample size T_0 in Equation (6) completes the Bayesian specification necessary to evaluate the SDF in Equation (1).

However, there is substantial uncertainty about the identity of asset pricing factors and

thus a reasonable choice of risk factors is up for debate (e.g., Harvey et al. (2015) report 316 potentially relevant risk factors). Thus, model uncertainty effectively gives rise to a vast model space of candidate factor models $\mathcal{M} = \{M_1, \dots, M_P\}$, considering $P = 2^K$ possible combinations of risk factors.

One additional, rather obvious challenge is that characterizing the SDF by a set of selected risk factors in M_p is not sufficient: there may be multiple reasonable ways to construct the risk factor returns f_t that populate a given model M_p. In other words, a risk factor such as the size factor may be measured in multiple defensible ways, and consequently, result in multiple proxies f^j to capture systematic exposure to a conceptually identical risk factor.

To illustrate how Bayesian model comparison alleviates the uncertainty with respect to choosing an appropriate proxy, consider the following setup: suppose, the asset pricing model in Equation (2) is populated by the market factor and one additional risk factor, i.e., $K = 2$. Further, suppose there are two reasonable ways to construct the additional risk factor, resulting in two different proxies f^1 and f^2. The two resulting models M_1 and M_2 thus differ only in the data preprocessing choices used to construct the additional risk factor. The relevant question is: which of the two models is more likely given the data? From a Bayesian perspective, a first step towards model comparison relies on the marginal model likelihood, denoted as

$$\pi\left(\mathcal{D}|M_p\right) = \int_{\theta,\Sigma} \mathcal{L}\left(\mathcal{D}|\theta, \Sigma, M_p\right) \pi\left(\theta|\Sigma, D, M_p\right) \pi\left(\Sigma|\mathcal{D}, M_p\right) d\theta d\Sigma. \tag{7}$$

The marginal likelihood effectively measures how well model M_p explains the observed data $\mathcal{D}$, integrating over all possible parameter values weighted by their posterior probabilities. Averaging across a posteriori plausible parameter values is in stark contrast to classical model selection criteria based on point estimates. To accomodate the multitude of preprocessing choices, we refer to $\mathcal{D}$ as the set $\{R, \mathcal{F}, Z\}$. $\mathcal{F}$ contains the union of all risk factor return time series f_k^j for all $k = 1, \dots, K$ risk factors and the preprocessing

choices indexed by $j = 1, \ldots, n_k$, where n_k is the number of available data preprocessing paths for the k-th risk factor. A model M_1 with a relatively higher marginal likelihood is considered more plausible than a competing model M_2 given the data. We show in the Appendix that for the conditional linear asset pricing model in Equation (2), the marginal likelihood under model M_p is

$$\pi\left(\mathcal{D}|M_p\right) = \frac{\Gamma_N\left(\frac{T+T_0-1}{2}\right)}{\Gamma_N\left(\frac{T_0-1}{2}\right)}\left(\frac{T_0}{T+T_0}\right)^{KN/2}\frac{\det\left(S^{\text{prior}}\right)^{(T_0-1)/2}}{\det\left(S^{\text{post}}\right)^{(T+T_0-1)/2}}, \tag{8}$$

where $\Gamma_N(\cdot)$ is the multivariate gamma function and

$$\begin{aligned} S^{\text{post}} = \frac{T_0}{T}S^{\text{prior}} &+ (R - W\theta^{\text{post}})'(R - W\theta^{\text{post}}) \\ &+ \frac{T_0}{T}(\theta^{\text{post}} - \theta^{\text{prior}})'\left(W'W\right)(\theta^{\text{post}} - \theta^{\text{prior}}). \end{aligned} \tag{9}$$

The first component of S^{post} measures the sum of squared errors resulting from the CAPM. The second component corresponds to the sum of squared pricing errors and shows how well the posterior model fits the observed returns. The last component penalizes the distance between the posterior and prior means of the regression coefficients and reflects how much the data has updated the prior beliefs.

To conduct model comparison between models M_1 and M_2 boils down to computing the (log) marginal likelihood difference $\log\left(\pi\left(\mathcal{D}|M_1\right)/\pi\left(\mathcal{D}|M_2\right)\right)$. Multiple components drive the evaluation of the two models. First, recall that the choice of proxy f^1 and f^2 affects the ex-ante Sharpe ratio of the risk factor returns $Sh(f^1)^2$ and $Sh(f^2)^2$, respectively, and thus the shrinkage intensity T_0 in Equation (6). A higher ex-ante Sharpe ratio leads to a higher hypothetical sample size T_0 and thus more shrinkage of the posterior regression coefficients towards the prior. Second, the choice of the proxy affects the fit of the model to the data through the sum of squared pricing errors in the second component of S^{post}. Third, the choice of proxy affects how much the data updates prior beliefs through the last component of S^{post}. If we assume for simplicity, that the two proxies for the same

underlying risk factor exhibit the same Sharpe ratio, i.e., $Sh(f^1)^2 = Sh(f^2)^2$. Then, the shrinkage intensity T_0 is the same across both models, and the log marginal likelihood difference between the two models M_1 and M_2 boils down to

$$\frac{T_0 + T - 1}{2} \log \left(\det \left(S_1^{\text{post}} \right) / \det \left(S_2^{\text{post}} \right) \right) \tag{10}$$

where S_1^{post} and S_2^{post} are the posterior scale matrices from Equation (9) under models M_1 and M_2, respectively. As a result, the model that achieves a better fit of the data while requiring less updating from the prior beliefs is preferred. In light of model complexity, our prior choices and the Bayesian paradigm explicitly penalize more complex models in multiple ways. First, more complex models with many risk factors tend to exhibit higher ex-ante Sharpe ratios, leading to a higher hypothetical sample size T_0 and thus stronger shrinkage towards the Capital asset pricing model. Second, more complex models tend to require more updating from the prior beliefs to fit the data well, which is penalized through the last component of S^{post}. In fact, deviating from the static CAPM is only considered useful if the resulting posterior model achieves a substantially better fit of the data.

Selecting the best-performing model according to the log marginal likelihood and implicitly discarding all competing specifications as inferior or irrelevant would neglect the uncertainty surrounding model and data preprocessing choices. Instead, a Bayesian perspective offers a coherent alternative that explicitly acknowledges conceptual uncertainty (Draper, 1995). Bayesian Model Averaging (BMA) integrates over all candidate models, weighting each by its posterior probability given the data. The approach thus effectively summarizes the uncertainties about the *model parameters*, the *model*, and the *data preprocessing choices*. Before stating BMA for the problem under consideration, we establish the necessary notation. We denote the generic index set $\mathcal{A}_p$, $p = 1, \dots, P$, labels the risk factors associated with model M_p, s.t. $|\mathcal{A}_p| \in \{0, 1, \dots, K\}$ where $|x|$ denotes the cardinality of

a set x. E.g., if M_p is the model only consisting of the market factor, then $|\mathcal{A}_p| = 1$.[8] Let n_k, $k = 1, \dots, K$, denote the number of construction possibilities (paths) of the k-th risk factor, whereas $q_p = \prod_{i \in \mathcal{A}_p} n_i$, $p = 1, \dots, P$, denotes the number of possible path combinations for model M_p. Last but not least, $\mathcal{Q} = \{Q_{pj}\}_{p=1,\dots,P, j=1,\dots,q_p}$ is the set of path combinations, s.t. Q_{pj} denotes the j-th path combination for model M_p. Then, the posterior probability for model $M_{\tilde{p}}$ with risk factors constructed according to $Q_{\tilde{p}\tilde{j}}$, conditional on the data $\mathcal{D}$ is given by Bayes' theorem:

$$\pi\left(M_{\tilde{p}}, Q_{\tilde{p}\tilde{j}} | D\right) = \frac{\pi\left(\mathcal{D} | M_{\tilde{p}}, Q_{\tilde{p}\tilde{j}}\right) \pi\left(M_{\tilde{p}}, Q_{\tilde{p}\tilde{j}}\right)}{\sum_{p=1}^{P} \sum_{j=1}^{q_p} \pi\left(\mathcal{D} | M_p, Q_{pj}\right) \pi\left(M_p, Q_{pj}\right)}. \tag{11}$$

Here, $\pi\left(\mathcal{D} | M_p, Q_{pj}\right)$ denotes the marginal likelihood conditional on the selected risk factors in M_p and the data preprocessing choices given by Q_{pj}, $\pi\left(M_{\tilde{p}}, Q_{\tilde{p}\tilde{j}}\right)$ is the joint prior probability assigned to model $M_{\tilde{p}}$ with data preprocessing choices given by $Q_{\tilde{p}\tilde{j}}$, and the denominator is the marginal likelihood of the raw data across all models and data preprocessing choices.[9] By integrating over both model and conceptual uncertainty, BMA yields predictions and inferences that reflect the full extent of data preprocessing choices in the modeling process.

2.3 Integrating conceptual uncertainty

In order to account for both uncertainty about risk factor selection and uncertainty about conceptual uncertainty by allowing for different data preprocessing steps in risk factor construction, we augment the conditional asset pricing model in Equation (3) in the following way

$$r_t = \alpha(z_{t-1}) + \sum_{k=1}^{K} \delta_k \sum_{j=1}^{n_k} \gamma_k^j \beta_k^j (z_{t-1}) f_{k,t+1}^j + \varepsilon_t, \tag{12}$$

[8] In our empirical application, every model always contains the market factor. The effective number of possible permutations in the model space is therefore $P = 2^{K-1}$.

[9] The exact prior probability $\pi\left(M_p, Q_{pj}\right)$ could be derived from the prior on δ and γ_k, $k = 1, \dots, K$, detailed in Section 2.3.

where $f^j_{k,t+1}$ denote the returns of risk factor k at time $t+1$, $j = 1, \dots, n_k$ is an index for the different data preprocessing choices and n_k is the number of available data preprocessing paths for the k-th risk factor. Selection of risk factors is indicated by a vector $\delta = (\delta_1, \dots, \delta_K)$ where $\delta_k \in \{0, 1\}$ for $k = 1, \dots, K$. The vector $\gamma_k = (\gamma^1_k, \dots, \gamma^{n_k}_k)'$, where $\gamma^j_k \in \{0, 1\}$ and $\sum_{j=1}^{n_k} \gamma^j_k = 1$, indicates the selected data preprocessing path of the k-th risk factor. The factor loadings associated with the kth factor are collected in the $N \times 1$ vector $\beta^i_k(z_{t-1}) = \beta^i_{k,0} + \beta^i_{k,1} z_{t-1}$. Note that the vector γ_k imposes a strong structure on the asset pricing model shown in Equation (12) as only one data preprocessing path is selected for each risk factor. In other words, we recognize that $f^1_{k,t}, \dots, f^{n_k}_{k,t}$ are different proxies for the same underlying risk factor and thus only one of them should be included in the model at a time. The imposed structure is economically meaningful as it avoids the inclusion of multiple, potentially highly correlated proxies for the same underlying risk factor. While Model (12) effectively nests the canonical conditional asset pricing model in Equation (2) as a special case (by setting $\delta_k = 1$ for all k and fixing $\gamma^j_k = 1$ for a specific i), the path-augmented model allows to explicitly account for conceptual uncertainty in risk factor construction. Our explicit restriction on γ^j_k also distinguishes our approach from classical machine learning procedures in the spirit of Gu et al. (2020), which would typically include all available proxies for a given risk factor and rely on penalization or cross-validation procedures to avoid overfitting. In that sense, both our approach and a setup which allows to populate multiple or all elements of γ_k simultaneously would thus imply a highly non-linear mapping from characteristics to test assets' excess return. The key difference is that our approach relies on the guidance imposed by the academic profession on "how we do things" to select appropriate data preprocessing choices, resulting in a set of interpretable models that can be used to derive economically meaningful conclusions.

Conditional on δ and γ, we specify the prior on the factor loadings analogously to Section 2.1. As prior for δ we set $P(\delta_k = 1|p_0) = p_0$ independently for $k = 1, \dots, K$ (Bernoulli prior). Instead of fixing p_0, we assign a uniform distribution on $[0, 1]$ as a

hyperprior to $p_0 \sim \mathcal{B}(1,1)$ where $\mathcal{B}(\alpha_0, \beta_0)$ denotes a beta distribution with parameters $\alpha_0 = 1$ and $\beta_0 = 1$. As a result, the prior inclusion probabilitiy for any risk factor is $E(p_0) = 1/2$. The prior for $\gamma_k \sim \text{Multinoulli}\left(\frac{1}{n_k}, \dots, \frac{1}{n_k}\right)$ independently for $k = 1, \dots, K$. The multinoulli distribution describes the possible results of a random variable that can take on one of K possible categories, each with prior probability $\frac{1}{n_k}$. The resulting posterior distribution is not of a known form but can efficiently be approximated by means of MCMC algorithms. We provide a description of the efficient MCMC sampler to generate draws from the path-augmented model in the Appendix.

3 Risk factor library with conceptual uncertainty

We construct a comprehensive library of risk factor returns that explicitly accounts for conceptual uncertainty by varying data preprocessing choices in the construction of each risk factor. This first-of-its-kind library includes a large number of possible specifications for each risk factor over our entire sample period, rather than just one specification, which ignores the multitude of preprocessing choices. We make the entire risk factor database publicly available as a part of the open-source Tidy Finance project (Scheuch et al., 2024), and host the data at app-download-center.cloud.sdu.dk. In this section, we specify the data sources and the various data preprocessing choices we consider.

We consider risk factor return time series for 50 cross-sectional sorting variables, which have been used extensively in the asset pricing literature (see, e.g., Hou et al., 2014; Harvey et al., 2015; Chen and Zimmermann, 2022) and feature a variety of underlying economic mechanisms. In Appendix B, we list these 50 sorting variables in Table A1 and document their construction in detail in the Internet Appendix IA-1. All sorting variables are constructed from data from the Center for Research in Security Prices (CRSP) and Compustat.[10] As we require that all risk factor returns are available over the entire sample period, we start our sample in July 1973 and end in December 2022.

[10]To be precise, we use monthly returns from the old CRSP tape (i.e., SIZ) to be consistent with other data providers. As discussed in Schwarz et al. (2025), this choice is not likely to impact major asset pricing findings.

We construct the risk factor return time series using portfolio sorts by varying eleven influential methodological decisions suggested by Walter et al. (2024). Before diving into the details of the preprocessing choices, we provide a brief overview of the overall procedure for portfolio sorting. Consider a cross-section of stocks and a sorting variable $V_{s,t}^k$, which is related to risk factor k, observed for stock s at the end of month t. To conduct portfolio sorts, we split the cross-section of stocks in each month into portfolios. The time-varying breakpoints for each portfolio are defined by the quantiles of $V_{s,t}^k$. In each portfolio, we weight the individual stock returns to form the portfolio return. In a final step, the portfolios are aggregated into a risk factor by going long the portfolio with the highest (lowest) sorting variable and taking a short position in the opposite extreme portfolio to form a long-short portfolio return $f_{k,t+1}^j$ for risk factor k where i denotes an index encoding the specific data preprocessing choices used in the portfolio sort. Finally, we repeat this procedure for each month in our sample to obtain the time series of risk factor returns.

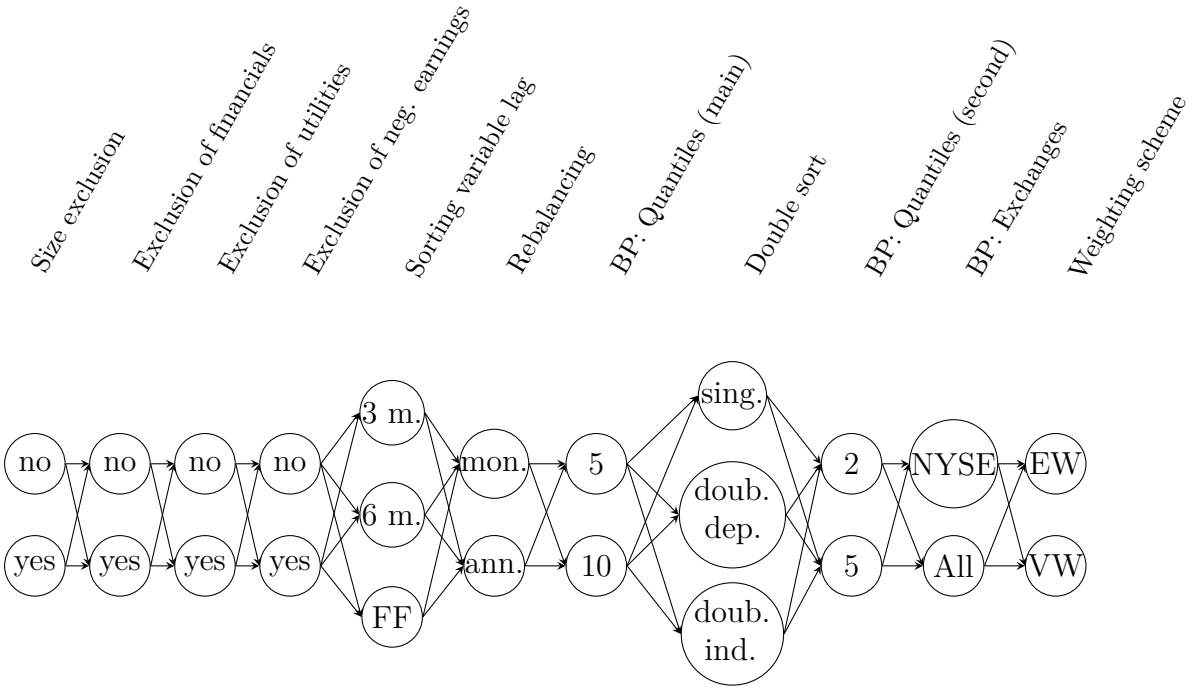
Multiple plausible data preprocessing choices vary at each of the eleven methodological decisions involved in portfolio sorts. The flowchart in Figure 2 shows each decision as a decision fork, which has a fixed set of possible choices. The first four forks concern the sample construction: excluding stocks with market capitalization smaller than the 20^{th} percentile of NYSE stocks, excluding financials, excluding utilities, and excluding firm-months with negative earnings.[11] The ensuing seven forks pertain to portfolio construction: the lag of the sorting variables,[12] the portfolio rebalancing frequency, the number of main portfolios, whether to consider a double sort[13], the number of secondary portfolios for

[11]The size restriction is implemented based on the time-varying market equity distribution of stocks traded on NYSE. In particular, we take the 20^{th} NYSE percentile in each month to classify stocks as small or large. If the preprocessing choice is "yes", we exclude the small stocks.

[12]A market participant cannot actively trade on information that is not observable at the beginning of each month. Therefore sorting variables have to be lagged in order to avoid a look-ahead bias. Lagging is particularly important for accounting information, which is only released months after the accounting date.

[13]A double sort creates portfolios along two dimensions. We consider both dependent and independent double sorts on market equity, in addition to single sorts. In an independent sort, the primary sorting variable's quantiles are computed independent of market equity. In dependent double sorts, the main sorting variable's quantiles are computed conditional on market equity.

Figure 2: Flowchart of preprocessing choices. This flowchart illustrates the decision forks and corresponding choices encountered during a portfolio sort, as outlined in Walter et al. (2024). The size exclusion is at the 20$^{\text{th}}$ NYSE percentile as detailed in the main text. We consider sorting variable lags of three and six months (i.e., *3 m.* and *6 m.*), as well as the procedure outlined in Fama and French (1993) (i.e., *FF*). We rebalance the portfolios either monthly (i.e., *mon.*) or annually (i.e., *ann.*). We abbreviate the breakpoints with *BP* as used in referring to the quantiles for the main sorting variable, the secondary quantiles in the double sort on size, and the selection of stocks that are included in the computation of the breakpoints (where *NYSE* refers to stocks traded on the New York Stock Exchange). We either aggregate individual returns to portfolio returns by equal weighting (i.e., *EW*) or value weighting (i.e., *VW*).



double sorts, the set of exchanges incorporated in the computation of a sorting variable's breakpoints (i.e., quantiles), and the weighting scheme for aggregating individual returns. Three decisions are kept constant: we never exclude stocks with negative book equity, always implement a two-year stock-age exclusion, and include all stocks, regardless of their stock price, in line with findings by Walter et al. (2024), documenting that these decision forks have the smallest impact on risk factor returns.

A path i in Figure 2 constitutes all data preprocessing steps to construct a time series of returns $f^j_{k,t+1}$ for risk factor k. Sampling over these forks results in $n_k = 3,840$

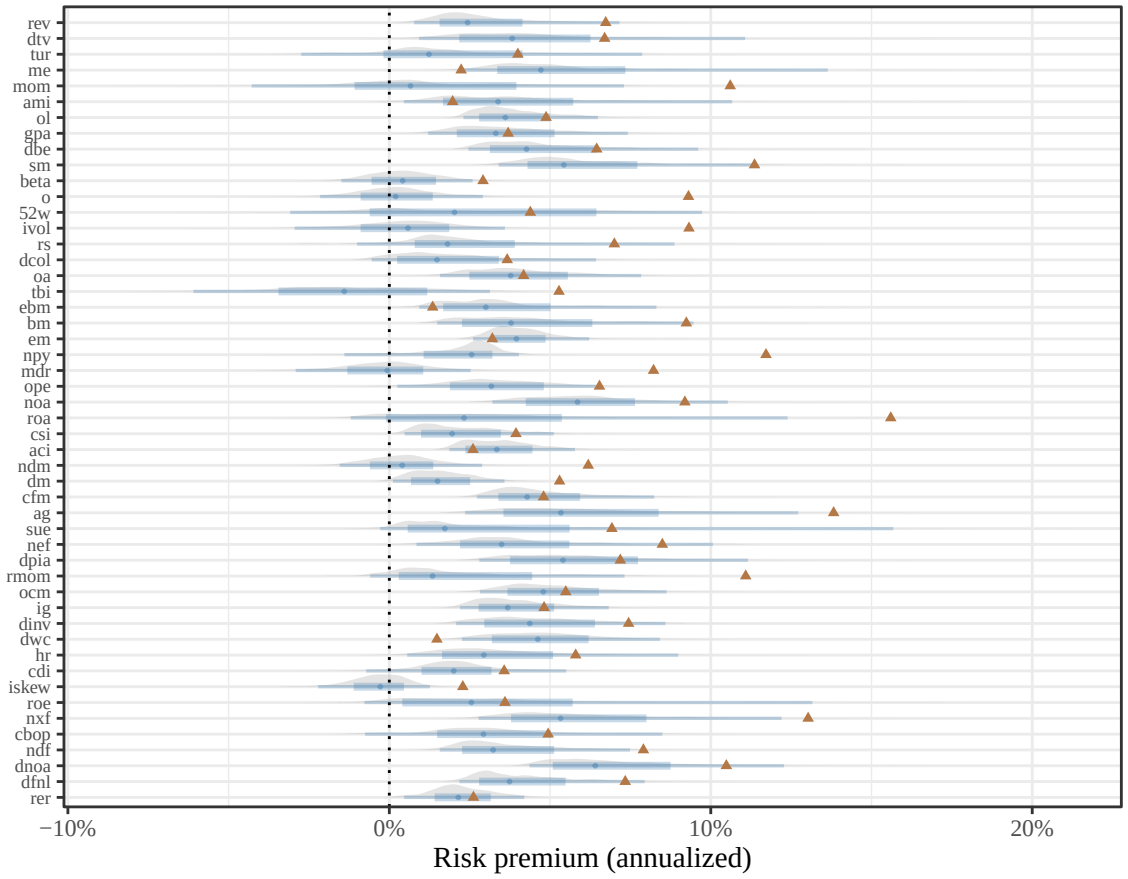
construction possibilities (paths) per sorting variable. Two sorting variables have fewer paths due to their construction. In particular, market equity (ME) features 768 paths, as market equity constitutes the secondary sorting variable such that double sorts for ME would be redundant. Earnings to market equity as a sorting variable only features $1,920$ paths as the varriable is only defined for positive earnings.

Before we turn to the estimation, we illustrate the variation in the long-short strategies induced by varying the preprocessing choices.[14] In particular, Figure 3 shows the distribution of annualized factor risk premia for each sorting variable as a ridge plot. Each observation is the annualized time-series average return of the long-short strategy. In line with Walter et al. (2024), we observe significant variation across specifications. Many variables even cross zero, implying that based on the same sorting variable, data preprocessing choices lead to sign switches of estimated risk premia. We also illustrate the risk premium estimates of the *published risk factors* (we obtain these estimates from Open Source Asset Pricing Chen and Zimmermann, 2022) that most closely align with our risk factor set as dots in the figure. While the majority of risk premia estimates are nested, clear positive outliers exist. The tendency of published risk factors to be in the right tail of possible constructions is in line with results from Hasler (2023).

While the average premia already indicate significant conceptual variation, the empirical estimation of asset pricing models is primarily concerned with risk factors' ability to span the stochastic discount factor. To this end, time-series correlations likely indicate the relevant variation induced by data preprocessing choices. Figure 4 shows the substantial variation within each sorting variable across data preprocessing choices. Most sorting variables have variants that show low time-series correlations with each other, some even being negatively correlated. To provide a further reference, we correlate our path-augmented strategies with the published risk factors and show the average of these correlations as brown dots. The variation around these averages documents that fixing choices eliminates a rich set of possibilities for identifying the SDF. Ultimately, these

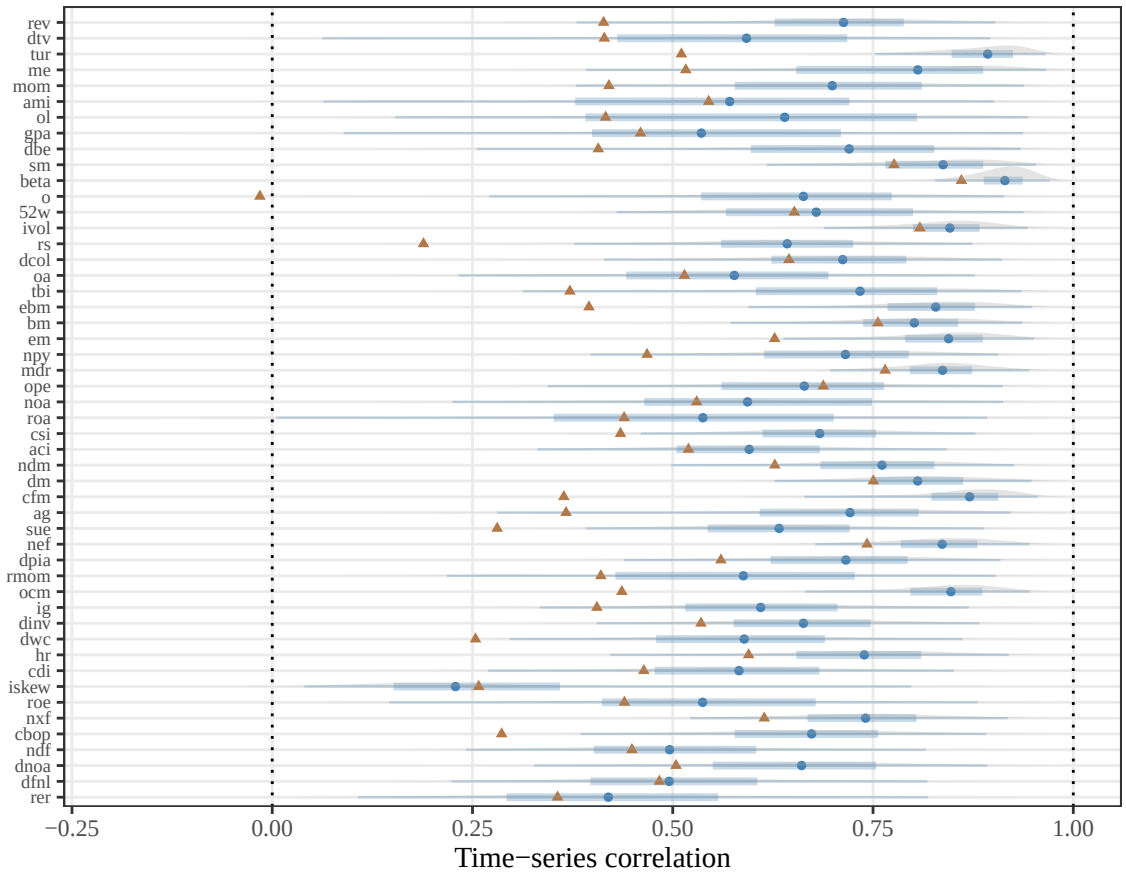
[14] We provide detailed summary statistics of all constructed risk factor returns in the Appendix, see Table A2.

Figure 3: Risk premia. The ridge plots indicate the distribution of annualized risk premia (i.e., the time-series average of long-short return differentials) in percent for a given sorting variable across all data preprocessing choices illustrated in Figure 2. The brown triangles indicate the risk premia of the published version of each risk factor. The estimates incorporate data from July 1973 until December 2022.



differences in risk premia and their time-series variation can alter the composition of the efficient frontier and, thus, the SDF due to conceptual uncertainty.

Figure 4: Time-series correlations. Average time-series correlations of long-short strategies arising from different preprocessing choices for the same sorting variable (blue dots) and underlying distribution thereof (blue bands). The brown triangles indicate the average correlation of our long-short time series with the published risk factor time series. The estimates incorporate data from July 1973 until December 2022.



4 Path-augmented SDF

In this section, we first build on our derivations in Section 2 by providing additional implementation details to estimate the stochastic discount factor (see Section 4.1). Equipped with our estimates, we discuss the posterior risk factor inclusion probabilities in two dimensions (see Section 4.2). On the one hand, we analyze which sorting variables are included in the path-augmented SDF. Here, we also compare our posterior inclusion probabilities to a non-path-augmented version of each risk factor, which we label published factors. On the other hand, we investigate which data preprocessing choices are selected for the included sorting variables.

Afterwards, we evaluate the performance of the path-augmented SDF in three tests. First, we conduct traditional cross-sectional asset pricing tests against several benchmark models (see Section 4.3). These benchmarks include two candidates implied by Bayesian model averaging without path augmentation. These two models with 106 published risk factor time series once include conditioning information (denoted *cond. SDF*) and once without conditioning variables (denoted *uncond. SDF*). The last benchmark is the static Fama and French (1993) three-factor model (denoted *FF3*). In a second evaluation of the path-augmented model, we assess the extent to which conceptual uncertainty contributes to total uncertainty (see Section 4.4). Finally, we demonstrate the impact of path augmentation on improving portfolio choice compared to the previously mentioned three benchmarks (see Section 4.5).

4.1 Estimating the path-augmented SDF

As outlined in Section 2, we estimate the path-augmented conditional stochastic discount factor (which we refer to as *PA cond. SDF*) defined in Equation (12). We have already discussed the construction of the comprehensive risk factor library with conceptual uncertainty in Section 3. We use these $K = 51$ potential risk factors $f_{k,t}^{j}$ in the estimation of the path-augmented SDF, i.e., the 50 risk factors created from varying the preprocessing choices alongside the market excess return, always included as a baseline risk factor. Ad-

ditionally, the estimation requires selecting conditioning variables z_{t-1}, test assets r_t, and the shrinkage parameter τ. As conditioning information, we select three macroeconomic predictors z_{t-1} from Goyal and Welch (2008): *dividend yield (dy)*, *earnings price ratio (ep)*, and *long-term yield (lty)*, which govern potential dynamics in mispricing and factor loadings. As test assets r_t, we use the $N = 106$ anomaly return time-series from Open Source Asset Pricing (Chen and Zimmermann, 2022) that are not already included in our set of $K = 51$ potential risk factors.[15] Finally, we set $\tau = 2$ in the spirit of Barillas and Shanken (2018) and Avramov et al. (2023). Our entire sample spans from July 1973 to December 2022, resulting in $T = 593$ monthly observations.[16]

We use a Markov chain Monte Carlo (MCMC) sampler to draw from the posterior distribution of the model parameters θ, Σ, δ, and γ_k.[17] We compute 30,000 draws from the posterior distribution from our MCMC-sampler for the path-augmented SDF and remove the first 2,000 observations as an appropriate burn-in period after careful convergence checks. To assess the robustness of our results, we generate the MCMC draws from seven independent chains. All results reported in the following sections are robust across the different chains, indicating convergence of the MCMC sampler.

In addition to the path-augmented SDF in Equation (12), we estimate two Bayesian model averaging benchmark models based on the 50 published risk factor time series. The benchmark model does not account for conceptual uncertainty in the sense that it fixes the data preprocessing choices and does not augment the model space with the selection parameter γ_k. We estimate two variants of the benchmark model: a conditional SDF that includes the same conditioning variables as the path-augmented model (denoted *cond. SDF*) and an unconditional SDF that does not include any conditioning variables (denoted *uncond. SDF*).

[15] For evaluating the path-augmented SDF in out-of-sample tests, we also consider alternative sets of test assets. We discuss these details in the respective sections.

[16] In our out-of-sample analyses, we restrict the estimation period to have a dedicated evaluation period. We discuss these details in the respective sections.

[17] We acknowledge that priors for some of the data preprocessing choices may have a justified preference. Acknowledging such associations with an informative prior structure could presumably strengthen the analysis even further.

4.2 Posterior inclusion probabilities

Given the MCMC draws from the posterior distribution of the path-augmented SDF, we can compute the marginal posterior inclusion probabilities δ for each risk factor and each data preprocessing choice γ_k. We investigate these posterior inclusion probabilities to understand which sorting variables and preprocessing choices are deemed relevant for pricing assets when accounting for model and conceptual uncertainty.

Figure 5 shows the marginal posterior inclusion probabilities $\delta|\mathcal{D}$ of the 50 sorting variables for the conditional model. The posterior inclusion probabilities are based on the entire sample period from July 1973 to December 2022. The blue dots indicate the posterior inclusion probabilities of the path-augmented factor model, while the brown dots indicate the inclusion probabilities for the benchmark model only based on published risk factors. Conditional models select on average 31 risk factors. *Long-term reversal, trading volume*, and *size* are included in virtually all models. The results are remarkebly stable across different values of τ. For $\tau = 1.5$, on average 32.8 risk factor are included, while for $\tau = 2.5$, on average 29.3 risk factors are included. As expected, an increase in τ leads to a more parsimonious model as the penalty for including additional risk factors increases. However, the overall pattern of inclusion probabilities remains similar across different values of τ. The average shrinkage intensity $\frac{T_0}{T_0+T} = 0.80$ indicates a strong penalty on pricing errors, which is in line with previous findings in the literature (see, e.g., Avramov et al., 2023).

Comparing the path-augmented model to the benchmark model, we observe that the path-augmented model tends to select a richer set of risk factors being considered relevant for pricing assets. However, the average number of included risk factors remains similar between the two models, suggesting that while different risk factors are deemed relevant, the overall model complexity does not increase substantially.

Similarly, Figure 6 shows the marginal posterior probabilities of the different data preprocessing choices for the conditional model. The figure illustrates that the most common

theme is to retain the original sample as much as possible: not excluding financials, utilities or small firms achieves high posterior probability mass. At the same time, the models have a tendency to rely on equal-weighted portfolio sorts.

Figure 5: Sorting variables' posterior inclusion probabilities. This figure shows the posterior inclusion probabilities of the 50 sorting variables for the path-augmented conditional model. As discussed above, the market factor is included by default. The blue dots indicate the posterior inclusion probabilities of the path-augmented risk factors, the brown triangles indicate those of the published risk factors, and the grey bars indicate their differences. The posterior incorporates the entire sample period from July 1973 until December 2022 and is estimated based on 106 anomaly portfolios as test assets.

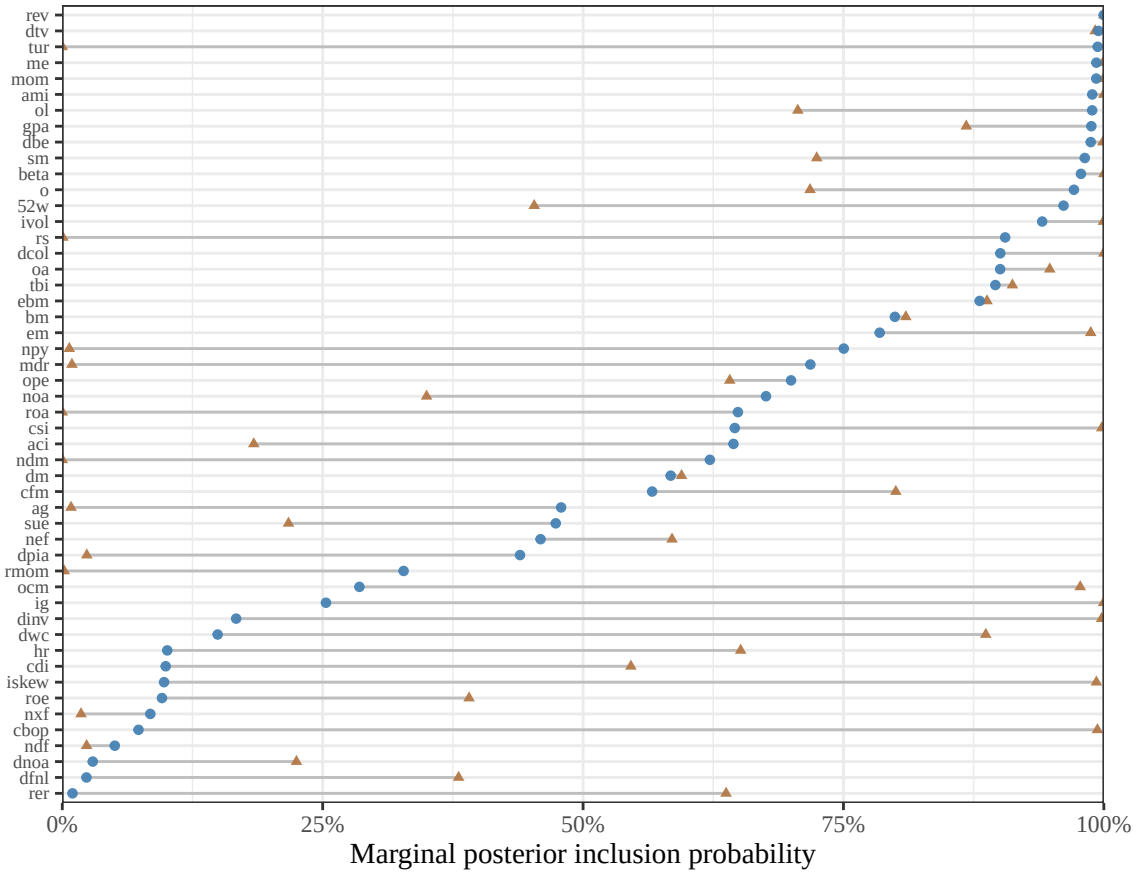
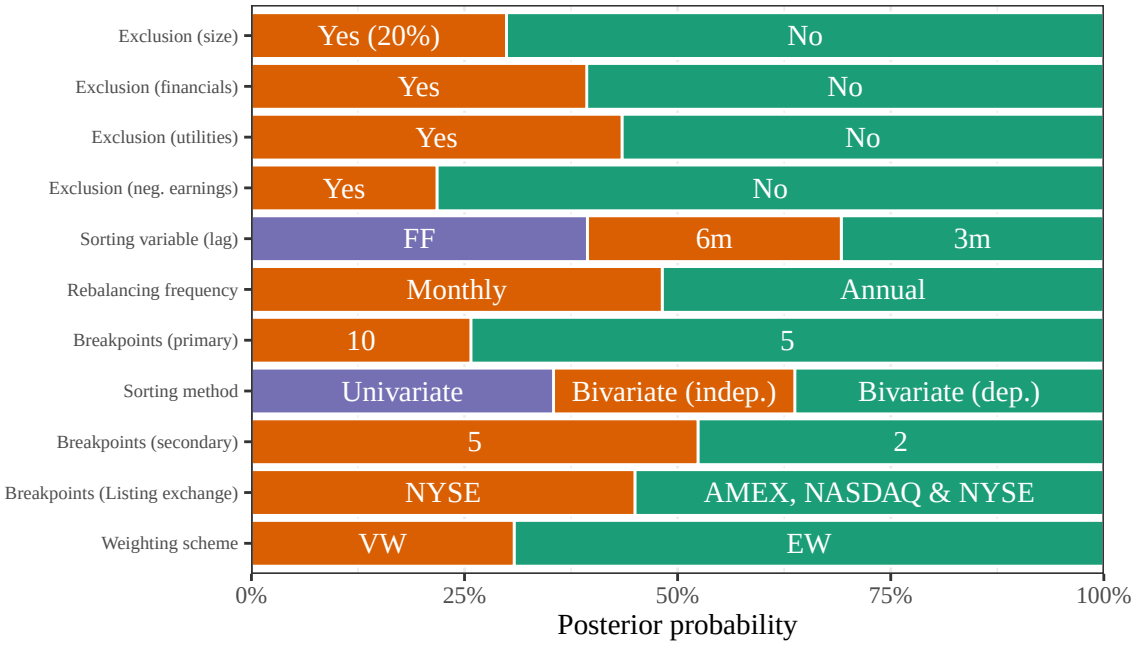


Figure 6: Posterior data preprocessing choices. This figure shows the marginal posterior probabilities for different choices at the data preprocessing forks introduced in Figure 2. These probabilities arise from our path-augmented BMA SDF with conditioning infromation with 50 risk factors (i.e., excluding the market factor) using 106 anomaly portfolios as test assets. The posterior incorporates the entire sample period from July 1973 until December 2022.



4.3 Cross-sectional performance tests

A key challenge in empirical asset pricing is whether models with good in-sample performance can generate accurate out-of-sample predictions. In this section, we conduct traditional cross-sectional asset pricing tests to assess the performance of our path-augmented BMA SDF with conditioning information. The performance of a candidate model is its ability to generate predictions that closely match the realized returns of the considered test assets.

All model performance tests start from the respective model's expected return. Conditional on a model M_p, its risk factor preprocessing path Q_{pj}, and the data input $\mathcal{D}$, the model's expected posterior predictive return is given by

$$\mathbb{E}(r_t|M_p, Q_p j, D) = w_t \theta^{\text{post}}, \tag{13}$$

where the vector w_t collects the right-hand side variables (i.e., it is the t-th row of W) and θ^{post} (see Equation 5) is the model's posterior mean of the parameter θ. Recall Equation (3), which is the stacked version of Equation (2). In terms of a conditional market model, w_t contains an intercept, the expected market risk premium, the conditioning information z_t, and all interactions with z_t. Then, θ^{post} contains the α and β estimated from our data $\mathcal{D}$.

In this paper, we introduce the path-augmented BMA SDF and evaluate its performance later in this section. The posterior expected return of the path-augmented BMA SDF is given by

$$\mathbb{E}(r_t|\mathcal{D}) = \sum_{p=1}^{P} \sum_{j=1}^{q_p} \pi(M_p, Q_{pj}|\mathcal{D}) \mathbb{E}(r_t|M_p, Q_{pj}, \mathcal{D}). \tag{14}$$

Equation (14) resembles an intuive weighted average of model-implied conditional posterior predictive returns of Equation (13) introduced above. The first component $\pi(M_p, Q_{pj}|\mathcal{D})$ is the posterior probability of a model and the construction choices of the model's factors. We introduced this probability in Equation (11) above. The second component is the

expected return implied by each specific model's posterior distribution. Summing up the product of both components over models and data preprocessing choices gives the expected return of the path-augmented BMA SDF. Specifically, we sum over j from 1 to q_p, which explicitly accounts for the q_p different data preprocessing paths considered for each model M_p, before we sum over the P models.

Next, we will discuss the setup of our cross-sectional tests. In Section 4.3.1, we outline the various scenarios used to assess out-of-sample performance and benchmark models. Then, we show the actual performance of the path-augmented BMA SDF and its competitors in Section 4.3.2.

4.3.1 Performance test setup

Our cross-sectional tests are out-of-sample in two dimensions. Temporally, we separate the estimation window from the evaluation period, and, cross-sectionally, we estimate posterior probabilities using 106 anomaly portfolios but evaluate performance on distinct asset sets. Additionally, we test the performance of several benchmark models using two evaluation metrics. All of these ingredients in our tests are discussed below.

Out-of-sample periods. Overfitting is a significant concern with all models that offer considerable flexibility. Given our large model space of our path-augmented BMA SDF, many parameters have to be estimated which raises overfitting concerns. Hence, we carefully split our estimation from the evaluation period. All our cross-sectional tests only report out-of-sample evidence. We denote the split between the periods by the length of the estimation window. We either use 50% of the sample period (denoted *T/2*) or two-thirds (denoted *2/3T*). Specifically, our full sample spans from July 1973 to December 2022, with the T/2- and 2/3T-cutoffs being January 1998 and 2006, respectively. All models' parameters are estimated over these periods and evaluated out of sample.

Out-of-sample assets. Our second out-of-sample dimension is the set of test assets. For all SDFs implied by Bayesian model averaging, the estimation relies on the 106 anomaly returns as the priced assets. The posterior probabilities $\pi(M_p, Q_{pj}|\mathcal{D})$ in Equation (14)

are estimated from these test assets. However, a useful SDF has to demonstrate its pricing abilities across all assets, i.e., we also seek external validation of the SDF. Hence, we evaluate each model's performance with two additional groups of assets not used to generate the posterior probabilities. In the additional tests, we use 49 industry portfolios and 25 size-and-book-to-market portfolios from Kenneth French's homepage.

Benchmark models. We evaluate our path-augmented BMA SDF with conditioning information (*PA Cond. SDF*) against several competitor models. First, we remove the path augmentation by only considering $n_k = 1$ possible ways of constructing the risk factors. We call this non-path-augmented version using only the published versions of each risk factor factors the conditional BMA SDF (*Cond. SDF*). Then, we take the published risk factors to estimate an unconditional version of the BMA SDF (*Uncond. SDF*). An unconditional model is achieved by replacing z_t with a vector of ones such that mispricing and vactor loadings are constant. Finally, we also employ the classical Fama and French (1993) three-factor model (*FF3*). While only three risk factors populate the three-factor model, all BMA-implied models consider the aforementioned 51 risk factors.

To be precise, we always use the 106 anomaly portfolios as test assets over the indicated estimation period to derive the posterior probabilitiy of the three models implied by BMA. Then, we use the estimation window to calibrate the factor exposures of the test assets used for evaluation. Our two conditional models have time-varying expected returns for each asset due to the interaction of factor exposures with conditioning variables. Finally, we evaluate the expected return predictions of the candidate model using the indicated evaluation period and test assets. This progression of benchmarks allows us to isolate the marginal contribution of (1) Bayesian model averaging versus a static factor models, (2) time-varying expected returns via conditioning information, and (3) path augmentation across data preprocessing choices.

Evaluation metrics. We evaluate each model's performance with two metrics. On the one hand, we compute the relative root mean square error (denoted *rel RMSE*), comparing the model's RMSE to a naive extrapolation of the assets' historical average returns. On

the other hand, we compute the out-of-sample R^2. We aggregate the respective measures across the test assets using the arithmetic average.

4.3.2 Performance test results

The design developed in Section 4.3.1 allows us to rigorously assess whether path augmentation provides genuine out-of-sample improvements. Having implemented all tests, Table 1 shows the results of these out-of-sample performance evaluations.

The headline result is that, indeed, integrating conceptual uncertainty provides better out-of-sample return predictions. Across the two estimation-evaluation splits and all three test asset groups, both performance metrics uniformly agree that the path-augmented BMA SDF with conditioning information (i.e., *PA cond. SDF*) outperforms all benchmarks. We consistently find the lowest root mean squared errors and the highest out-of-sample R^2 for the path-augmented BMA SDF. Thus, allowing for variation in data preprocessing choices improves the pricing abilities from SDFs implied by Bayesian model averaging relative to using only the single path resulting in the published version of each risk factor.

The performance improvements captured by path augmentation are comparable to the jump introduced by the state-of-the-art Bayesian model benchmarks. The sparse three-factor model of Fama and French (1993) (with and without conditioning information) consistently underperforms the richer BMA models. These findings align with prior results from Avramov et al. (2023) and Bryzgalova et al. (2023), who document that BMA-implied SDFs outperform static factor models in out-of-sample tests.

We also compare models with and without conditioning information for all models. First, we confirm prior evidence from Avramov et al. (2023) that allowing for conditioning information improves pricing performance relative to BMA-implied unconditional models. Second, we also study the role of conditioning information in the context of path-augmented BMA SDFs. It is ex ante unclear whether previous results would extend to the much richer set of risk factors resulting from varying preprocessing choices. Here, we find that the path-augmented conditional BMA SDF outperforms its unconditional version, i.e.,

PA cond. SDF outperforms *PA uncond. SDF* across all tests. Hence, we conclude that conditioning information remains relevant even when integrating conceptual uncertainty into the SDF.

Overall, the BMA-implied models suggested by the literature provided a new best practice. However, these models take an extreme stance on risk factor construction by fixing all data preprocessing choices. The results from Table 1 show that fixing these choices is not optimal. Allowing for conceptual uncertainty in the data preprocessing choices to construct risk factors yields superior cross-sectional pricing performance.

Table 1: Model performance: Pricing. This table presents out-of-sample model performance tests for our path-augmented conditional model (PA cond. SDF) and three benchmark models. These benchmarks include BMA-implied conditional (Cond. SDF) and unconditional models (Uncond. SDF) with static factors, alongside the classical Fama and French (1993) three-factor model (FF3). As shown in the two panels, we use two time splits between the estimation and evaluation period. Our sample spans from July 1973 to December 2022, with the 1/2- and 2/3-cutoffs being January 1998 and 2006, respectively. We use three sets of test assets: 106 anomaly returns (used to estimate the BMA models), 49 industry portfolios, and 25 size and book-to-market portfolios. The test statistics are the relative root mean squared error (rel RMSE), which subtracts a naive extrapolation of historical average returns from each model's RMSE, and the out-of-sample R^2_{OOS} (in percent).

	1/2 T		2/3 T	
	rel RMSE	R^2_{OOS}	rel RMSE	R^2_{OOS}
Anomaly returns				
cond. PA SDF (BMA)	**4.603**	**0.308**	**3.984**	**0.808**
PA SDF (BMA)	4.604	0.269	3.988	0.628
cond. SDF (BMA)	4.605	0.222	3.988	0.616
SDF (BMA)	4.605	0.222	3.989	0.554
FF3	4.610	0.000	4.001	0.000
CAPM	4.610	0.000	4.001	0.000
Industry portfolios				
cond. PA SDF (BMA)	**7.128**	**0.236**	**6.997**	**0.171**
PA SDF (BMA)	7.128	0.222	6.999	0.123
cond. SDF (BMA)	7.131	0.137	6.998	0.153
SDF (BMA)	7.132	0.132	6.998	0.135
FF3	7.136	0.000	7.003	0.000
CAPM	7.136	0.000	7.003	0.000
Size and value sorts				
cond. PA SDF (BMA)	**6.156**	**0.494**	**6.008**	**0.537**
PA SDF (BMA)	6.157	0.442	6.012	0.411
cond. SDF (BMA)	6.162	0.273	6.013	0.350
SDF (BMA)	6.163	0.267	6.015	0.312
FF3	6.171	0.000	6.024	0.000
CAPM	6.171	0.000	6.024	0.000

4.4 Variance decomposition

Our asset pricing framework accounts for two distinct sources of uncertainty: model uncertainty and conceptual uncertainty. We next investigate how these two sources contribute to the overall predictive investment risk. First, we denote the posterior predictive variance for the test asset returns r_t and given model M_p with data preprocessing choices Q_{pj} implied by Model (12) as $\mathbb{V}\text{ar}\left(r_t|M_p, Q_{pj}\right)$.

Then, the posterior predictive distribution of the test asset returns r_t resulting from the path-augmented BMA SDF is given by integrating across the space of all models M_p and data preprocessing choices Q_{pj} such that

$$\mathbb{V}\text{ar}_{\text{BMA}}(r_t) = \overbrace{\mathbb{E}(\mathbb{V}\text{ar}(r_t|\mathcal{M}, \mathcal{Q}))}^{\text{Average model specific risk}} + \mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q})) \tag{15}$$

The average model specific risk $\mathbb{E}(\mathbb{V}\text{ar}(r_t|\mathcal{M}, \mathcal{Q})) = \sum_{p=1}^{P}\sum_{j=1}^{q_p}\pi(M_p, Q_{pj}|\mathcal{D})\mathbb{V}\text{ar}(r_t|M_p, Q_{pj})$ captures the average conditional variance of future returns given a specific model and data preprocessing path. Model specific risk naturally is of substantial magnitude in financial return prediction due to the inherent stochasticity of asset returns and limited predictability.

The second term in Equation (15), $\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q}))$, captures the uncertainty about the conditinal expected returns induced by model disagreement (MD) and conceptual uncertainty (CU). The variation is driven by differences in the conditional expected returns $\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q})$ across models and data preprocessing paths:

$$\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q})) = \underbrace{\mathbb{V}\text{ar}(\mathbb{E}(\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q})|\mathcal{M}))}_{\text{Model disagreement (MD)}} + \underbrace{\mathbb{E}(\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M}, \mathcal{Q})|\mathcal{M}))}_{\text{Conceptual unvertainty (CU)}}, \tag{16}$$

Model disagreement quantifies the variation in expected returns due to model uncertainty after integrating out variation from data preprocessing steps. Specifically, the model

disagreement term is given by

$$\begin{aligned} MD &= \mathbb{V}\text{ar}(\mathbb{E}(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})|\mathcal{M})) = \mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M})) \\ &= \sum_{p=1}^{P} \pi(M_p|\mathcal{D}) \left((\mathbb{E}(r_t|M_p) - \mathbb{E}(r_t))(\mathbb{E}(r_t|M_p) - \mathbb{E}(r_t))'\right), \end{aligned} \tag{17}$$

where $\mathbb{E}(r_t|M_p) = \sum_{j=1}^{q_p} \frac{\pi(M_p,Q_{pj}|\mathcal{D})}{\pi(M_p|\mathcal{D})} \mathbb{E}(r_t|M_p,Q_{pj})$ is the conditional expected return implied by model M_p after integrating across all data preprocessing choices Q_p, $\mathbb{E}(r_t) = \sum_{p=1}^{P} \pi(M_p|\mathcal{D})\mathbb{E}(r_t|M_p)$ is the posterior preditcive expected return, and $\pi(M_p|\mathcal{D}) = \sum_{j=1}^{q_p} \pi(M_p,Q_{pj}|\mathcal{D})$ is the posterior model probability of model M_p. Intuitively speaking, MD measures how much the conditional expected returns differ across models. A high value of MD indicates that different models imply substantially different expected returns, thus indictaing increased predictive investment risk.

The final term in Equation (16) captures conceptual uncertainty, i.e., the variation in conditional expected returns due to different data preprocessing paths *within* a given model. CU is given by

$$CU = \mathbb{E}(\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})|\mathcal{M})) = \sum_{p=1}^{P} \pi(M_p|\mathcal{D})\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})|M_p), \tag{18}$$

where

$$\begin{aligned} \mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})|M_p) &= \mathbb{E}\left(\left(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q}) - \mathbb{E}(r_t|M_p)|M_p\right)\left(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q}) - \mathbb{E}(r_t|M_p)|M_p\right)'\right) \\ &= \sum_{j=1}^{q_p} \frac{\pi(M_p,Q_{pj}|\mathcal{D})}{\pi(M_p|\mathcal{D})} \left((\mathbb{E}(r_t|M_p,Q_{pj}) - \mathbb{E}(r_t|M_p))(\mathbb{E}(r_t|M_p,Q_{pj}) - \mathbb{E}(r_t|M_p))'\right) \end{aligned} \tag{19}$$

Substantial variation due to different preprocessing choices implies high uncertainty with respect to the actual predicted expected return. Integrating out conceptual uncertainty therefore increases the implied predictive investment risk. We derive analytical representations of the predictive moments $\mathbb{E}(r_t|M_p,Q_{pj})$ and $\mathbb{V}\text{ar}(r_t|M_p,Q_{pj})$ in the Appendix.

Table 2: Variance decomposition. This table shows the decomposition of the variance in the predictive return distribution. We evaluate the variance decomposition for the entire sample period, from July 1973 to December 2022, for the path-augmented BMA SDF with conditioning information. Each number represents the share of variance associated with each component, i.e., model disagreement and conceptual uncertainty.

Size and value sorts	Industry portfolios	Anomaly returns
39.16	38.35	37.11

To quantify the relevance of the individual components in our sample, we compute $\frac{tr(CU)}{(Var(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})))} = \frac{\sum_{i=1}^{N}\lambda_i(CU)}{\sum_{i=1}^{N}\lambda_i(\mathbb{V}\text{ar}(\mathbb{E}(r_t|\mathcal{M},\mathcal{Q})))}$, the proportion of total variance contributed by conceptual uncertainty spread across all principal components for our full sample and conditional on the vector of macroeconomics predictors $z_t = \bar{z}$, the time-series average as a representative economic scenario. To compute the conditional posterior predictive moments, we use the MCMC output from our estimation procedure, thus effectively only considering the models visited by MCMC and implicitly assigning a zero posterior model probability to models and data preprocessing paths not visited by the sampler.

The results are shown in Table 2. We find that conceptual uncertainty contributes, on average, 46% to the total conditional expected return variation of the predictive return distribution. This indicates that the choice of data preprocessing steps is a relevant source of uncertainty in the predictive return distribution.

4.5 Portfolio choice

In the previous section, we document that conceptual uncertainty stemming from the many possible preprocessing choices is a sizable component of total uncertainty. Prior work shows that ignoring uncertainty leads to poorer portfolio choices when ignoring model uncertainty. In this part, we investigate whether ignoring conceptual uncertainty yields worse investment outcomes.

We derive mean-variance efficient tangency and global minimum variance portfolios for our suggested path-augmented BMA SDF with conditioning information (PA cond. SDF).

We exclusively focus on annualized Sharpe ratios as our measure of model performance. In addition to the benchmarks introduced above, we assess the impact of path augmentation by comparing the PA SDF against the classical capital asset pricing model (CAPM). We provide evidence for in- and out-of-sample periods for different splits between estimation and evaluation periods.

Sharpe ratio results are presented in Table 3. Out-of-sample, the path augmentation leads to significantly better performance for the tangency portfolio, irrespective of the time split. The roughly twofold improvement is economically impactful and shows that ignoring conceptual uncertainty hurts investors. In relation to the previous section, the contribution of conceptual uncertainty to the total uncertainty is a relevant driver of model performance. We also find that path augmentation leads to significantly better performances for the global minimum variance portfolio's Sharpe ratios. In particular, the overfitting-sensitive out-of-sample performance shows a clear preference for the path-augmented BMA SDF with conditioning information.

Table 3: Model performance: Portfolio choice. Annualized Sharpe ratios for the mean-variance efficient tangency portfolio and the global minimum variance portfolio for different splits between in- and out-of-sample evidence. In particular, the full sample spans July 1973 until December 2022 (see column T), with the 1/2- (columns 1/2T) and 2/3-cutoffs (columns 2/3T) being January 1998 and 2006, respectively. EST corresponds to the in-sample period Sharpe ratio, and OOS to the out-of-sample period Sharpe ratio. Out-of-sample Sharpe ratios are computed based on the posterior predictive return distributions stemming from different factor models.

	1/2T		**2/3T**		**T**
	EST	OOS	EST	OOS	EST
Efficient tangency portfolio					
PA cond. SDF	8.600	4.319	7.856	3.765	5.743
Cond. SDF	5.870	1.881	5.327	1.364	3.722
CAPM	0.477	0.436	0.405	0.553	0.456
FF3	1.442	0.727	1.399	0.526	1.075
Global minimum variance portfolio					
PA cond. SDF	3.038	1.942	3.185	1.284	2.430
Cond. SDF	3.415	1.256	2.895	0.589	1.997
FF3	1.216	0.703	1.309	0.445	0.952

5 Conclusion

Even *if* the set of risk factors populating the stochastic discount factor were known, there would be no clear mapping from such a stochastic discount factor as a scientific concept to a feasible implementation. Instead, data preprocessing, i.e., transforming raw data into suitable proxies for any risk factor via portfolio sorts as a nonparametric method, always entails arbitrary yet defensible choices. As a result, conceptual uncertainty may harm financial decisions, assessments of investment risks, or a proper understanding of the risk-return trade-off.

In this paper, we propose a Bayesian model averaging procedure for conditional linear stochastic discount factor models by explicitly accounting for conceptual uncertainty. We derive a feasible MCMC sampling scheme to identify relevant data preprocessing choices and to estimate the path-augmented stochastic discount factor. We show empirically that the shape of the efficient frontier depends substantially on a narrow set of generally plausible data preprocessing choices for portfolio sorts. Thus, the question is no longer: which risk factors matter? Rather, we must also take a stance on the question: how should such risk factors be constructed? Instead of fixing the data preprocessing choices, we allow for a range of plausible choices and estimate the path-augmented stochastic discount factor. The resulting set of plausible stochastic discount factor representations is vast and orders of magnitude larger than the number of observable atoms in the universe. The results are statistically and economically significant. We find that the path-augmented stochastic discount factor outperforms the state-of-the-art Bayesian model averaging procedure regarding cross-sectional asset pricing performance. Additionally, we find that conceptual uncertainty contributes substantially to the total uncertainty in the predictive return distribution. Finally, we show that the path-augmented stochastic discount factor leads to improved portfolio choice decisions.

Thus, the message of this paper is clear: instead of arbitrarily discarding alternative data preprocessing choices, conceptual uncertainty is a relevant determinant and should be incorporated into the financial decision-making process.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1):259–299.

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

Avramov, D., Cheng, S., Metzker, L., and Voigt, S. (2023). Integrating factor models. *The Journal of Finance*, 78(3):1593–1646.

Avramov, D. and Chordia, T. (2006). Predicting stock returns. *Journal of Financial Economics*, 82(2):387–415.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680.

Balakrishnan, K., Bartov, E., and Faurel, L. (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics*, 50(1):20–41.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, 99(2):427–446.

Bali, T. G., Engle, R. F., and Murray, S. (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Ball, R., Gerakos, J., Linnainmaa, J. T., and Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics*, 121(1):28–45.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18.

Barbee Jr, W. C., Mukherji, S., and Raines, G. A. (1996). Do sales–price and debt–equity

explain stock returns better than book–market and firm size? *Financial Analysts Journal*, 52(2):56–60.

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

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics*, 12(1):129–156.

Belo, F. and Lin, X. (2012). The inventory growth spread. *The Review of Financial Studies*, 25(1):278–313.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43(2):507–528.

Blitz, D., Huij, J., and Martens, M. (2011). Residual momentum. *Journal of Empirical Finance*, 18(3):506–521.

Boudoukh, J., Michaely, R., Richardson, M., and Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance*, 62(2):877–915.

Brennan, M. J., Chordia, T., and Subrahmanyam, A. (1998). Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics*, 49(3):345–373.

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

Cartea, A., Jin, Q., and Shi, Y. (2025). The limited virtue of complexity in a noisy world. *Available at SSRN*.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Schaumburg, E. (2020). Characteristic-sorted portfolios: Estimation and inference. *The Review of Economics and Statistics*, 102(3):531–551.

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

Chen, M., Hanauer, M. X., and Kalsbach, T. (2024). Design choices, machine learning, and the cross-section of stock returns. *Available at SSRN*.

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

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

Cooper, M. J., Gulen, H., and Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance*, 63(4):1609–1651.

Daniel, K. and Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, 61(4):1605–1643.

Datar, V. T., Naik, N. Y., and Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, 1(2):203–219.

Davis, J. L., Fama, E. F., and French, K. R. (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance*, 55(1):389–406.

De Bondt, W. F. and Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, 40(3):793–805.

Desai, H., Rajgopal, S., and Venkatachalam, M. (2004). Value-glamour and accruals mispricing: One anomaly or two? *The Accounting Review*, 79(2):355–385.

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk? *The Journal of Finance*, 53(3):1131–1147.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97.

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

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

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84.

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

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Ferson, W. E. and Harvey, C. R. (1999). Conditioning variables and the cross section of stock returns. *The Journal of Finance*, 54(4):1325–1360.

Foster, G., Olsen, C., and Shevlin, T. (1984). Earnings releases, anomalies, and the behavior of security returns. *The Accounting Review*, 59(4):574–603.

Gao, X. and Ritter, J. R. (2010). The marketing of seasoned equity offerings. *Journal of Financial Economics*, 97(1):33–52.

George, T. J. and Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, 59(5):2145–2176.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2):436–451.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, page 1121–1152.

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

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

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

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

Harvey, C. R., Liu, Y., and Zhu, H. (2015). … and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hasler, M. (2023). Looking under the hood of data-mining. *Available at SSRN 4279944*.

Hasler, M. (2025). Is the value premium smaller than we thought? *Critical Finance Review (forthcoming)*.

Hellum, O., Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2025). The power of the common task framework. *Available at SSRN 5242901*.

Hirshleifer, D., Hou, K., Teoh, S. H., and Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, 38:297–331.

Hou, K., Xue, C., and Zhang, L. (2014). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Hribar, P. and Collins, D. W. (2002). Errors in estimating accruals: Implications for empirical research. *Journal of Accounting Research*, 40(1):105–134.

Jagannathan, R. and Wang, Z. (1996). The conditional capm and the cross-section of expected returns. *The Journal of Finance*, 51(1):3–53.

Jegadeesh, N. and Livnat, J. (2006). Revenue surprises and stock returns. *Journal of Accounting and Economics*, 41(1-2):147–171.

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

Kelly, B., Malamud, S., and Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503.

Kessler, S., Scherer, B., and Harries, J. P. (2020). Value by design? *Journal of Portfolio Management*, 46(2):25–43.

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

Lakonishok, J., Shleifer, A., and Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance*, 49(5):1541–1578.

Lev, B. and Nissim, D. (2004). Taxable income, future earnings, and equity values. *The Accounting Review*, 79(4):1039–1074.

Lyandres, E., Sun, L., and Zhang, L. (2008). The new issues puzzle: Testing the investment-based explanation. *The Review of Financial Studies*, 21(6):2825–2855.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johanneson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., Abudy, M. M., and Others (2024). Non-standard errors. *The Journal of Finance*, 79(3):2339–2390.

Mitton, T. (2022). Methodological variation in empirical corporate finance. *The Review of Financial Studies*, 35(2):527–575.

Novy-Marx, R. (2011). Operating leverage. *Review of Finance*, 15(1):103–134.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1–28.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Pastor, L. and Stambaugh, R. F. (1999). Costs of equity capital and model mispricing. *The Journal of Finance*, 54(1):67–121.

Penman, S. H., Richardson, S. A., and Tuna, I. (2007). The book-to-price effect in stock returns: accounting for leverage. *Journal of Accounting Research*, 45(2):427–467.

Richardson, S. A., Sloan, R. G., Soliman, M. T., and Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics*, 39(3):437–485.

Roll, R. (1977). A critique of the asset pricing theory's tests part i: On past and potential testability of the theory. *Journal of Financial Economics*, 4(2):129–176.

Scheuch, C., Voigt, S., and Weiss, P. (2023). *Tidy Finance with R*. Chapman and Hall/CRC, 1st edition. https://tidy-finance.org/r/.

Scheuch, C., Voigt, S., Weiss, P., and Frey, C. (2024). *Tidy Finance with Python*. Chapman and Hall/CRC, 1st edition. https://tidy-finance.org/python.

Schwarz, P., Walter, D., and Weiss, P. (2025). Rewriting crsp's history: Impact of altered monthly returns on asset pricing. *Available at SSRN 5074864*.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3):289–315.

Soebhag, A., Van Vliet, B., and Verwijmeren, P. (2024). Non-standard errors in asset pricing: Mind your sorts. *Journal of Empirical Finance*, 78:101517.

Stambaugh, R. F. (1982). On the exclusion of assets from tests of the two-parameter model: A sensitivity analysis. *Journal of Financial Economics*, 10(3):237–268.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712.

Thomas, J. K. and Zhang, H. (2002). Inventory changes and future returns. *Review of Accounting Studies*, 7(2):163–187.

Titman, S., Wei, K. J., and Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, 39(4):677–700.

Walter, D., Weber, R., and Weiss, P. (2024). Methodological uncertainty in portfolio sorts. *Available at SSRN 4164117*.

Xing, Y. (2008). Interpreting the value effect through the q-theory: An empirical investigation. *The Review of Financial Studies*, 21(4):1767–1795.

Appendix

A Bayesian asset pricing model

A.1 Posterior distribution

Consider the Bayesian multivariate linear regression $R = W\theta + E$ where $R \in \mathbb{R}^{T\times N}$ is the matrix of observed test asset returns, $W \in \mathbb{R}^{T\times P}$ is the design matrix with $P = (K+1)(L+1)$, $\theta \in \mathbb{R}^{P\times N}$ is the matrix of regression coefficients, and $E \in \mathbb{R}^{T\times N}$ is the matrix of error terms. The rows of E are assumed to be independently and identically distributed as multivariate normal with mean zero and covariance matrix Σ. Then, the likelihood of the data $\mathcal{D}$ is

$$\mathcal{L}(\mathcal{D}|\theta) = (2\pi)^{-TN/2} \det(\Sigma)^{-T/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(R - W\theta)^{\top}(R - W\theta)\right)\right). \tag{20}$$

The canonical conjugate prior for (θ, Σ) is the Normal-Inverse-Wishart distribution where

$$\Sigma \sim \mathcal{IW}\left(\nu^{\text{prior}}, S^{\text{prior}}\right) \text{ and } \mathrm{vec}(\theta)\,|\Sigma \sim \mathcal{N}(\mathrm{vec}\left(\theta^{\text{prior}}\right), \Sigma \otimes {\Lambda^{\text{prior}}}^{-1}) \tag{21}$$

with ν^{prior} degrees of freedom, $N \times N$ scale matrix S^{prior}, prior mean θ^{prior}, and $P \times P$ prior covariance matrix ${\Lambda^{\text{prior}}}^{-1}$. The corresponding prior probability density functions are given by

$$\pi(\theta|\Sigma) \propto \det(\Sigma)^{-P} \det\left(\Lambda^{\text{prior}}\right)^{N} \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(\theta - \theta^{\text{prior}}\right)'\Lambda^{\text{prior}}\left(\theta - \theta^{\text{prior}}\right)\right)\right)$$

$$\pi(\Sigma) = \frac{\det\left(S^{\text{prior}}\right)^{\nu^{\text{prior}}/2}}{2^{\nu^{\text{prior}}N/2}\Gamma_N\left(\frac{\nu^{\text{prior}}}{2}\right)} \det(\Sigma)^{-(\nu^{\text{prior}}+N+1)/2} \exp\left(-\tfrac{1}{2}\,\mathrm{tr}\left(\Sigma^{-1}S^{\text{prior}}\right)\right)$$

We obtain the joint posterior distribution of (θ, Σ) and the data $\mathcal{D}$ by combining the

likelihood in Equation (20) with the prior distributions in Equation (21)

$$\begin{aligned}\pi(\theta, \Sigma|\mathcal{D}) &= \mathcal{L}\left(\mathcal{D}|\theta, \Sigma\right) \pi(\theta|\Sigma)\pi(\Sigma) \\ &\propto |\Sigma|^{-\frac{T+P+\nu^{\text{prior}}+N+1}{2}} \exp\left(-\tfrac{1}{2}\text{tr}\left(\Sigma^{-1}Z\right)\right)\end{aligned} \tag{22}$$

where

$$\begin{aligned}Z &= (R - W\theta)'(R - W\theta) + (\theta - \theta^{\text{prior}})'\Lambda^{\text{prior}}(\theta - \theta^{\text{prior}}) + S^{\text{prior}}. \\ &= \theta' A\theta - 2\theta' B + S^{\text{prior}} + R'R + \theta'_{\text{prior}}\Lambda^{\text{prior}}\theta_{\text{prior}}\end{aligned} \tag{23}$$

with $A = W'W + \Lambda^{\text{prior}}$ and $B = W'R + \Lambda^{\text{prior}}\theta_{\text{prior}}$. To obtain a more explicit representation of the posterior distribution, we next complete the square in Z by making use of the standard result that for conformable matrices A and B it holds that $\theta' A\theta - 2\theta' B = (\theta - A^{-1}B)'A(\theta - A^{-1}B) - B'A^{-1}B$. Define $\Lambda^{\text{post}} = A^{-1} = \left(W'W + \Lambda^{\text{prior}}\right)^{-1}$, $\theta^{\text{post}} = A^{-1}B = {\Lambda^{\text{post}}}^{-1}\left(W'R + \Lambda^{\text{prior}}\theta^{\text{prior}}\right)$ and $S^{\text{post}} = S^{\text{prior}} + R'R + {\theta^{\text{prior}}}'\Lambda^{\text{prior}}\theta^{\text{prior}} - {\theta^{\text{post}}}'\Lambda^{\text{post}}\theta^{\text{post}}$ to get

$$Z = S^{\text{post}} + (\theta - \theta^{\text{post}})'\Lambda^{\text{post}}(\theta - \theta^{\text{post}}) \tag{24}$$

Combining all terms from Equation (22) and defining $\nu^{\text{post}} = \nu^{\text{prior}} + T$ yields

$$\pi(\theta, \Sigma|\mathcal{D}) \propto \det\left(\Sigma\right)^{-\frac{\nu^{\text{post}}+P+N+1}{2}} \exp\left(-\tfrac{1}{2}\text{tr}\left(\Sigma^{-1}\left(S^{\text{post}} + (\theta - \theta^{\text{post}})'\Lambda^{\text{post}}(\theta - \theta^{\text{post}})\right)\right)\right) \tag{25}$$

which is the well-known form of the Normal-Inverse-Wishart posterior distribution. More precisely, the posterior distribution is given by

$$\begin{aligned}p(\theta, \Sigma|\mathcal{D}) \propto &\det\left(\Sigma\right)^{-P/2} \exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}\left((\theta - \theta^{\text{post}})^{\text{T}}\Lambda^{\text{post}}(\theta - \theta^{\text{post}})\right)\right)\right) \\ &\cdot \det\left(\Sigma\right)^{-(\nu^{\text{post}}-P-1)/2} \cdot \exp\left(-\frac{1}{2}\text{tr}\left(S^{\text{post}}\Sigma^{-1}\right)\right)\end{aligned} \tag{26}$$

with multivariate normal conditional posterior for θ and Inverse-Wishart marginal posterior

for Σ

$$\theta|\Sigma, \mathcal{D} \sim \mathcal{N}\left(\text{vec}\left(\theta^{\text{post}}\right), \Sigma \otimes {\Lambda^{\text{post}}}^{-1}\right)$$
$$\Sigma|\mathcal{D} \sim \mathcal{IW}\left(\nu^{\text{post}}, S^{\text{post}}\right)$$

In Section 2.1 we specify the prior hyperparameters as

$$\begin{aligned}
\Lambda^{\text{prior}} &= \frac{T_0}{T} W'W, \\
\theta^{\text{prior}} &= \left[0_{N\times(L+1)}, \beta_0^{\text{prior}}, 0_{N\times K(L+1)}\right]. \\
\nu^{\text{prior}} &= T_0 - (K+1)(L+1) - 1 \\
S^{\text{prior}} &= \frac{T_0}{T}\left(R - W\theta^{\text{prior}}\right)'\left(R - W\theta^{\text{prior}}\right)
\end{aligned}$$

Plugging in these prior hyperparameters yields

$$\begin{aligned}
\Lambda^{\text{post}} &= W'W + \Lambda^{\text{prior}} = \frac{T_0 + T}{T} W'W, \\
\theta^{\text{post}} &= \left(\Lambda^{\text{post}}\right)^{-1}\left(\Lambda^{\text{prior}}\theta^{\text{prior}} + W'R\right) = \frac{T_0}{T+T_0}\theta^{\text{prior}} + \frac{T}{T+T_0}(W'W)^{-1}W'R, \\
\nu^{\text{post}} &= \nu^{\text{prior}} + T = T_0 + T - (K+1)(L+1), \\
S^{\text{post}} &= \frac{T_0 + T}{T} R'R - {\theta^{\text{post}}}'\Lambda^{\text{post}}\theta^{\text{post}}.
\end{aligned}$$

where we used that $S^{\text{prior}} + R'R + {\theta^{\text{prior}}}'\Lambda^{\text{prior}}\theta^{\text{prior}} - {\theta^{\text{post}}}'\Lambda^{\text{post}}\theta^{\text{post}} = \frac{T_0+T}{T}R'R - {\theta^{\text{post}}}'\Lambda^{\text{post}}\theta^{\text{post}}$.

A.2 Marginal Likelihood

The marginal likelihood is defined as the probability of the data under the model, integrating over all unknown parameters with respect to their prior distributions. In the specific case of the Bayesian multivariate linear regression model with Normal-Inverse-Wishart prior, the marginal likelihood has a closed-form expression (see, e.g., the web appendix to Giannone et al., 2015)

$$
\begin{aligned}
p(R|W) &= \int\int p(R|W,\theta,\Sigma)p(\theta|\Sigma)p(\Sigma)d\theta d\Sigma \\
&= \pi^{-(NT)/2}\frac{\Gamma_N\left(\frac{\nu^{\text{post}}}{2}\right)}{\Gamma_N\left(\frac{\nu^{\text{prior}}}{2}\right)}\frac{\det\left(\Lambda^{\text{prior}\,-1}\right)^{-N/2}}{\det\left(\Lambda^{\text{post}\,-1}\right)^{-N/2}}\frac{\det\left(S^{\text{prior}}\right)^{-\nu^{\text{prior}}/2}}{\det\left(S^{\text{post}}\right)^{-\nu^{\text{post}}/2}}.
\end{aligned} \tag{27}
$$

Conditional on the prior specification detailed in Section 2.1 the marginal likelihood simplifies to the following expression

$$
p(R|W) = \pi^{-(NT)/2}\frac{\Gamma_N\left(\frac{\nu^{\text{post}}}{2}\right)}{\Gamma_N\left(\frac{\nu^{\text{prior}}}{2}\right)}\left(\frac{T_0}{T+T_0}\right)^{KN/2}\frac{\det\left(S^{\text{prior}}\right)^{\nu^{\text{prior}}/2}}{\det\left(S^{\text{post}}\right)^{\nu^{\text{post}}/2}}. \tag{28}
$$

A.3 Posterior predictive moments

Let $w_* = [1, z_*', f_*', \psi_*']$ denote the row vector collecting the right-hand side variables used for prediction, where f_* contains the risk factors of model M_p constructed according to Q_{ip}. Then, using the prior described in Section 2.1, the posterior predictive expectation is

$$
\mathbb{E}(r_t|M_p, Q_{pj}) = \mathbb{E}(w_*\theta) + \mathbb{E}(\varepsilon_t) = w_*\mathbb{E}(\theta) = w_*\theta^{\text{post}},
$$

where θ^{post} is the posterior expectation of θ.

In order to derive the posterior predictive covariance, it is convenient to use the following formulation of the model

$$
r_t = (I_N \otimes w_*)\text{vec}(\theta) + \varepsilon_t.
$$

Then, the **posterior predictive covariance** is

$$
\begin{aligned}
\mathbb{V}\text{ar}(r|M_p, Q_{pj}) &= \mathbb{V}\text{ar}((I_N \otimes w_*)\text{vec}(\theta)) + \mathbb{V}\text{ar}(\varepsilon_t) \\
&= (I_N \otimes w_*)\mathbb{V}\text{ar}(\text{vec}(\theta))(I_N \otimes w_*)' + \overbrace{\underbrace{\mathbb{V}\text{ar}(\mathbb{E}(\varepsilon_t|\Sigma))}_{=0} + \underbrace{\mathbb{E}(\mathbb{V}\text{ar}(\varepsilon_t|\Sigma))}_{\mathbb{E}(\Sigma)}}^{\text{law of total variance}} \\
\vdots \quad \mathbb{V}\text{ar}(\text{vec}(\theta)) &= \overbrace{\underbrace{\mathbb{V}\text{ar}(\underbrace{\mathbb{E}(\text{vec}(\theta)|\theta^{\text{post}}, \Lambda^{\text{post}}, \Sigma)}_{=\theta^{\text{post}}})}_{=0} + \underbrace{\mathbb{E}(\underbrace{\mathbb{V}\text{ar}(\text{vec}(\theta)|\theta^{\text{post}}, \Lambda^{\text{post}}, \Sigma)}_{=\Sigma\otimes\Lambda^{\text{post}^{-1}}})}_{\mathbb{E}(\Sigma)\otimes 7\Lambda^{\text{post}}0^{-1}}}^{\text{law of total variance}} \\
&= (I_N \otimes w_*)(\frac{1}{\nu^{\text{post}} - N - 1} S^{\text{post}} \otimes \Lambda^{\text{post}^{-1}})(I_N \otimes w_*)' + \frac{1}{\nu^{\text{post}} - N - 1} S^{\text{post}} \\
&= \frac{1}{\nu^{\text{post}} - N - 1}(S^{\text{post}} \otimes \underbrace{w_* \Lambda^{\text{post}^{-1}} w_*'}_{\text{scalar}}) + \frac{1}{\nu^{\text{post}} - N - 1} S^{\text{post}} \\
&= \frac{1}{\nu^{\text{post}} - N - 1} S^{\text{post}}(w_* \Lambda^{\text{post}^{-1}} w_*' + 1),
\end{aligned}
$$

The third equality uses the fact, that the expectation of $X \sim \mathcal{IW}(S, \nu)$, where X is $(p \times p)$-dimensional, is $\mathbb{E}(X) = \frac{1}{\nu - p - 1} S$. The second to the last equality follows from the mixed-product property of the Kronecker product.

A.4 MCMC algorithm

A.4.1 Full conditional posterior distributions

The full conditional posterior distribution of the indicator variable δ_j follows a Bernoulli distribution, with success probability

$$
P(\delta_j = 1|R, \delta_{\setminus j}, \gamma) = \frac{P(R|\delta_j = 1, \delta_{\setminus j}, \gamma)p_0}{P(R|\delta_j = 1, \delta_{\setminus j}, \gamma)p_0 + P(R|\delta_j = 0, \delta_{\setminus j}, \gamma)(1 - p_0)}, \tag{29}
$$

where $P(R|\delta_j = 1, \delta_{\setminus j}, \gamma)$ denotes the marginal likelihood of the corresponding model.

The full conditional posterior distribution of γ_j is Multinoulli, with probabilities:

$$P(\gamma_{ji} = 1|R, \delta, \gamma_{\setminus j}) = \frac{P(R|\gamma_{ji} = 1, \delta, \gamma_{\setminus j})q_{ji}}{\sum_{s=1}^{n_j} P(R|\gamma_{js} = 1, \delta, \gamma_{\setminus j})q_{js}}, \quad \text{for } i = 1, \ldots, n_j. \tag{30}$$

The full conditional posterior distribution of the prior inclusion probability p_0 follows a Beta distribution

$$p_0|\delta \sim \mathcal{B}\left(s_1 + \sum_{j=1}^{K} \delta_j, s_2 + K - \sum_{j=1}^{K} \delta_j\right). \tag{31}$$

A.4.2 Gibbs Sampling Algorithm

1. Initialize $\delta^{(s=1)}$, $\gamma^{(s=1)}$, and $p_0^{(s=1)}$.
2. For $j = 1, \ldots, K$:
   (a) Update $\delta_j^{(s)}$ using one of the following:
   - **Gibbs update:** draw from the full conditional posterior.
   - **Metropolis update:** propose $\delta_j^{*(s)} = 1 - \delta_j^{(s-1)}$ and accept with probability
   $$p^* = \frac{P(R|\delta_j = \delta_j^{*(s)}, \delta_{\setminus j}^{(s^*)}, \gamma^{(s^*)})P(\delta_j = \delta_j^{*(s)}|p_0^{(s^*)})}{P(R|\delta_j = \delta_j^{(s-1)}, \delta_{\setminus j}^{(s^*)}, \gamma^{(s^*)})P(\delta_j = \delta_j^{(s-1)}|p_0^{(s^*)})},$$
   where the superscript (s^*) refers to the latest available state.

   (b) Draw $\gamma_j^{(s)}$ from its full conditional posterior.
3. Draw $p_0^{(s)}$ from its full conditional posterior.
4. Repeat steps 2 and 3 until convergence of the chain.

Remark on step 2a: the Metropolis update (step 2.a.ii) is used instead of the Gibbs update (2.a.i) because it results in more frequent changes in the state of $\delta_j^{(s)}$, which may accelerate exploration of the posterior space.

A key computational challenge in the MCMC algorithm arises from the vast number of

possible paths that must be evaluated for each risk factor. To mitigate this, the marginal likelihoods $P\Big(R \mid \gamma_{ji}^{(s)} = 1, \delta^{(s)}, \gamma_{\setminus j}^{(s)}\Big)$, $\quad i = 1, \ldots, n_j$, can be computed in parallel for fixed values of $\delta^{(s)}$ and $\gamma_{\setminus j}^{(s)}$. The algorithm is written in `C++`, leveraging `Rcpp`, `RcppArmadillo`, and `RcppParallel` to achieve efficient and scalable parallel computation of the marginal likelihoods.

B Variable construction

In this section, we show the 50 risk factors grouped according to the underlying mechanism following the outline in Table A1. All returns are from the "old" SIZ version of CRSP. See Schwarz et al. (2025), for a detailed discussion on the new CRSP version. We relegate all variable construction details to the Internet Appendix IA-1.

Table A1: Sorting variables. In this table, we show the 50 sorting variables used along with their abbreviations and reference papers (in parentheses). We group the variables into seven groups depending on the underlying economic mechanism by following Hou et al. (2020).

	Description (Reference Paper)
Group:	**Financing**
CDI	Composite debt issuance (Lyandres et al., 2008)
CSI	Composite share issuance (Daniel and Titman, 2006)
DBE	Change in common equity (Richardson et al., 2005)
DCOL	Change in current operating liabilities (Richardson et al., 2005)
DFNL	Change in financial liabilities (Richardson et al., 2005)
NDF	Net debt financing (Baker and Wurgler, 2006)
NEF	Net equity financing (Baker and Wurgler, 2006)
NXF	Net external financing (Baker and Wurgler, 2006)
Group:	**Intangibles**
HR	Hiring rate (Belo and Lin, 2012)
OL	Operating leverage (Novy-Marx, 2011)
RER	Real-estate ratio (Titman et al., 2004)

	Description (Reference Paper)
Group:	**Investment**
ACI	Abnormal corporate investment (Titman et al., 2004)
AG	Asset growth (Cooper et al., 2008)
DNOA	Change in net operating assets (Hirshleifer et al., 2004)
DPIA	Change in property, plant, and equipment (Lyandres et al., 2008)
DWC	Change in net non-cash working capital (Richardson et al., 2005)
IG	Investment growth (Xing, 2008)
DINV	Inventory changes (Thomas and Zhang, 2002)
NOA	Net operating assets Hirshleifer et al. (2004)
OA	Operating accruals (Sloan, 1996)
Group:	**Momentum**
MOM	Return momentum (Fama and French, 1996)
RMOM	Residual momentum (Blitz et al., 2011)
RS	Revenue surprise (Jegadeesh and Livnat, 2006)
SUE	Standardized unexpected earnings (Foster et al., 1984)
52W	52-week high (George and Hwang, 2004)
Group:	**Profitability**
CBOP	Cash-based operating profitability (Ball et al., 2016)
GPA	Gross profits to assets (Novy-Marx, 2013)
O	O-score (Ohlson, 1980; Dichev, 1998)
OPE	Operating profits to book equity (Fama and French, 2015)
ROA	Return on assets (Balakrishnan et al., 2010)
ROE	Return on equity (Hou et al., 2014)
TBI	Taxable income to book income (Lev and Nissim, 2004)
Group:	**Size**
ME	Logarithm of market equity (Banz, 1981)
Group:	**Trading frictions**
AMI	Amihud illiquidity measure (Amihud, 2002)
BETA	Beta relative to the market (Fama and MacBeth, 1973)

	Description (Reference Paper)
DTV	Dollar trading volume (Brennan et al., 1998)
ISKEW	Idiosyncratic skewness (Bali et al., 2016)
IVOL	Idiosyncratic volatility (Ang et al., 2006)
MDR	Maximum daily return (Bali et al., 2011)
TUR	Share turnover (Datar et al., 1998)
Group:	**Valuation**
BM	Book equity to market equity (Daniel and Titman, 2006)
CFM	Cash flow to market equity (Lakonishok et al., 1994)
DM	Debt to market equity (Bhandari, 1988)
EBM	Enterprise book equity to market equity (Penman et al., 2007)
EM	Earnings to market equity (Basu, 1983)
NDM	Net debt to market equity (Penman et al., 2007)
NPY	Net payout yield (Boudoukh et al., 2007)
OCM	Operating cash flow to market equity (Desai et al., 2004)
REV	Long-term reversal (De Bondt and Thaler, 1985)
SM	Sales to market equity (Barbee Jr et al., 1996)

Table A2: Summary statistics. For each sorting variable we report the median Sharpe ratio (SR, annualized), the distribution of annualized risk premia (in percent) across all data-preprocessing choices (median, 25th and 75th percentiles) and the published risk premium replicated by Open Source Asset Pricing (OSAP). We also report the median and central intervals (25th, 75th) of time-series correlations across variants and the average correlation with the OSAP replication (Published). The sample period ranges from July 1973 until December 2022.

		Risk premia (annualized %)				Time-series correlations			
	SR	Median	Q25	Q75	Published	Median	Q25	Q75	Published
52w	0.138	2.03%	-0.11%	5.47%	0.04%	0.68	0.57	0.80	0.65
aci	0.529	3.35%	2.58%	4.11%	0.03%	0.60	0.50	0.68	0.52
ag	0.558	5.34%	3.93%	7.22%	0.14%	0.72	0.61	0.81	0.37

ami	0.288	3.38%	2.01%	4.84%	0.02%	0.57	0.38	0.72	0.54
beta	0.022	0.42%	-0.28%	1.12%	0.03%	0.91	0.89	0.94	0.86
bm	0.293	3.79%	2.65%	5.20%	0.09%	0.80	0.74	0.86	0.76
cbop	0.387	2.93%	1.90%	4.23%	0.05%	0.67	0.58	0.76	0.29
cdi	0.346	2.01%	1.32%	2.76%	0.04%	0.58	0.48	0.68	0.46
cfm	0.323	4.29%	3.62%	5.34%	0.05%	0.87	0.82	0.91	0.36
csi	0.226	1.96%	1.19%	3.06%	0.04%	0.68	0.61	0.75	0.43
dbe	0.411	4.27%	3.38%	5.64%	0.06%	0.72	0.60	0.83	0.41
dcol	0.171	1.49%	0.58%	2.72%	0.04%	0.71	0.62	0.79	0.65
dfnl	0.716	3.74%	2.99%	4.96%	0.07%	0.50	0.40	0.61	0.48
dinv	0.570	4.37%	3.33%	5.70%	0.07%	0.66	0.58	0.75	0.54
dm	0.115	1.50%	0.88%	2.21%	0.05%	0.81	0.75	0.86	0.75
dnoa	0.842	6.41%	5.41%	7.98%	0.10%	0.66	0.55	0.75	0.50
dpia	0.630	5.40%	4.14%	6.95%	0.07%	0.72	0.62	0.79	0.56
dtv	0.306	3.82%	2.54%	5.27%	0.07%	0.59	0.43	0.72	0.41
dwc	0.660	4.62%	3.56%	5.71%	0.01%	0.59	0.48	0.69	0.25
ebm	0.213	3.00%	2.02%	3.95%	0.01%	0.83	0.77	0.88	0.40
em	0.341	3.96%	3.40%	4.59%	0.03%	0.84	0.79	0.89	0.63
gpa	0.323	3.32%	2.37%	4.49%	0.04%	0.54	0.40	0.71	0.46
hr	0.326	2.94%	1.97%	4.30%	0.06%	0.74	0.65	0.81	0.59
ig	0.528	3.68%	3.00%	4.65%	0.05%	0.61	0.52	0.71	0.41
iskew	-0.055	-0.28%	-0.86%	0.27%	0.02%	0.23	0.15	0.36	0.26
ivol	0.036	0.58%	-0.43%	1.49%	0.09%	0.85	0.80	0.88	0.81
mdr	-0.004	-0.06%	-0.92%	0.71%	0.08%	0.84	0.80	0.87	0.77
me	0.338	4.72%	3.66%	6.00%	0.02%	0.81	0.65	0.89	0.52
mom	0.047	0.66%	-0.52%	3.03%	0.11%	0.70	0.58	0.81	0.42
ndf	0.601	3.23%	2.49%	4.50%	0.08%	0.50	0.40	0.60	0.45
ndm	0.035	0.40%	-0.32%	1.05%	0.06%	0.76	0.68	0.83	0.63
nef	0.319	3.49%	2.59%	4.71%	0.08%	0.84	0.78	0.88	0.74
noa	0.648	5.86%	4.69%	6.95%	0.09%	0.59	0.46	0.75	0.53

npy	0.279	2.56%	1.64%	3.03%	0.12%	0.72	0.61	0.79	0.47
nxf	0.623	5.33%	4.17%	7.08%	0.13%	0.74	0.67	0.80	0.61
o	0.022	0.20%	-0.55%	0.94%	0.09%	0.66	0.54	0.77	-0.02
oa	0.517	3.77%	2.84%	4.95%	0.04%	0.58	0.44	0.69	0.51
ocm	0.372	4.79%	3.95%	5.91%	0.05%	0.85	0.80	0.89	0.44
ol	0.351	3.61%	3.01%	4.50%	0.05%	0.64	0.39	0.80	0.42
ope	0.381	3.18%	2.25%	4.34%	0.07%	0.66	0.56	0.76	0.69
rer	0.320	2.15%	1.62%	2.85%	0.03%	0.42	0.29	0.56	0.36
rev	0.222	2.44%	1.79%	3.48%	0.07%	0.71	0.63	0.79	0.41
rmom	0.159	1.35%	0.56%	3.07%	0.11%	0.59	0.43	0.73	0.41
roa	0.217	2.33%	0.51%	4.23%	0.16%	0.54	0.35	0.70	0.44
roe	0.263	2.55%	0.91%	4.63%	0.04%	0.54	0.41	0.68	0.44
rs	0.211	1.81%	1.07%	3.00%	0.07%	0.64	0.56	0.73	0.19
sm	0.409	5.43%	4.58%	6.96%	0.11%	0.84	0.77	0.89	0.78
sue	0.212	1.73%	0.82%	3.70%	0.07%	0.63	0.54	0.72	0.28
tbi	-0.105	-1.40%	-2.86%	0.53%	0.05%	0.73	0.60	0.83	0.37
tur	0.077	1.23%	0.32%	2.85%	0.04%	0.89	0.85	0.92	0.51

Internet Appendix

This is the Internet Appendix to "Uncertainty everywhere: Integrating conceptual uncertainty in the stochastic discount factor"

IA-1 Variable construction

In this section, we show the construction details for the risk factors. We group them by their underlying mechanism as shown in Table A1 in the Appendix of the main paper.

IA-1.1 Financing

CDI - Composite Debt Issuance

Following Lyandres et al. (2008), we compute composite debt issuance as the logarithmic growth rate of book debt (i.e., current debt, DLC, plus long-term debt, DLTT) over the last five years, i.e.,

$$\text{CDI}_t = \log(\text{DLC}_t + \text{DLTT}_t) - \log(\text{DLC}_{t-60} + \text{DLTT}_{t-60})$$

We replace the missing values of DLC and DLTT with zero individually. Additionally, the sum of DLC and DLTT must be larger than zero, and is set to missing otherwise.

CSI - Composite Share Issuance

Following Daniel and Titman (2006), we measure composite share issuance from CRSP data as the difference between the change in market equity (ME) (i.e., the shares outstanding, shrout, times the absolute value of their price, altprc, in million USD) and the cumulative log return (where r is a simple return) in each month from t to $t-60$, i.e.,

$$\text{CSI}_t = \log(ME_t/ME_{t-60}) - \sum_{\tau=t-60}^{t} \log(1+r_\tau)$$

DBE - Change in Common Equity

Following Richardson et al. (2005), we calculate the change in common equity as the ratio of common equity (CEQ) over one year scaled by lagged assets (AT), i.e.,

$$DBE_t = \frac{CEQ_t - CEQ_{t-12}}{AT_{t-12}}$$

DCOL - Change in Current Operating Liabilities

Following Richardson et al. (2005), we measure the change in current operating liabilities as the change in current operating liabilities (i.e., current liabilities, LCT, minus current debt, DLC) over one year scaled by lagged assets (AT), i.e.,

$$DCOL_t = \frac{(LCT_t - DLC_t) - (LCT_{t-12} - DLC_{t-12})}{AT_{t-12}}$$

We replace the missing values of DLC with 0.

DFNL - Change in Financial Liabilities

Following Richardson et al. (2005), we measure the change in financial liabilities (i.e., the sum of current debt, DLC, plus long-term debt, DLTT, plus preferred stocks, PSTK) over one year scaled by lagged assets (AT), i.e.,

$$DFNL_t = \frac{(DLC_t + DLTT_t + PSTK_t) - (DLC_{t-12} + DLTT_{t-12} + PSTK_{t-12})}{AT_{t-12}}$$

We replace the missing values of DLC, DLTT and PSTK with 0, if at least one of them is given.

NDF - Net Debt Financing

Following Baker and Wurgler (2006), we measure net debt financing as cash proceeds

from the issuance of long-term debt (DLTIS) minus cash payments for long-term debt reductions (DLTR) plus net changes in current debt (DLCCH) scaled by the average total assets (AT) over the current and the last year, i.e.,

$$NDF_t = \frac{DLTIS_t - DLTR_t + DLCCH_t}{\frac{1}{2}(AT_t + AT_{t-12})}$$

We replace the missing values of DLCCH with zero.

NEF - Net Equity Financing

Following Baker and Wurgler (2006), we measure net equity financing as proceeds from the sale of common and preferred stocks (SSTK) minus payments for the repurchase of common and preferred stocks (PRSTKC) minus cash payments for dividends (DV) scaled by the average total assets (AT) over the current and the last year, i.e.,

$$NEF_t = \frac{SSTK_t - PRSTKC_t - DV_t}{\frac{1}{2}(AT_t + AT_{t-12})}$$

NXF - Net External Financing

Following Baker and Wurgler (2006), we measure net external financing as the sum of net debt financing (NDF) and net equity financing (NEF) as defined above.

IA-1.2 Intangibles

HR - Hiring Rate

Following Belo and Lin (2012), we measure the hiring rate as the change in the number of employees (EMP) over one year scaled by the average number of employees, i.e.,

$$HR_t = \frac{EMP_t - EMP_{t-12}}{\frac{1}{2}(EMP_t + EMP_{t-12})}$$

We exclude firms with a hiring rate of zero.

OL - Operating Leverage

Following Novy-Marx (2011), we measure operating leverage as cost of goods sold (COGS) plus selling, general, and administrative expenses (XSGA), both scaled by current total assets (AT), i.e.,

$$OL_t = \frac{COGS_t + XSGA_t}{AT_t}$$

RER - Real-Estate Ratio

Following Titman et al. (2004), we measure the real-estate ratio in two subperiods differently.

(a) Prior to 1983, the real-estate ratio is the sum of buildings (PPENB) and capital leases (PPENLS) scaled by net property, plant, and equipment (PPENT), i.e.,

$$RER_t = \frac{PPENB_t + PPENLS_t}{PPENT_t}$$

(b) After the end of 1983, the real-estate ratio is the sum of buildings at cost (FATB) and leases at cost (FATL), both divided by gross property, plant, and equipment (PPEGT), i.e.,

$$RER_t = \frac{FATB_t + FATL_t}{PPEGT_t}$$

Subsequently, we winsorize the real-estate ratios in each fiscal year at the first and 99th percentile. The industry-adjusted real-estate ratio is obtained by subtracting the industry average real-estate ratio from each stock-specific real-estate ratio. We use 2-digit SIC codes to assign stocks to industries. We always require at least five observations to calculate the industry average each year.

IA-1.3 Investment

ACI - Abnormal Corporate Investment

Following Titman et al. (2004), we measure abnormal corporate investments using the variable CE, which is capital expenditures (CAPX) scaled by sales (SALE), then ACI is given by CE scaled by CE from the previous three years minus one, i.e.,

$$ACI_t = \frac{CE_t}{\frac{1}{3}(CE_{t-12} + CE_{t-24} + CE_{t-36})} - 1$$

We follow Hou et al. (2020) and exclude stocks with sales below 10 million dollars.

AG - Asset Growth

Following Cooper et al. (2008), we measure asset growth as the change in total assets (AT) over one year scaled by lagged assets (AT), i.e.,

$$AG_t = \frac{AT_t - AT_{t-12}}{AT_{t-12}}$$

DNOA - Change in Net Operating Assets

Following Hirshleifer et al. (2004), we measure net operating assets (defined as NOA^{int}) as

$$NOA_t^{int} = (AT_t - CHE_t) - (AT_t - DLC_t - DLTT_t - MIB_t - PSTK_t - CEQ_t)$$

where AT corresponds to total assets, CHE to cash and short-term investments, DLC to current liabilities, DLTT to long-term debt, MIB to minority interests, PSTK to the value of preferred stocks, and CEQ to common equity. Missing values in DLC, DLTT, MIB, and PSTK are set to zero. Then, the change in net operating assets is the difference

between net operating assets of fiscal year t and fiscal year $t-12$ scaled by lagged assets (AT), i.e.,

$$DNOA_t = \frac{NOA_t^{int} - NOA_{t-12}^{int}}{AT_{t-12}}$$

DPIA - Change in Property, Plant, Equipment, and Inventory

Following Lyandres et al. (2008), we measure the change in property, plant, equipment, and inventory by combining the change in gross property, plant, and equipment (PPEGT) with the annual change in inventory (INVT) and scaling this sum by lagged total assets (AT), i.e.,

$$DPIA_t = \frac{(PPEGT_t - PPEGT_{t-12}) + (INVT_t - INVT_{t-12})}{AT_{t-12}}$$

DWC - Change in Net Non-Cash Working Capital

Following Richardson et al. (2005), we measure non-cash working capital as

$$WC_t = (ACT_t - CHE_t) - (LCT_t - DLC_t)$$

where ACT corresponds to current assets, CHE to cash, LCT to current liabilities, and DLC to short-term debt. We set missing values of DLC to zero. Then, the change in net non-cash working capital corresponds to the change of WC from fiscal year t to fiscal year $t-1$ scaled by lagged total assets (AT), i.e.,

$$DWC_t = \frac{WC_t - WC_{t-12}}{AT_{t-12}}$$

IG - Investment Growth

Following Xing (2008), we measure investment growth as the annual change in capital

expenditures (CAPX) scaled by lagged capital expenditures, i.e.,

$$IG_t = \frac{CAPX_t - CAPX_{t-12}}{CAPX_{t-12}}$$

DINV - Inventory Changes

Following Thomas and Zhang (2002), we measure the change in inventory as the annual change in inventories (INVT) scaled by average total assets (AT) over the current and the last year, i.e.,

$$DINV_t = \frac{INVT_t - INVT_{t-12}}{\frac{1}{2}(AT_t + AT_{t-12})}$$

NOA - Net Operating Assets

Following Hirshleifer et al. (2004), we measure net operating assets (defined as NOA^{int}) as

$$NOA_t^{int} = (AT_t - CHE_t) - (AT_t - DLC_t - DLTT_t - MIB_t - PSTK_t - CEQ_t)$$

where AT corresponds to total assets, CHE to cash and short-term investments, DLC to current liabilities, DLTT to long-term debt, MIB to minority interests, PSTK to the value of preferred stocks, and CEQ to common equity. Missing values in DLC, DLTT, MIB, and PSTK are set to zero. Then, net operating assets is the net operating assets of fiscal year t scaled by lagged assets (AT), i.e.,

$$NOA_t = \frac{NOA_t^{int}}{AT_{t-12}}$$

OA - Operating Accruals

Following Sloan (1996) (before 1988) and Hribar and Collins (2002) (after 1988), we measure operating accruals until 1988 as

$$OA_t = \frac{(\Delta ACT_t - \Delta CHE_t) - (\Delta LCT_t - \Delta DLC_t - \Delta TXP_t) - DP_t}{AT_{t-12}}$$

where ACT is current assets, CHE cash, LCT current liabilities, DLC short-term debt, TXP taxes payable, and DP depreciation and amortization. Moreover, we replace the missing values of DLC and TXP with zero. Δ represented changes over one year. From 1988 onwards, we measure operating accruals as

$$OA_t = \frac{NI_t - OANCF_t}{AT_{t-12}}$$

where NI is net income and OANCF corresponds to net cash flow from operations.

IA-1.4 Momentum

MOM - Return Momentum

Following Fama and French (1996), we measure return momentum from monthly CRSP returns (where r is a simple return) as the cumulative return from month $t-11$ to month $t-1$ skipping the most recent month, i.e.,

$$\text{MOM}_t = \exp\left(\sum_{\tau=t-11}^{t-1} \log(1+r_\tau)\right) - 1$$

RMOM - Residual Momentum

Following Blitz et al. (2011), we define the 11-month residual momentum (RMOM) in each month and for each stock as cumulative residual returns from month $t-1$ to month $t-11$, scaled by the standard deviation of residual returns over the same time horizon. Residual returns ϵ are obtained in each month from regressing monthly simple, excess stock returns

from month $t-1$ to month $t-36$ on the Fama and French (1993) three-factor model. Throughout these rolling regressions, we always require 36 monthly returns.

RS - Revenue Surprise

Following Jegadeesh and Livnat (2006), we measure the change revenues per share (DRPS) from quarterly Compustat data as quarterly sales (SALEQ) scaled by the product of quarterly shares outstanding (CHSPRQ) times their cumulative adjustment factor (AJEXQ), i.e.,

$$DRPS_q = \frac{SALEQ_q}{CSHPRQ_q \cdot AJEXQ_q} - \frac{SALEQ_{q-4}}{CSHPRQ_{q-4} \cdot AJEXQ_{q-4}}$$

Then, the revenue surprise corresponds to the change in revenues per share over the last four quarters scaled by the standard deviation of the change in revenues per share over the last eight quarters. We require at least six quarterly observations for this rolling standard deviation and remove infinite values. Finally, the earnings announcement date has to be after the fiscal quarter end date.

SUE - Standardized Unexpected Earnings

Following Foster et al. (1984), we measure unexpected earnings per share (UES) from quarterly Compustat data as the change in split-adjusted earnings (which are equal to quarterly earnings per share, EPSPXQ, scaled by the correction factor for shares outstanding, AJEXQ) per share from its value four quarters ago

$$UES_q = \frac{EPSPXQ_q}{AJEXQ_q} - \frac{EPSPXQ_{q-4}}{AJEXQ_{q-4}}$$

Then, standardized unexpected earnings are defined as unexpected earnings per share divided by the standard deviation of unexpected earnings per share over the previous eight quarters. We require at least six quarterly observations for this rolling standard deviation. Finally, the earnings announcement date has to be before the fiscal quarter

end date.

52W - 52-Week High

Following George and Hwang (2004), we measure the 52-week high as month t's closing split-adjusted stock price scaled by the highest daily split-adjusted stock price over the previous twelve months. We only consider stock-month pairs with more than three valid price observations.

IA-1.5 Profitability

CBOP - Cash-Based Operating Profitability

Following Ball et al. (2016), we measure cash-based operating profitability

$$\begin{aligned} CBOP_t =& (REVT_t - COGS_t - XSGA_t + XRD_t \\ & - \Delta RECT_t - \Delta INVT_t - \Delta XPP_t + \Delta DRC_t \\ & + \Delta DRLT_t + \Delta AP_t + \Delta XACC_t)/AT_t \end{aligned}$$

where REVT is total revenue, COGS are cost of goods sold, XSGA are selling, general, and administrative expenses, and XRD are R&D expenses. Moreover, Δ refers to one-year changes, which we measure for accounts receivable (RECT), inventory (INVT), prepaid expenses (XPP), deferred revenues (the sum of current, DRC, and long-term, DRLT), trade accounts payable (AP), and accrued expenses (XACC). We set the missing values of XRD and all missing changes to zero.

GPA - Gross Profits to Assets

Following Novy-Marx (2013), we measure gross profits to assets as total revenues (REVT) minus cost of goods sold (COGS) scaled by current total assets (AT), i.e.,

$$GPA_t = \frac{REVT_t - COGS_t}{AT_t}$$

O - Ohlson's O-score

Following Ohlson (1980), we measure O-score with the following linear relation

$$\begin{aligned} O_t = &- 1.32 - 0.407 \cdot \log(AT_t) + 6.03 \cdot \frac{DLC_t + DLTT_t}{AT_t} \\ &- 1.43 \cdot \frac{ACT_t - LCT_t}{AT_t} + 0.076 \cdot \frac{LCT_t}{AT_t} \\ &- 1.72 \cdot 1_{LT_t > AT_t} - 2.37 \cdot \frac{NI_t}{AT_t} - 1.83 \cdot \frac{PI_t + DP_t}{LT_t} \\ &+ 0.285 \cdot 1_{NI_t<0 \ \& \ NI_{t-12}<0} - 0.521 \cdot \frac{NI_t - NI_{t-12}}{|NI_t| + |NI_{t-12}|} \end{aligned}$$

where AT corresponds to total assets, DLC to short-term debt, DLTT to long-term debt, ACT to current assets, LCT to current liabilities, LT to total liabilities, PI to pretax income, DP to depreciation and amortization, and NI to net income. We winsorize all variables except for dummy variables at the first and 99^{th} percentile of their respective distribution.

OPE - Operating Profits to Book Equity

Following Fama and French (2015), we measure operating profits to book equity as

$$OPE_t = \frac{REVT_t - COGS_t - XSGA_t - XINT_t}{BE_t}$$

where REVT corresponds to total revenues, COGS to cost of goods sold, XSGA to selling, general and administrative expenses XINT to interest expenses, and BE to book equity. The definition of book equity is (a) the book equity of shareholders plus (b) deferred taxes and investment tax credit (Compustat item TXDITC or TXDB + ITCB if TXDITC is

unavailable) minus (c) the book value of preferred stock. Regarding (a), the book equity of shareholders is shareholders' equity (SEQ), or the sum of common equity (CEQ) and the par value of preferred stock (PSTK), or, if all previous items are unavailable, total assets (AT) minus total liabilities (LT). Regarding (b), deferred taxes (TXDITC) or investment tax credits (TXDB plus ITCB) if deferred taxes is unavailable. Regarding (c), the book value of preferred stock corresponds in the following order either to the redemption value (PSTKRV), or the liquidation value (PSTKL), or if all previous items are unavailable to the par value (PSTK). We replace missing values of items (b) and (c) with zero. Moreover, missing values in COGS, XSGA, and XINT are set to zero.

ROA - Return on Assets

Following Balakrishnan et al. (2010), we measure return on assets as quarterly income before extraordinary items (IBQ) relative to lagged quarterly total assets (ATQ), i.e.,

$$ROA_q = \frac{IBQ_q}{ATQ_{q-1}}$$

Finally, the earnings announcement date has to be before the fiscal quarter end date.

ROE - Return on Equity

Following Hou et al. (2014), we measure return on equity as quarterly income before extraordinary items (IBQ) relative to lagged book equity (BEQ)

$$ROE_q = \frac{IBQ_q}{BEQ_{q-1}}$$

Quarterly book equity (BEQ) is computed as the (a) book equity of shareholders plus (b) balance sheet deferred taxes and investment tax credit minus (c) the book value of preferred stock. Regarding (a), we measure book equity of shares holders by shareholders' equity (SEQQ), or the sum of common equity (CEQQ) and the par value of preferred stock (PSTKQ), or if all previous items are unavailable by total assets (ATQ) minus

total liabilities (LTQ). Regarding (b), the balance sheet deferred taxes and investment tax credit is TXDITCQ, or TXDBQ, or zero if both are missing. Regarding (c), the book value of preferred stocks corresponds to PSTKRQ, or PSTKQ, or zero if both are unavailable. Finally, the earnings announcement date has to be before the fiscal quarter end date.

TBI - Taxable Income to Book Income

Following Green et al. (2013), we measure taxable income to book income as pretax income (PI) scaled by net income (NI)

$$TBI_t = \frac{PI_t}{NI_t}$$

We require positive pretax and net income.

IA-1.6 Size

ME - Size

Following Fama and French (1992), we measure size as the natural logarithm of the market equity. We obtain market equity data from CRSP by multiplying the shares outstanding (shrout * 1,000, as they are given in thousands) with the corresponding share price (the absolute value of altprc)), i.e.,

$$ME_t = \log\left(shrout_t * 1000 * |altprc_t|\right)$$

IA-1.7 Trading frictions

AMI - Amihud Illiquidity Measure

Following Amihud (2002), we measure the illiquidity from daily return-to-volume (RV), which is the absolute daily return (r) scaled by the daily dollar trading volume (product

of vol and prc), i.e.,

$$RV_t = \frac{|r_d|}{prc_t * vol_t}$$

Then, the Amihud illiquidity measure corresponds to the average return to volume estimate over the last six months. We require at least 50 observations for this average and adjust the trading volume of NASDAQ stocks according to Ga and Ri (2010).

BETA - Beta Relative to the Market

Following Fama and MacBeth (1973), we measure the market beta β_1 from monthly CRSP data. Specifically, we run the following time series regression over the previous five years

$$r_t^e = \alpha + \beta_1 \cdot (MKT_t - R_t^f) + u_t$$

Moreover, we require at least 24 monthly observations for the regression above.

DTV - Dollar Trading Volume

Following Brennan et al. (1998), we measure dollar trading volume from daily CRSP data as the average dollar trading volume (product of vol and prc) from month $t-6$ to month $t-1$ (with at least 50 observations). Moreover, we adjust dollar trading volume from NASDAQ according to Ga and Ri (2010).

ISKEW - Idiosyncratic Skewness Relative to the Fama and French (1993) three-factor model

Following Bali et al. (2016), we regress daily excess returns in each month on the three Fama and French (1993) factor model (with a minimum of 15 daily observations), i.e.,

$$r_t^e = \alpha + \beta_1 \cdot (MKT_t - R_t^f) + \beta_2 \cdot SMB_t + \beta_3 \cdot HML_t + u_t$$

Then, idiosyncratic skewness is measured as the skewness of residuals u_t.

IVOL - Idiosyncratic Volatility Relative to the Fama and French (1993) model

Follow Ang et al. (2006), we compute idiosyncratic volatility relative to the Fama and French (1993) three-factor model as the volatility of residuals from the following regression (with a minimum of 15 daily observations)

$$r_t^e = \alpha + \beta_1 \cdot (MKT_t - R_t^f) + \beta_2 \cdot SMB_t + \beta_3 \cdot HML_t + u_t$$

MDR - Maximum Daily Return

Following Bali et al. (2011), we measure the maximum daily (excess) return in each month from daily CRSP data (with a minimum of at least 15 return observations).

TUR - Share Turnover

Following Datar et al. (1998), we measure the daily share turnover of each stock as the number of shares traded (vol) scaled by the number of shares outstanding (shrout), i.e.,

$$T_t = \frac{vol}{shrout}$$

Then, the share turnover for each firm in each month t is the average daily share turnover over the previous six months (with at least 50 observations). Moreover, we adjust the trading volume of NASDAQ stocks according to Gao and Ritter (2010).

IA-1.8 Valuation

BM - Book Equity to Market Equity

Following Davis et al. (2000), we measure the book-to-market ratio as book equity (BE) from Compustat divided by market equity (ME) from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD),

i.e.,

$$BM_t = \frac{BE_t}{ME_t}$$

The definition of book equity is (a) the book equity of shareholders plus (b) deferred taxes and investment tax credit (Compustat item TXDITC or TXDB + ITCB if TXDITC is unavailable) minus (c) the book value of preferred stock. Regarding (a), the book equity of shareholders is shareholders' equity (SEQ), or the sum of common equity (CEQ) and the par value of preferred stock (PSTK), or, if all previous items are unavailable, total assets (AT) minus total liabilities (LT). Regarding (b), deferred taxes (TXDITC) or investment tax credits (TXDB plus ITCB) if deferred taxes is unavailable. Regarding (c), the book value of preferred stock corresponds in the following order either to the redemption value (PSTKRV), or the liquidation value (PSTKL), or if all previous items are unavailable to the par value (PSTK). We replace missing values of items (b) and (c) with zero.

CFM - Cash Flow to Market Equity

Following Lakonishok et al. (1994), we measure the cash-flow-to-price ratio as income before extraordinary items (IB) plus depreciation (DP) scaled by market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$CFM_t = \frac{IB_t + DP_t}{ME_t}$$

We exclude all stocks with negative cash flows.

DM - Debt to Market Equity

Following Bhandari (1988), we measure the debt-to-market ratio as short-term debt (DLC) plus long-term debt (DLTT) from Compustat divided by market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price,

altprc, in million USD), i.e.,

$$DM_t = \frac{DLC_t + DLTT_t}{ME_t}$$

We exclude stocks with missing DLC and DLTT observations.

EBM - Enterprise Book Equity to Market Equity

Following Penman et al. (2007), we measure enterprise book equity scaled by market equity as net debt plus book equity scaled by net debt plus market equity

$$EBM_t = \frac{(DLTT_t + DLC_t + PSTK_t + DVPA_t - TSTKP_t) - CHE_t + BE_t}{(DLTT_t + DLC_t + PSTK_t + DVPA_t - TSTKP_t) - CHE_t + ME_t}$$

where DLTT corresponds to long-term debt, DLC to current liabilities, PSTK to the value of preferred stock, DVPA to preferred dividends in arrears, TSTKP to preferred treasury stock, CHE to cash and short-term investments, and ME to the market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD). Lastly, in this case, book equity BE is computed as common equity (CEQ) plus TSTKP minus DVPA. Missing observations in DVPA and TSTKP are set to zero. Additionally, we require that the sum of net debt and book equity, as well as the sum of net debt plus market equity, are positive.

EM - Earnings to Market Equity

Following Basu (1983), we measure the earnings-to-price ratio as income before extraordinary items (IB) from Compustat divided by market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$EM_t = \frac{IB_t}{ME_t}$$

We exclude firms with negative earnings.

NDM - Net Debt to Market Equity

Following Penman et al. (2007), we measure net debt to price as net debt from Compustat scaled by market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$NDM_t = \frac{(DLTT_t + DLC_t + PSTK_t + DVPA_t - TSTKP_t) - CHE_t}{ME_t}$$

where DLTT corresponds to long-term debt, DLC to current liabilities, PSTK to the value of preferred stock, DVPA to preferred dividends in arrears, TSTKP to preferred treasury stock, CHE to cash and short-term investments. Lastly, missing observations in DVPA and TSTKP are set to zero.

NPY - Net Payout Yield

Following Boudoukh et al. (2007), we measure the net payout yield as net payouts from Compustat scaled by market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$\begin{aligned} NPY_t = &[(DVC_t + PRSTKC_t + \Delta PSTKRV_t \cdot 1_{\Delta PSTKRV<0}) \\ &- (SSTK_t - \Delta PSTKRV_t \cdot 1_{\Delta PSTKRV>0})]/ME_t \end{aligned}$$

where DVC are dividends from common stock, PRSTKC is the purchase of common and preferred stock, PSTKRV is the value of the net number of preferred stocks outstanding, and SSTK reflects the sale of common and preferred stocks. $1_{\Delta PSTKRV<0}$ is a dummy

variable that has value one if the annual change in PSTKRV is negative and zero otherwise. Moreover, we exclude stocks with negative net payouts.

OCM - Operating Cash Flow to Market Equity

Following Desai et al. (2004), we measure the ratio of operating cash flows to price as operating cash flows (OC) from Compustat divided by the market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$OCM_t = \frac{OC_t}{ME_t}$$

Before 1988, we measure OC as funds from operations (FOPT) minus the change in working capital (item WCAP). Thereafter, we measure OC as net cash flows from operating activities (OANCF). Moreover, we exclude firms with negative operating cash flows.

REV - Long-Term Reversal

Following De Bondt and Thaler (1985), we measure the long-term reversal effect by the cumulative returns from month $t-60$ to month $t-13$, i.e.,

$$\text{REV}_t = \exp\left(\sum_{\tau=t-60}^{t-13} \log(1+r_\tau)\right) - 1$$

SM - Sales to Market Equity

Following Barbee Jr et al. (1996), we measure the sales-to-price ratio (SM) as sales (SALE) from Compustat divided by the market equity from CRSP (i.e., the fiscal-year end's shares outstanding, shrout, times the absolute value of their price, altprc, in million USD), i.e.,

$$SM_t = \frac{SALE_t}{ME_t}$$