Empirical Asset Pricing with Bayesian Regression Trees[*]

Drew Creal[†] and Jaeho Kim[‡]

January 5, 2021

Abstract

Portfolio sorts are a popular technique used in finance to study the cross-section of expected returns. However, existing methods are typically limited to including one or two variables at a time, making it difficult to disentangle which characteristics are the most important. We address this problem by developing a new Bayesian factor model with regression tree priors that selects across a large space of characteristics simultaneously. We apply our methods to an unbalanced panel of currency returns. We find that the interest rate differential and FX volatility are the primary drivers of currencies' betas. Portfolios constructed from the model exhibit a higher Sharpe ratio than the carry trade and can be achieved using information on the VIX, aggregate capital ratio of financial intermediaries, and the global interest rate differential.

JEL Classification Codes: C11, C58, G12, G15, F31.

Keywords: risk premia; machine learning; portfolio sorts; currency carry trade

---

* We thank Cynthia Wu as well as seminar participants at Indiana Univ. for helpful comments.
† Department of Economics, University of Notre Dame; dcreal@nd.edu.
‡ Department of Economics, University of Oklahoma; jaeho@ou.edu.

1 Introduction

Empirical research in asset pricing focuses on characterizing the trade-off between risk and expected returns in the cross-section of assets, how this risk-return trade-off changes through time, and whether it can be summarized by a small number of known risk factors. Financial economists commonly study these questions using portfolios of returns sorted by one or two characteristics; a technique known as portfolio sorts. Portfolio sorts allow economists to forecast expected returns cross-sectionally. Assets with a given characteristic, say an equity's book-to-market ratio or size as measured by market capitalization, can have a higher or lower expected return because the characteristics proxy for an asset's beta. Betas can change through time as the assets' characteristics change. Portfolio sorts play an important role statistically as well because they reduce the number of parameters to be estimated and eliminate idiosyncratic noise by taking an a-priori weighted sample average of the raw data. They are also a simple non-parametric estimator of the conditional expectation function.

When financial economists sort returns into portfolios, many characteristics appear to be important, especially when the sort is conducted separately, variable-by-variable. It is often unclear which variables are ultimately the most important for explaining the cross-section of returns. For example, in our application to currency trades, the finance literature has documented several characteristics that appear to be relevant predictors. These variables include the interest rate differential as in Lustig et al. (2011), momentum as in Menkhoff et al. (2012b) and Burnside et al. (2011), real exchange rates as in Menkhoff et al. (2017) and Chernov and Creal (2021), and macroeconomic fundamentals like unemployment and inflation rates as in Dahlquist and Hasseltoft (2020) and Colacito et al. (2019), etc. As multiple variables are documented to explain the cross-section of returns, econometric methods that perform variable selection are needed to find characteristics that best explain assets' betas.

Compounding the difficulty of understanding the risk-return tradeoff is the presence of economy-wide factor risk premia embedded in the stochastic discount factor. Factor risk prices may vary over time as functions of economy-wide variables; variables that are not asset specific characteristics. While performing a variable search over characteristics driving betas, asset pricing models also need to account for these global predictors. Examples of variables that may predict currency returns include the VIX, global credit spread, and the global slope of the yield curve.

We develop a Bayesian factor model based on regression trees that simultaneously selects the best characteristics that determine the asset's betas while also selecting the best global predictors for the factor risk prices. In a regression tree, data is sorted into groups (which we interpret as portfolios) by splitting the space of characteristics sequentially according to a binary decision tree, see Breiman et al. (1984). After the splitting process groups data into clusters (or portfolios), the conditional expectation function is approximated locally by the sample mean or a regression on the cluster-level data. In our application, we develop a regression tree model that searches over split points for assets' betas as a function of their characteristics. After groups/portfolios are formed, it approximates the conditional expectation function and covariance by a factor model at the group level. Our model also includes a flexible specification for the factor risk prices. We use a spike and slab prior to find the set of economy wide variables that drive aggregate time series predictability in returns.

While the clustering of assets into portfolios via regression trees is more involved than the standard portfolio sort procedure, there are several important benefits. First, a regression tree can form portfolios that are more flexible than a traditional portfolio sort used in finance. Second, each split of a regression tree only increases the total number of groups by one while each new split in a portfolio sort has a multiplicative impact. In a portfolio sort, the number of groups quickly explodes with each new split if more than one characteristic is used during the sort. And, the number of observations assigned to any group decreases dramatically. Consequently, very few characteristics and split points can be considered jointly during a portfolio sort, typically only one or two characteristics.

The methods developed in our paper contribute to a growing literature in finance and financial econometrics that develops factor models targeted for individual assets; see, e.g. Gagliardini et al. (2016) and Kelly et al. (2019). Factor models in finance are traditionally applied to portfolios sorted on characteristics, see Cattaneo et al. (2020). Information in the data is potentially thrown out. Traditional econometric methods applied to portfolios do not account for the variable selection problem used for creating the portfolios. While our model clusters assets into portfolios, the model must be defined for individual assets. Using individual assets is required to make Bayesian model comparisons across characteristics sensible. Similar to Kelly et al. (2019), our model explains assets' betas as a function of their characteristics. Our paper also complements a literature using machine learning techniques to flexibly model forecasting variables, select factors, or model expected returns non-parametrically including Gu et al. (2020a), Gu et al. (2020b), Gu et al. (2020b), and Feng et al. (2020). The idea of applying regression trees or more generally a random forest to asset returns was also introduced by Moritz and Zimmermann (2016).

Our model uses a prior over regression trees to select the characteristics that determine assets' betas. Bayesian regression trees, as developed by Chipman et al. (1998) and Denison et al. (1998), are a popular non-parametric statistics and machine learning method that can be viewed as the Bayesian analogue to the random forest, see Hastie et al. (2009). Bayesian regression trees are traditionally applied to cross-sectional data under an *i.i.d.* assumption conditional on the covariates. A notable exception is the time series model of Taddy et al. (2011). As a methodological contribution, we extend standard Bayesian regression tree models to unbalanced panels of data with a factor structure and conditional heteroskedasticity. There are multiple ways a regression tree can be defined for unbalanced panels. Our regression tree prior shares some similarities with Bayesian regression tree models from the statistics literature but we make modifications to connect it to the financial economics literature using portfolio sorts. Our prior uses a hierarchical structure to first place a distribution over the number P of portfolios or groups (nodes of the tree). It then performs binary splits of the characteristic space until P portfolios are formed. This prior over the regression tree shares similarities with the work of Denison et al. (1998) and Wu et al. (2007). There are two key differences between our prior and the methods in these papers. First, our prior assigns the probability of splitting any existing region into two new regions as a function of the number of data points within the region. Regions with more data are more likely to be split, which acts to regularize the estimator. Second, we incorporate the idea of placing a prior over the vector of characteristics that can be selected as in Linero (2018). We extend our initial model based on a single regression tree to a sum of trees model in the spirit of Chipman et al. (2010). Conceptually, this extension is straightforward but the economic interpretation of the sorting process in terms of forming portfolios is not as direct. Similar to Chipman et al. (1998) and Denison et al. (1998), we use reversible jump Markov chain Monte Carlo (MCMC) to estimate the model. We use a particle Gibbs sampler as in Creal and Tsay (2015) to estimate the stochastic volatility factors.

In our application, we investigate the risk-return trade-off of currency excess returns. We find that the interest rate differential and idiosyncratic FX volatility are the primary characteristics driving betas in the FX market. In contrast, a traditional portfolio sort that utilizes one variable at a time suggests that idiosyncratic FX volatility is not important at all while other characteristics significantly contribute to the cross sectional variation of currency excess returns. Another important result is that a higher Sharpe ratio than the carry trade strategy can be obtained. The high Sharpe ratio is due to the the predictive powers of the VIX, aggregate capital ratio of financial intermediaries, and global interest rate differential for the risk factors of the FX market. Also, our portfolio function offers profitable investment strategies with higher excess returns than that of the carry trade by constructing new portfolios sorted by the interest rate differential.

2 A classical model of portfolio sorts

In this section, we build a model of portfolio sorts where the number of portfolios being formed is fixed and the characteristics used for sorting are chosen by the researcher. Our goal is to establish a benchmark model for unbalanced panels of returns that mimics how portfolio sorts are used in the finance literature. We extend this model in Section 3 to include variable selection via regression trees.

2.1 Model

We assume an unbalanced panel of data with n_t assets at date t in the cross-section for $t = 1, \ldots, T$. At every date, each excess return r_{it} for $i = 1, \ldots, n_t$ is assigned to one of P portfolios based on an $L \times 1$ vector of observable characteristics $\mathbf{x}_{i,t-1}$ known at least one period ahead of the return.

We model the excess returns as

$$r_{it} = \sum_{p=1}^{P} \mathbb{1}\left(d_{it} = p\right)\left[\alpha_p + \boldsymbol{\beta}_p' \tilde{\mathbf{f}}_t\right] + \varepsilon_{it}, \quad (1)$$

$$(d_{1t}, \ldots, d_{n_t,t}) = \text{finance-sort}\left(\mathbf{q}_{1:L}, C_{1:L}, \mathbf{X}_{t-1}\right), \quad (2)$$

where the idiosyncratic shocks follow $\varepsilon_{it} \sim \mathrm{N}\left(0, \sigma_i^2\right)$ and d_{it} for $i = 1, \ldots, n_t$ are indicator variables describing to which one of the P portfolios that asset i is assigned. We let $\tilde{\mathbf{f}}_t$ denote a $K \times 1$ vector of demeaned risk factors with conditional mean $\mathbb{E}_{t-1}\left(\tilde{\mathbf{f}}_t\right) = 0$. The intercepts α_p and factor loadings $\boldsymbol{\beta}_p$ are assumed to be common parameters for those assets assigned to the same portfolio. The sorting function (2) takes the matrix of characteristics $\mathbf{X}_{t-1}$ as inputs and deterministically assigns the portfolio indicator variables given user-chosen tuning parameters $\mathbf{q}_{1:L} = (\mathbf{q}_1, \ldots, \mathbf{q}_L)$, $C_{1:L} = (C_1, \ldots, C_L)$. We describe the sort function (2) and it's tuning parameters used in the finance literature in detail in Section 2.2.

In this model, the factors $\tilde{\mathbf{f}}_t$ may be latent or observable. If they are latent, the factor loadings $\boldsymbol{\beta}_p$ and the covariance matrix of the factors $\tilde{\mathbf{f}}_t$ must be restricted to achieve identification; see e.g. Geweke and Zhou (1996). If the factors are observable, like the shocks to macroeconomic fundamentals as in Chen et al. (1986), then no restrictions on the loadings $\boldsymbol{\beta}_p$ or on the covariance matrix of $\tilde{\mathbf{f}}_t$ are required unless the factors are tradeable assets, i.e. a contemporaneous linear combination of returns. In this case, an internal consistency condition must be imposed on the model to ensure that the same linear combination of data is not used twice.

The model (1)-(2) captures two key ideas that originally motivated the use of portfolios by Black et al. (1972) and Fama and MacBeth (1973). First, the betas of individual assets are time-varying as a function of their characteristics through their portfolio indicators $(d_{1t}, \ldots, d_{n_t,t})$. And, second, the parameters of different assets are pooled together. This reduces the total number of parameters in the model and potentially improves the precision with which they are estimated.

2.2 Portfolio sort function

We follow Bali et al. (2016) in designing the portfolio sort function in equation (2), which they build off Black et al. (1972) and Fama and MacBeth (1973). For simplicity, we discuss the case where all the characteristics are continuous-valued. In a portfolio sort, the indicator variables d_{it} are deterministic functions of three inputs: (i) the observable characteristics $\mathbf{X}_{t-1}$; (ii) the number of regions that each characteristic is split into C_ℓ for $\ell = 1, \ldots, L$; (iii) the vectors of quantile values $\mathbf{q}_\ell = (q_{\ell,0}, q_{\ell,1}, \ldots, q_{\ell,C_\ell})'$ for $\ell = 1, \ldots, L$ that determine the breakpoints for separating assets into different regions. The total number of portfolios being formed is $P = \prod_{\ell=1}^{L} C_\ell$. As is standard in finance, the number of portfolios P is constant through time. And, we follow the convention that the first and last quantiles are defined as $q_{\ell,0} = 0$ and $q_{\ell,C_\ell} = 1$, respectively. By default, each value C_ℓ must be greater than or equal to one. If $C_\ell = 1$, then the conditional mean of returns does not change as a function of the ℓ-th characteristic.

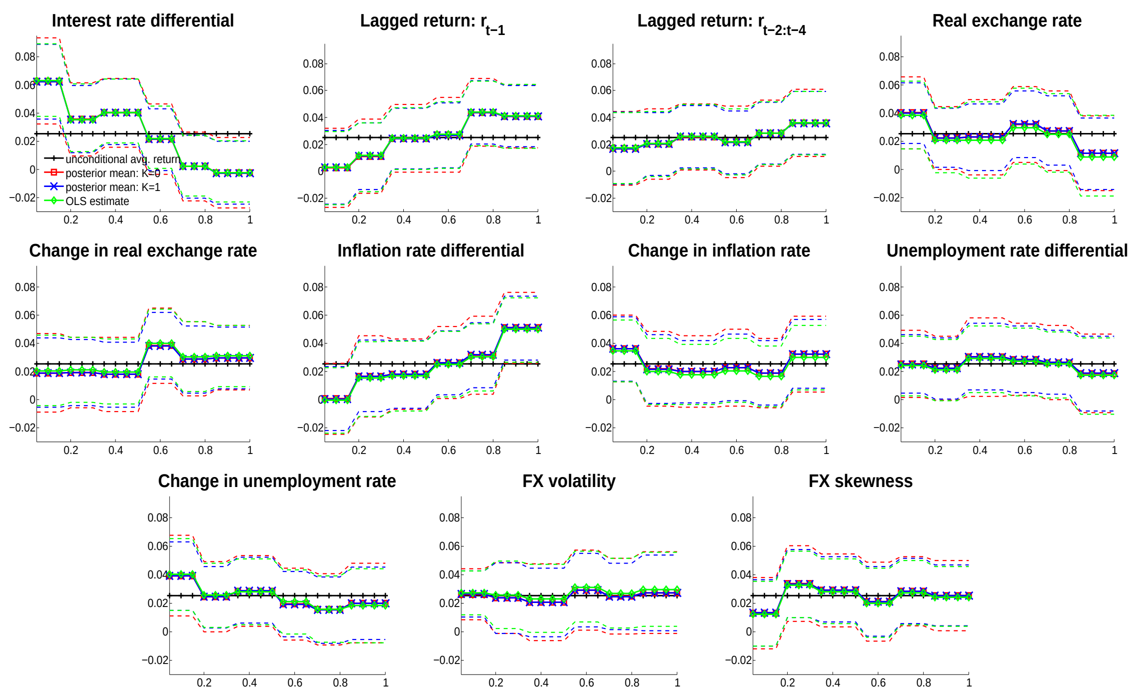
Let $\mathcal{X}_t = \mathbb{R}^L$ denote the space of characteristics at date t, which we assume is the same across time. The portfolio sort function (2) breaks $\mathcal{X}_t$ into non-overlapping sub-regions $\mathcal{X}_t = \mathcal{X}_{1t} \cup \mathcal{X}_{2t} \cup \ldots \cup \mathcal{X}_{Pt}$. The breakpoints are found by separately inverting the empirical cumulative distribution function of each characteristic at each of the pre-fixed quantile values $c_{\ell,t,j} = \widehat{F}_{\ell,t}^{-1}(q_{\ell,j})$ for $j = 0, \ldots, C_\ell$. By construction, the first and last breakpoints are always $c_{\ell,t,0} = -\infty$ and $c_{\ell,t,C_\ell} = \infty$ since $q_{\ell,0} = 0$ and $q_{\ell,C_\ell} = 1$. In finance, the vector $\mathbf{q}_\ell$ is typically chosen as a set of evenly spaced values. For example, with $L = 1$ characteristic and with $P = 5$ portfolios, a standard choice for the quantiles would be $\mathbf{q}_1 = (0, 0.2, 0.4, 0.6, 0.8, 1)'$. With $L = 2$ characteristics and with $C_1 = 3$ and $C_2 = 2$, there are $P = 6$ portfolios and a standard choice for the quantiles would be $\mathbf{q}_1 = (0, 1/3, 2/3, 1)'$ and $\mathbf{q}_2 = (0, 0.5, 1)'$. The indicator variables d_{it} are then assigned using the breakpoints as

$$d_{it} \quad = \quad j, \qquad \text{if } \mathbf{x}_{i,t-1} \in \mathcal{X}_{j,t-1} \qquad j = 1, \ldots, P \tag{3}$$

where each sub-region $\mathcal{X}_{j,t-1}$ is defined as

$$\mathcal{X}_{j,t-1} \quad = \quad \cap_{\ell=1}^{L} \left\{ c_{\ell,t-1,j_\ell^*-1} < x_{\cdot,\ell,t-1} \leq c_{\ell,t-1,j_\ell^*} \right\} \tag{4}$$

Figure 1: Expected returns: classical portfolio sort



The dashed lines are 5% and 95% robust confidence intervals for OLS estimates and 5% and 95% percentiles for the Bayesian posterior distributions. The univariate portfolio sort is performed by equally spaced split points to make 6 portfolios. $(L = 1, P = 6)$

We use $x_{\cdot,\ell,t-1}$ to denote characteristic ℓ at date $t-1$ and j_ℓ^* equal to the quantile value of variable ℓ used to define portfolio j.

A defining feature of the portfolio sort function is that a relatively low-dimensional set of choice variables $\{\mathbf{q}_\ell, C_\ell\}_{\ell=1}^L$ are applied to all the dates and determine all the breakpoints. Only one or two characteristics at most are typically considered ($L = 1, 2$) in a portfolio sort. This is a key feature that we attempt to overcome in Section 3 by developing a new Bayesian model with a regression tree prior.

2.3 Application to currency excess returns

We implement the model in equations (1)-(2) with known characteristics and compare it to the standard portfolio sorting procedure as in Bali et al. (2016). We make two points. First, the portfolio sort model (1)-(2) estimated on individual assets gives similar estimates as the standard approach that conditions only on P portfolios. And, several different characteristics look to be important when portfolio sorts are performed one variable at a time.

We analyze the log monthly excess return to an investor who borrows U.S. dollars at the U.S. risk-free rate and buys a foreign risk-free asset. Our unbalanced panel data contains a total of 47 countries plus the Euro area with the cross-section size n_t ranging from a minimum of 9 to a maximum of 34 assets. In this exercise, 6 currency portfolios are formed based on each of 11 country-specific characteristics in the data set, i.e., $L = 1$ and $P = 6$. The 11 country-specific characteristics are the interest rate differential between the U.S. and a foreign counterparty country, two past (1 month and 3 month) returns, the real exchange rate, the change in the real exchange rate, the cross-country unemployment rate differential and its change; the cross-country inflation rate differential and its change; idiosyncratic FX volatility and skewness. Further details about the sources of the data and how each variable is constructed can be found in Section 5.1 and the online appendix.

First, we construct $P = 6$ portfolios using univariate sorts for 11 separate characteristics, which produce a balanced panel. The unconditional return of each portfolio is estimated as the time series mean with confidence intervals calculated from Newey-West standard errors. With the same portfolio-level data, we compare the frequentist unconditional returns to a simple Bayesian linear model with a constant, normal errors, and no factors ($K = 0$). Lastly, we estimate the unconditional returns from the Bayesian model defined in (1)-(2). Instead of using only the P portfolios, this factor model utilizes the entire unbalanced panel of all individual currency returns. We assume one latent factor ($K = 1$) in the model and estimate it by imposing the identifying restrictions from Geweke and Zhou (1996).

In Figure 1, the posterior means from the model (1)-(2) and the OLS estimates that condition on P portfolios are almost identical in the sense that the same conclusions would have been drawn for each characteristic variable. Four characteristics out of 11 appear to have a strong monotonic relationship with the unconditional portfolio return. The interest rate differential has the strongest relationship. For the portfolio with the lowest interest rate differential (i.e., relatively high foreign interest rates compared to the U.S.), the return is estimated to be 6% per annum. The one month lagged return, the real exchange rate, and the inflation rate differential also display meaningful monotonic relationships. Additionally, the portfolios sorted by the lagged 3 month return and the change in unemployment rates also show similar monotonic but much weaker associations with those characteristics. The fact that many predictors appear to be important motivates our main econometric problem that we solve in Section 3. In Section 5, we show that characteristics that do not appear to be important for univariate sorts can be important when the variable selection problem is jointly estimated.

3 Models that search over characteristics

In this section, we build a factor model with time-varying factor risk premia for unbalanced panels of returns. We use regression tree priors to stochastically select the best characteristics for explaining movements in assets' betas.

3.1 Factor models with regression tree priors

The excess return of an asset r_{it} is modeled by the conditionally linear factor model

$$
\begin{aligned}
r_{it} &= \sum_{p=1}^{P} \mathbb{1}\left(d_{it}=p\right)\left[\alpha_{p,t}+\boldsymbol{\beta}_p'\left(\mathbf{f}_t-\boldsymbol{\mu}_{f,t}\right)\right]+\varepsilon_{it} \qquad \varepsilon_{it} \sim \mathrm{N}\left(0,\sigma_{it}^2\right), \qquad (5)\\
\mathbf{f}_t &= \boldsymbol{\mu}_{f,t}+\mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathrm{N}(0,\boldsymbol{\Sigma}_{f,t}). \qquad (6)
\end{aligned}
$$

where $\mathbf{f}_t$ denotes a $K \times 1$ vector of risk factors with conditional mean $\mathbb{E}_{t-1}\left(\mathbf{f}_t\right) = \boldsymbol{\mu}_{f,t}$ and conditional covariance matrix $\mathbb{V}_{t-1}\left(\mathbf{f}_t\right) = \boldsymbol{\Sigma}_{f,t}$; $\boldsymbol{\beta}_p = [\beta_{1,p},\ldots,\beta_{K,p}]'$ is a $K \times 1$ vector of betas of portfolio p; d_{it} is a portfolio index variable determined from L asset characteristics in $\mathbf{X}_{t-1}$.

Let $r_{f,t}$ denote the one-period risk-free rate. Consider a one-period stochastic discount factor

$$M_{t,t+1} = -\frac{1}{r_{f,t}} - \frac{1}{r_{f,t}} \mathbf{b}_t (\mathbf{z}_t)' (\mathbf{f}_{t+1} - \boldsymbol{\mu}_{f,t+1}) . \tag{7}$$

that is assumed to price all assets. Here, the factor risk prices $\mathbf{b}_t (\mathbf{z}_t)$ may be functions of a $Z \times 1$ vector of observable variables $\mathbf{z}_t$. We write the dynamics of $\mathbf{b}_t (\mathbf{z}_{t-1})$ as

$$\mathbf{b}_{t+1} (\mathbf{z}_t) = \mathbf{b}_0 + \mathbf{b}_z \mathbf{z}_t + \mathbf{b}_{z^2} \mathbf{z}_t^2 \tag{8}$$

where $\mathbf{z}_t^2 = (\mathbf{z}_t \odot \mathbf{z}_t)$. We include $\mathbf{z}_t^2$ in (8) to consider any non-linear relationships between $\mathbf{b}_{t+1}$ and $\mathbf{z}_t$.

The conditional moment condition from asset pricing theory

$$\mathbb{E}_t [M_{t,t+1} r_{i,t+1}] = 0 \tag{9}$$

together with the stochastic discount factor in (7) imposes restrictions on the intercepts $\alpha_{p,t}$ in the model. By assuming the stochastic discount factor in (7) is capable of pricing all assets and the risk factors $\mathbf{f}_t$ are tradable portfolios, one can show that

$$\alpha_{p,t} = \boldsymbol{\beta}_p' \boldsymbol{\lambda}_t (\mathbf{z}_{t-1}) , \tag{10}$$

$$\boldsymbol{\mu}_{f,t} = \boldsymbol{\lambda}_t (\mathbf{z}_{t-1}) = \boldsymbol{\Sigma}_{f,t} \mathbf{b}_t (\mathbf{z}_{t-1}) , \tag{11}$$

where $\boldsymbol{\lambda}_t (\mathbf{z}_{t-1})$ are the $K \times 1$ vector of time-varying factor risk premia. See Section 1 of the online appendix for the derivation. These factor risk premia create an additional source of forecastability for returns, separate from the characteristics $\mathbf{X}_{t-1}$. Unlike the characteristics $\mathbf{X}_{t-1}$ that determine the assets' betas, the variables $\mathbf{z}_{t-1}$ driving economy wide risk are not asset specific.

We use a conjugate normal prior over the factor loadings

$$\boldsymbol{\beta}_p \sim \mathrm{N} \left(\underline{\boldsymbol{\mu}}_\beta , \underline{\mathbf{V}}_\beta \right) , \tag{12}$$

which allows us to marginalize out $\boldsymbol{\beta}_p$ in its joint density with the data. To ensure the prior covers a plausible range of values for $\boldsymbol{\beta}_p$, we use OLS estimates to set its prior-hyper parameters as

$$\underline{\boldsymbol{\mu}}_{\beta} = \widehat{\boldsymbol{\beta}}_{OLS}, \qquad \underline{\mathbf{V}}_{\beta} = NT \times \widehat{\mathbf{V}}_{\boldsymbol{\beta},OLS} \tag{13}$$

where $NT = \sum_{t=1}^{T} n_t$ is the total number of observations in the unbalanced panel data; $\widehat{\boldsymbol{\beta}}_{OLS}$ is the OLS estimate of $\boldsymbol{\beta}$ and $\widehat{\mathbf{V}}_{\boldsymbol{\beta},OLS}$ is the variance estimate of $\widehat{\boldsymbol{\beta}}_{OLS}$ from the linear regression

$$r_{it} = \beta_0 + \boldsymbol{\beta}'\mathbf{f}_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim i.i.d(0, \sigma^2). \tag{14}$$

We multiply $\widehat{\mathbf{V}}_{\boldsymbol{\beta},OLS}$ by NT to imitate the amount of estimation uncertainty for a hypothetical case where only one observation is used during OLS estimation. This prior covers a wide range of values for $\boldsymbol{\beta}$. Because the prior distributions over other model parameters except d_{it} are standard, we provide them in Appendix A.

We introduce stochastic volatility to the idiosyncratic errors ε_{it} as

$$\log\left(\sigma_{it}^2\right) = \mu_{\sigma,i} + \phi_{\sigma,i}\left(\log\left(\sigma_{i,t-1}^2\right) - \mu_{\sigma,i}\right) + \eta_{\sigma,i,t}, \qquad \eta_{\sigma,i,t} \sim \mathrm{N}\left(0, \sigma_{\sigma,i}^2\right). \tag{15}$$

The time-varying covariance matrix for the factor error $\mathbf{u}_t$ is assumed to follow

$$\boldsymbol{\Sigma}_{f,t} = \mathbf{M}_t^{-1}\mathbf{G}_t\mathbf{M}_t, \tag{16}$$

where

$$\mathbf{M}_t = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{2,1,t} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{K,1,t} & m_{K,2,t} & \dots & 1 \end{bmatrix} \quad \mathbf{G}_t = \begin{bmatrix} g_{1,t} & 0 & \dots & 0 \\ 0 & g_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g_{K,t} \end{bmatrix},$$

$$\log\left(g_{k,t}\right) = \mu_{g,k} + \phi_{g,k}\left(\log\left(g_{k,t-1}\right) - \mu_{g,k}\right) + \eta_{g,k,t}, \qquad \eta_{g,k,t} \sim \mathrm{N}\left(0, \sigma_{g,k}^2\right),$$

for $k = 1, 2, ..., K$. Each latent variable in $\mathbf{M}_t$ follows a random walk process with error $e_{k,\ell,t} \sim \mathrm{N}\left(0, \sigma_{e,k,\ell}^2\right)$ for $k = 2, 3, ..., K$ and $\ell = 1, 2, .., k-1$.

We employ a spike and slap prior over $\mathbf{b}_z$ and $\mathbf{b}_{z^2}$ in (8) to select relevant predictors for the time-varying risk prices. Let $\widetilde{\mathbf{b}} = [\mathbf{b}_z, \mathbf{b}_{z^2}]$ denote a $K \times 2Z$ matrix of the risk price coefficients. We write $b_{k,z}$ for each coefficient in $\widetilde{\mathbf{b}}$ with the subscripts denoting the k-th factor and the z-th variable. For $k = 1, \ldots, K$ and $z = 1, \ldots, 2Z$, the prior over $b_{k,z}$ is given by the following mixture of distributions:

$$b_{k,z} \sim s_{k,z} \mathrm{N}\left(0, \underline{v}_b\right) + (1 - s_{k,z}) \mathbb{1}\left(b_{k,z} = 0\right) \quad (17)$$

$$s_{k,z} \sim \text{Bernoulli}\left(\pi_{s,k}\right) \quad (18)$$

$$\pi_{s,k} \sim \text{Beta}\left(\underline{\gamma}_{s,1}, \underline{\gamma}_{s,2}\right) \quad (19)$$

where $\mathbb{1}\left(b_{k,z} = 0\right)$ is a function that takes 1 if $b_{k,z} = 0$ and takes 0 otherwise.

3.2 Tree Prior

Our prior distribution over the portfolio indicator variables d_{it} is built through a regression tree. First, we provide an intuitive algorithmic description of the prior because the formal probabilistic model described below is notationally complicated. The tree prior sequentially splits the characteristic space $\mathcal{X}_t$ for $t = 1, \ldots, T$ into P non-overlapping sub-regions. It starts by drawing one of the L characteristics $x_{\cdot, t-1, \ell}$ and randomly splits this variable at one of a finite number of pre-set quantile values $\mathbf{q}_\ell = \left\{q_{\ell,1}, \ldots, q_{\ell,\overline{Q}_\ell}\right\}$ where $\overline{Q}_\ell$ is the total number of the quantile values that can be selected for variable ℓ. The set of values $\mathbf{q}_\ell$ are chosen by the user to equally sub-divide the $[0, 1]$ interval with the end points set at $q_{\ell,1} > 0$ and $q_{\ell,\overline{Q}_\ell} < 1$. The prior is a uniform distribution whose support is the quantiles $q_{\ell,1}$ through $q_{\ell,\overline{Q}_\ell}$. Suppose the quantile $q_{\ell,j}$ was drawn. Then, the prior inverts the empirical distribution function to obtain the breakpoints $c_{\ell,t,j} = \widehat{F}_{\ell,t}^{-1}\left(q_{\ell,j}\right)$ for $t = 1, \ldots, T$ similar to the procedure used in Section 2.2. Given the breakpoints, this splits the characteristic space $\mathcal{X}_t$ at every date t into two non-overlapping sub-regions $\mathcal{X}_{1t}$ and $\mathcal{X}_{2t}$. At the first iteration or decision node, there are only two sources of randomness, which are the characteristic $x_{\cdot, \ell, t-1}$ and the quantile value $q_{\ell,j}$ selected for splitting.

The algorithm then proceeds sequentially in three steps:

(i) randomly draw an existing sub-region (node of the tree) to break into two new sub-regions;

(ii) draw one of the L characteristics $x_{\cdot,\ell,t-1}$ along which to split the existing region;

(iii) given the chosen sub-region and characteristic, draw one of the quantile values $q_{\ell,j}$ for splitting the sub-region into two new regions.

The process continues for $j = 1, \ldots, P-1$ until P sub-regions are created. From these, we can assign the indicators d_{it} as a deterministic function of the break points and the empirical CDF's. The name 'regression tree' comes from the fact that the sequential splitting process can be depicted as a binary decision tree.

A key feature of our model is that we are placing a prior on a relatively low dimensional set of variables to determine the value of a higher dimensional set of portfolio indicators. This is similar to the standard portfolio sort model used in finance as in Section 2.2. While the portfolio sort function in Section 2.2 forms a hyper-lattice, a regression tree forms a set of "hyper-rectangles." A regression tree can form portfolios that are more irregular and potentially more flexible than a traditional portfolio sort used in finance. Statistically, each split of a regression tree only increases the total number of regions by one while each additional split of the portfolio sort function (2) is multiplicative. With each split, the number of parameters that get added and the number of data points per sub-region is substantially different across the two methods. Another key difference between the two methods is that in a regression tree the order in which variables are split matters. Conversely, for the portfolio sort method used in finance, the order does not matter.

Our prior over the indicator variables d_{it} has three main goals. First, many finance researchers may opt to fix P at a given value. Therefore, we start by placing a prior over P and then forming a tree consistent with this value, as in Denison et al. (1998) and Wu et al. (2007). A user can fix P a priori if desired.

Second, we want the prior to split the characteristic space more heavily in regions where there is more data. Informally, such a prior can regularize the overall estimator because it lowers the probability of breaking the characteristic space $\mathcal{X}_t$ into many sub-regions with vary few data points in them. To do this, we assign the probability of splitting a region proportional to the amount of data in that region. This feature of our tree prior is different from existing tree priors in the literature. Finally, we want to assign higher/lower probability to those variables we think are more/less important. We introduce a vector $\boldsymbol{\pi}_x$ of probabilities associated with each of the characteristics that we estimate. The idea of introducing the vector $\boldsymbol{\pi}_x$ of selection probabilities is similar to Linero (2018).

To form P portfolios, there must be a total of $P-1$ splits or decisions in the tree. These are determined sequentially for $j = 1, \ldots, P-1$. Conditional on the covariates $\mathbf{X}_{0:T-1}$, the prior is naturally defined as a sequence of discrete random variables whose support is the set of indices that determine the regions/variables/quantiles where splits can occur. As the splitting process evolves, the prior must keep track of the set of possible splits that are feasible at each iteration and update the probabilities of each possible split combination. If the splitting process were totally unconstrained, some potential split combinations could lead to regions with no data contained in them. For Bayesian analysis, this implies that the only information about the conditional mean in this region comes from the prior. Many researchers may prefer to guarantee a minimum number of observations in each portfolio, which is common for Bayesian tree models; see, e.g. Chipman et al. (1998). When constraints like these are imposed, the probabilities of selecting different splits must adapt across iterations to reflect the constraints. Consequently, the prior needs to calculate the probabilities of splitting each region/variable/quantile recursively because they depend on the splits that have occurred up to that point. Our method allows for such a restriction. For example, in our empirical application with currency excess returns, we impose a restriction that the overall proportion that all portfolios contain at least one observation at time t should be larger than 0.95 over the entire sample period.

As the outcome space is naturally discrete, we place a prior over the indices to each of the relevant objects. We denote the set of indices to the regions which can be split at iteration j as $\mathcal{I}_j^r$ for $j = 1, \ldots, P-1$. The set $\mathcal{I}_j^r$ is indexed by j because it must increase over the iterations as more splits occur. After splitting the regions at iteration $P-1$, there are a total of P regions. The P regions are indexed by $\mathcal{I}^r = \mathcal{I}_P^r$. We also denote the set of indices to all intermediate regions generated by the splitting process as $\mathcal{I}_j^m$ for $j = 1, \ldots, P$. These auxiliary indices allow us to keep track of the entire history of the splitting process and are useful when implementing our proposed MCMC algorithm. After generating the P regions, we have $\mathcal{I}^m = \mathcal{I}_P^m$.

We denote by $\mathcal{I}^x$ and $\mathcal{I}^q_\ell$ for $\ell = 1, \ldots, L$ the set of indices to the characteristics $\mathbf{x}_{\ell,t-1}$ and quantiles $\mathbf{q}_\ell$ that can be chosen. The probabilities of selecting individual elements of these sets will change across iterations. The set $\mathbf{q}_\ell$ is chosen a-priori by the researcher to split up the $[0, 1]$ interval. In theory, we could place a continuous uniform prior over $[0, 1]$ and draw the quantile accordingly. As each characteristic's empirical distribution function is discrete from conditioning on $\mathbf{X}_{0:T-1}$, we subdivide the $[0, 1]$ interval using a finite number of values. This makes the MCMC algorithm easier and more efficient to implement. A natural choice is to split the $[0, 1]$ interval into evenly spaced sub-intervals of size $1/(\overline{Q}_\ell + 1)$. For example, we split the interval as $q_{\ell,0} = 0$ and $q_{\ell,j} = q_{\ell,j-1} + \frac{1}{\overline{Q}+1}$ until $q_{\ell,\overline{Q}_\ell+1} = 1$. This process generates $\overline{Q}_\ell$ quantile values that can be selected in estimation. The last quantile value $q_{\ell,\overline{Q}_\ell} = \frac{\overline{Q}_\ell}{\overline{Q}_\ell+1}$ is always less than 1. There are two key reasons for choosing an a-priori set of values for $\mathbf{q}_\ell$. First, it is comparable to how portfolio sorts are conducted in finance. Secondly, it allows the prior to be well-defined for an unbalanced panel of data, as the same split point will be applied to all dates.

Formally, our prior distribution is a joint distribution over the number of portfolios formed P, the indices into the region $i^r \in \mathcal{I}^r_j$ that are split, the indices of the characteristics $i^x \in \mathcal{I}^x$ that are split, and the indices of the quantiles $i^q_\ell \in \mathcal{I}^q_\ell$ that determine the split point. This joint distribution can be decomposed as

$$p\left(i^r_{1:P-1}, i^x_{1:P-1}, i^q_{1:P-1}, P|\mathbf{X}_{0:T-1}\right) = p\left(i^r_{1:P-1}, i^x_{1:P-1}, i^q_{1:P-1}|P, \mathbf{X}_{0:T-1}\right) p\left(P|\mathbf{X}_{0:T-1}\right)$$

The indicator variables d_{it} are deterministic functions of these indices. For the remainder of this section, we drop dependence on the covariates $\mathbf{X}_{0:T-1}$.

We start with a distribution over the number of portfolios:

$$P \sim 1 + \text{binomial}\left(\underline{p} - 1, \pi_p\right) \tag{20}$$

where $\underline{p}$ is a user-specified maximum number of portfolios to form and π_p controls the expected number of portfolios.

Given a draw of P from (20), the prior draws split points for $j = 1, \ldots, P - 1$. At the initial iteration $j = 1$, the prior starts with the initial index to the region $i^r = 0$.

The portfolio indicators are initialized as $d_{it} = 1$ for all i, t. It then draws an index to one of the characteristics

$$i^x \sim \text{Categorical}\left(\boldsymbol{\pi}_x | \mathcal{I}^x\right)$$

where $\pi_{x,\ell}$ is the probability of selecting characteristic ℓ and $\boldsymbol{\pi}_x = (\pi_{x,1}, \ldots, \pi_{x,L})'$. Conditional on the variable selected, the prior draws an index uniformly from the set of feasible choices

$$i^q \sim \text{Categorical}\left(\boldsymbol{\pi}_{q,\ell} | i^x = \ell, \mathcal{I}_\ell^q\right)$$

with $\boldsymbol{\pi}_{q,\ell}$ a $\overline{Q}_\ell \times 1$ vector with equal probabilities.

At the end of the iteration, the prior updates the set of indices to existing regions by creating new sets, $\mathcal{I}_2^r$ and $\mathcal{I}_2^m$. The first index set $\mathcal{I}_2^r$ includes two new indices

$$\begin{aligned} \bar{i}^r &= 2i^r + 1 \\ \underline{i}^r &= 2i^r + 2. \end{aligned}$$

The old index i^r which indicates which region is split previously is included in $\mathcal{I}_2^m$. We follow the simple method of indexing regions developed by Wu et al. (2007). This index variable is helpful during our proposed MCMC algorithm.

With $i^x = \ell$ and $i^q = k$, we calculate the breakpoints $c_{\ell,t-1} = \widehat{F}_{\ell,t-1}^{-1}(q_{\ell,k})$ and update the portfolio indicators as

$$d_{it} = \begin{cases} d_{it}, & \text{if} \quad x_{i,\ell,t-1} \leq c_{\ell,t-1} \\ j+1, & \text{otherwise} \end{cases}$$

All observations are assigned to one of two portfolios. After drawing $\{i^r = 0, i^x = \ell, i^q = k\}$, we set $i_1^r = 0$, $i_1^x = \ell$, and $i_1^q = k$.

We use $\mathcal{D}_j$ to denote the sigma algebra that contains all previous split decisions up to and including iteration j. Then, the prior continues for $j = 2, \ldots, P-1$ with the following steps

- **Draw a region to split:** Draw an index $i^r \in \mathcal{I}_j^r$ to a region that will be split into two new regions

$$i^r \sim \text{Categorical}\left(\boldsymbol{\pi}_r | \mathcal{I}_j^r, \mathcal{D}_{j-1}\right)$$

The vector of probabilities $\boldsymbol{\pi}_r$ is determined as follows. Let n_{i^r} equal the number of observations in the region with index $i^r \in \mathcal{I}_j^r$. The probability of splitting any individual region is proportional to the number of data points in that region

$$\pi_{r,i^r} = \frac{n_{i^r}}{NT}$$

- **Draw a characteristic to split:** Conditional on splitting region $i^r = r$, draw an index $i^x \in \mathcal{I}^x = \{1, 2, ..., L\}$ to one of the characteristics

$$i^x \sim \text{Categorical}\left(\bar{\boldsymbol{\pi}}_x | \mathcal{I}^x, i^r = r, \mathcal{D}_{j-1}\right) \tag{21}$$

Let $\bar{\pi}_{x,i^x}$ denote the i^x-th element of $\bar{\boldsymbol{\pi}}_x$ which represents the probability of choosing variable i^x. We assume that $\bar{\pi}_{x,i^x}$ depends on the conditional information $\{i^r = r, \mathcal{D}_{j-1}\}$

$$\bar{\pi}_{x,i^x} \propto \pi_{x,i^x} \mathbb{1}\left(i^x | i^r = r\right) \tag{22}$$

where $\mathbb{1}\left(i^x | i^r\right)$ is an indicator function that is equal to one if at least one split quantile value is feasible for variable i^x in region i^r, i.e. if any quantile values of variable i^x lead to new regions with the minimum observation count satisfied. The probability π_{x,i^x} is the unconditional probability of choosing variable i^x.

- **Draw a split point:** Conditional on splitting region $i^r = r$ and characteristic $i^x = \ell$, draw an index $i^q \in \mathcal{I}_\ell^q$ to one of the quantile values

$$i^q \sim \text{Categorical}\left(\bar{\boldsymbol{\pi}}_q | i_j^x = \ell, i_j^r = r, \mathcal{I}_\ell^q, \mathcal{D}_{j-1}\right) \tag{23}$$

where

$$\bar{\pi}_{q,i^q} \propto \mathbb{1}\left(i^q | i^x = \ell, i^r = r\right) \tag{24}$$

The probability $\bar{\pi}_{q,i^q}$ is the i^q-th element of $\bar{\boldsymbol{\pi}}_q$. This is a conditionally uniform distribution, where any indices that lead to two new regions that violate the required minimum observation count have zero probability. After drawing $\{i^r = r, i^x = \ell, i^q = k\}$, we set $i^r_j = r$, $i^x_j = \ell$, and $i^q_j = k$. We also update $\mathcal{D}_j$ accordingly for the next iteration.

- **Update the indices:** Update the set of indices to regions that can be split at the next iteration $\mathcal{I}^r_{j+1}$. Specifically, we start with $\mathcal{I}^r_j \subset \mathcal{I}^r_{j+1}$, add to it two new indices

$$\begin{aligned} \bar{i}^r &= 2i^r + 1 \\ \underline{i}^r &= 2i^r + 2 \end{aligned}$$

and delete the current index i^r. We also update $\mathcal{I}^m_{j+1}$ as $\mathcal{I}^m_{j+1} = \{\mathcal{I}^m_j, i^r\}$. With $i^x = \ell$ and $i^q = k$, we calculate the breakpoints $c_{\ell,t-1} = \widehat{F}^{-1}_{\ell,t-1}(q_{\ell,k})$ and update the portfolio indicators for those observations within the region that was split as

$$d_{it} = \begin{cases} d_{it}, & \text{if} \quad x_{i,\ell,t-1} \le c_{\ell,t-1} \\ j+1, & \text{otherwise} \end{cases}$$

At the end of the j-th iteration, all observations are assigned to one of $j+1$ portfolios.

At the end of the prior, each observation is assigned to one of P non-overlapping regions via the indicators d_{it}. There are also two sets of indices, $\mathcal{I}^r_P$ and $\mathcal{I}^m_P$. $\mathcal{I}^r_P$ represents regions that contain P portfolios and $\mathcal{I}^m_P$ represents $P-1$ regions that contain decision rules. The constraints on the probabilities (22)-(24) ensure that no combination of indices $\{i^r, i^x, i^q\}$ is selected that leads to new two regions being created without a minimum number of observations in them. As the cross-section sample sizes n_t grow, these constraints are less likely to bind meaning that draws from (21) will have probability $\boldsymbol{\pi}_x = \{\pi_{x,1}, \pi_{x,2}, ..., \pi_{x,L}\}'$ and (23) is a uniform distribution with $\boldsymbol{\pi}_q = \{\frac{1}{Q_l}, \frac{1}{Q_l}, ..., \frac{1}{Q_l}\}$.

In the proposed regression tree model (or the portfolio sort function in Section 2.2), the conditional mean is only a function of the rank statistics of the characteristics $\mathbf{X}_{t-1}$. This implies that two different characteristics $\mathbf{x}_{\cdot,\ell_1,t-1}$ and $\mathbf{x}_{\cdot,\ell_2,t-1}$ that result in the same ranking for all time periods are equivalent predictors. This rules out creating "new" characteristics from existing characteristics through monotonic transformations that do not re-order the variables.

3.3 Model extension

We extend the model in the previous section by incorporating a sum-of-trees prior and systematic pricing errors. As shown by Chipman et al. (2010), the sum-of-trees prior can add considerable flexibility to the conditional expectation function of r_{it} defined by the single tree prior. However, a direct comparison of the sum-of-trees prior with the classical portfolio sorts is no longer straightforward. In this model extention, we also consider systematic pricing errors that are not captured by the stochastic discount factor. This extended model allows us to directly estimate the pricing errors of individual returns.

We begin by elaborating on the basic form of the conditionally linear factor model based on the sum-of-trees prior. Suppose that there are H independent regression trees each of which is denoted by $\mathcal{T}_\tau\left(d_{\tau,i,t}, \tilde{\boldsymbol{\beta}}_\tau\right)$. The excess return is modeled by the sum of the H independent regression trees as

$$\begin{aligned}
r_{it} &= \sum_{\tau=1}^{H} \mathcal{T}_\tau\left(d_{\tau,i,t}, \tilde{\boldsymbol{\beta}}_\tau\right) + \varepsilon_{it} \qquad \varepsilon_{i,t} \sim \mathrm{N}\left(0, \sigma_{it}^2\right), \qquad (25)\\
\mathcal{T}_\tau\left(d_{\tau,i,t}, \tilde{\boldsymbol{\beta}}_\tau\right) &= \sum_{p=1}^{P_\tau} \mathbb{1}\left(d_{\tau,i,t} = p\right)\left[\alpha_{\tau,p,t} + \boldsymbol{\beta}_{\tau,p}'\left(\mathbf{f}_t - \boldsymbol{\mu}_{f,t}\right)\right]\\
\alpha_{\tau,p,t} &= \beta_{0,\tau,p} + \boldsymbol{\beta}_{\tau,p}'\boldsymbol{\Sigma}_{f,t}\mathbf{b}_t\left(\mathbf{z}_{t-1}\right)
\end{aligned}$$

where P_τ is the number of regions generated by the τ-th regression tree, $\tilde{\boldsymbol{\beta}}_\tau = \{\boldsymbol{\beta}_{0,\tau}, \boldsymbol{\beta}_\tau\}$, $\boldsymbol{\beta}_{0,\tau} = \{\beta_{0,\tau,1}, ..., \beta_{0,\tau,P_\tau}\}$, and $\boldsymbol{\beta}_\tau = \{\boldsymbol{\beta}_{\tau,1}, ..., \boldsymbol{\beta}_{\tau,P_\tau}\}$. We designate $\boldsymbol{\beta}_{\tau,p} = [\beta_{1,\tau,p}, \ldots, \beta_{K,\tau,p}]'$ to a $K \times 1$ vector of betas in portfolio p of regression tree τ. In this model, different regression trees can explain different portions of the data, which is the reason why the overall fitting capacities of the model are significantly improved. Different from the restricted model in equation (10) which assumes that the conditional expected return is an exact function of the factor loadings and the factor risk premia, an additional term is introduced to capture the pricing errors. If $\sum_{\tau=1}^{H} \mathbb{1}\left(d_{\tau,i,t}=p\right)\beta_{0,\tau,p}$ varies with any asset characteristics, we can conclude that the factor loadings and the factor risk premia derived by the stochastic discount factor are not enough to explain the cross sectional variation in the unbalanced panel of returns.

We assume that all the regression trees share a common parameter vector $\boldsymbol{\pi}_x$ but have independent portfolio decision rules along with region count P_τ. The same tree prior in the previous section is applied to each regression tree. According to this setup, the posterior distribution of $\boldsymbol{\pi}_x$ is decided by how frequently individual characteristics are selected in total in all the decision trees.

We control for over-fitting that may arise in the sum-of-trees model by imposing the following shrinkage prior over the factor loadings

$$\tilde{\boldsymbol{\beta}}_{\tau,p} \sim \mathrm{N}\left(\underline{\boldsymbol{\mu}}_\beta/H, \underline{\mathbf{V}}_\beta/H\right) \tag{26}$$

where $\tilde{\boldsymbol{\beta}}_{\tau,p}=[\beta_{0,\tau,p},\beta_{1,\tau,p},\ldots,\beta_{K,\tau,p}]'$. We set $\underline{\boldsymbol{\mu}}_\beta=\widehat{\tilde{\boldsymbol{\beta}}}_{OLS}$ and $\underline{\mathbf{V}}_\beta=NT\times\widehat{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}}$ where $\widehat{\tilde{\boldsymbol{\beta}}}_{OLS}$ is the OLS estimate of $\tilde{\boldsymbol{\beta}}_{\tau,p}$ and $\widehat{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}}$ is the variance estimate of $\widehat{\tilde{\boldsymbol{\beta}}}_{OLS}$ from the linear factor model in equation (14). It is a counterpart of the prior used by the single tree model. As the number of the regression trees increases, each decision tree plays the role of a weak learner for the data by being automatically adjusted by this prior.

4 Estimation

This section outlines the Markov chain Monte Carlo (MCMC) algorithms used to estimate the factor models with the single tree and the sum-of-trees priors. Step-by-step details of the MCMC algorithm are given in Appendix B.

4.1 Markov chain Monte Carlo algorithms

First, we focus on estimation of the single tree model. We collect the variables defining the regression tree and the other model parameters together as

$$\mathcal{L} = \left\{ i^x_{1:P-1}, i^r_{1:P-1}, i^q_{1:P-1}, P \right\}, \ \boldsymbol{\theta}_f = \left(\mathbf{S}, \mathbf{B}, \boldsymbol{\pi}_s, \boldsymbol{\Sigma}_{f,1:T}, \boldsymbol{\Gamma}_f \right), \ \boldsymbol{\theta}_r = \left(\boldsymbol{\beta}, \boldsymbol{\pi}_x, \sigma^2_{1:N,1:T}, \boldsymbol{\Gamma}_r \right),$$

where $\mathbf{S} = \{s_{1,1}, \ldots, s_{1,2Z}, \ldots, s_{K,1}, \ldots, s_{K,2Z}\}$; $\mathbf{B} = \{\mathbf{b}_0, \mathbf{b}_z, \mathbf{b}_{z^2}\}$; $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_P\}$; $\boldsymbol{\pi}_s = \{\pi_{s,1}, \ldots, \pi_{s,K}\}$; $\boldsymbol{\Gamma}_r$ and $\boldsymbol{\Gamma}_f$ are the sets of parameters associated with the scholastic volatilities for ε_{it} and $\mathbf{u}_t$; $\sigma^2_{1:N,1:T}$ represents all the time-varying variances for the idiosyncratic errors. The target posterior density of the model is given by

$$\begin{aligned} p\left(\mathcal{L}, \boldsymbol{\theta}_r, \boldsymbol{\theta}_f | \mathbf{D}_{1:T}\right) \ \propto \ & p\left(\mathbf{r}_{1:T} | \mathcal{L}, \boldsymbol{\theta}_r, \mathbf{f}_{1:T}, \mathbf{X}_{0:T-1}\right) \ p\left(\mathbf{f}_{1:T} | \boldsymbol{\theta}_f, \mathbf{z}_{0:T-1}\right) \\ \times \ & p\left(\boldsymbol{\beta} | \mathcal{L}\right) p\left(\mathcal{L} | \boldsymbol{\pi}_x, \mathbf{X}_{0:T-1}\right) p\left(\boldsymbol{\pi}_x\right) \\ \times \ & p\left(\mathbf{B} | \mathbf{S}\right) p\left(\mathbf{S} | \boldsymbol{\pi}_s\right) p\left(\boldsymbol{\pi}_s\right) \\ \times \ & p\left(\sigma^2_{1:N,1:T} | \boldsymbol{\Gamma}_r\right) p\left(\boldsymbol{\Gamma}_r\right) p\left(\boldsymbol{\Sigma}_{f,1:T} | \boldsymbol{\Gamma}_f\right) p\left(\boldsymbol{\Gamma}_f\right) \end{aligned}$$

where $\mathbf{r}_{1:T}$ and $\mathbf{f}_{1:T}$ represent the return and factor observations from $t = 1$ to $t = T$; $\mathbf{D}_{1:T}$ represents all available data including $\mathbf{X}_{0:T-1}$ and $\mathbf{z}_{0:T-1}$. The MCMC algorithm consists of the following steps

1. Conditional on $\{\mathbf{B}, \boldsymbol{\Gamma}, \mathbf{f}_{1:T}, \mathbf{z}_{0:T-1}\}$, draw the stochastic covariance matrix for the factors $\boldsymbol{\Sigma}_{f,1:T}$ using the particle Gibbs sampler in Appendix B.

2. Conditional on $\{\boldsymbol{\Sigma}_{f,1:T}, \mathbf{f}_{1:T}, \mathbf{z}_{0:T-1}\}$, sequentially draw each latent variable of $\mathbf{S}$ using the single-move Gibbs sampling algorithm in Appendix B. Let $\tilde{\mathbf{z}}_{t-1}$ denote $[1,\ \mathbf{z}'_{t-1},\ \mathbf{z}^{2\prime}_{t-1}]$. Given $\mathbf{S}$, draw $\mathbf{B}^*$ using the linear regression model

$$\mathbf{f}_t = \overline{\mathbf{Z}}^*_{t-1} \mathbf{B}^* + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathrm{N}\left(0, \boldsymbol{\Sigma}_{f,t}\right).$$

where $\overline{\mathbf{Z}}^*_{t-1}$ is a matrix composed of the columns of $\overline{\mathbf{Z}}_{t-1} = \boldsymbol{\Sigma}_{f,t} \left(\tilde{\mathbf{z}}'_{t-1} \otimes \mathbf{I}_K \right)$ that correspond to non-zero coefficients in $\mathbf{b}_t\left(\mathbf{z}_{t-1}\right)$. $\mathbf{B}^*$ represents the non-zero coefficient vector.

3. Conditional on $\mathbf{S}$, draw $\boldsymbol{\pi}_s$.

4. Conditional on $\{\mathcal{L}, \boldsymbol{\beta}, \mathbf{r}_{1:T}, \mathbf{f}_{1:T}\}$, draw the stochastic volatilities for the idiosyncratic errors $\sigma^2_{1:N,1:T}$ using the particle Gibbs sampler in Appendix B. For fast computation, this step can be parallelized across each entity i.

5. Conditional on $\left\{\sigma^2_{1:N,1:T}, \mathbf{r}_{1:T}, \mathbf{f}_{1:T}, \mathbf{X}_{0:T-1}\right\}$, draw a posterior sample of $\mathcal{L}$ using the Metropolis-Hastings algorithm in Appendix B and calculate new portfolio indicators.

6. Conditional on $\left\{\mathcal{L}, \sigma^2_{1:N,1:T}, \mathbf{r}_{1:T}, \mathbf{f}_{1:T}\right\}$, draw the factor loadings $\boldsymbol{\beta}_p$ using the linear factor model

$$r_{it} = \boldsymbol{\beta}_p' \mathbf{f}_t + \varepsilon_{it} \qquad \varepsilon_{it} \ \sim \ \mathrm{N}\left(0, \sigma^2_{it}\right)$$

   for $p = 1, \ldots, P$.

7. Conditional on $\{\mathcal{L}\}$, draw the vector of probabilities $\boldsymbol{\pi}_x$ using the Metropolis-Hastings algorithm in Appendix B.

8. Conditional on $\sigma^2_{1:N,1:T}$, draw $\boldsymbol{\Gamma}_r$.

9. Conditional on $\boldsymbol{\Sigma}_{1:T}$, draw $\boldsymbol{\Gamma}_f$.

We repeat the above steps until convergence is achieved. During the MCMC algorithm, we compute and store posterior samples of $\boldsymbol{\lambda}_t(\mathbf{z}_{t-1}) = \boldsymbol{\Sigma}_{f,t}\mathbf{b}(\mathbf{z}_{t-1})$ and $\alpha_{p,t} = \boldsymbol{\beta}_p'\boldsymbol{\Sigma}_{f,t}\mathbf{b}(\mathbf{z}_{t-1})$ to recover their posterior distributions.

We employ an iterative Bayesian backfitting MCMC algorithm for the estimation of the sum-of-trees model as in Chipman et al. (2010). Let $\mathcal{L}_\tau$ denote the decision rules of regression tree τ. Also, let $\mathcal{L}_{\neq\tau} = \mathcal{L} \backslash \mathcal{L}_\tau$ and $\tilde{\boldsymbol{\beta}}_{\neq\tau} = \tilde{\boldsymbol{\beta}} \backslash \tilde{\boldsymbol{\beta}}_\tau$ where $\mathcal{L} = \{\mathcal{L}_1, \ldots, \mathcal{L}_H\}$ and $\tilde{\boldsymbol{\beta}} = \left\{\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_H\right\}$. The posterior sampling for $\mathcal{L}_\tau$ and $\tilde{\boldsymbol{\beta}}_\tau$ is sequentially performed conditioning on $\widehat{\varepsilon}_{\tau,i,t} = r_{it} - \sum_{\forall\kappa\neq\tau} \mathcal{T}_\tau\left(d_{\kappa,i,t}, \tilde{\boldsymbol{\beta}}_\kappa\right)$ as

- Repeat step 5 and 6 for $\tau = 1, 2, ..., H$.

5 Conditional on $\left\{\mathcal{L}_{\neq\tau}, \tilde{\boldsymbol{\beta}}_{\neq\tau}, \sigma^2_{1:N,1:T}, \mathbf{r}_{1:T}, \mathbf{f}_{1:T}, \mathbf{X}_{0:T-1}\right\}$, compute $\widehat{\varepsilon}_{\tau,i,t}$ for all return observations. Draw regression tree $\mathcal{L}_\tau = \left\{i^r_{\tau,1:P_\tau-1}, i^x_{\tau,1:P_\tau-1}, i^q_{\tau,1:P_\tau-1}, P_\tau\right\}$ using the MH algorithm in Appendix B and calculate new portfolio indicators. In the MH algorithm, $\widehat{\varepsilon}_{\tau,i,t}$ is treated as a new dependent variable.

6 Conditional on $\left\{\mathcal{L}_{\neq\tau}, \mathcal{L}_\tau, \sigma^2_{1:N,1:T}, \mathbf{r}_{1:T}, \mathbf{f}_{1:T}\right\}$, draw $\tilde{\boldsymbol{\beta}}_{\tau,p}$ using the linear factor model

$$\hat{\varepsilon}_{\tau,i,t} = \beta_{0,\tau,p} + \boldsymbol{\beta}'_{\tau,p}\mathbf{f}_t + \varepsilon_{it} \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \sigma^2_{it}\right)$$

for $p = 1, \ldots, P_\tau$.

All the other steps of the MCMC algorithm for the sum-of-trees model are identical to those of the single tree model.

4.2 Estimation of marginal functions

To investigate how asset attributes are linked to the conditional expected return, we estimate its marginal function with respect to each characteristic. Let $r_{\cdot,t}$, $d_{\tau,\cdot,t}$, $x_{\cdot,\ell,t-1}$ denote random variables for the realized counterparts, $r_{i,t}$, $d_{\tau,i,t}$ and $x_{i,\ell,t-1}$ for $\tau = 1, \ldots, H$ and $\ell = 1, \ldots, L$. We write $\bar{\mu}_t\left(\widehat{F}_{\ell,t-1}\left(x_{\cdot,\ell,t-1}\right) = q_\ell\right)$ as the expected value of $r_{\cdot,t}$ conditioning on $x_{\ell,\cdot,t-1}$ and $\mathbf{z}_{t-1}$. For $q_\ell \in \left\{\frac{1}{1+\overline{Q}_\ell}, \ldots, \frac{\overline{Q}_\ell}{1+\overline{Q}_\ell}\right\}$, the conditional expectation is given by

$$\begin{aligned}\bar{\mu}_t\left(q_\ell\right) &= \left[\sum_{\tau=1}^{H}\sum_{p=1}^{P_\tau}\mathsf{P}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right)\beta_{0,\tau,p}\right] + \left[\sum_{\tau=1}^{H}\sum_{p=1}^{P_\tau}\mathsf{P}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right)\boldsymbol{\beta}'_{\tau,p}\right]\boldsymbol{\lambda}_t\left(\mathbf{z}_{t-1}\right) \\ &= \bar{\beta}_{0,t}\left(q_\ell\right) + \bar{\boldsymbol{\beta}}'_t\left(q_\ell\right)\boldsymbol{\lambda}_t\left(\mathbf{z}_{t-1}\right), \end{aligned} \tag{27}$$

where $\mathsf{P}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right) = \mathsf{P}\left(d_{\tau,\cdot,t} = p \mid \widehat{F}_{\ell,t-1}\left(x_{\cdot,\ell,t-1}\right) = q_\ell\right)$ represents the conditional probability for portfolio p at date t. We note that for the regression tree model in Section 3.1, $H = 1$ and the marginal function for the pricing error term $\bar{\beta}_{0,t}\left(q_\ell\right)$ is restricted to be 0.

Let $\mathbf{x}_{i,:,t-1} = [x_{i,1,t-1}, x_{i,2,t-1}, ..., x_{i,L,t-1}]'$ denote a $L \times 1$ vector of the characteristics for asset i. The conditional probability for portfolio p is approximated via

$$\widehat{\mathsf{P}}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right) = \left(\sum_{i=1}^{n_t} \mathbb{1}\left(\mathcal{L}_\tau(x_{i,\ell,t-1} = c_{\ell,t-1}, \mathbf{x}_{i,\neq\ell,t-1} | \mathbf{X}_{t-1}) = p\right)\right) \frac{1}{n_t} \tag{28}$$

where $c_{\ell,t-1} = \widehat{F}_{\ell,t-1}^{-1}(q_\ell)$, $d_{\tau,i,t} = \mathcal{L}_\tau(\mathbf{x}_{i,:,t-1}|\mathbf{X}_{t-1})$ represents the portfolio sort function of regression tree τ, and $\mathbf{x}_{i,\neq\ell,t-1} = \mathbf{x}_{i,:,t-1} \backslash x_{i,\ell,t-1}$ are all the characteristics of asset i except $x_{i,\ell,t-1}$. The integration targets

$$\begin{aligned} \mathsf{P}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right) &= \int \mathbb{1}\left(d_{\tau,\cdot,t} = p | x_{\ell,\cdot,t-1} = c_{\ell,t-1}, \mathbf{x}_{\cdot,\neq\ell,t-1}\right) \\ &\times \; p\left(\mathbf{x}_{\cdot,\neq\ell,t-1} | x_{\cdot,\ell,t-1} = c_{\ell,t-1}\right) d\mathbf{x}_{\cdot,\neq\ell,t-1} \end{aligned}$$

where $\mathbf{x}_{\cdot,\neq\ell,t-1}$ is the random variable that is the realized counterpart of $\mathbf{x}_{i,\neq\ell,t-1}$. The marginal functions for the unconditional mean, beta, and pricing error of r_{it} are respectively calculated as

$$\bar{\mu}(q_\ell) \approx \frac{1}{T}\sum_{t=1}^{T} \bar{\mu}_t\left(q_\ell\right), \quad \overline{\boldsymbol{\beta}}(x_\ell = q) \approx \frac{1}{T}\sum_{t=1}^{T} \overline{\boldsymbol{\beta}}_t\left(q_\ell\right), \quad \overline{\beta}_0(q_\ell) \approx \frac{1}{T}\sum_{t=1}^{T} \overline{\beta}_{0,t}\left(q_\ell\right) \tag{29}$$

For the Sharpe ratio, we first define the marginal function for the undiversified idiosyncratic volatility in a portfolio as

$$\overline{\sigma}_{v,t}^2(q_\ell) \approx \frac{1}{H}\sum_{\tau=1}^{H}\left[\sum_{p=1}^{P_\tau} \widehat{\mathsf{P}}_{q_\ell}\left(d_{\tau,\cdot,t} = p\right) \overline{\sigma}_{\tau,p,t}^2\right], \; \overline{\sigma}_{\tau,p,t}^2 \approx \frac{1}{n_{\tau,p,t}^2} \sum_{\forall i: d_{\tau,i,t} = p} \sigma_{it}^2 \tag{30}$$

where $n_{\tau,p,t}$ is the number of returns in portfolio p of regression tree τ at date t. Accordingly, the marginal function for the total portfolio variance is defined by

$$\overline{\sigma}_t^2(q_\ell) \approx \bar{\boldsymbol{\beta}}_t'\left(q_\ell\right) \boldsymbol{\Sigma}_{f,t} \bar{\boldsymbol{\beta}}_t\left(q_\ell\right) + \overline{\sigma}_{v,t}^2(q_\ell). \tag{31}$$

We compute the unconditional Sharpe ratio as

$$\overline{SR}(q_\ell) \approx \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\overline{\mu}_t(q_\ell)}{\sqrt{\overline{\sigma}_t^2(q_\ell)}}\right). \tag{32}$$

5 Application: Currency trades

A number of authors argue that alternative variables predict the cross-section of currency returns. Our goal is to shed light on which country-specific characteristics are meaningful predictors of the cross-section of currency returns using our proposed models. Another important goal is to find which economy-wide variables can explain factor risk premia.

We let S_{it} denote the spot nominal exchange rate for country i in month t, defined as the number of U.S. dollars per unit of foreign currency. The monthly annualized log-excess return from the perspective of a U.S. investor who buys a foreign one-month risk-free asset with U.S. dollars borrowed at the U.S. one month risk-free rate is given by $r_{it} = 12 \times \left(\Delta s_{it} - (i_{t-1} - i^*_{i,t-1})\right)$ where $\Delta s_{it} = s_{it} - s_{i,t-1}$. i_{t-1} and $i^*_{i,t-1}$ are the U.S. and foreign one month interbank borrowing rates.[1]

Our main empirical model for r_{it} is the factor model with a single tree prior in equation (5). We set the total number of MCMC iterations to be 400,000 and discard the first 5,000 posterior samples as the burn-in period. Posterior distributions are obtained by retaining only every 10-th iteration for thinning. During estimation, we set $\bar{Q}_\ell = 19$ such that $\mathbf{q}_\ell = \{0.05, 0.1, \ldots, 0.95\}$ for $\ell = 1, 2, ..., L$. In equation (8), all the explanatory variables ($\mathbf{z}_{t-1}$ and $\mathbf{z}^2_{t-1}$) are standardized prior to estimation so that $\mathbf{b}_0$ directly captures the unconditional mean of $\mathbf{b}_t(\mathbf{z}_{t-1})$. The priors for the parameters of the model are provided in Appendix A.

5.1 Data

In our empirical analysis, we use monthly forward and spot exchange rates, and interbank borrowing rate data for 47 countries and the Euro from January 1985 to April

[1] Under covered interest parity, the interest rate differential and the forward discount are related such that $i_t - i^*_{it} \approx fr_{it} - s_{it}$ where $fr_{it} = \log(FR_{it})$ and FR_{it} is the forward nominal exchange rate. If covered interest parity holds, r_{it} can be approximated by $s_{it} - fr_{i,t-1}$ which is the log excess return of buying a foreign currency in the forward market in month $t-1$ and selling in the spot market in month t. We replace any missing values of r_{it} with $12 \times (fr_{i,t-1} - s_{i,t-1})$ whenever possible to obtain more observations.

2019. The countries included are: Australia, Austria, Belgium, Bulgaria, Canada, Chile, China, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hong Kong, Hungary, Iceland, India, Indonesia, Ireland, Israel, Italy, Japan, Kuwait, Lithuania, Malaysia, Mexico, Netherlands, New Zealand, Norway, Philippines, Poland, Portugal, Russia , Saudi Arabia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, United Kingdom, and Vietnam. The currency data are collected from Barclays and Reuters via Datastream. Our selection of countries follows the work of Lustig et al. (2011), Menkhoff et al. (2012b), Menkhoff et al. (2017), and Dahlquist and Hasseltoft (2020).

5.1.1 Country-specific characteristics

We collected data on eleven country-specific characteristics. Most of these characteristics are taken from the finance literature. The first variable is the interest rate differential (IRD) between the U.S. and a foreign counterparty country. The investment strategy based on the IRD is referred to as the carry trade. See Lustig et al. (2011).

To consider momentum strategies (buying currencies that have recently been performing relatively well and short-selling currencies with relatively poor performance), we use the one-month lagged excess return (L1R) and the two-month lagged cumulative excess return for 3 months (L3R). See Burnside et al. (2011) and Menkhoff et al. (2012b). The periods used to compute the two past returns are not overlapping.

We employ the real exchange rate (RFX) and the 3 month change in the real exchange rate (CRFX). See Menkhoff et al. (2017). We define RFX such that an increase in the exchange rate means an appreciation of a foreign currency and depreciation of the U.S. dollar. CRFX is included to capture the momentum in RFX.

We consider four macroeconomic variables. These include the cross-country unemployment rate differential (URD) and inflation rate differential (IFD) between the U.S. and a foreign counterpart country. We also include the 12 month change in the unemployment rate (CUR) and the 3 month change in the inflation rate (CIF) to capture economic momentum. See Nucera (2017), Dahlquist and Hasseltoft (2020), and Colacito et al. (2019).

In our work, we introduce idiosyncratic FX volatility (FXV) and skewness (FXS) variables that were not used in previous studies. The linear factor model proposed by Verdelhan (2018) is employed to compute the two variables.

5.1.2 Observed risk factors

Finance researchers often like to define their (tradeable) factors as an exact linear combination of returns. Following Lustig and Verdelhan (2007), Lustig et al. (2011), and Verdelhan (2018), we use the dollar and carry risk factors as observed factors in our model, i.e., $\mathbf{f}_t = [f_{DF,t}, f_{CF,t}]'$. These studies show that the cross-section of the currency excess returns are largely explained by the dollar factor (DF) and carry factor (CF). We calculate the dollar factor by taking the cross-sectional average of all available excess returns in each month and compute the carry factor by using the difference between the two average returns of the highest IRD portfolio and the lowest IRD portfolio in each month.[2]

5.1.3 Variables driving factor risk prices

We consider a total of 10 economy-wide explanatory variables for the factor risk premia. Many of these variables are motivated by the literature. We include the VIX and global FX volatility (GFXV) in our model to check if they can explain the risk premia for the dollar and carry factors. Chinn and Frankel (2019) argue that there exists a close link between currency excess returns and the VIX due to the "safe haven" characterization of the U.S. dollar. Menkhoff et al. (2012a) argue that currency returns are sensitive to global FX volatility (GFXV). These authors and Cenedese et al. (2014) examine whether the high excess returns of the carry trade strategy are compensations for bearing risk by using GFXV.

[2] As the observed risk factors are linear combinations of currency returns, we exclude two countries in each month to avoid using the same data information twice during estimation.

To measure FX market liquidity, we use the TED spread and the rate of change in the aggregate capital ratio of financial intermediaries (ΔCR). Brunnermeier et al. (2008) argue that the TED spread explains currency returns while Mancini et al. (2013) and Karnaukh et al. (2015) demonstrate that FX market liquidity is reflected in the TED spread; see also Adrian et al. (2011), Adrian et al. (2014), and Brunnermeier and Pedersen (2009). ΔCR is taken from He et al. (2017), who show that capital ratios of financial intermediaries can explain the cross-sectional variations in equity excess returns.

We include in our analysis the rate of change for the IMF primary commodity price index: non-fuel primary commodity prices and energy prices (ΔcPI) to explain the FX risk premia. Bakshi and Panayotov (2013) provide evidence that the global commodity price predicts carry trade returns. Rossi (2013) show that commodity price returns are closely associated with exchange rate predictability for commodity exporting countries. As noted by Chen and Rogoff (2003), these countries are also the countries that often show up in the currency portfolio composed of countries with high interest rates.

The global interest rate differential ($GIRD$) is another explanatory variable that we consider for the FX risk factors. GIRD is computed by the average interest rate differential between U.S. and other developed countries. Lustig et al. (2014) provide evidence that GIRD has a predictive power for the dollar factor.

The last two aggregate predictors that we consider directly reflect the state of the U.S. economy. Those variables are the rate of change for the FRB trade weighted U.S. dollar index (ΔDI) and the U.S. term spread (TS). Avdjiev et al. (2019) show that a strengthening of the U.S. dollar adversely affects banks' balance sheets, which leads to a reduction in banks' risk bearing capacity. Chen and Tsang (2013) employ yield curve information to examine exchange rate predictability. Information from the yield curve has predictive power for future exchange rates because it contains market's expectations about future economic fundamentals. Estrella and Mishkin (1998) show that TS has a strong predictive power for U.S. economic expansions and recessions.

Lastly, we include lagged dollar (DF_{t-1}) and carry factors (CF_{t-1}) in our empirical

Table 1: Posterior estimates of $\boldsymbol{\pi}_x$: single regression tree prior

Parameter	Mean	SD	[5%	95%]	Parameter	Mean	SD	[5%	95%]
π_{IRD}	**0.23**	**0.09**	**0.10**	**0.40**	$\pi_{\Delta IF}$	0.05	0.05	0.00	0.15
π_{L1R}	0.06	0.05	0.00	0.17	π_{URD}	**0.15**	**0.08**	**0.04**	**0.30**
π_{L3R}	0.06	0.05	0.00	0.16	$\pi_{\Delta UR}$	0.05	0.05	0.00	0.16
π_{RFX}	**0.10**	**0.06**	**0.02**	**0.22**	π_{FXV}	**0.15**	**0.08**	**0.04**	**0.30**
$\pi_{\Delta RFX}$	0.06	0.05	0.00	0.16	π_{FXS}	0.05	0.05	0.00	0.14
π_{IFD}	0.05	0.05	0.00	0.14					

Note: Posterior mean, standard deviation, and 90% highest posterior density intervals for the vector $\boldsymbol{\pi}_x$ of model selection probabilities from the single tree model.

analysis to capture persistence in the risk factors.

5.2 Empirical Results

5.2.1 Marginal return function

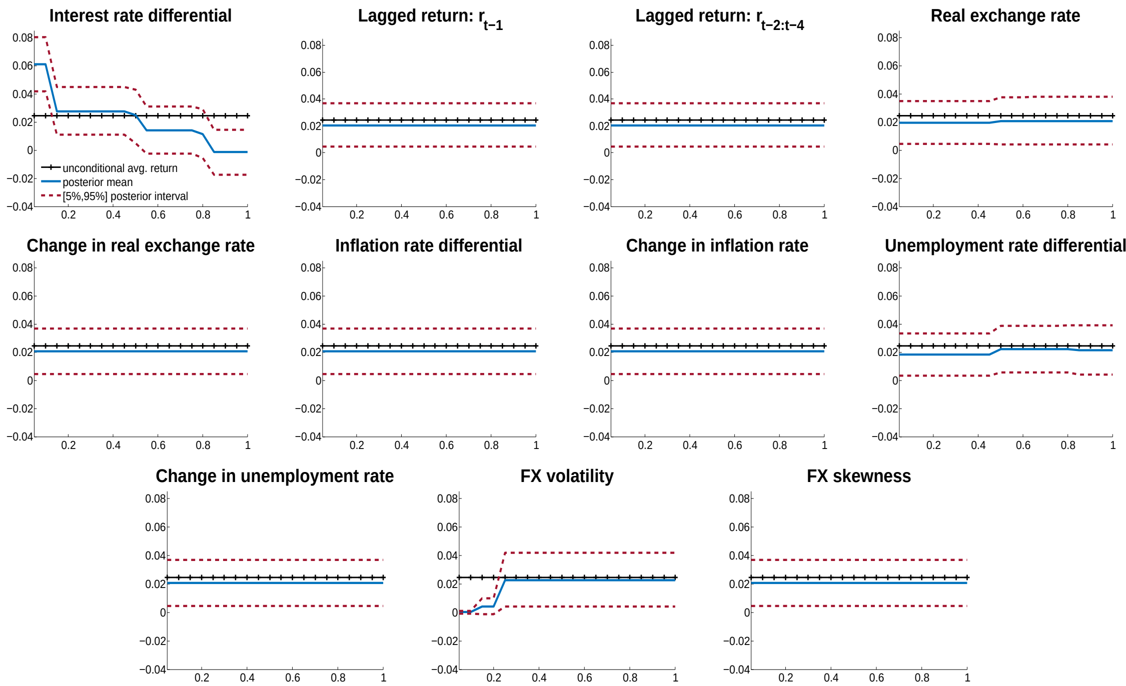
In Table 1, we report the posterior mean and standard deviation of the vector of selection probabilities $\boldsymbol{\pi}_x$. IRD, RFX, URD, and FXV are the only four characteristics selected by the model with a single regression tree prior. The posterior distributions of π_{IRD}, π_{RFX}, π_{URD}, and π_{FXV} noticeably shift from their prior distributions.

Figure 2 depicts the model-based return function in equation (29). If the regression tree does not contain some characteristics, the marginal functions formulated in Section 4.2 corresponding to those characteristics do not vary over $\mathbf{q}_\ell$. This makes it clear which characteristics drive the cross-section of expected returns.

The first variable we inspect is IRD. The marginal return function with respect to IRD is similar to that of the traditional portfolio sort in Figure 1. It indicates that a strong relationship between the excess return and the IRD is a robust feature of the currency return data. A portfolio with relatively high interest rate currencies (low IRD) leads to high excess returns.

Another noticeable characteristic among those selected is FXV. On average, the currencies with relatively low FX volatility provide excess returns substantially lower

Figure 2: Marginal return function

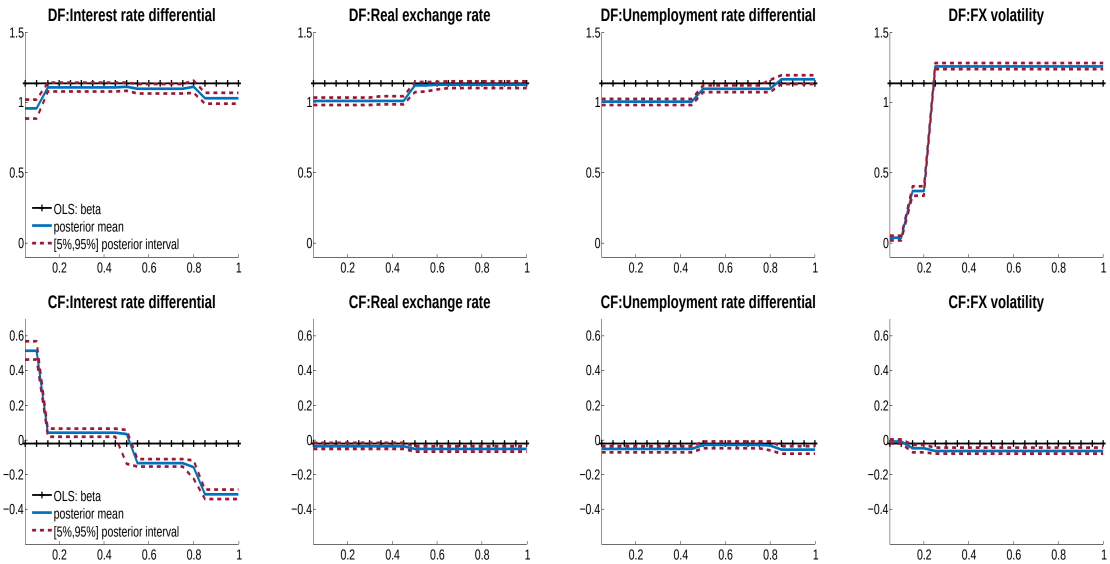


Note: The marginal return function is estimated by the posterior samples of the model parameters based on equation (29).

than the unconditional average return. RFX and URD are other variables selected by the model. However, their impacts on the marginal return function are negligible. The marginal return function for other characteristics that are not selected by the model is constant around the unconditional average return.

A drawback of the traditional portfolio sort is that even though only a small number of key characteristics account for a large portion of the cross sectional variation in excess returns in the population, it could falsely indicate that irrelevant characteristics are important. A misleading result can arise if irrelevant characteristics are correlated with the key characteristics. Our proposed Bayesian approach resolves the issue by simultaneously considering all characteristics at once and by letting only selected characteristics affect the marginal return function. The result in Figure 2 clearly demonstrates this point.

Figure 3: Marginal beta function



Notes: In the first row, we plot the marginal beta functions for the dollar factor. In the second row, we plot the marginal beta functions for the carry factor.

Figure 3 reports the marginal beta functions for the four selected characteristics. The result tells us that the excess returns associated with the dollar and carry factors mostly depend on FXV and IRD, respectively. Currency excess returns whose empirical quantile of FXV (q_{FVX}) is less than 20% are rarely exposed to the dollar risk factor. Lustig et al. (2011) interpret the dollar factor as a level factor in the FX market in the sense that the degree to which individual currency excess returns are exposed to the dollar risk factor are identical. In contrast, our result provides evidence that the currencies exposed to relatively low FX volatility are not exposed to the dollar factor at all.

Figure 3 also shows that the excess return significantly varies over IRD. A low value of IRD is positively associated with the carry factor, and a high value of IRD is negatively associated with the carry factor. On the contrary, the marginal beta return functions for RFX and URD rarely vary even though they are selected by the model.

Table 2: Frequency of being included in the low FXV portfolios

portfolio	country							
	Hong Kong	Saudi Arabia	Kuwait	China	Singapore	India	Vietnam	Denmark
$q_{FXV} \leq 0.1$	277	253	80	83	3	2	38	13
$0.1 < q_{FXV} \leq 0.2$	22	9	177	79	75	68	28	50

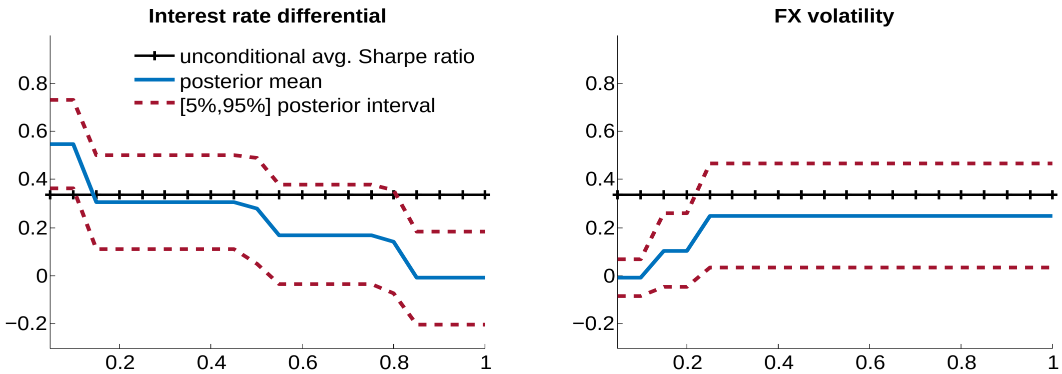
Note: Only countries that are selected more than 60 times in the two FXV portfolios are reported.

To identify which countries belong to the two groups such that $q_{FXV} \leq 0.1$ and $0.1 < q_{FXV} \leq 0.2$, we store the corresponding country names and count how many times those counties are classified in the low FXV portfolios. Over the full sample period, a total of 41 countries are sorted into either one of the two FXV portfolios at least once. Among them, eight countries frequently show up in the low FXV portfolios.

Table 2 exhibits the names of the eight countries and the frequencies of being included in the two FXV portfolios. The new finding in Table 2 is that the underlying exchange rate regimes of the countries are well captured by FXV. The three most frequently selected countries are Hong Kong, Saudi Arabia, and Kuwait, and they all have fixed exchange rates. The Hong Kong Dollar and Saudi Arabia Riyal are pegged to the U.S. dollar. The Kuwait Dinar was pegged to the U.S. dollar up until May 2007 at which time the peg was switched to a basket of currencies. China uses a carefully managed exchange rate system to improve their FX risk management. Other countries such as Singapore and Vietnam use a composite exchange rate anchor like Kuwait. Denmark pegs their currency to the Euro. While India uses a floating exchange rate system now, the Indian Rupee was pegged to the U.S. dollar before December 1991. We emphasize that the countries included in Table 2 are typically included when analyzing foreign exchange rates and currency excess returns (e.g., Lustig et al. (2011), Menkhoff et al. (2012b), Menkhoff et al. (2017), Filippou et al. (2018), Verdelhan (2018)). To the best of our knowledge, no previous study shows that FXV influences the exposure to the dollar risk factor and is linked to their exchange rate regime.

We plot the marginal Sharpe ratio functions with respect to the selected characteristics in Table 4. A significant amount of variation in the marginal Sharpe ratio function is observed for IRD. For the lowest IRD portfolio, the posterior mean of

Figure 4: Marginal Sharpe ratio function



Note: Estimated model-based Sharpe ratio calculated according to (32).

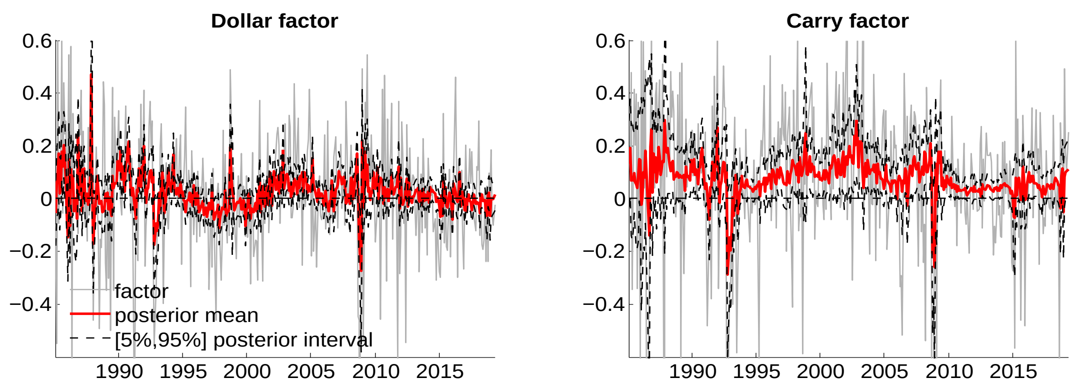
the annualized Sharpe ratio obtained from the model is 0.55 and the interval between the 5% and 95% posterior percentiles does not contain zero and the unconditional average Sharpe ratio. Also, the highest IRD portfolio provides a Sharpe ratio much lower than the unconditional average Sharpe ratio. For the lowest FXV portfolio, the Sharpe ratio is estimated to be approximately 0 while the Sharpe ratio of the remaining FXV portfolios are not significantly different from the unconditional average Sharpe ratio.

5.2.2 Risk factors

Figure 5 displays the posterior estimates of the time-varying factor risk premia $\boldsymbol{\lambda}_t$ along with the observed dollar and carry factors. The posterior distributions of the risk premia track the risk factors well.

The dollar risk premium frequently switches its sign while the carry risk premium remains positive most of the time. Changes in sign of the dollar risk premium can generate additional fruitful investment opportunities. A representative example is the dollar carry trade proposed by Lustig et al. (2014). They extract a signal for the sign switch of the dollar risk premium from the average difference between the interest rate of developed countries and the U.S. interest rate (or the average forward

Figure 5: Estimated factor risk premia



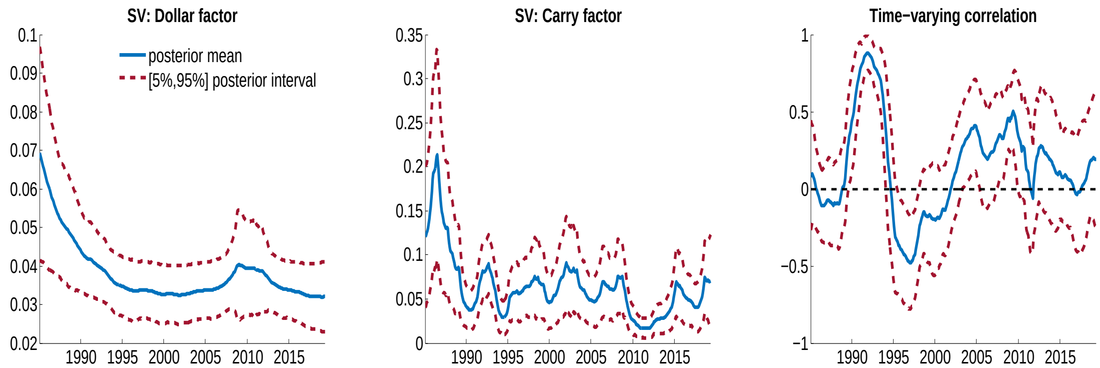
Note: Observed dollar and carry risk factors together with the estimated factor risk premia $\boldsymbol{\lambda}_t$ *and 95% highest posterior density intervals.*

discount rate), and use the signal to determine when to take long or short positions for the U.S. dollar and a basket of all other currencies.

The time-varying risk premium for the carry factor has been documented by numerous studies, including Lustig and Verdelhan (2007), Lustig et al. (2011), and Menkhoff et al. (2012a). These studies claim that there are predictable variations in the carry factor and attempt to find meaningful predictors. Mulder and Tims (2018) and Suh (2019) use global FX volatility to capture time periods during which the price of carry risk is low or even negative to fully exploit investment opportunities associated with carry risk. However, these special versions of the carry trade may not be as beneficial as the dollar carry trade and also may be difficult to implement in practice because the low or negative price of the carry factor lasts for only short periods of time. Our result in Figure 5 provides implicit evidence regarding alternative investment strategies in the FX literature.

Figure 6 reports the posterior estimates for the time-varying covariance matrix of the two risk factors. The conditional volatility of the carry factor is more volatile than the dollar factor. The posterior mean of the correlation between the shocks to the two factors varies by switching its sign. Before 1990 and from the mid 1990s to the early 2000s the conditional correlation was negative. There were sharp increases

Figure 6: Time-varying covariance matrix



Note: Estimated stochastic volatility for the dollar and carry factors and their conditional correlation. 95% highest posterior density intervals.

in the correlation around the recessions of 1990 and early 2000s. Since 2004, the correlation has been mostly positive. Boudoukh et al. (2018) argue that the excess returns of dollar-beta-sorted portfolios developed by Verdelhan (2018) are not reliable because the dollar and carry factors are conditionally correlated. However, Figure 6 shows that although the correlation varies over time, the correlation was substantially reduced to 0 during the late 2000 and its posterior interval contains 0 since then.

The posterior estimates for the parameters of the risk prices $\mathbf{b}_t(\mathbf{z}_{t-1})$ are reported in Table 3. The posterior mean and standard deviations of $\mathbf{b}_0$ are estimated to be 0.42 and 0.27 for the dollar factor while they are estimated to be 1.29 and 0.27 for the carry factor. Recall that all economic-wide explanatory variables in Equation (8) are standardized before estimation so that $\mathbf{b}_0$ corresponds to the unconditional mean of $\mathbf{b}_t(\mathbf{z}_{t-1})$. The posterior estimates of $\mathbf{b}_0$ are low explaining why the dollar factor is more prone to switch signs than the carry factor.

The results in Table 3 show that three economy-wide explanatory variables are significant for explaining the dollar risk factor. To better find variables that explain $\mathbf{b}_t(\mathbf{z}_{t-1})$ well, we mark in bold all parameters that have $\mathsf{P}(s=1)$ larger than 0.8.

Table 3 shows that a decline in ΔCR leads to a higher expected return of the dollar factor. However, the negative marginal effect of ΔCR on the dollar factor shrinks to

Table 3: Posterior estimates of $\mathbf{b}(\mathbf{z}_{t-1})$

	Dollar factor					Carry factor				
	$p(b\|s=1)$					$p(b\|s=1)$				
Parameter	Mean	SD	[5%	95%]	$\mathsf{P}(s=1)$	Mean	SD	[5%	95%]	$\mathsf{P}(s=1)$
b_0	**0.42**	**0.27**	**-0.03**	**0.87**	**1.00**	**1.29**	**0.27**	**0.87**	**1.76**	**1.00**
b_{VIX}	-0.63	0.41	-1.34	0.06	0.66	0.88	0.36	0.33	1.49	0.66
b_{VIX^2}	0.76	0.40	0.09	1.44	0.73	-0.43	0.35	-1.00	0.14	0.32
b_{GFXV}	0.12	0.37	-0.47	0.76	0.55	-0.37	0.30	-0.87	0.13	0.32
b_{GFXV^2}	0.01	0.42	-0.68	0.71	0.55	0.49	0.39	-0.16	1.14	0.34
b_{TED}	-0.04	0.41	-0.70	0.63	0.56	-0.57	0.35	-1.17	-0.03	0.38
b_{TED^2}	-0.76	0.43	-1.50	-0.07	0.78	0.36	0.46	-0.38	1.13	0.31
$b_{\Delta CR}$	**-1.10**	**0.29**	**-1.57**	**-0.63**	**0.98**	-0.07	0.23	-0.45	0.31	0.26
$b_{\Delta CR^2}$	**0.80**	**0.33**	**0.27**	**1.34**	**0.88**	0.46	0.28	0.01	0.95	0.39
$b_{\Delta cPI}$	0.31	0.29	-0.17	0.80	0.60	0.58	0.24	0.18	0.99	0.61
$b_{\Delta cPI^2}$	-0.27	0.37	-0.88	0.34	0.61	-0.64	0.30	-1.13	-0.14	0.51
b_{GIRD}	**-0.81**	**0.38**	**-1.41**	**-0.17**	**0.82**	-0.11	0.44	-0.78	0.60	0.33
b_{GIRD^2}	-0.59	0.43	-1.30	0.15	0.63	-0.67	0.28	-1.13	-0.23	0.59
b_{TS}	0.17	0.38	-0.44	0.77	0.57	-0.51	0.32	-1.04	-0.01	0.39
b_{TS^2}	0.46	0.28	0.01	0.91	0.65	0.42	0.25	0.02	0.82	0.41
$b_{\Delta DI}$	-0.42	0.54	-1.31	0.42	0.57	0.15	0.44	-0.55	0.95	0.28
$b_{\Delta DI^2}$	-0.26	0.44	-1.02	0.46	0.56	-0.57	0.37	-1.22	-0.03	0.42
b_{DF}	-0.51	0.54	-1.44	0.29	0.57	0.22	0.42	-0.47	0.97	0.27
b_{DF^2}	0.04	0.45	-0.69	0.79	0.55	0.21	0.41	-0.43	0.88	0.31
b_{CF}	**0.92**	**0.31**	**0.42**	**1.44**	**0.94**	-0.17	0.24	-0.57	0.21	0.27
b_{CF^2}	0.31	0.28	-0.13	0.77	0.58	-0.40	0.18	-0.69	-0.11	0.55

Note: All explanatory variables are standardized before estimation. $\mathsf{P}(s=1)$ *is the posterior probability of a corresponding variable being included in the model.* s *is an indicator variable for whether or not to include a corresponding variable.* $p(b|s=1)$ *represents the posterior distribution conditioning on* $s=1$. *The model parameters whose posterior* $\mathsf{P}(s=1)$ *is larger than 0.8 are marked in bold.*

zero as ΔCR increases. Adrian et al. (2011) claim that tighter funding constraints of financial intermediaries ($\Delta CR < 0$) amplify the impact of fundamental shocks, which is compensated by a higher risk premium. Also the result may reflect the U.S. dollar's special role as a safe heaven currency. If CR falls unexpectedly at $t-1$, there will be high net demand for the U.S. dollar. Thus, the foreign currencies depreciate at $t-1$ and expected to appreciate in the following periods after the shock is gone. Our result is consistent with Krohn (2019) who relates funding constraints in the FX market with the dollar factor.

Another important variable for the dollar factor is $GIRD$ which is negatively related with the dollar risk premium. Lustig et al. (2014) and Verdelhan (2018) employ $GIRD$ to identify the sign of the dollar risk price. Verdelhan (2018) shows that there exists a locally priced global shock and the local risk prices are associated with country-specific volatilities. For instance, if the U.S. has a relatively high risk price compared to foreign countries, the U.S. dollar appreciates and foreign currencies depreciate when a bad global shock arises. In order to compensate for the risk of a foreign currency depreciation, a positive expected return of the dollar factor should be granted. Conversely, when the U.S. has a relatively low risk price compared to foreign countries, foreign currencies appreciate instead of depreciate in the case of a bad global shock. This can generate a negative expected return of the dollar factor because it offers a hedge opportunity to the bad shock. Verdelhan (2018) uses country-specific interest rates which inversely move with country-specific volatilities to measure a relative risk price between U.S. and foreign countries. According to our variable definition, a high value of $GIRD$ means a relatively high U.S. interest rate (low U.S. specific volatility and low U.S. risk price) compared to foreign interest rates. The negative marginal effect of $GIRD$ on the dollar risk price holds within Verdelhan (2018)'s framework.

The lagged carry factor CF_{t-1} also has a high posterior probability of being included in the model. It implies that some variables that predict the carry factor but are not included in our model can have a predictive power for the dollar factor.

Other variables do not appear to be important as much as ΔCR, $GIRD$, and CF_{t-1} for the dollar factor. The conditional posterior distributions of the other variables are either close to zero or contain 0 with wide posterior intervals. Also, their posterior probability of being included in the model is all lower than 0.8.

The result in Table 3 indicates that the prediction of the carry factor is more difficult than that of the dollar factor. This is illustrated by the relatively large interval estimates for the carry factor shown in the second panel of Figure 5 and the relatively low posterior probabilities of predictors being included in the model in Table 3.

Across all predictors, VIX has the strongest forecasting power for the carry factor. The absolute value of the conditional posterior mean (0.88) is the largest, its interval estimate does not contain 0, and the posterior inclusion probability (0.66) is the highest. Many previous studies such as Lustig et al. (2011), Menkhoff et al. (2012a), and Lettau et al. (2014) show that the carry factor is positively associated with global volatility and uncertainty. We can think of the result in Table 3 in relation to such a conclusion. A high value of VIX indicates high economic or financial uncertainty, and a rise in uncertainty can be reflected in the FX market via a high expected return or high risk price of the carry factor. However, different from Menkhoff et al. (2012a), we find no direct effect of $GFXV$ on the carry factor after controlling for VIX.

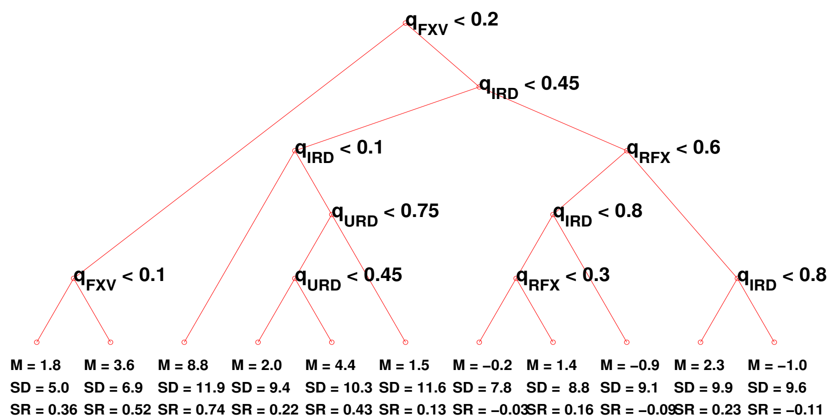
The second variable that has a relatively strong foresting power for the carry factor is ΔcPI. As for VIX, its interval estimate does not contain zero. And the posterior inclusion probability (0.61) is the second highest. A possible reason for the positive marginal impact of ΔcPI on the carry risk premium would be that the currencies of commodity exporting countries often show up in the high interest rate portfolios and a high ΔcPI predicts the appreciation of such commodity currencies.

5.2.3 Portfolio sort function

This section analyzes alternative portfolio strategies and their risk compensations based on the estimated model. Figure 7 shows the portfolio sort function estimated by the proposed model. Following Chipman et al. (1998), we report the portfolio sort function that has the lowest sum of squared residuals in the measurement equation. At each internal node, we report a variable and a quantile value used by the corresponding split rule, and at each terminal node, we report the unconditional return, standard deviation, and Sharpe ratio (SR) for the resulting portfolio. The calculations in Figure 7 are somewhat different from those of the previous section in that means, standard deviations, and Sharpe ratios are calculated from sample moments instead of model-based estimates.[3]

[3] If there are sufficiently many returns in portfolio p, the model-based and actual returns coincide because the idiosyncratic shocks are completely diversified.

Figure 7: Portfolio sort function

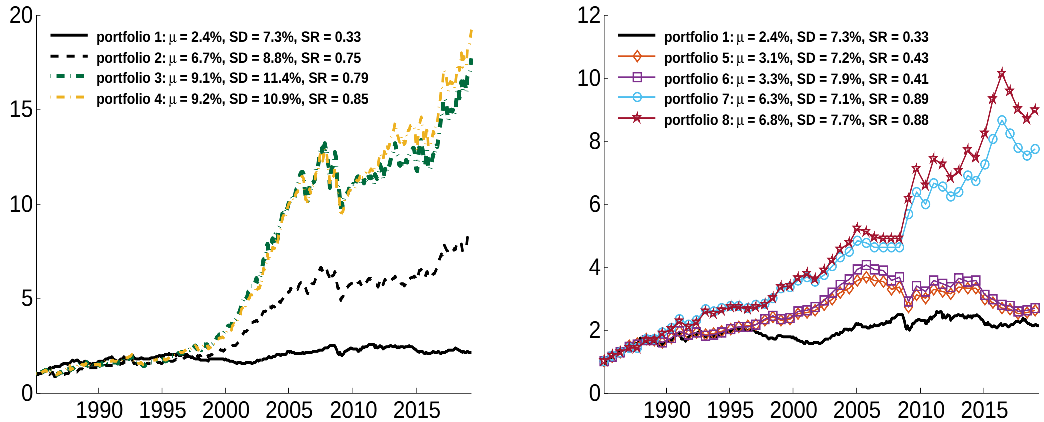


Note: The estimated portfolio sort function generates equally weighted portfolios. M and SD represent the unconditional annual return and the annualized standard deviation of portfolio returns. SR represents the Sharpe ratio. q_{IRD}, q_{FXV}, q_{RFX}, q_{URD} represent the empirical CDFs of the interest rate differential (IRD), the idiosyncratic FX volatility (FXV), the real exchange rate (RFX), and the unemployment rate differential (URD).

The two portfolios constructed by $\{q_{FXV} \leq 0.1\}$ and $\{0.1 < q_{FXV} \leq 0.2\}$ provide 1.8% and 3.6% returns per annum respectively. The highest portfolio return generated by the portfolio sort function is 8.8% per annum and the corresponding portfolio is constructed by $\{q_{FXV} > 0.2, q_{IRD} \leq 0.1\}$. The portfolios sorted by $\{q_{FXV} > 0.2, q_{RFX} < 0.6, 0.8 < q_{IRD}\}$ and $\{q_{FXV} > 0.2, q_{RFX} > 0.6, 0.8 < q_{IRD}\}$ give very similar negative returns −1.0% and −0.9%, which means that RFX does not matter in obtaining the portfolios with large negative returns.

We compare the cumulative returns of 8 portfolios in Figure 8. *Portfolio 1* represents an equally weighted portfolio that contains all currency excess returns. *Portfolio 2* reflects the conventional carry trade strategy that uses equally spaced split points of IRD. For *Portfolio 2*, we assume $P = 6$. The returns of *Portfolio 1* and *Portfolio 2* are identical to the returns of the dollar and carry factors respectively. *Portfolio 3* and *Portfolio 4* are constructed by the estimated portfolio sort function. *Portfolio 3* is the

Figure 8: Cumulative portfolio return



Note: Portfolio 1- long positions on all assets; Portfolio 2- long positions on currencies such that $\{q_{IRD} \leq 1/6\}$ *and short positions on assets such that* $\{q_{IRD} > 5/6\}$*; Portfolio 3- long positions on assets such that* $\{q_{FXV} > 0.2, q_{IRD} \leq 0.1\}$ *and short positions on assets such that* $\{q_{FXV} > 0.2, q_{IRD} > 0.8\}$*; Portfolio 4- long positions on assets such that* $\{q_{IRD} \leq 0.1\}$ *and short positions on assets such that* $\{q_{IRD} > 0.8\}$*; Portfolio 5- long positions on all assets if* $GIRD_{t-1} < 0$ *and short positions on all assets if* $GIRD_{t-1} > 0$*; Portfolio 6- long positions on assets such that* $\{q_{FXV} > 0.2\}$ *if* $GIRD_{t-1} < 0$ *and short positions on assets such that* $\{q_{FXV} > 0.2\}$ *if* $GIRD_{t-1} > 0$*; Portfolio 7- long positions on all assets if* $\lambda_{DF,t} > 0$ *and short positions on all assets if* $\lambda_{DF,t} < 0$ *; Portfolio 8- long positions on assets such that* $\{q_{FXV} > 0.2\}$ *if* $\lambda_{DF,t} > 0$ *and short positions on assets such that* $\{q_{FXV} > 0.2\}$ *if* $\lambda_{DF,t} < 0$

long-short portfolio constructed by using the two portfolios that offer the highest and lowest returns in Figure 7. *Portfolio 4* is constructed similarly with *Portfolio 3* but we employ the split rules only associated with IRD for *Portfolio 4*. As in Lustig et al. (2014), we use $GIRD_{t-1}$ to construct the dollar carry factor. In month t, the return of the dollar carry factor (*Portfolio 5*) proposed by Lustig et al. (2014) is the average return of all currency assets if $GIRD_{t-1} < 0$ and is the negative average return of all currency assets if $GIRD_{t-1} > 0$. *Portfolio 6* also uses the sign of $GIRD_{t-1}$ for portfolio construction. However, as in Verdelhan (2018), we separate currency returns that are highly and lowly exposed to the dollar risk factor based on the estimated portfolio sort function in Figure 7. *Portfolio 6* only reflects the returns of the currency assets that are highly exposed to the dollar risk factor. *Portfolio 7* and *Portfolio 8*

Table 4: Posterior estimates of $\boldsymbol{\pi}_x$: sum-of-trees model

Parameter	Mean	SD	[5%	95%]	Parameter	Mean	SD	[5%	95%]
π_{IRD}	**0.34**	**0.09**	**0.18**	**0.49**	$\pi_{\Delta IF}$	0.02	0.02	0.00	0.07
π_{L1R}	0.02	0.02	0.00	0.07	π_{URD}	**0.16**	**0.07**	**0.07**	**0.29**
π_{L3R}	0.02	0.02	0.00	0.06	$\pi_{\Delta UR}$	0.02	0.02	0.00	0.07
π_{RFX}	**0.09**	**0.05**	**0.03**	**0.18**	π_{FXV}	**0.21**	**0.07**	**0.10**	**0.34**
$\pi_{\Delta RFX}$	0.02	0.02	0.00	0.06	π_{FXS}	0.02	0.02	0.00	0.07
π_{IFD}	**0.07**	**0.04**	**0.02**	**0.14**					

Note: Posterior mean, standard deviation, and 90% highest posterior density intervals for the vector $\boldsymbol{\pi}_x$ of model selection probabilities from the sum-of-trees model.

are similar to *Portfolio 5* and *Portfolio 6* respectively, but we use the posterior mean of $\lambda_{DF,t}$ estimated by our proposed model to infer the sign of the dollar risk price.

Two conclusions can be reached from Figure 8. First, we can obtain higher unconditional excess returns using unequally spaced split points of the IRD while maintaining a similar or higher Sharpe ratio compared to the conventional carry trade (*Portfolio 2* v.s. *Portfolio 4*). Second, the portfolios that use $\lambda_{DF,t}$ as a signal for the negative dollar risk price instead of $GIRD_{t-1}$ offer higher unconditional excess returns and higher Sharpe ratios than the conventional dollar carry trades (*Portfolio 5* and *Portfolio 6*) and carry trade (*Portfolio 2*).[4]

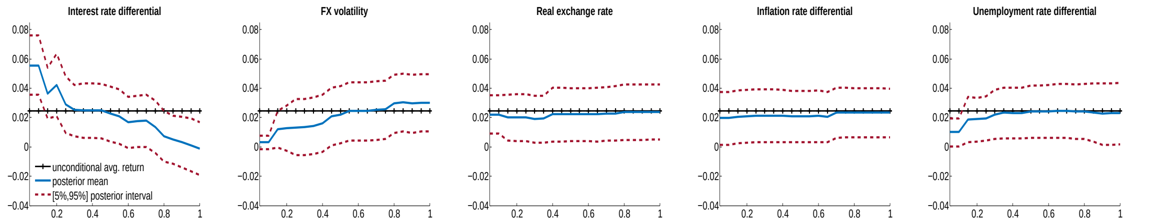
5.2.4 Sum-of-tree model

We also estimated the sum-of-trees model in (25). During estimation, we assume $H = 20$ and $\underline{p} = 5$, which means that each regression tree can generate 5 portfolios at most.

Table 4 shows that a total of 5 characteristics are selected by the model which include IRD, RFX, IFD, URD, and FXV. The posterior means for $\boldsymbol{\pi}_x$ in Table 4 suggest that IRD and FXV are more frequently used than any other characteristics including RFX, IFD, and URD in the sum-of-trees model.

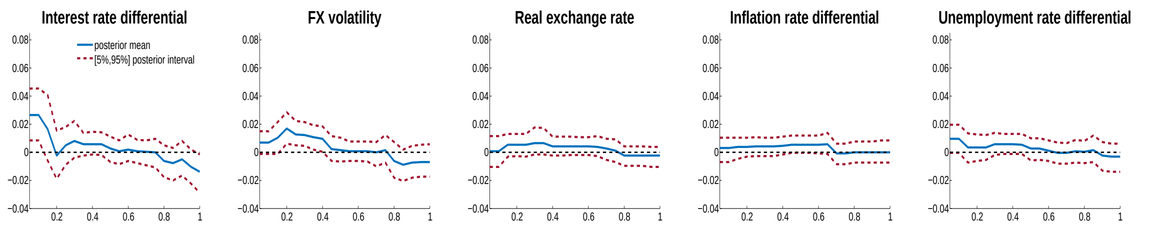
[4] Even though our portfolio sort model is estimated by the full sample, this evidence suggests that there exists a room for constructing profitable portfolios by building a good prediction model for the dollar factor.

Figure 9: Marginal return function associated with the risk factors



Note: In this figure, the marginal return function is estimated based on $\bar{\boldsymbol{\beta}}_t' \left(x_{\ell,t-1} = q\right) \boldsymbol{\lambda}_t \left(\mathbf{z}_{t-1}\right)$ *in equation* (27).

Figure 10: Marginal return function associated with the pricing error



Note: In this figure, the marginal return function is estimated based on $\bar{\beta}_{0,t} \left(x_{\ell,t-1} = q\right)$ *in equation* (27).

The marginal return functions associated with the risk factors are reported in Figure 9. As in the single tree model, the expected return significantly varies with IRD. Moreover, the marginal function of FXV shows a strong monotonic relation with the expected return. When the quantile value of FXV is near 0, the expected return is also close to 0. Other selected characteristics do not appear to be important for the cross section of the currency returns. The results in Figure 9 are consistent with those of the single tree model.

Figure 10 shows the marginal functions for the systematic pricing error. None of the selected characteristics except IRD has a strong monotonic relation with the pricing error. The pricing error being around 2 % when $q_{IRD} < 0.1$ means that the observed carry factor used in the model does not fully explain the risk compensation associated with IRD.

We do not report the posterior estimates for $\mathbf{b}(\mathbf{z}_{t-1})$ from the sum-of-trees model, as the corresponding results are qualitatively and quantitatively similar to those of the single tree model. The empirical evidence presented in this section clearly demonstrates that the results from the single tree model are robust to the alternative model setup, and IRD and FXV are the most important characteristics driving the cross sectional variation in the currency returns.

6 Conclusion

We developed a Bayesian factor model for unbalanced panel data whose intercept and factor loadings have regression tree priors. Our methods simultaneously estimate the factor risk premia while searching over the set of characteristics that determine the assets' betas. Our study demonstrates empirically that characteristics selected by traditional portfolio sorts do not coincide with those selected by our model. This discordance can arise because the classical portfolio sort approach cannot analyze all characteristics at once in a unified framework.

The benefits of the proposed model have been showcased by applying it to an unbalanced panel of currency returns. Our empirical analysis based on the proposed model provides strong evidence that the interest rate differential and idiosyncratic FX volatility are meaningful predictors for the cross-section of currency returns. New evidence of the time-series predictability for the FX risk factors is also provided. The predictability that implies time-varying factor risk premium stems from important economy-wide variables such as the VIX, aggregate capital ratio of financial intermediaries, and the global interest rate differential.

References

Adrian, Tobias, Erkko Etula, and Jan JJ Groen, "Financial amplification of foreign exchange risk premia.," *European Economic Review*, 2011, *55* (3), 354–370.

_ , _ , and Tyler Muir, "Financial intermediaries and the cross-section of asset returns.," *The Journal of Finance*, 2014, *69* (6), 2557–2596.

Avdjiev, Stefan, Wenxin Du, Catherine Koch, and Hyun Song Shin, "The dollar, bank leverage, and deviations from covered interest parity.," *American Economic Review: Insights*, 2019, *1* (2), 193–208.

Bakshi, Gurdip and George Panayotov, "Predictability of currency carry trades and asset pricing implications.," *Journal of Financial Economics*, 2013, *110* (1), 139–163.

Bali, Turan G., Robert F. Engle, and Scott Murray, *Empirical Asset Pricing: the Cross Section of Stock Returns*, Hoboken, NJ: Wiley, 2016.

Black, Fischer, Michael Jensen, and Myron Scholes, "The capital asset pricing model: some empirical findings," in Michael Jensen, ed., *Studies in the Theory of Capital Markets*, 1972, pp. 986–999.

Boudoukh, Jacob, Matthew P Richardson, Ashwin K Thapar, and Franklin Wang, "Is there a Dollar Risk Factor?," 2018. Unpublished manuscript.

Breiman, Leo, Jerome Friedman, R. A. Olshen, and Charles J. Stone, *Classification and Regression Trees.*, New York, NY: Wadsworth, 1984.

Brunnermeier, Markus K and Lasse Heje Pedersen, "Market liquidity and funding liquidity.," *The Review of Financial Studies*, 2009, *22* (6), 2201–2238.

Brunnermeier, Markus K., Stefan Nagel, and Lasse H. Pedersen, *Carry Trades and Currency Crashes* April

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo, "Carry trade and momentum in currency markets," *Annu. Rev. Financ. Econ.*, 2011, *3* (1), 511–535.

Cattaneo, Matias D., Richard H. Crump, Max H. Farrell, and Ernst Schaumburg, "Characteristic sorted portfolios: estimation and inference," *The Review of Economics and Statistics*, 2020, *102* (3), 531–551.

Cenedese, Gino, Lucio Sarno, and Ilias Tsiakas, "Foreign exchange risk and the predictability of carry trade returns.," *Journal of Banking and Finance*, 2014, *42*, 302–313.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, "Economic Forces and the Stock Market," *The Journal of Business*, 1986, *59* (3), pp. 383–403.

Chen, Yu-chin and Kenneth Rogoff, "Commodity currencies.," *Journal of International Economics*, 2003, *60* (1), 133–160.

_ and Kwok Ping Tsang, "What does the yield curve tell us about exchange rate predictability?," *The Review of Economics and Statistics*, 2013, *95* (1), 185–205.

Chernov, Mikhail and Drew Creal, "The PPP view of Multihorizon Currency Risk Premiums," *The Review of Financial Studies*, 2021. Forthcoming.

Chinn, Menzie D and Jeffrey A Frankel, "A third of a century of currency expectations data: the carry trade and the risk premium.," 2019. Unpublished manuscript.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch, "Bayesian CART model search," *Journal of the American Statistical Association*, 1998, *93* (443), 935–960.

_ , _ , and _ , "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, 2010, *4* (1), 266–298.

Colacito, Riccardo, Steven J Riddiough, and Lucio Sarno, "Business cycles and currency returns," Technical Report, National Bureau of Economic Research 2019.

Creal, Drew D. and Ruey S. Tsay, "High-dimensional dynamic stochastic copula models.," *Journal of Econometrics*, 2015, *189* (2), 335–345.

Dahlquist, Magnus and Henrik Hasseltoft, "Economic momentum and currency returns," *Journal of Financial Economics*, April 2020, *136* (1).

Denison, David G. T., Bani K. Mallick, and Adrian F. M Smith, "A Bayesian CART algorithm," *Biometrika*, 1998, *85* (2), 363–377.

Estrella, Arturo and Frederic S Mishkin, "Predictability of currency carry trades and asset pricing implications.," *The Review of Economics and Statistics*, 1998, *80* (1), 45–61.

Fama, Eugene F. and James D. MacBeth, "Risk, return, and equilibrium: empirical tests.," *Journal of Political Economy*, 1973, *81* (3), 607–636.

Feng, Gavin, Stefano Giglio, and Dacheng Xiu, "Taming the factor zoo: a test of new factors," *The Journal of Finance*, 2020, *75* (3), 1327–1370.

Filippou, Ilias, Arie E Gozluklu, and Mark P Taylor, "Global political risk and currency momentum," *Journal of Financial and Quantitative Analysis*, 2018, *53* (5), 2227–2259.

Gagliardini, Patrick, Elisa Ossola, and Olivier. Scaillet, "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 2016, *84* (3), 187–229.

Geweke, John F and Goufu Zhou, "Measuring the pricing error of the arbitrage pricing theory," *The Review of Financial Studies*, 1996, *9* (2), 557–587.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Autoencoder asset pricing models.," *Journal of Econometrics*, 2020. forthcoming.

_ , _ , and _ , "Empirical asset pricing via machine learning.," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, second ed., New York, NY: Springer Press, 2009.

He, Zhiguo, Bryan Kelly, and Asaf Manela, "Intermediary asset pricing: New evidence from many asset classes.," *Journal of Financial Economics*, 2017, *126* (1), 1–35.

Karnaukh, Nina, Angelo Ranaldo, and Paul Söderlind, "Understanding FX liquidity.," *The Review of Financial Studies*, 2015, *28* (11), 3073–3108.

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

Krohn, Ingomar, "Time-Varying Global Dollar Risk in Currency Markets.," 2019. Unpublished manuscript.

Lettau, Martin, Matteo Maggiori, and Michael Weber, "Conditional risk premia in currency markets and other asset classes.," *Journal of Financial Economics*, 2014, *114* (2), 197–225.

Linero, Antonio R., "Bayesian regression trees for high-dimensional prediction and variable selection," *Journal of the American Statistical Association*, 2018, *113* (522), 626–636.

Lustig, Hanno and Adrien Verdelhan, "The cross-section of foreign currency risk premia and consumption-growth risk.," *American Economic Review*, 2007, *97* (1), 89–117.

_ , Nikolai Roussanov, and Adrien Verdelhan, "Common risk factors in currency markets.," *The Review of Financial Studies*, 2011, *24* (11), 3731–3777.

_ , _ , and _ , "Countercyclical currency risk premia.," *Journal of Financial Economics*, 2014, *111* (3), 527–553.

Mancini, Loriano, Angelo Ranaldo, and Jan Wrampelmeyer, "Liquidity in the foreign exchange market: Measurement, commonality, and risk premiums.," *The Journal of Finance*, 2013, *68* (5), 1805–1841.

Menkhoff, Lukas, Lucio Sarno, Maik Schmeling, and Andreas Schrimpf, "Carry trades and global foreign exchange volatility," *The Journal of Finance*, 2012, *67* (2), 681–718.

_ , _ , _ , and _ , "Currency momentum strategies," *Journal of Financial Economics*, 2012, *106* (3), 660–684.

_ , _ , _ , and _ , "Currency value," *The Review of Financial Studies*, 2017, *30* (2), 416–441.

Moritz, Benjamin and Tom Zimmermann, "Tree-based conditional portfolio sorts: the relation between past and future stock returns," 2016. Working paper, Ludwig Maximilian University of Munich.

Mulder, Arjen and Ben Tims, "Conditioning carry trades: Less risk, more return.," *Journal of International Financial Markets, Institutions and Money*, 2018, *85*, 1–19.

Nucera, Federico, "Unemployment fluctuations and the predictability of currency returns," *Journal of Banking and Finance*, 2017, *84*, 88–106.

Rossi, Barbara, "Exchange rate predictability.," *Journal of Economic Literature*, 2013, *51* (4), 1063–1119.

Suh, Sangwon, "Unexploited currency carry trade profit opportunity.," *Journal of International Financial Markets, Institutions and Money*, 2019, *58*, 236–254.

Taddy, Matthew A., Robert B. Gramacy, and Nicholas G. Polson, "Dynamic trees for learning and design," *Journal of the American Statistical Association*, 2011, *106* (493), 109–123.

Verdelhan, Adrien, "The share of systematic variation in bilateral exchange rates," *The Journal of Finance*, 2018, *73* (1), 375–418.

Wu, Yuhong, Hkon Tjelmeland, and Mike West, "Bayesian CART: Prior specification and posterior simulation.," *Journal of Computational and Graphical Statistics*, 2007, *16* (1), 44–66.

Appendix A Prior distributions

- The prior over the risk prices is

$$b_{k,z}|s_{k,z} = 1 \quad \sim \quad \mathrm{N}(0, \underline{v}_b), \tag{A.1}$$

where $\underline{v}_b = 1$, and

$$\pi_{s,k} \quad \sim \quad \mathrm{Beta}\left(\underline{\gamma}_{s,1}, \underline{\gamma}_{s,2}\right) \tag{A.2}$$

where $\underline{\gamma}_{s,1} = \underline{\gamma}_{s,2} = 0.1$.

- The prior over σ_{it}^2 is

$$\begin{aligned}
\mu_{\sigma,i} &\sim \mathrm{N}\left(\log\left(v\hat{a}r(y_i)\right), 3^2\right) && \text{(A.3)}\\
\sigma_{\sigma,i}^2 &\sim \mathrm{IG}\left(3, 3 \times 0.01^2\right) && \text{(A.4)}\\
\phi_{\sigma,i} &\sim \mathrm{N}\left(0.9, 0.2^2\right) && \text{(A.5)}
\end{aligned}$$

where $v\hat{a}r(y_i)$ is the sample variance for asset i's returns. We assume weakly informative priors to reflect the prior belief that the change in σ_{it}^2 is persistent and gradual over time.

- The prior over $\mathbf{\Sigma}_{f,t}$ is

$$\begin{aligned}
\mu_{g,k} &\sim \mathrm{N}\left(\log\left(v\hat{a}r(f_k)\right), 3^2\right) && \text{(A.6)}\\
\sigma_{g,k}^2 &\sim \mathrm{IG}\left(3, 3 \times 0.01^2\right) && \text{(A.7)}\\
\phi_{g,k} &\sim \mathrm{N}\left(0.9, 0.2^2\right) && \text{(A.8)}\\
\sigma_{e,2,1}^2 &\sim \mathrm{IG}\left(3, 3 \times 0.01^2\right) && \text{(A.9)}
\end{aligned}$$

where $v\hat{a}r(f_k)$ is the sample variance for factor k for $k = \{DF, CF\}$. We assume weakly informative priors for the parameters to reflect the prior belief that the change in $\mathbf{\Sigma}_{f,t}$ is persistent and gradual over time.

- The prior over the number of portfolios P is

$$P \quad \sim \quad 1 + \mathrm{binomial}(\underline{p} - 1, \underline{\pi_p}). \tag{A.10}$$

where $\underline{p} = 11$ and $\underline{\pi_p} = 0.5$. The assumed prior restricts the maximum number of portfolios to be 11 in the stochastic portfolio sort model with a single tree prior. The mean and variance of the prior are $\mathbb{E}(P) = 1 + 10 \times 0.5$ and $\mathbb{V}(P) = 10 \times 0.5^2$. In the sum-of-trees model, we assume $\underline{p} = 5$ and $\underline{\pi_p} = 0.5$ for each regression tree.

- The prior over $\boldsymbol{\pi}_x$ is

$$\boldsymbol{\pi}_x \quad \sim \quad \text{Dirichlet}(\kappa_1, \kappa_2, \ldots, \kappa_L). \tag{A.11}$$

  We assume the asymmetric Dirichlet prior where $\kappa_\ell = 1$ for $\ell = 1, 2, ..., L$.

- We impose a restriction on the minimum observation count in the Bayesian estimation. If any of portfolios have no return more often than 90% of the entire sample period, the corresponding regression tree is rejected by the MCMC algorithm.

Appendix B MCMC algorithm

Appendix B.1 Drawing the regression tree

Our proposed tree prior repeatedly generates two sub-portfolios from one randomly selected portfolio until the total number of portfolios reaches P. We refer to the selected portfolio where a split occurs as a parent portfolio and to the two sub-portfolios as child portfolios. Let $\mathbf{I}^r = i^r_{1:P-1}$, $\mathbf{I}^x = i^x_{1:P-1}$, $\mathbf{I}^q = i^q_{1:P-1}$, and $\mathbf{I} = \{\mathbf{I}^r, \mathbf{I}^x, \mathbf{I}^q\}$. Conditioning on $\mathbf{I}$ and P, we employ a Metropolis-Hastings (MH) algorithm to locally change the given portfolio forming structure. We consider four candidate proposals; *grow* ($P^* = P + 1$), *merge* ($P^* = P - 1$), *change* ($P^* = P$) and *swap* ($P^* = P$) where P^* is a candidate portfolio count. For the *grow* proposal, we randomly select one existing portfolio and create two new sub-portfolios from it. For the *merge* proposal, we randomly select a pair of two child portfolios that were generated from a parent portfolio and merge them together. For the *change* proposal, we randomly select one existing split rule and change it to new one. For the *swap* proposal, we swap two split rules in a randomly selected pair of parent and child portfolios.

As in the MH algorithms for a Bayesian CART model or a BART model, our MH algorithm yields a reversible Markov chain, because every step that moves from the original split rules, $\mathbf{I}$, to new split rules, $\mathbf{I}^*$, has a counterpart move from $\mathbf{I}^*$ to $\mathbf{I}$. The reversible Markov chain is required for any valid MH algorithm.

The acceptance probability for our proposed MH algorithm is given as:

$$\alpha \quad = \quad \min\left[1, \frac{p\left(\mathbf{I}^*, P^*|\mathbf{D}_{1:T}\right)}{q\left(\mathbf{I}^*, P^*|\mathbf{I}, P\right)} \frac{q\left(\mathbf{I}, P|\mathbf{I}^*, P^*\right)}{p\left(\mathbf{I}, P|\mathbf{D}_{1:T}\right)}\right]$$

where $q(.)$ is the candidate generating density; $\{\mathbf{I}^*, P^*\}$ is the set of new split rules; $\{\mathbf{I}, P\}$ is the set of original split rules; $\mathbf{D}_{1:T}$ represents all available data. In the above acceptance probability and all the densities used in the MH algorithm, all redundant model parameters for the MH algorithm are omitted for the sake of notational simplicity. Also, we assume that the number of quantile values to be selected is same for all characteristics. (i.e., $\bar{Q}_\ell = \bar{Q}$) Computation details for the implementation of the MH algorithm are provided in the online-appendix.

Appendix B.1.1 Grow Proposal

For the *grow* move, we randomly select one existing portfolio in $\{\mathbf{I}, P\}$ and split it into two child portfolios. First we draw $i^{r*} \in \mathcal{I}_P^r$ from Categorical $(\boldsymbol{\pi}_r|...)$ where $\pi_{r,i^r} = \frac{n_{i^r}}{N_{r,1:T}}$ is an element of $\boldsymbol{\pi}_r$ and represents the probability of selecting region i^r. The term $N_{r,1:T}$ in the denominator of π_{r,i^r} represents the total number of observations in all regions that can be further split and the term n_{i^r} in the numerator of π_{r,i^r} represents the number of observations in region i^r that can be further split. Conditioning on i^{r*}, we move to the next step to draw one of characteristics. A new index $i^{x*} \in \mathcal{I}_P^x$ is drawn from Categorical $(\bar{\boldsymbol{\pi}}_x|...)$ where $\bar{\pi}_{x,i^x} = \frac{\pi_{x,i^x}\mathbb{1}(i^x|i^{r*})}{\sum_{\ell=1}^{L}\pi_{x,\ell}\mathbb{1}(i^x=\ell|i^{r*})}$ is an element of $\bar{\boldsymbol{\pi}}_x$ and represents the probability of selecting variable i^x. The probability π_{x,i^x} is the unconditional probability of selecting variable i^x without information on $\mathbf{I}^r$. The indicator function $\mathbb{1}(i^x = \ell|i^{r*})$ takes one if there exists at least one quantile value of variable ℓ in region i^{r*} that meets the condition for the minimum observation count in portfolios and takes zero otherwise.

Suppose that i^{r*} and i^{x*} are selected for *grow* proposal. To draw an optimal split point, $i^{q*} \in \mathcal{I}_P^q$, we evaluate the target posterior density at all possible quantile values of variable i^{x*}:

$$p(i^{q*}|i^{x*}, i^{r*}, \mathbf{I}, P^*, \mathbf{D}_{1:T}) \quad = \quad \frac{\left(\prod_{\forall d} p\left(\mathbf{r}_d|i^{q*}, i^{x*}, i^{r*}, \mathbf{I}, P^*\right)\right)^{\frac{\mathbb{1}(i^{q*}|i^{x*}, i^{r*})}{N_{i^q}}}}{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p\left(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}, P^*\right)\right)^{\frac{\mathbb{1}(i^q=k|i^{x*}, i^{r*})}{N_{i^q}}}\right]}$$

where $\mathbf{r}_d$ is the collection of return data included in each region; $N_{i^q} = \sum_{k=1}^{\bar{Q}} \mathbb{1}(i^q = k|i^{x*}, i^{r*})$ is the number of possible quantile values that meet the required minimum observation count. Note that any split rule that violates the required minimum observation count has a zero weight due to the indicator function $\mathbb{1}(i^q|i^{x*}, i^{r*})$.

The candidate generating density of *grow* move is given by:

$$q\left(\mathbf{I}^*, P^*|\mathbf{I}, P\right) \quad = \quad \frac{\left(\prod_{\forall d} p\left(\mathbf{r}_d|i^{q*}, i^{x*}, i^{r*}, \mathbf{I}, P^*\right)\right)^{\frac{1}{N_{i^q}}}}{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p\left(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}, P^*\right)\right)^{\frac{\mathbb{1}(i^q=k|i^{x*}, i^{r*})}{N_{i^q}}}\right]} \frac{\pi_{x,i^{x*}}}{N_{i^x}} \frac{n_{i^{r*}}}{N_{r,1:T}} p\left(grow\right)$$

where $p(grow)$ is the probability of choosing the *grow* proposal in the MH algorithm; $N_{i^x} = \sum_{\ell=1}^{L} \pi_{x,\ell}\mathbb{1}(i^x = \ell|i^{r*})$. Many terms of the first ratio of the acceptance probability are canceled out:

$$\frac{p\left(\mathbf{I}^{*},P^{*}|\mathbf{D}_{1:T}\right)}{q\left(\mathbf{I}^{*},P^{*}|\mathbf{I},P\right)}\propto\frac{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d}p\left(\mathbf{r}_{d}|i^{q}=k,i^{x*},i^{r*},\mathbf{I},P^{*}\right)\right)\frac{\mathbb{1}(i^{q}=k|i^{x*},i^{q*})}{N_{i^{q}}}\right]}{\left(\prod_{\forall d}p\left(\mathbf{r}_{d}|i^{q*},i^{x*},i^{r*},\mathbf{I},P^{*}\right)\right)\frac{1}{N_{i_{q}}}}\frac{N_{i^{x}}}{\pi_{x,i^{x*}}}\frac{N_{r,1:T}}{n_{i^{r*}}}\frac{1}{p\left(grow\right)}$$

$$\times\left(\prod_{\forall d}p\left(\mathbf{r}_{d}|i^{q*},i^{x*},i^{r*},\mathbf{I},P^{*}\right)\right)\frac{1}{N_{i_{q}}}\frac{\pi_{x,i^{x*}}}{N_{i^{x}}}\frac{n_{i^{r*}}}{N_{r,1:T}}p(\mathbf{I}|P^{*})p(P^{*})$$

$$=\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d}p\left(\mathbf{r}_{d}|i^{q}=k,i^{x*},i^{r*},\mathbf{I},P^{*}\right)\right)\frac{\mathbb{1}\left(i^{q}=k|i^{x*},i^{q*}\right)}{N_{i^{q}}}\right]p\left(\mathbf{I}|P\right)p\left(P^{*}\right)\frac{1}{p\left(grow\right)}.$$

Note that $p(\mathbf{I}|P)=p(\mathbf{I}|P^{*})$.

To compute the second ratio of the acceptance probability, we need to define the candidate generating density for the *merge* proposal. In the *merge* move of the MH algorithm, we randomly draw a pair of two child portfolios from a uniform distribution. Therefore, conditional on $\{\mathbf{I}^{*},P^{*}\}$, the density for obtaining $\{\mathbf{I},P\}$ via the *merge* proposal is given by:

$$q\left(\mathbf{I},P|\mathbf{I}^{*},P^{*}\right)\quad=\quad\frac{p(merge)}{N_{merge}}$$

where $p(merge)$ is the probability of choosing the *merge* move in the MH algorithm; N_{merge} is the number of pairs of two child portfolios that can be merged conditional on $\{\mathbf{I}^{*},P^{*}\}$. Accordingly, the second ratio of the acceptance probability is easily defined as:

$$\frac{p(\mathbf{I},P|\mathbf{D}_{1:T})}{q(\mathbf{I},P|\mathbf{I}^{*},P^{*})}\propto\left[\prod_{\forall d}p(\mathbf{r}_{d}|\mathbf{I},P)\right]p\left(\mathbf{I}|P\right)p\left(P\right)\frac{N_{merge}}{p(merge)}.$$

The minimum acceptance probability is given as below:

$$\frac{p\left(\mathbf{I}^{*},P^{*}|\mathbf{D}_{1:T}\right)}{q\left(\mathbf{I}^{*},P^{*}|\mathbf{I},P\right)}\frac{q\left(\mathbf{I},P|\mathbf{I}^{*},P^{*}\right)}{p\left(\mathbf{I},P|\mathbf{D}_{1:T}\right)}$$

$$=\frac{\frac{1}{N_{i^{q}}}\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d}p\left(\mathbf{r}_{d}|i^{q}=k,i^{x*},i^{r*},\mathbf{I},P^{*}\right)\right)\mathbb{1}\left(i^{q}=k|i^{x*},i^{r*}\right)\right]}{\left[\prod_{\forall d}p(\mathbf{r}_{d}|\mathbf{I},P)\right]}\frac{p(P+1)}{p(P)}\frac{1}{N_{merge}}\frac{p(merge)}{p(grow)}.$$

Note that the acceptance probability compares the likelihood value of the original split rules and the average likelihood value obtained by integrating out i^{q} from its joint density with $\mathbf{r}_{1:T}$. If the new split rule is accepted, we set $\mathbf{I}^{q}=\{i_{1}^{q},i_{2}^{q},...,i_{P-1}^{q},i^{q*}\}$, $\mathbf{I}^{x}=\{i_{1}^{x},i_{2}^{x},...,i_{P-1}^{x},i^{x*}\}$, and $\mathbf{I}^{r}=\{i_{1}^{r},i_{2}^{r},...,i_{P-1}^{r},i^{r*}\}$ for the next MCMC iteration. Also we modify the region and portfolio indices accordingly.

Appendix B.1.2 Merge Proposal

The *merge* move randomly selects a pair of two child portfolios that stem from a same parent portfolio and merge them together. ($P^* = P - 1$) Suppose that $\tilde{i}^r \in \mathcal{I}_P^r$ is a randomly selected odd index number for a region that has a sister region. Then the region index of the parent portfolio is $\tilde{i}_p^r = \frac{\tilde{i}^r - 1}{2}$ and the region index of its sister portfolio which shares the same parent portfolio is $\tilde{i}^r + 1$. When merging the two portfolios, the number of observations of the merged portfolio becomes $n_{\tilde{i}_p^r} = n_{\tilde{i}^r} + n_{\tilde{i}^r+1}$. We use $\tilde{i}_p^x$ and $\tilde{i}_p^q$ to denote the split variable and split quantile value in the region $\tilde{i}_p^r$.

Let $\mathbf{I}^* = \{\mathbf{I}^{q*}, \mathbf{I}^{x*}, \mathbf{I}^{r*}\}$ denote the set of split rules after merging the two selected portfolios. The target density evaluated at $\mathbf{I}^*$ is proportional to:

$$p(\mathbf{I}^*, P^*|\mathbf{D}_{1:T}) \propto \left[\prod_{\forall d} p\left(\mathbf{r}_d|\mathbf{I}^*, P^*\right)\right] p\left(\mathbf{I}^*|P^*\right) p\left(P^*\right).$$

The first ratio of the acceptance probability is easily derived as:

$$\frac{p(\mathbf{I}^*, P^*|\mathbf{D}_{1:T})}{q(\mathbf{I}^*, P^*|\mathbf{I}, P)} \propto \left[\prod_{\forall d} p\left(\mathbf{r}_d|\mathbf{I}^*, P^*\right)\right] p\left(\mathbf{I}^*|P^*\right) p\left(P^*\right) \frac{N_{merge}}{p(merge)}$$

where N_{merge} is the number of pairs of two child portfolios that can be selected in the *merge* move.

The candidate generating density for moving from $\{\mathbf{I}^*, P^*\}$ to $\{\mathbf{I}, P\}$ via the *grow* move is:

$$q(\mathbf{I}, P|\mathbf{I}^*, P^*) = \frac{\prod_{\forall d} p(\mathbf{r}_d|\tilde{i}_p^q, \tilde{i}_p^x, \tilde{i}_p^r, \mathbf{I}^*, P) \frac{\mathbb{1}\left(\tilde{i}_p^q|\tilde{i}_p^x, \tilde{i}_p^r\right)}{N_{i^q}}}{\left[\sum_{k=1}^{\bar{Q}} \left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}_p^x, \tilde{i}_p^r, \mathbf{I}^*, P)\right) \frac{\mathbb{1}\left(i^q = k|\tilde{i}_p^x, \tilde{i}_p^r\right)}{N_{i^q}}\right]} \frac{\pi_{x,\tilde{i}_p^x}}{N_{i^x}} \frac{n_{\tilde{i}_p^r}}{N_{\tilde{r},1:T}} p\left(grow\right)$$

where $N_{i^q} = \sum_{k=1}^{\bar{Q}} \mathbb{1}\left(i^q = k|\tilde{i}_p^x, \tilde{i}_p^r\right)$; $N_{i^x} = \sum_{\ell=1}^{L} \pi_{x,\ell} \mathbb{1}\left(i^x = \ell|\tilde{i}_p^r\right)$; $N_{\tilde{r},1:T}$ is the total number of observations in all regions that can further split given $\{\mathbf{I}^*, P^*\}$; and

$$\frac{p(\mathbf{I}, P|\mathbf{D}_{1:T})}{q(\mathbf{I}, P|\mathbf{I}^*, P^*)} \propto \left[\sum_{k=1}^{\bar{Q}} \left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}_p^x, \tilde{i}_p^r, \mathbf{I}^*, P)\right) \frac{\mathbb{1}\left(i^q = k|\tilde{i}_p^x, \tilde{i}_p^r\right)}{N_{i^q}}\right] p(I^*|P) p(P) \frac{1}{p(grow)}.$$

Therefore, the minimum acceptance probability is:

$$\frac{\left[\prod_{\forall d} p\left(\mathbf{r}_d|\mathbf{I}^*, P^*\right)\right]}{\frac{1}{N_{i^q}}\left[\sum_{k=1}^{\bar{Q}} \left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}_p^x, \tilde{i}_p^r, \mathbf{I}^*, P)\right) \mathbb{1}\left(i^q = k|\tilde{i}_p^x, \tilde{i}_p^r\right)\right]} N_{merge} \frac{p(P-1)}{p(P)} \frac{p(grow)}{p(merge)}$$

where $\frac{1}{N_{i^q}}\left[\sum_{k=1}^{\bar{Q}-1} \left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}_p^x, \tilde{i}_p^r, \mathbf{I}^*, P)\right) \mathbb{1}\left(i^q = k|\tilde{i}_p^x, \tilde{i}_p^r\right)\right]$ is the average likelihood obtained by integrating out i^q from the joint density of $\mathbf{D}_{1:T}$ and i^q. If the *merge* proposal is accepted, we set $\mathbf{I}^q = \mathbf{I}^q_{-\tilde{i}^q_p}$, $\mathbf{I}^x = \mathbf{I}^x_{-\tilde{i}^x_p}$, and $\mathbf{I}^r = \mathbf{I}^r_{-\tilde{i}^r_p}$ for the next MCMC iteration where $\mathbf{I}^q_{-\tilde{i}^q_p}$, $\mathbf{I}^x_{-\tilde{i}^x_p}$, and $\mathbf{I}^r_{-\tilde{i}^r_p}$ represent the collection of the split rules except one in the region $\tilde{i}^r_p$. Also we modify the region indices accordingly.

Appendix B.1.3 Change proposal

In the *change* move, we randomly select one of existing split rules and replace it with new one. Therefore, this proposal does not change the portfolio count. (i.e., $P^* = P$) Let $i^{r*} \in \mathcal{I}^m$ denote a randomly selected region that has two two child regions and $\mathbf{I}^*_{-i^{r*}}$ denote the set of all split rules except one associated with the region i^{r*}. We use $\tilde{i}^x$ and $\tilde{i}^q$ to denote the original split variable and split quantile value in the selected region i^{r*}.

To replace the split rule $\{\tilde{i}^q, \tilde{i}^x\}$ in the region i^{r*}, we first draw i^{x*} from Categorical($\bar{\boldsymbol{\pi}}_x|...$) where $\bar{\pi}_{x,i^x} = \frac{\pi_{x,i^x}\mathbb{1}(i^x)}{N_{i^x}}$ and $N_{i^x} = \sum_{\ell=1}^{L} \mathbb{1}\left(i^x = \ell\right)$. The indicator function $N_{i^x} = \sum_{\ell=1}^{L} \mathbb{1}\left(i^x = \ell\right)$ checks if there exists at least one feasible quantile value of i^x which can meet the condition of the observation count for all generated portfolios. Note that $\mathbb{1}\left(i^x\right) \neq \mathbb{1}\left(i^x|i^{r*}\right)$ because the latter function checks the condition for the minimum observation count only for the region i^{r*}. Next, we draw i^{q*} from the following candidate generating distribution:

$$q(i^{q*}|i^{x*}, i^{r*}, \mathbf{I}_{-i^{r*}}, P^*, \mathbf{D}_{1:T}) = \frac{\left(\prod_{\forall d} p(\mathbf{r}_d|i^{q*}, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P^*)^{\frac{\mathbb{1}(i^{q*}|i^{x*})}{N_{i^q}}}\right)}{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P^*)^{\frac{\mathbb{1}(i^q=k|i^{x*})}{N_{i^q}}}\right)\right]}$$

where $\mathbb{1}\left(i^q|i^{x*}\right)$ is the indicator function determining if all portfolios generated by the *change* move meet the required minimum observation count; $N_{i^q} = \sum_{k=1}^{\bar{Q}} \mathbb{1}\left(i^q = k|i^{x*}\right)$ represents the number of feasible split points. Note that $\mathbb{1}\left(i^q|i^{x*}\right) \neq \mathbb{1}\left(i^q|i^{x*}, i^{r*}\right)$ because the latter function checks the condition for the observation count only for the region i^{r*}. The candidate generating density for the *change* move is given by:

$$q(\mathbf{I}^*, P^*|\mathbf{I}, P) = \frac{\left(\prod_{\forall d} p(\mathbf{r}_d|i^{q*}, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P^*)\right)^{\frac{\mathbb{1}(i^{q*}|i^{x*})}{N_{i^q}}}}{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P^*)\right)^{\frac{\mathbb{1}(i^q=k|i^{x*})}{N_{i^q}}}\right]} \frac{\pi_{x,i^{x*}}}{N_{i^x}} \frac{1}{N_{change}} p(change)$$

where $p(change)$ is the probability of selecting the *change* move; N_{change} is the number of regions that can be selected in the *change* move; $\frac{1}{N_{change}}$ is the probability of selecting a region i^{r*}; $\frac{\pi_{x,i^{x*}}}{N_{i^x}}$ is the probability of selecting variable i^{x*}; $\mathbf{I}^*$ is the set of new split rules; $P^* = P$.

The first ratio of the acceptance probability is:

$$\frac{p(\mathbf{I}^*, P^*|\mathbf{D}_{1:T},)}{q(\mathbf{I}^*, P^*|\mathbf{I}, P)} \propto \left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P^*)\frac{\mathbb{1}\left(i^q = k|i^{x*}\right)}{N_{i^q}}\right)\right] p(\mathbf{I}^*|P^*, i^{q*}, i^{x*})p(P^*)\frac{N_{change}}{p(change)}.$$

where $p(\mathbf{I}^*|P^*, i^{q*}, i^{x*}) = p(\mathbf{I}^*|P^*)\frac{N_{i^q}}{\mathbb{1}(i^{q*}|i^{x*})}\frac{N_{i^x}}{\pi_{x,i^{x*}}}$.

The candidate generating density for $\mathbf{I}, P$ conditional on $\mathbf{I}^*, P^*$ is:

$$q(\mathbf{I}, P|\mathbf{I}^*, P^*) = \frac{\left(\prod_{\forall d} p(\mathbf{r}_d|\tilde{i}^q, \tilde{i}^x, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P)\right)\frac{\mathbb{1}\left(\tilde{i}^q|\tilde{i}^x\right)}{N_{\tilde{i}^q}}}{\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}^x, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P)\right)\frac{\mathbb{1}\left(i^q = k|\tilde{i}^x\right)}{N_{\tilde{i}^q}}\right]}\frac{\pi_{x,\tilde{i}^x}}{N_{i^x}}\frac{p(change)}{N_{change}}$$

where $\mathbb{1}\left(i^q|\tilde{i}^x\right)$ is an indicator function determining if all generated portfolios meet the required minimum observation count; $N_{\tilde{i}^q} = \sum_{k=1}^{\bar{Q}} \mathbb{1}\left(i^q = k|\tilde{i}^x\right)$ represents the number of feasible split points conditional on $\tilde{i}^x$. This means that:

$$\frac{p(\mathbf{I}, P|\mathbf{D}_{1:T})}{q(\mathbf{I}, P|\mathbf{I}^*, P^*)} \propto \left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}^x, i^{r*}, \mathbf{I}^*_{-i^{r*}}, P)\right)\frac{\mathbb{1}\left(i^q = k|\tilde{i}^x\right)}{N_{\tilde{i}^q}}\right] p(\mathbf{I}|P, \tilde{i}^q, \tilde{i}^x)p(P)\frac{N_{change}}{p(change)}$$

where $p(\mathbf{I}|P, \tilde{i}^q, \tilde{i}^x) = p(\mathbf{I}|P)\frac{N_{\tilde{i}^q}}{\mathbb{1}\left(\tilde{i}^q|\tilde{i}^x\right)}\frac{N_{i^x}}{\pi_{x,\tilde{i}^x}}$.

Finally, the minimum acceptance probability for the *change* move is given by:

$$\frac{\frac{1}{N_{i^q}}\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, i^{x*}, i^{r*}, \mathbf{I}^*_{-i^{r*}})\right)\mathbb{1}\left(i^q = k|i^{x*}\right)\right]}{\frac{1}{N_{\tilde{i}^q}}\left[\sum_{k=1}^{\bar{Q}}\left(\prod_{\forall d} p(\mathbf{r}_d|i^q = k, \tilde{i}^x, i^{r*}, \mathbf{I}^*_{-i^{r*}})\right)\mathbb{1}\left(i^q = k|\tilde{i}^x\right)\right]}\frac{p(\mathbf{I}^*|P, i^{q*}, i^{x*})}{p(\mathbf{I}|P, \tilde{i}^q, \tilde{i}^x)}.$$

Note that only portfolios that are generated from the region i^{r*} are affected by the *change* proposal. Therefore, when computing the acceptance probability in practice, it is sufficient to take into account the changes in the prior and the likelihood only for the affected portfolios.

Appendix B.1.4 Swap proposal

In the *swap* move, we randomly select a pair of parent and child regions both of which contain split rules. Then the two selected split rules are exchanged to create new portfolios. Let N_{swap} denote the number of such parent-child region pairs. Also we use $i_p^{r*} \in \mathcal{I}^m$ and $i_c^{r*} \in \mathcal{I}^m$ to denote the selected parent region index and the selected child region index, respectively. We draw $\{i_p^{r*}, i_c^{r*}\}$ from a discrete uniform distribution. Therefore, the candidate generating density for the *swap* move is

$$q(\mathbf{I}^*, P^*|\mathbf{I}, P) = \frac{1}{N_{swap}} p(swap)$$

where $\mathbf{I}^*$ is the set of new split rules after exchanging the two split rules in the parent-child region pair; $p(swap)$ is the probability of selecting the *swap* move.

The first ratio of the acceptance probability is given by:

$$\frac{p(\mathbf{I}^*, P^*|\mathbf{D}_{1:T})}{q(\mathbf{I}^*, P^*|\mathbf{I}, P)} \propto \left(\prod_{\forall d} p(\mathbf{r}_d|\mathbf{I}^*, P^*)\right) p(\mathbf{I}^*|P^*) p(P^*) \frac{N_{swap}}{p(swap)},$$

and the second ratio of the acceptance probability is given by:

$$\frac{p(\mathbf{I}, P|\mathbf{D}_{1:T})}{q(\mathbf{I}, P|\mathbf{I}^*, P^*)} \propto \left(\prod_{\forall d} p(\mathbf{r}_d|\mathbf{I}, P)\right) p(\mathbf{I}|P) p(P) \frac{N^*_{swap}}{p(swap)}.$$

Because $P^* = P$ and $N_{swap} = N^*_{swap}$, the minimum acceptance probability for the *swap* proposal is given by:

$$\frac{\left(\prod_{\forall d} p(\mathbf{r}_d|\mathbf{I}^*, P^*)\right) p(\mathbf{I}^*|P)}{\left(\prod_{\forall d} p(\mathbf{r}_d|\mathbf{I}, P)\right) p(\mathbf{I}|P)}.$$

Note that only portfolios that are generated from the region i_p^{r*} are affected by the *swap* proposal. Therefore, when computing the acceptance probability in practice, it is sufficient to take into account the changes in the prior and the likelihood only for the affected portfolios.

Appendix B.2 Drawing $\boldsymbol{\pi}_x$

We draw $\boldsymbol{\pi}_x = \{\pi_{x,1}, \pi_{x,2}, ..., \pi_{x,L}\}$ using a MH algorithm. The target posterior density conditional on $\mathbf{I}$ and P is proportional to

$$p(\boldsymbol{\pi}_x|\mathbf{I}, P, \mathbf{r}_{1:T}) \quad \propto \quad \prod_{j=1}^{P-1} p(i_j^x|i_j^r, \mathcal{D}_{j-1}) p(\boldsymbol{\pi}_x)$$

where $\mathcal{D}_{j-1}$ represents the sigma algebra that contains all previous split decisions up to iteration $j-1$, $p(i_j^x|i_j^r, \mathcal{D}_{j-1}) = \frac{\pi_{x,i_j^x}}{\sum_{\ell=1}^{L} \pi_{x,\ell}\mathbb{1}\left(i^x=\ell|i_j^r\right)}$, and

$$\begin{aligned} p(\boldsymbol{\pi}_x) &= \frac{1}{B(\kappa)} \prod_{\ell=1}^{L} \pi_{x,\ell}^{\kappa_\ell - 1}, \\ B(\kappa) &= \frac{\prod_{\ell=1}^{L} \Gamma(\kappa_\ell)}{\Gamma(\sum_{\ell=1}^{L} \kappa_\ell)}. \end{aligned}$$

The parameter $\kappa = \{\kappa_1, \kappa_2, ..., \kappa_L\}$ is the vector of prior-hyperparameters for $\boldsymbol{\pi}_x$.

We generate the candidate value of $\boldsymbol{\pi}_x$ from the following Dirichlet distribution

$$\boldsymbol{\pi}_x^* \sim \text{Dir}(\tilde{\kappa})$$

where $\boldsymbol{\pi}_x^* = \{\pi_{x,1}^*, \pi_{x,2}^*, ..., \pi_{x,L}^*\}$, $\tilde{\kappa} = \{\tilde{\kappa}_1, \tilde{\kappa}_2, ..., \tilde{\kappa}_L\}$ and $\tilde{\kappa}_\ell = \kappa_\ell + \sum_{j=1}^{P-1} \mathbb{1}\left(i_j^x = \ell\right)$, which is the posterior distribution of $\boldsymbol{\pi}_x$ when no constraint is imposed. The candidate draw $\boldsymbol{\pi}_x^*$ is accepted or rejected according to the acceptance probability

$$\min \left[1, \frac{p\left(\boldsymbol{\pi}_x^*|\mathbf{I}, P, \mathbf{r}_{1:T}\right)}{q\left(\boldsymbol{\pi}_x^*\right)} \frac{q\left(\boldsymbol{\pi}_x\right)}{p\left(\boldsymbol{\pi}_x|\mathbf{I}, P, \mathbf{r}_{1:T}\right)}\right]$$

where $q(.)$ is the candidate generating density; $\boldsymbol{\pi}_x^*$ is the candidate probabilities; $\boldsymbol{\pi}_x$ is the original probabilities. The minimum acceptance probability can be shown to be

$$\frac{p\left(\boldsymbol{\pi}_x^*|\mathbf{I}, P, \mathbf{r}_{1:T}\right)}{q\left(\boldsymbol{\pi}_x^*\right)} \frac{q\left(\boldsymbol{\pi}_x\right)}{p\left(\boldsymbol{\pi}_x|\mathbf{I}, P, \mathbf{r}_{1:T}\right)} = \prod_{j=1}^{P-1} \frac{\left(\sum_{\ell=1}^{L} \pi_{x,\ell}\mathbb{1}\left(i^x = \ell|i_j^r\right)\right)}{\left(\sum_{\ell=1}^{L} \pi_{x,\ell}^*\mathbb{1}\left(i^x = \ell|i_j^r\right)\right)}.$$

Note that when the constraint is not binding at all, the minimum acceptance probability becomes 1.

Appendix B.3 Drawing the factor risk prices

To draw the posterior samples of the model parameters in $\mathbf{b}_t\left(\mathbf{z}_{t-1}\right)$, we need to derive their joint posterior distribution. The target posterior density for the MCMC step is given below

$$p\left(\mathbf{B}, \mathbf{S}|\mathbf{D}_{1:T}, \mathcal{L}, \boldsymbol{\theta}_r, \boldsymbol{\theta}_f\right) \propto \left[\prod_{t=1}^{T} p\left(\mathbf{r}_t, \mathbf{f}_t|\mathbf{D}_{1:t-1}, \mathcal{L}, \mathbf{B}, \mathbf{S}, \mathcal{L}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{f,t}, \sigma_t^2\right)\right] p(\mathbf{B}|\mathbf{S})p(\mathbf{S}|\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s) \tag{B.1}$$

$$\propto \left[\prod_{t=1}^{T} p\left(\mathbf{f}_t|\mathbf{z}_{t-1}, \mathbf{B}, \mathbf{S}, \boldsymbol{\Sigma}_{f,t}\right)\right] p(\mathbf{B}|\mathbf{S})p(\mathbf{S}|\boldsymbol{\pi}_s)p(\boldsymbol{\pi}_s) \tag{B.2}$$

where $\mathbf{r}_t = \{r_{1,t}, r_{2,t}, ..., r_{n_t,t}\}$ and $\sigma_t^2 = \{\sigma_{1,t}^2, \sigma_{2,t}^2, ..., \sigma_{n_t,t}^2\}$. We note that the model restricted by the stochastic discount factor can be reduced to

$$\begin{aligned} r_{it} &= \sum_{p=1}^{P} \mathbb{1}\left(d_{it}=p\right)\left[\boldsymbol{\beta}_p' \boldsymbol{\lambda}_t\left(\mathbf{z}_{t-1}\right) + \boldsymbol{\beta}_p'\left(\mathbf{f}_t - \boldsymbol{\lambda}_t\left(\mathbf{z}_{t-1}\right)\right)\right] + \varepsilon_{it} && \text{(B.3)} \\ &= \sum_{p=1}^{P} \mathbb{1}\left(d_{it}=p\right)\boldsymbol{\beta}_p'\mathbf{f}_t + \varepsilon_{it}, && \text{(B.4)} \end{aligned}$$

and similarly the unrestricted model can be reduced to

$$r_{it} = \sum_{p=1}^{P} \mathbb{1}\left(d_{it}=p\right)\left[\beta_{0,p} + \boldsymbol{\beta}_p'\mathbf{f}_t\right] + \varepsilon_{it}. \tag{B.5}$$

The above equations demonstrates that r_{it} is irrelevant to the posterior of $\mathbf{B}$ and $\mathbf{S}$ conditioning on $\mathbf{f}_t$. The same conclusion can be applied to each regression tree in the sum-of-trees model.

Consider the following equation for the observed risk factors

$$\mathbf{f}_t = \tilde{\mathbf{Z}}_{t-1}\mathbf{B} + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_{f,t}) \tag{B.6}$$

where $\tilde{\mathbf{Z}}_{t-1} = \Sigma_{f,t}(\tilde{\mathbf{z}}_{t-1}' \otimes \mathbf{I}_K)$ and $\mathbf{B} = \mathrm{vec}\left([\mathbf{b}_0,\ \mathbf{b}_z,\ \mathbf{b}_{z^2}]\right)$. Let $\tilde{\mathbf{Z}}_{t-1}^*$ denote the columns of $\tilde{\mathbf{Z}}_{t-1}$ that are relevant to $\mathbf{b}_t\left(\mathbf{z}_{t-1}\right)$ conditioning on $\mathbf{S}$. We use $\mathbf{B}^*$ to denote the $(K_b \times 1)$ coefficient vector corresponding to $\tilde{\mathbf{Z}}_{t-1}^*$:

$$\mathbf{f}_t = \tilde{\mathbf{Z}}_{t-1}^*\mathbf{B}^* + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}_{f,t}). \tag{B.7}$$

By vertically stacking observations for all t, we obtain the matrix equation given below:

$$\mathbf{F} = \tilde{\mathbf{Z}}^*\mathbf{B}^* + \mathbf{u}, \qquad \varepsilon \sim \mathrm{N}\left(0, \boldsymbol{\Sigma}\right) \tag{B.8}$$

where $\mathbf{F} = [\mathbf{f}_1', \ldots, \mathbf{f}_T']'$ and $\boldsymbol{\Sigma}$ is the block diagonal matrix of $\{\boldsymbol{\Sigma}_{f,1}, \ldots, \boldsymbol{\Sigma}_{f,T}\}$. The prior distribution of $\mathbf{B}^*$ is given by:

$$\boldsymbol{\Lambda}^* \sim \mathrm{N}\left(0, \underline{\mathbf{V}}_{B^*}\right) \tag{B.9}$$

where $\underline{\mathbf{V}}_{B^*} = diag(v_b, v_b, ..., v_b)$ and v_b is a prior hyper-parameter.

The marginal likelihood used to generate each latent variable of $\mathbf{S}$ is derived by integrating out $\mathbf{B}^*$

from the joint density of $\mathbf{B}^*$ and $\mathbf{f}$ as:

$$\begin{aligned} p(\mathbf{F}|\tilde{\mathbf{Z}},\mathbf{S},\mathbf{\Sigma}) &= (2\pi)^{-\frac{KT}{2}} \left(\prod_{t=1}^{T} |\mathbf{\Sigma}_t|\right)^{-\frac{1}{2}} \upsilon_{\lambda}^{-\frac{K_b}{2}} |\overline{\mathbf{V}}_{B^*}|^{\frac{1}{2}} \\ & \quad \exp\left[-\frac{1}{2}\left(-\overline{\boldsymbol{\mu}}'_{B^*}\overline{\mathbf{V}}_{B^*}^{-1}\overline{\boldsymbol{\mu}}_{B^*} + \mathbf{F}'\mathbf{\Sigma}^{-1}\mathbf{F}\right)\right]. \end{aligned} \tag{B.10}$$

where $\overline{\mathbf{V}}_{B^*} = \left(\underline{\mathbf{V}}_{B^*}^{-1} + \tilde{\mathbf{Z}}^{*\prime}\mathbf{\Sigma}^{-1}\tilde{\mathbf{Z}}^*\right)^{-1}$ and $\overline{\boldsymbol{\mu}}_{B^*} = \overline{\mathbf{V}}_{B^*}\left(\tilde{\mathbf{Z}}^{*\prime}\mathbf{\Sigma}^{-1}\mathbf{F}\right)$.

Also, the above result indicates that:

$$\mathbf{B}^*|\mathbf{F},\tilde{\mathbf{Z}},\mathbf{S},\mathbf{\Sigma} \sim \mathrm{N}\left(\overline{\boldsymbol{\mu}}_{B^*}, \overline{\mathbf{V}}_{B^*}\right). \tag{B.11}$$

For $\forall k \in \{1,2,...,K\}$ and $\forall z \in \{1,2,...,2Z\}$, we draw $s_{k,z}$ using the following posterior distribution:

$$\begin{aligned} Pr[s_{k,z}=0|...] &= \frac{p(\mathbf{F}|\tilde{\mathbf{Z}},\mathbf{\Sigma},\mathbf{S}_{\neq k,z},s_{k,z}=0)}{p(\mathbf{F}|\tilde{\mathbf{Z}},\mathbf{\Sigma},\mathbf{S}_{\neq k,z},s_{k,z}=0)+p(\mathbf{F}|\tilde{\mathbf{Z}},\mathbf{\Sigma},\mathbf{S}_{\neq k,z},s_{k,z}=1)} \\ Pr[s_{k,z}=1|...] &= 1-Pr[s_{k,z}=0|...] \end{aligned} \tag{B.12}$$

where $\mathbf{S}_{\neq k,z}$ represents all latent variables in $\mathbf{S}$ except $s_{k,z}$. After updating all latent variables in $\mathbf{S}$, we draw $\mathbf{B}^*$ from $\mathrm{N}\left(\overline{\boldsymbol{\mu}}_{B^*}, \overline{\mathbf{V}}_{B^*}\right)$ and set all other parameters of $\mathbf{B}$ that are not included in $\mathbf{B}^*$ to zero.

For $\forall k \in \{1,2,...,K\}$, we draw $\pi_{s,k}$ using the following posterior beta distribution:

$$\pi_{s,k}|\mathbf{S},\ldots \sim \mathrm{Beta}\left(\overline{\gamma}_{s,1},\overline{\gamma}_{s,2}\right) \tag{B.13}$$

where $\overline{\gamma}_{s,1} = \underline{\gamma}_{s,1} + \sum_{z=1}^{2Z} \mathbb{1}\left(s_{k,z}=1\right)$ and $\overline{\gamma}_{s,2} = \underline{\gamma}_{s,2} + \sum_{z=1}^{2Z} \mathbb{1}\left(s_{k,z}=0\right)$.

Appendix B.4 Drawing the stochastic volatility of the idiosyncratic shock

This appendix derives a particle Gibbs (PG) sampling algorithm to generate $h_{i,t} = \log\left(\sigma_{it}^2\right)$ for $i = 1,2,...,N$ from its posterior distribution.

Appendix B.4.1 Sequential Monte Carlo method

Suppose that we have an approximate filtering density for $h_{i,t-1}$:

$$p\left(h_{i,t-1}|r_{i,1:t-1}\right) \approx \sum_{c=1}^{C} \hat{\omega}_{t-1}^{(c)} \delta_{\left(h_{i,t-1}^{(c)}\right)} \tag{B.14}$$

where $\{h_{i,t-1}^{(c)}\}_{c=1}^{C}$ and $\{\hat{\omega}_{t-1}^{(c)}\}_{c=1}^{C}$ denote C particles of $h_{i,t-1}$ and their corresponding normalized importance weights at time $t-1$. The Dirac measure for each particle in $\{h_{i,t-1}^{(c)}\}_{c=1}^{C}$ is denoted by $\delta_{\left(h_{i,t-1}^{(c)}\right)}$. In the filtering density, we suppress other model parameters for notation simplicity.

For forward filtering, an ancestor index denoted by $a_t^{(c)}$ is drawn using the following importance weights prior to sampling a new particle at time t:

$$\mathsf{P}\left(a_t^{(c)} = j\right) = \hat{\omega}_{t-1}^{(j)}, \tag{B.15}$$

where $c = 1, 2, ..., C$. Then, an importance distribution is used to generate a new particle $h_{i,t}^{(c)}$ conditional on $h_{i,t-1}^{(a_t^{(c)})}$, which is the $a_t^{(c)}$-th particle of the particle swarm at time $t-1$. In the particle Gibbs sampling algorithm, we employ the transition density for $h_{i,t}$ which is denoted by $p(h_{i,t}|h_{i,t-1})$. The corresponding un-normalized importance weight is given by

$$\bar{\omega}_t^{(c)} \quad = p\left(r_{i,t}|h_{i,t}^{(c)}\right), \tag{B.16}$$

and the normalized importance weight is given by $\hat{\omega}_t^{(c)} = \frac{\bar{\omega}_t^{(c)}}{\sum_{o=1}^{C} \bar{\omega}_t^{(o)}}$. At time t, therefore, we draw $\{a_t^{(c)}\}_{c=1}^{C}$ with $\{\hat{\omega}_{t-1}^{(c)}\}_{c=1}^{C}$ and generate $\{h_{i,t}^{(c)}\}_{c=1}^{C}$ from its transition density. Then, we compute $\{\hat{\omega}_t^{(c)}\}_{c=1}^{C}$ for the next time period. Continuously performing the particle sampling steps up to time T, we can collect C different volatility paths. This algorithm is referred to as a sequential Monte Carlo (SMC) method.

Appendix B.4.2 Particle Gibbs sampler with ancestor sampling

Among the C particle paths of stochastic volatility generated by the SMC method, we sample one particular volatility path known as a reference particle trajectory. There are several ways to sample the reference particle trajectory. In Bayesian estimation, we employ the ancestor sampling algorithm of Lindsten et al. (2014). Let $h_{i,t}^{*}$ denote $h_{i,t}$ generated at the previous MCMC iteration.

The ancestor sampling method updates the ancestor index for $h_{i,t}^{*}$ using the importance weight proportional to $\bar{w}_{t-1|t}^{(c)} = p(h_{i,t}^{*}|h_{t-1}^{(c)})\hat{\omega}_{t-1}^{(c)}$ for $c = 1, 2, ..., C$. By treating $h_{i,t}^{*}$ as the C-th particle, we draw its ancestor index based on the following normalized importance weight

$$\hat{w}_{t-1|t}^{(c)} = \frac{\bar{w}_{t-1|t}^{(c)}}{\sum_{o=1}^{C} \bar{w}_{t-1|t}^{(o)}}. \tag{B.17}$$

The SMC algorithm combined with this additional step refers to the conditional sequential Monte Carlo (CSMC) algorithm. A detailed exposition regarding its implementation is provided below in the outline of the proposed PG algorithm.

CSMC Algorithm

i) For $t = 1$, set $\{h_{i,1}^{(c)} = 1\}_{c=1}^{C}$ and $\{\hat{\omega}_1^{(c)} = \frac{1}{C}\}_{c=1}^{C}$.

- Iterate steps ii-1), ii-2), ii-3), ii-4), and ii-5) for $t = 2, 3, ..., T$.

ii-1) Draw the ancestor index, $\{a_t^{(c)}\}_{c=1}^{C-1}$ with the normalized probabilities, $\{\hat{\omega}_{t-1}^{(c)}\}_{c=1}^{C}$.

ii-2) Draw the ancestor index, $a_t^{(C)}$ using the normalized weight $\hat{\omega}_{t-1|t}^{(c)} = \frac{\bar{\omega}_{t-1|t}^{(c)}}{\sum_{o=1}^{C} \bar{\omega}_{t-1|t}^{(o)}}$ where

$$\bar{\omega}_{t-1|t}^{(c)} = \frac{1}{\sqrt{2\pi\sigma_{\sigma,i}^2}} \exp\left(-\frac{(h_{i,t}^{*} - \mu_{\sigma,i} - \phi_{\sigma,i}(h_{i,t-1}^{(c)} - \mu_{\sigma,i}))^2}{2\sigma_{\sigma,i}^2}\right) \hat{\omega}_{t-1}^{(c)}.$$

ii-3) Draw $h_{i,t}^{(c)}$ from $p(h_{i,t}|h_{i,t-1}^{(a_t^{(c)})})$ for $c = 1, 2, ..., C-1$.

ii-4) Calculate the unnormalized weight

$$\bar{\omega}_t^{(c)} = \frac{1}{\sqrt{2\pi\sigma_{i,t}^{2(c)}}} \exp\left(-\frac{\varepsilon_{i,t}^2}{2\sigma_{i,t}^{2(c)}}\right)$$

and the normalized weight $\hat{\omega}_t^{(c)} = \frac{\bar{\omega}_t^{(c)}}{\sum_{o=1}^{C} \bar{\omega}_t^{(o)}}$ for $c = 1, 2, ..., C$ where $\sigma_{i,t}^{2(c)} = \exp\left(h_{i,t}^{(c)}\right)$.

- Perform step iii-1), and iii-2) for $t = T$.

iii-1) Draw b_T^{*} with $\{\hat{\omega}_T^{(c)}\}_{c=1}^{C}$ and construct $b_{1:T-1}^{*}$ using $b_{t-1}^{*} = a_t^{(b_t^{*})}$ for $t = T, T-1, ..., 2$.

iii-2) Construct a new reference particle trajectory $h_{i,1:T}^{*} = \{h_{i,1}^{(b_1^{*})}, h_{i,2}^{(b_2^{*})}, \ldots, h_{i,T}^{(b_T^{*})}\}$.

The MCMC algorithm use the newly generated particle trajectory $h_{i,1:T}^{*}$ as the new posterior sample of $h_{i,1:T}$. We perform the SMC algorithm for $i = 1, 2, ..., N$ using the parallel computation built in Matlab to reduce the computational cost.

Conditioning on $h_{i,1:T}$, the posterior sampling for $\{\mu_{\sigma,i}, \phi_{\sigma,i}, \sigma_{\sigma,i}^2\}$ is standard. So, we omit the posterior sampling steps in this appendix.

Appendix B.5 Drawing the stochastic covariance matrix of the risk factors

We apply the PG algorithm to estimate $\mathbf{\Sigma}_{f,1:T}$. The corresponding CSMC algorithm follows the same steps explained in the previous section. The unnormalized importance weight to draw $\{a_t^{(c)}\}_{c=1}^{C-1}$ is given by

$$\bar{w}_{t-1}^{(c)} = (2\pi)^{-\frac{K}{2}} \det\left(\mathbf{\Sigma}_{f,t-1}^{(c)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\mathbf{f}_{t-1} - \boldsymbol{\lambda}_{t-1}^{(c)}\left(\mathbf{z}_{t-2}\right)\right)' \mathbf{\Sigma}_{f,t-1}^{(c)-1}\left(\mathbf{f}_{t-1} - \boldsymbol{\lambda}_{t-1}^{(c)}\left(\mathbf{z}_{t-2}\right)\right)\right) \tag{B.18}$$

where $\mathbf{\Sigma}_{f,t-1}^{(c)}$ is the c-th particle of $\mathbf{\Sigma}_{f,t-1}$, $\boldsymbol{\lambda}_{t-1}^{(c)}\left(\mathbf{z}_{t-2}\right) = \mathbf{\Sigma}_{f,t-1}^{(c)}\mathbf{b}_{t-1}\left(\mathbf{z}_{t-2}\right)$, and $\mathbf{\Sigma}_{f,t-1}^{(c)-1}$ is the inverse of $\mathbf{\Sigma}_{f,t-1}^{(c)}$. We generate $\mathbf{\Sigma}_{f,t}^{(c)}$ for $c = 1, 2, ..., C$ using the transition equations of $\log(g_{k,t})$ and $m_{k_1,k_2,t}$ defined in equation (16). The unnormalized importance weigh is based on the nonlinear factor equation given by

$$\mathbf{f}_t = \boldsymbol{\lambda}_t(\mathbf{z}_{t-1}) + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathrm{N}(0, \mathbf{\Sigma}_{f,t}) \tag{B.19}$$

where $\boldsymbol{\lambda}_t(\mathbf{z}_{t-1}) = \mathbf{\Sigma}_{f,t}\mathbf{b}_t\left(\mathbf{z}_{t-1}\right)$ and $\mathbf{\Sigma}_{f,t} = (\mathbf{M}_t^{-1})\mathbf{G}_t(\mathbf{M}_t^{-1})'$. Conditioning on $\{m_{k,\ell,t}^*, hg_{k,t}^*\}$, the unnormalized importance weight to draw $a_t^{(C)}$ is given by

$$\bar{w}_{t|t-1}^{(c)} = \left[\prod_{k=2}^{K}\prod_{\ell=1}^{k-1} p(m_{k,\ell,t}^*|m_{k,\ell,t-1}^{(c)})\right]\left[\prod_{k=1}^{K} p(hg_{k,t}^*|hg_{k,t-1}^{(c)})\right]\hat{w}_{t-1}^{(c)} \tag{B.20}$$

where $hg_{k,t} = log(g_{k,t})$,

$$p(m_{k,\ell,t}|m_{k,\ell,t-1}) = \frac{1}{\sqrt{2\pi\sigma_{e,k,\ell}^2}} \exp\left(-\frac{e_{k,\ell,t}^2}{2\sigma_{e,k,\ell}^2}\right), \tag{B.21}$$

$$p(hg_{k,t}|hg_{k,t}) = \frac{1}{\sqrt{2\pi\sigma_{g,k}^2}} \exp\left(-\frac{\eta_{g,k,t}^2}{2\sigma_{g,k}^2}\right), \tag{B.22}$$

$e_{k,\ell,t}^2 = m_{k,\ell,t} - m_{k,\ell,t-1}$, and $\eta_{g,k,t} = hg_{k,t} - \mu_{g,k} - \phi_{g,k}(hg_{k,t-1} - \mu_{g,k})$. The CSMC with the above importance weights generates the posterior sample of $\Sigma_{f,1:T}$.

Conditioning on $\Sigma_{f,1:T}$, the posterior sampling for $\{\mu_{g,k}, \phi_{g,k}, \sigma_{g,k}^2, \sigma_{e,k_1,k_2}^2\}$ is standard. So, we omit the posterior sampling steps in this appendix.

Appendix C Computational details

This section explains how to compute the acceptance probability of the MH algorithm to estimate a regression tree.

Appendix C.1 Computation of the marginal likelihood used in the MH algorithm

Suppose we condition on the regression tree $\mathcal{T}$. Let $n_{p,t}$ denote the number of observations in portfolio p at time t, i.e. with $d_{it} = p$. Grouping all the observations in portfolio p together, we get for the unrestricted model

$$\mathbf{r}_{p,t} \quad = \quad \tilde{\mathbf{F}}_{p,t}\tilde{\boldsymbol{\beta}}_p + \boldsymbol{\varepsilon}_{p,t}, \qquad \boldsymbol{\varepsilon}_{p,t} \sim \mathrm{N}\left(0, \boldsymbol{\Sigma}_{r,p,t}\right), \qquad p = 1, \ldots, P \qquad t = 1, \ldots, T$$

where $\tilde{\boldsymbol{\beta}}_p = [\beta_{0,p}, \boldsymbol{\beta}_p']'$, $\boldsymbol{\Sigma}_{r,p,t}$ is a $(n_{p,t} \times n_{p,t})$ diagonal matrix of all σ_{it}^2 included in portfolio p, $\tilde{\mathbf{F}}_{p,t} = [\iota,\ \mathbf{F}_{p,t}]$ is a $(n_{pt} \times (K+1))$ matrix with the factors $\mathbf{f}_t$ replicated in each of the rows

$$\iota \quad = \quad \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{F}_{p,t} = \begin{pmatrix} \mathbf{f}_t' \\ \vdots \\ \mathbf{f}_t' \end{pmatrix}.$$

For the unrestricted model, we only include $\mathbf{F}_{p,t}$ and $\boldsymbol{\beta}_p$ in the factor equation. Next, we stack all the time periods together.

$$\begin{pmatrix} \mathbf{r}_{p,1} \\ \vdots \\ \mathbf{r}_{p,T} \end{pmatrix} \quad = \quad \begin{pmatrix} \tilde{\mathbf{F}}_{p,1} \\ \vdots \\ \tilde{\mathbf{F}}_{p,T} \end{pmatrix} \tilde{\boldsymbol{\beta}}_p + \begin{pmatrix} \boldsymbol{\varepsilon}_{p,1} \\ \vdots \\ \boldsymbol{\varepsilon}_{p,T} \end{pmatrix}$$

$$\left(\mathbf{r}_p = \tilde{\mathbf{F}}_p + \boldsymbol{\varepsilon}_p\right)$$

The marginal likelihood is given by

$$\begin{aligned} p\left(\mathbf{r}_p | \mathcal{T}, \boldsymbol{\Sigma}_{r,p,1:T}, \mathbf{f}_{1:T}\right) \quad &= \quad \int p\left(\mathbf{r}_p, \tilde{\boldsymbol{\beta}}_p | \mathcal{T}, \boldsymbol{\Sigma}_{r,p,1:T}, \mathbf{f}_{1:T}\right) d\tilde{\boldsymbol{\beta}}_p \\ &= \quad \left(\frac{1}{2\pi}\right)^{-\frac{\sum_{t=1}^{T} n_{p,t}}{2}} \left(\prod_{t=1}^{T} |\boldsymbol{\Sigma}_{r,p,t}|^{-\frac{1}{2}}\right) \left|\underline{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}_p}\right|^{-\frac{1}{2}} \left|\overline{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}_p}\right|^{\frac{1}{2}} \\ &\times \quad \exp\left(-\frac{1}{2}\left(-\overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\beta}}_p}' \overline{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}_p}^{-1} \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\beta}}_p} + \underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}}' \underline{\mathbf{V}}_{\boldsymbol{\beta}}^{-1} \underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \sum_{t=1}^{T} \mathbf{r}_{p,t}' \boldsymbol{\Sigma}_{r,p,t}^{-1} \mathbf{r}_{p,t}\right)\right) \end{aligned}$$

where

$$
\begin{aligned}
\overline{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}_p}^{-1} &= \underline{\mathbf{V}}_{\boldsymbol{\beta}}^{-1} + \sum_{t=1}^{T} \tilde{\mathbf{F}}'_{p,t} \boldsymbol{\Sigma}_{r,p,t}^{-1} \tilde{\mathbf{F}}_{p,t}, \\
\overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\beta}}_p} &= \overline{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}_p} \left[\underline{\mathbf{V}}_{\boldsymbol{\beta}}^{-1} \underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \sum_{t=1}^{T} \tilde{\mathbf{F}}'_{p,t} \boldsymbol{\Sigma}_{r,p,t}^{-1} \mathbf{r}_{p,t} \right],
\end{aligned}
$$

$\underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ and $\underline{\mathbf{V}}_{\boldsymbol{\beta}}$ are the prior mean and variance of $\tilde{\boldsymbol{\beta}}_p$ or $\boldsymbol{\beta}_p$.

Appendix C.2 Calculating ratios of binomials

The prior distribution over the number of portfolios is give by:

$$
P \sim 1 + \text{binomial}\left(\underline{p} - 1, \pi_p\right).
$$

Let $P^* = P + 1$ denote the new portfolio count generated from the MH algorithm where P is the portfolio count accepted at the previous MCMC iteration. Then, the ratio of the prior density evaluated at P^* and P is given by:

$$
\begin{aligned}
\frac{p\left(P^*\right)}{p\left(P\right)} = \frac{p\left(P+1\right)}{p\left(P\right)} &= \frac{\frac{(\underline{p}-1)!}{P!((\underline{p}-1)-P)!} \pi_p^{P} \left(1-\pi_p\right)^{(\underline{p}-1)-P}}{\frac{(\underline{p}-1)!}{(P-1)!((\underline{p}-1)-(P-1))!} \pi_p^{P-1} \left(1-\pi_p\right)^{(\underline{p}-1)-(P-1)}} \\
&= \frac{(P-1)!(\underline{p}-P)!}{P!(\underline{p}-P-1)!} \frac{\pi_p}{1-\pi_p} \\
&= \frac{(\underline{p}-P)}{P} \frac{\pi_p}{1-\pi_p}.
\end{aligned}
$$

We note that $P! = P \times (P-1)!$. The above result indicates that the grow proposal of the MH algorithm is never accepted once the current portfolio count P has already reached to the maximum portfolio count $\underline{p}$. Another prior ratio for the portfolio count we use in the merge proposal of the MH algorithm is given by:

$$
\begin{aligned}
\frac{p\left(P^*\right)}{p\left(P\right)} = \frac{p\left(P-1\right)}{p\left(P\right)} &= \frac{\frac{(\underline{p}-1)!}{(P-2)!((\underline{p}-1)-(P-2))!} \pi_p^{P-2} \left(1-\pi_p\right)^{(\underline{p}-1)-(P-2)}}{\frac{(\underline{p}-1)!}{(P-1)!((\underline{p}-1)-(P-1))!} \pi_p^{P-1} \left(1-\pi_p\right)^{(\underline{p}-1)-(P-1)}} \\
&= \frac{(P-1)!(\underline{p}-P)!}{(P-2)!(\underline{p}-P+1)!} \frac{1-\pi_p}{\pi_p} \\
&= \frac{(P-1)}{(\underline{p}-P+1)} \frac{1-\pi_p}{\pi_p}.
\end{aligned}
$$

If $P = 1$, the merge move is never accepted in the MH algorithm according to the above prior ratio.