The Factorial Path-Dependent Market Model

Léo Parent

The Factorial Path-Dependent Market Model

Léo Parent
PRISM Sorbonne
Paris 1 Panthéon-Sorbonne University
leo.parent@etu.univ-paris1.fr

May 30th, 2024

Abstract

This article introduces the factorial path-dependent market (FPDM) model, a multivariate asset price dynamics model in which these dynamics are determined by a set of elementary factors. In this framework, both the factorial drift and factorial volatilities are conditioned by the past dynamics of the factorial portfolios, resulting in a model mostly path-dependent. Derived from this theoretical foundation, the paper subsequently designs a market generator positioned midway between parametric models based on strong assumptions and purely data-driven approaches. The aim is to combine the best of both worlds, offering a model capable of faithfully reproducing the empirical financial dynamics while maintaining a clear understanding of the financial phenomena driven by the simulated price paths. To evaluate the effectiveness of the proposed approach, a thorough out-of-sample assessment of the market generator is conducted based on the S&P500 investment universe.

Keywords: multivariate process, market generator, generation of market scenarios, complex dependence structures, nonlinear dependences, volatility modeling, path-dependent model.

JEL classification: C15, C22, C30, C32, C38, C53, G1, G10, G12, G14, G17.

1 Introduction

"Backtesting against synthetic datasets should be the preferred approach for developing tactical investment algorithms" [37]. This statement from Marco Lopez De Prado underscores a key point: the importance of synthetic financial series in a modern quantitative finance approach. Beyond the interest for strategy backtesting, synthetic data are increasingly used by the industry for other applications such as stress testing ([37], [63], [64]), obtaining risk metrics ([29]), generating conditional scenarios ([9]), etc. However, the consistency of these synthetic data-based approaches depends to a large extent on the ability to accurately model multivariate asset price dynamics. Yet, this modeling represents one of the most challenging problems in quantitative finance.

The difficulty stems from a twofold complexity: a complexity inherent in marginal price dynamics (where each asset price is considered independently of the others) and a complexity in the correlation structure of asset price dynamics. On the first aspect, the distributions of asset returns (and log-returns) at different time scales are non-Gaussian, exhibit heavy tails, and are generally asymmetric ([51], [28]). Moreover, asset price dynamics are path-dependent, leading to significant time-series features such as volatility clustering

and the Zumbach effect ([76], [35]). Regarding the correlation structure of asset price dynamics, it is itself dynamic and exhibits significant variability over time ([75]). Thus, a set of assets that are weakly correlated over a given period can be highly correlated over another period. Additionally, the joint distributions of returns for a significant portion of assets are non-elliptical ([25]), adding yet another layer of complexity to the modeling.

While classical multivariate dynamics models such as those based on multivariate geometric Brownian motion fail to capture these various empirical properties, more sophisticated approaches better suited to handle this complexity have emerged in recent years. In particular, models based on new machine learning methods have garnered particular interest. Among the most popular are models based on Restricted Boltzmann Machines ([44], [48]) and Generative Adversarial Networks ([48], [32], [56]), which have yielded very good results in terms of reproducing the various features characterizing financial time series. Nevertheless, these methods have also important limitations. Firstly, most of them are not suitable for handling the temporal dependence of large multi-dimensional data: as the data dimension increases, calibrating these models becomes impractical ([56]). However, in practice, asset portfolios are typically constructed from investment universes comprising several hundred to several thousand assets. Moreover, these models are typically black boxes. Therefore, while they can be very effective at reproducing the statistical properties of financial series, they do not provide an intelligible framework for the reproduced phenomena.

On the theoretical front, effectively modeling the dynamics of a price system goes beyond the technical question of "which model fits the data best?" Instead, it involves identifying the mechanisms through which the asset price system is formed and understanding the conditions that enable various historical market paths. In this respect, ambitious modeling should enable us to answer questions such as: "what causes a certain asset pair, which was weakly correlated at one time t, to become strongly correlated at another time t'?" In broader terms, it must provide to some extent a hermeneutic of market dynamics. However, the value of such an approach goes beyond the simple search for knowledge for its own sake. It creates conditions that allow for the avoidance, or at least the mitigation, of generalization error better than a model based solely on the brute reproduction of historical sample characteristics, and thereby potentially anticipates the possibility of unprecedented but non-zero probability events.

It is in this perspective that the present article is situated. Its objective is to propose a general model for asset price dynamics that can be adapted into a market generator, bridging the gap between classical Monte Carlo approaches based on parametric models with strong assumptions and purely data-driven approaches without theoretical a priori. The aim is to combine the best of both worlds by offering a model capable of faithfully reproducing the empirical financial dynamics in all their complexity while maintaining a clear understanding of the financial phenomena driven by the simulated price paths. Furthermore, thanks to the theoretical framework induced by the model structure, this approach aims to disentangle contingent statistical phenomena, which are merely expressions of specific historical realizations, from the structural mechanisms that generate the probability distribution of the asset price vector dynamics.

The article is structured as follows. Section 2 introduces the general framework of the Factorial Path-Dependent Market (FPDM) model and outlines the underlying logic of its structure. Section 3 derives a market generator from one specification of this model and presents a calibration method to the latter. Lastly, section 4 assesses this market generator and its calibration from various perspectives using market data.

2 General framework of the factorial Path-Dependent Market Model

2.1 General framework of the FPDM model

2.1.1 Elementary factor-based decomposition of asset price vector dynamics

Let us consider $\mathbf{P}$ as the random vector of prices of dimension $n \times 1$ for an investment universe composed of n assets. Adopting a similar approach to that proposed in [61], we assume that the dynamics of $\mathbf{P}$ are driven by a set of m factors denoted $\left\{\mathbf{F}_j\right\}_{1\leq j\leq m}$, which is composed of m_C common factors and n idiosyncratic factors. We will refer to this set as the elementary factor set. For all t, $\mathrm{d}\mathbf{P}_t$ is defined by the following multi-dimensional stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{P}_t = \mathbf{P}_t \odot (\mathbf{A}\mathrm{d}\mathbf{F}_t), \tag{1}$$

where $\mathbf{F}$ is the vector of elementary factors, and $\mathbf{A}$ is an $n \times m$ factor loadings matrix that represents the sensitivity of the assets to the elementary factors, respectively given by:

$$\mathbf{F}_t^\top = \left(\left(\mathbf{F}_t^{(C)}\right)^\top, \left(\mathbf{F}_t^{(I)}\right)^\top\right) \quad \text{and} \quad \mathbf{A}^\top = \left(\left(\mathbf{A}^{(C)}\right)^\top, \mathbf{I}_n\right),$$

with $\mathbf{F}^{(C)}$ being the m_C-dimensional vector of common elementary factors, $\mathbf{F}^{(I)}$ the n-dimensional vector of idiosyncratic elementary factors, $\mathbf{A}^{(C)}$ the $n \times m_C$ matrix of the sensitivity of the assets to the common elementary factors, and $\mathbf{I}_n$ is the n-dimensional identity matrix. Furthermore, we suppose that the dynamics of the elementary factors are given by:

$$\mathrm{d}\mathbf{F}_t = \boldsymbol{\mu}_t \mathrm{d}t + \sqrt{\boldsymbol{\Omega}}_t \mathrm{d}\mathbf{W}_t, \tag{2}$$

with $\mathbf{W}$ being an m-dimensional Brownian motion, $\boldsymbol{\mu}$ the drift vector of elementary factors of dimension $m \times 1$, and Ω a $m \times m$ diagonal matrix with diagonal elements corresponding to the variance process of the elementary factors. In this framework, the solution to the asset price vector is given by (see proof in appendix B.1):

$$\mathbf{P}_t = \mathbf{P}_0 \odot \exp \circ \left(\int_0^t \mathbf{A}\boldsymbol{\mu}_u - \frac{1}{2} \cdot \mathrm{diag}\left(\mathbf{A}\boldsymbol{\Omega}_u\mathbf{A}^\top\right) \mathrm{d}u + \int_0^t \mathbf{A}\sqrt{\boldsymbol{\Omega}}_u \mathrm{d}\mathbf{W}_u. \right).$$

Therefore, the instantaneous returns are expressed as a linear combination of elementary factor dynamics, with each factor associated with an independent source of randomness corresponding to the margin of $\mathbf{W}$. Furthermore, in this model, the drift vector and covariance matrix associated with instantaneous asset returns are entirely determined by the drift and volatility vector of the elementary factors. Indeed, at time t, the drift vector and covariance matrix of instantaneous asset returns are respectively defined as follows:[1]:

$$\boldsymbol{\mu}_t = \mathbf{A}\boldsymbol{\mu}_t \quad \text{and} \quad \boldsymbol{\Sigma}_t = \mathbf{A}\boldsymbol{\Omega}_t\mathbf{A}^\top.$$

[1] Because $\mathrm{d}t\mathrm{d}t = 0, \mathrm{d}t\mathrm{d}\mathbf{W}_t = \mathbf{0}_m$ and $\mathrm{d}\mathbf{W}_t\mathrm{d}\mathbf{W}_t^\top = \mathbf{I}_m\mathrm{d}t$:

$$\begin{aligned}(\mathrm{d}\mathbf{P}_t \oslash \mathbf{P}_t)(\mathrm{d}\mathbf{P}_t \oslash \mathbf{P}_t)^\top &= \left(\mathbf{A}\boldsymbol{\mu}_t\mathrm{d}t + \mathbf{A}\sqrt{\boldsymbol{\Omega}}_t\mathrm{d}\mathbf{W}_t\right)\left(\mathbf{A}\boldsymbol{\mu}_t\mathrm{d}t + \mathbf{A}\sqrt{\boldsymbol{\Omega}}_t\mathrm{d}\mathbf{W}_t\right)^\top \\ &= \mathbf{A}\boldsymbol{\Omega}_t\mathbf{A}^\top\mathrm{d}t.\end{aligned}$$

From the expression of the matrix of instantaneous asset returns and given $\mathbf{\Omega}_t$ is a diagonal matrix, it follows that the instantaneous correlation between the returns of the i-th and j-th assets in $\mathbf{P}$ is defined by:

$$(\mathbf{C}_t)_{i,j} = \frac{\sum_{k=1}^{m} (\mathbf{A})_{i,k} \cdot (\mathbf{A})_{j,k} \cdot (\mathbf{\Omega}_t)_{k,k}}{\sqrt{\sum_{k=1}^{m} (\mathbf{A})_{i,k}^2 \cdot (\mathbf{\Omega}_t)_{k,k}} \sqrt{\sum_{k=1}^{m} (\mathbf{A})_{j,k}^2 \cdot (\mathbf{\Omega}_t)_{k,k}}}. \tag{3}$$

This expression highlights one of the main advantages of the factorial form of the model: generating a dynamic correlation structure between asset returns from the dynamics of the elementary factors. More precisely, since the exposure of assets to the factor given by the elements of the matrix $\mathbf{A}$ is assumed to be constant, the correlation dynamics are solely driven by the changes in the factorial volatilities within the vector $\mathcal{V}$. Therefore, as illustrated in figure 1 through a simple example, the correlation structure between assets can undergo significant changes following movements in factorial volatilities. This property of the model is of major interest in explaining empirical phenomena, such as the abrupt increase in correlations between different assets that almost systematically accompanies an increase in market factor volatility.

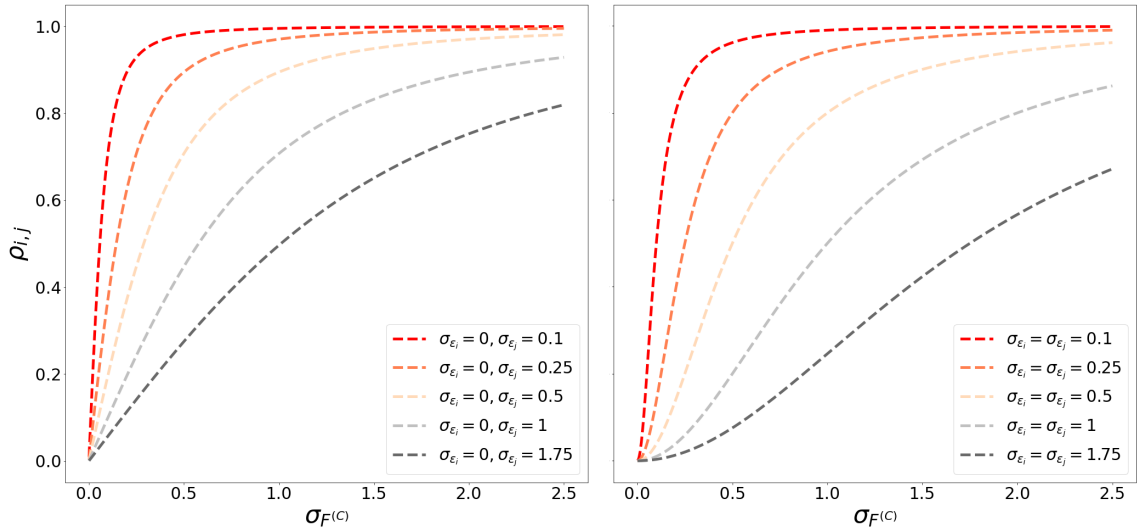


Figure 2: The values taken by $\rho_{i,j}$, the linear correlation coefficient between the instantaneous returns of two assets, as a function of the values taken by the factorial volatilities in the simple case where:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \text{diag}\left(\sqrt{\mathbf{\Omega}_t}\right)^\top = \begin{bmatrix} \sigma_{F^{(C)}} & \sigma_{\epsilon_i} & \sigma_{\epsilon_j} \end{bmatrix}$$

In this example, since both assets, i and j, are positively exposed to a common risk factor $F^{(C)}$, their correlation $\rho_{i,j}$ follows an increasing relationship with the volatility of this factor. However, the profile of the relationship between these two quantities is also strongly influenced by the levels of volatility of the idiosyncratic factors.

2.1.2 The information driven the dynamics of factorial drift and volatilities

In the FPDM model introduced in section 2.1.1, the dynamics of the factors that drive the price vector $\mathbf{P}$ depend on two major components: the vector of factorial drifts $\boldsymbol{\mu}$ and the factorial variances defined by the diagonal of $\mathbf{\Omega}$. However, these components themselves have dynamics that need to be defined. To this end, both can be viewed as functions that, given $\mathcal{I}_t$, a set of information available at time t, map to an m-dimensional vector. More specifically, $\boldsymbol{\mu} : \mathcal{I}_t \to \mathbb{R}^m$ and $\text{diag}_{M \to d}(\mathbf{\Omega}) : \mathcal{I}_t \to \mathbb{R}_+^m$. Adopting this approach, the identification of the information comprising $\mathcal{I}_t$ becomes a central question. First and foremost, this information can be categorized into two main components: endogenous information and

exogenous information. Here, endogenous information corresponds to the natural filtrations of $\mathbf{P}$ and $\mathbf{F}$, i.e., $\{\mathbf{P}_u\}_{u\leq t}$ and $\{\mathbf{F}_u\}_{u\leq t}$. Exogenous information is defined, on the other hand, as the complement of endogenous information, representing the set of information on which $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ depend but is not contained in $\{\mathbf{P}_u, \mathbf{F}_u\}_{u\leq t}$. The philosophy adopted by the FPDM model aligns with that of path-dependent volatility models ([33], [38], [39]), aiming to explain the dynamics of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ as much as possible through an endogenous manner. More specifically, we assume that at a given time t, all the relevant endogenous information on which $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ depend is contained in the following set of state variables:

$$\mathcal{I}_{1,t} = \left\{\tilde{\boldsymbol{\mu}}_t^{(k)}, \tilde{\boldsymbol{\Omega}}_t^{(k)}\right\}_{k=1}^{n_\tau},$$

where $\tilde{\boldsymbol{\mu}}_t^{(k)}$ and $\tilde{\boldsymbol{\Omega}}_t^{(k)}$ correspond to the following exponential weighting moving averages (EWMA):

$$\tilde{\boldsymbol{\mu}}_t^{(k)} = \frac{1}{\tau_k}\int_{-\infty}^{t} e^{-\frac{t-u}{\tau_k}}\mathrm{d}\mathbf{F}_u \quad \text{and} \quad \tilde{\boldsymbol{\Omega}}_t^{(k)} = \frac{1}{\tau_k}\int_{-\infty}^{t} e^{-\frac{t-u}{\tau_k}}\boldsymbol{\Omega}_u\mathrm{d}u,$$

whose dynamics are respectively defined by (see details in the appendix B.2):

$$\mathrm{d}\tilde{\boldsymbol{\mu}}_t^{(k)} = \frac{1}{\tau_k}\cdot\left(\mathrm{d}\mathbf{F}_t - \tilde{\boldsymbol{\mu}}_t^{(k)}\mathrm{d}t\right), \quad \text{and} \quad \mathrm{d}\tilde{\boldsymbol{\Omega}}_t^{(k)} = \frac{1}{\tau_k}\cdot\left(\boldsymbol{\Omega}_u - \tilde{\boldsymbol{\Omega}}_t^{(k)}\right)\mathrm{d}t.$$

Given that $\mathcal{I}_{1,t}$ constitutes a set of information that impact $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, we also have the following inclusion relation: $\mathcal{I}_{1,t} \subseteq \mathcal{I}_t$. However, in the considered approach, the EWMA estimators of the drift and covariance matrix of $\mathbf{F}$, represented by $\tilde{\boldsymbol{\mu}}_t^{(k)}$ and $\tilde{\boldsymbol{\Omega}}_t^{(k)}$ respectively, only impact $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ indirectly. Indeed, the causal relationship operates through transmission channels that consist of metrics associated with a set of factorial portfolios denoted by $\mathcal{Y}$ and defined as:

$$\mathcal{Y}(\mathcal{I}_t) = \left\{\mathbf{y}_{p,t}\right\}_{p=1}^{n_y},$$

with $\mathbf{y}_{p,t} \in \mathbb{R}^m$ and $\forall\ p,t : \|\mathbf{y}_{p,t}\|_1 = 1$. The modeling idea is that variations in factorial trends and volatilities arise from changes in anticipations about the future dynamics of factorial portfolios, anticipations which are formed based on the endogenous information $\mathcal{I}_{1,t}$. Furthermore, the composition of the factorial portfolios included in $\mathcal{Y}$ is itself a deterministic function of the information $\mathcal{I}_t$. This general form includes the particular case where $\mathcal{Y}$ ($\mathcal{Y}_t = \mathcal{Y}\ \forall\ t$) is invariant over time. In this specific case, it is clear that $\mathcal{Y}$ is independent of the information $\mathcal{I}$. Another interesting special case is when the entire set of information on which $\mathcal{Y}$ depends is contained in $\mathcal{I}_{1,t}$. In this configuration, $\mathcal{Y}$ is purely path-dependent. Still, the metrics of the factorial portfolios that impact $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ do not depend on the form taken by $\mathcal{Y}$. These will be of two types.

The first type, which we will refer to as *convolved dynamics features*, takes the form:

$$\hat{\mu}_t\left(\mathbf{y}_{p,t}, \boldsymbol{\delta}_p\right) = \sum_{k=1}^{n_\tau}\left(\boldsymbol{\delta}_p\right)_k \cdot \mathbf{y}_{p,t}^\top\tilde{\boldsymbol{\mu}}_t^{(k)} = \mathbf{y}_{p,t}^\top\int_{-\infty}^{t} g_p(t-u)\mathrm{d}\mathbf{F}_u, \tag{4}$$

where $\boldsymbol{\delta}_p \in \mathbb{R}^{n_y}$ and $g_p(s) = \sum_{k=1}^{n_\tau}\frac{(\boldsymbol{\delta}_p)_k}{\tau_k}e^{-\frac{s}{\tau_k}}$. These features thus correspond to stochastic convolutions with respect to $\mathrm{d}\mathbf{F}$ over the interval $]-\infty, t]$. It should be noted that the kernels $\{g_p\}_{p=1}^{n_y}$ on which these factors depend can be highly diverse. In the case where $\boldsymbol{\delta}_p \in \mathbb{R}_+^{n_y}$ (resp. $\boldsymbol{\delta}_p \in \mathbb{R}_-^{n_y}$), g_p is a positive, decreasing, convex function (resp. negative, increasing, concave) over $\mathbb{R}_+$. In contrast, when the coordinates of $\boldsymbol{\delta}_p$ are not of the same sign, g_p can be non-homogeneous and have variable sign over $\mathbb{R}_+$.

The second type of features impacting $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, that we will term as *historical volatility features*, takes the form:

$$\hat{\sigma}_t\left(\mathbf{y}_{p,t}, \mathbf{w}_p\right) = \sqrt{\sum_{k=1}^{n_\tau} \left(\mathbf{w}_p\right)_k \cdot \mathbf{y}_{p,t}^\top \tilde{\boldsymbol{\Omega}}_t^{(k)} \mathbf{y}_{p,t}} = \sqrt{\mathbf{y}_{p,t}^\top \left(\int_{-\infty}^{t} k_p(t-u)\boldsymbol{\Omega}_u \mathrm{d}u\right) \mathbf{y}_{p,t}} \tag{5}$$

where $\mathbf{w}_p \in \mathbb{R}_+^{n_y}, \|\mathbf{w}_p\|_1 = 1$ and $k_p(s) = \sum_{k=1}^{n_\tau} \frac{(\mathbf{w}_p)_k}{\tau_k} e^{-\frac{s}{\tau_k}}$. Therefore, the form of the kernels $\{k_p\}_{p=1}^{n_y}$ is much more constrained than that of $\{g_p\}_{p=1}^{n_y}$: they are necessarily positive, decreasing, convex functions over $\mathbb{R}_+$, and their integral is equal to 1. This distinction arises from the very nature of this second type of features, which correspond to averages of the realized volatility of factorial portfolios. Specifically, $\hat{\sigma}_t\left(\mathbf{y}_{p,T}, \mathbf{w}_p\right)^2$ is a moving average of the realized variance of the process $(\mathbf{y}_{p,T})^\top \mathrm{d}\mathbf{F}_t$ (not of $(\mathbf{y}_{p,t})^\top \mathrm{d}\mathbf{F}_t$).

These two types of features thus form the set

$$\mathcal{F}_t = \left\{\hat{\mu}_t\left(\mathbf{y}_{p,t}, \boldsymbol{\delta}_p\right), \hat{\sigma}_t\left(\mathbf{y}_{p,t}, \mathbf{w}_p\right)\right\}_{p=1}^{n_y}, \tag{6}$$

which integrates as follows into the causal chain of the FPDM model: $\mathcal{I}_t$ generates the set of factorial portfolios $\mathcal{Y}_t$, and the combination of $\mathcal{I}_t$ and $\mathcal{Y}_t$ defines the set of features $\mathcal{F}_t$ on which $\boldsymbol{\mu}_t$ and $\boldsymbol{\Omega}_t$ depend. More precisely, as will be exposed in section 2.2, the set $\mathcal{F}_t$ determines (either entirely or partially) $\boldsymbol{\Omega}_t$, and the triplet $(\boldsymbol{\Omega}_t, \mathcal{F}_t, \mathcal{Y}_t)$ in turn defines $\boldsymbol{\mu}_t$. The figure 3 summarizes the sequence of causal relationships in the FPDM model.

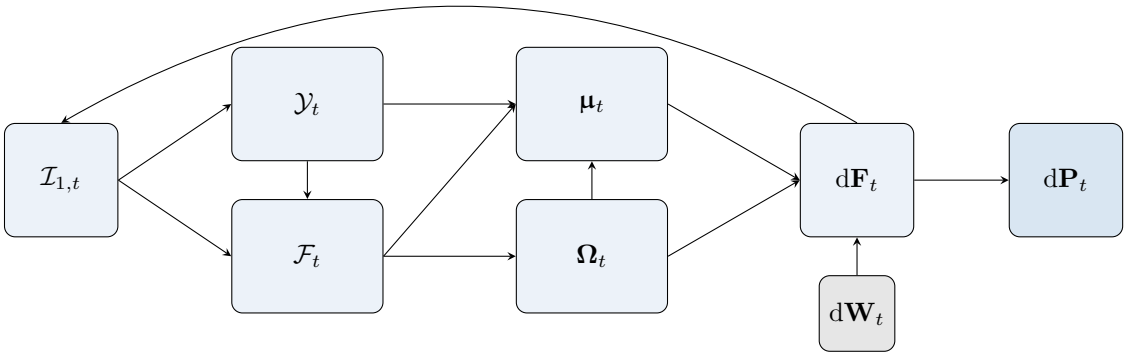


Figure 3: Causal diagram of the FPDM model in the case where $\mathcal{Y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$ depend solely on endogenous information. The endogenous information $\mathcal{I}_{1,t}$ defines the set of factorial portfolios $\mathcal{Y}_t$. The features constituting the set $\mathcal{F}_t$ are computed from the pair $(\mathcal{I}_{1,t}, \mathcal{Y}_t)$, and this set determines the value of the matrix of factorial variances $\boldsymbol{\Omega}_t$. In turn, $\boldsymbol{\Omega}_t, \mathcal{F}_t$ and $\mathcal{Y}_t$ determine the vector of factorial drifts $\boldsymbol{\mu}_t$. Finally, the pair of $\boldsymbol{\mu}_t$ and $\boldsymbol{\Omega}_t$, coupled with the random shocks modeled by $\mathrm{d}\mathbf{W}_t$, generates the dynamics $\mathrm{d}\mathbf{F}_t$ of the elementary factors, which ultimately leads to the variation in the vector of prices $\mathrm{d}\mathbf{P}_t$. Furthermore, the dynamics of the elementary factors produces a feedback effect by increasing the endogenous information and, through the same causal chain, modifying the state of the system in turn.

2.2 Volatilities and drifts of elementary factors

2.2.1 The factorial volatilities

The modeling approach adopted for factorial volatilities $\sqrt{\boldsymbol{\Omega}}$ presents significant analogies with the PDV model proposed by Guyon and Lekeufack ([39]). Similar to this model, volatility dynamics are considered as a primarily endogenous phenomenon with the addition and interaction of short-term exogenous dynamics. Accordingly, the matrix of factorial volatilities takes the following general form[2]:

$$\sqrt{\boldsymbol{\Omega}}_t = \operatorname{diag}\big(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right) \odot \mathbf{X}_t \cdot \mathcal{S}_t\big), \tag{7}$$

where $\boldsymbol{\mathcal{V}}$ is a deterministic function of the set of features $\mathcal{F}_t$ such as $\boldsymbol{\mathcal{V}} : \mathcal{F}_t \to \mathbb{R}_+^m$, $\mathbf{X}$ is an m-dimensional stochastic process responsible for capturing for capturing the exogenous dynamics of factorial volatilities, and $\mathcal{S}$ is a univariate stochastic process corresponding to a common factor sensitivity operator to random frictions.

More precisely, we define the path-dependent component of the vector of volatilities of the elementary factors $\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)$ as follows:

$$\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right) = \underbrace{\mathbf{b}_0}_{(1)} + \underbrace{\sum_{p=1}^{n_y} \mathbf{b}_{1,p} \cdot \hat{\mu}_t\left(\mathbf{y}_p, \boldsymbol{\delta}_p^{(\mathcal{V})}\right)}_{(2)} + \underbrace{\sum_{p=1}^{n_y} \mathbf{b}_{2,p} \cdot \hat{\sigma}_t\left(\mathbf{y}_p, \mathbf{w}_p^{(\mathcal{V})}\right)}_{(3)}, \tag{8}$$

where:

- (1) is an intercept vector such as $\mathbf{b}_0 \in \mathbb{R}_+^m$.
- (2) is the component of $\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)$ attributable to the convolved dynamics features (equation 4), with $\mathbf{b}_{1,p} \in \mathbb{R}^m$.
- (3) is the component of $\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)$ attributable to the historical volatility features (equation 5), with $\mathbf{b}_{2,p} \in \mathbb{R}_+^m$ and $\sum_{p=1}^{n_y} \mathbf{b}_{2,p} \leq \mathbf{1}_m$.

This formalization may be viewed as a generalization of the PDV model proposed by Guyon and Lekeufack ([39]) in a factorial and multivariate framework. Indeed, first, the 4-factor PDV ([39]) model is a specific case of 8 (see appendix C.1), second, even the general form of 8 exhibits an analogous three-block structure. However, the generalization induced by 8 entails substantial specificities that need to be detailed.

Firstly, regarding component (2), where its counterpart in the Guyon and Lekeufack model corresponds to a moving average of past returns (up to a multiplicative constant) for the considered asset, (2) is a linear combination of stochatic convolution of factorial portfolio dynamics. This aims to capture the impact of past dynamics of the factorial portfolios belonging to $\mathcal{Y}_t$ on the level of factorial volatilities. The well-known leverage effect and the strong Zumbach effect ([76], [35]), respectively defined as the negative relationship between past returns and spot volatility, and the dependence of the volatility process on the historical price path, can thus be modeled through this component. In addition to these effects already captured in a univariate framework by the PDV model of Guyon and Lekeufack, (2) allows for the consideration of

[2] Or equivalently:

$$\boldsymbol{\Omega}_t = \mathcal{S}_t^2 \cdot \left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right) \odot \mathbf{X}_t\right)\left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right) \odot \mathbf{X}_t\right)^\top \odot \mathbf{I}_m.$$

other effects specific to multidimensional and factorial modeling. In particular, the past dynamics of one elementary factor can impact the volatility level of another elementary factor. By extension, the volatility level of an asset can be conditioned by the past dynamics of other assets. Concretely, this transmission mechanism allows for the modeling of empirical phenomena such as the difference in amplitude of the leverage effect between individual assets and indices that group a set of assets ([16]).

This type of transmission mechanism is also enabled by component (3), which is a linear combination of historical volatility features associated with different factor portfolios. It is akin to the "historical volatility factor" in the Guyon and Lefeufack model. However, unlike the latter, (3) does not solely depend on the volatility of an elementary factor based on its own past trajectory (the past trajectory of the volatility of this elementary factor) but also depends on the past trajectories of volatilities of other elementary factors. This structure allows for capturing, in addition to the volatility feedback and volatility clustering phenomena already enabled by the Guyon and Lefeufack model, effects specific to multi-asset dynamics, such as volatility spillover effects ([41], [6], [26]). Thus, an increase in the volatility level of a subset of the investment universe can propagate to other assets through this relationship.

In addition to the path-dependent component $\boldsymbol{\mathcal{V}}(\mathcal{F}_t)$ just defined, as per 7, the factorial volatilities are functions of the vector $\mathbf{X}$ whose purpose is to capture dynamics of exogenous origin in volatility. The simplest case is, of course, when $\mathbf{X}$ is invariant (typically $\forall\ t : \mathbf{X}_t = \mathbf{1}_m$), and therefore, the factorial volatilities are purely path-dependent. Beyond this particular case, we adopt and generalize once again the idea proposed by Guyon and Lekeufack by specifying $\mathbf{X}$ as a mean-reverting process. More precisely, following [59] adapted to the considered multivariate framework, we define $\mathbf{X}$ as an m-dimensional exponential Ornstein-Uhlenbeck process, such that:

$$\mathbf{X}_t = \exp \circ (\mathbf{Y}_t) \quad \text{with} \quad \mathrm{d}\mathbf{Y}_t = \boldsymbol{\Upsilon}\left(\bar{\mathbf{Y}} - \mathbf{Y}_t\right)\mathrm{d}t + \boldsymbol{\Psi}\mathrm{d}\mathbf{B}_t, \tag{9}$$

where $\boldsymbol{\Upsilon}$ is a $m \times m$ transition matrix, $\bar{\mathbf{Y}}$ is the unconditional expectation of $\mathbf{Y}$, $\boldsymbol{\Psi}$ is a $m \times m$ scatter matrix, and $\mathbf{B}$ is a m-dimensional Brownian motion independent of $\mathbf{W}$. As shown in [54], the conditional distribution of $\mathbf{Y}_{t+s}$ given $\mathbf{Y}_t$ is normal at all times, such that:

$$\mathbf{Y}_{t+s}|\mathbf{Y}_t \sim \mathcal{N}\left(\bar{\mathbf{Y}} + \mathrm{e}^{-\boldsymbol{\Upsilon}\cdot s}\left(\mathbf{Y}_t - \bar{\mathbf{Y}}\right),\ \mathrm{vec}^{-1}_{m\times m}\left(\left(\boldsymbol{\Upsilon}\oplus\boldsymbol{\Upsilon}\right)^{-1}\left(\mathbf{I}_m - e^{-(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)\cdot s}\right)\mathrm{vec}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)\right)\right).$$

Therefore $\mathbf{X}_{t+s}$ conditional on $\mathbf{X}_t$ follows the log-normal distribution:

$$\mathbf{X}_{t+s}|\mathbf{X}_t \sim \mathcal{LN}\left(\bar{\mathbf{Y}} + \mathrm{e}^{-\boldsymbol{\Upsilon}\cdot s}\left(\mathbf{Y}_t - \bar{\mathbf{Y}}\right),\ \mathrm{vec}^{-1}_{m\times m}\left(\left(\boldsymbol{\Upsilon}\oplus\boldsymbol{\Upsilon}\right)^{-1}\left(\mathbf{I}_m - e^{-(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)\cdot s}\right)\mathrm{vec}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)\right)\right).$$

Furthermore, the asymptotic distribution of $\mathbf{X}$ is given by:

$$\lim_{s\to+\infty}\mathbf{X}_{t+s}|\mathbf{X}_t \sim \mathcal{LN}\left(\bar{\mathbf{Y}},\ \mathrm{vec}^{-1}_{m\times m}\left(\left(\boldsymbol{\Upsilon}\oplus\boldsymbol{\Upsilon}\right)^{-1}\mathrm{vec}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)\right)\right).$$

Several remarks can be made regarding the specification of parameters defining the dynamics of $\mathbf{X}$. First, if we assume that the path-dependent component exhausts the structural relationships linking the volatilities of the elementary factors, it is consistent then for $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Psi}$ to be both diagonal matrices. Additionally, [39] and [59] tend to show that when adopting a primarily path-dependent calibration approach, exogenous dynamics of volatility appear as a short-term phenomenon (intraday). In this context, $\mathbf{Y}$ mean-reverts

very fast toward $\bar{\mathbf{Y}}$, implying high values for elements of $\boldsymbol{\Upsilon}$.

Finally, the last component of the factorial volatilities $\mathcal{S}$ is a univariate process that defines a common sensitivity to all factors. This allows for capturing the homothetic dynamics of the matrix of factorial volatilities. Like $\mathbf{X}$, aside from the special case of constant $\mathcal{S}$, it is coherent to model it using a mean-reverting process. Thus, the choice of an exponential Ornstein-Uhlenbeck process remains meaningful in this context. An alternative is to model $\log(\mathcal{S})$ through a continuous-time ARMA (autoregressive-moving average) process ([19], [20], [22]). Due to the multitude of approaches included in this family of models, this choice provides a high degree of flexibility in capturing various types of dynamics, particularly in terms of temporal dependence relationships.

2.2.2 The drifts of the elementary factors

The modeling of the drift vector of the elementary factors proposed in this section aims to achieve a dual purpose. On one hand, it aims to incorporate the insights provided by financial literature, and on the other hand, to be flexible enough to capture potential market-specific patterns that may not be accounted for by standard approaches. In line with this objective, we propose to define this drift as the following linear combination:

$$\boldsymbol{\mu}_t = \sum_{p=1}^{n_y} \boldsymbol{\beta}_{p,t} \cdot \Gamma_{p,t}, \tag{10}$$

where Γ_p is the performance factor associated with portfolio $\mathbf{y}_{p,t}$, and $\beta(\mathbf{y}_{p,t})$ is the vector of sensitivities of the elementary factors defined by:

$$\boldsymbol{\beta}_{p,t} = \frac{\mathrm{Cov}\left(\mathrm{d}\mathbf{F}_t, \mathbf{y}_{p,t}^\top \mathrm{d}\mathbf{F}_t\right)}{\mathrm{Var}\left(\mathbf{y}_{p,t}^\top \mathrm{d}\mathbf{F}_t\right)} = \frac{\boldsymbol{\Omega}_t \mathbf{y}_{p,t}}{\mathbf{y}_p^\top \boldsymbol{\Omega}_t \mathbf{y}_{p,t}}. \tag{11}$$

The form 10, although very general in its current state, frames the modeling of $\boldsymbol{\mu}$ by defining the drifts of the elementary factors based on their respective exposures to the portfolios included in $\mathcal{Y}$. Therefore, the composition of $\mathcal{Y}$ plays a crucial role in this framework. Furthermore, it is clear from 10 that the significance of the various portfolios contained in $\mathcal{Y}$ on $\boldsymbol{\mu}$ is contingent upon the forms taken by the functions Γ_p. Here, according to our dual objective outlined in the introduction, we consider the following specification:

$$\Gamma_{p,t} = \underbrace{\lambda_p^{(S)}\left(\sigma_t(\mathbf{y}_{p,t})\right)}_{(1)} + \underbrace{\lambda_p^{(P)}\left(\hat{\sigma}_t(\mathbf{y}_{p,t}, \mathbf{w}_p)\right)}_{(2)} + \underbrace{\hat{\mu}_t\left(\mathbf{y}_{p,t}, \boldsymbol{\delta}_p\right)}_{(3)} + \underbrace{\mathcal{E}_{p,t}}_{(4)}, \tag{12}$$

where:

- (1) corresponds to the instantaneous risk premium of the portfolio $y_{p,t}$, which is a function of the value taken by $\sigma_t\left(\mathbf{y}_{p,t}\right)$, the spot volatility of this portfolio, such that $\lambda_p^{(S)} : \mathbb{R}_+ \to \mathbb{R}$.
- (2) corresponds to the premium compensating for the historical risk of the portfolio $y_{p,t}$ measured by $\hat{\sigma}_t\left(\mathbf{y}_{p,t}, \mathrm{w}_p\right)$ (equation 5), such that $\lambda_p^{(P)} : \mathbb{R}_+ \to \mathbb{R}$.
- (3) corresponds to the feedback effect of the past dynamics of $\mathbf{y}_{p,t}$ on its current drift deined by equation 4).
- (4) corresponds to the residual component of the performance factor associated with portfolio $\mathbf{y}_{p,t}$, not explained by the first three components.

This modeling approach has the double advantage of being highly flexible while preserving a clear understanding of the factors influencing the vector of factor drifts. For instance, as demonstrated in appendix C.2, the proposed model can be specified to align with the theoretical coordinates of the capital asset pricing model (CAPM) ([72], [52]). In addition to a strict adoption of this type of specification based on a strong theoretical framework, those can serve as a baseline for integrating other components of $\boldsymbol{\mu}$. This allows for a combination of a theoretical model with a data-driven approach. Moreover, the combination of the four components that form a performance factor enables the modeling of complex drift patterns and different natures.

Firstly, components (1) and (2) allow linking the volatility levels of factorial portfolios - instantaneous volatilities in the case of (1) and past volatilities in the case of (2) - with drifts. However, the relationship between the level of volatility and expected returns is both common in financial modeling ([70], [40]) and extensively documented [34]. Furthermore, the very general definition given to $\lambda_p^{(S)}(.)$ and $\lambda_p^{(P)}(.)$ enables the consideration of a variety of ways to model this relationship, including for instance linear and polynomial forms, as proposed in [40]. Regarding the dual risk premium structure employed, comprising a spot volatility premium and a historical volatility premium, it may seem unconventional at first glance. Nevertheless, in practice, this specific form allows for the modeling of various important empirical phenomena documented in financial literature. One such example is the market behaviors studied by French et al. ([34]), where the expected returns of stocks are positively related to predictable volatility through autoregressive (AR) models and negatively related to unexpected changes in volatility. In the proposed framework, considering y_p as a factorial portfolio representative of the equity market, this phenomenon can be captured by jointly specifying $\lambda_p^{(S)}$ as an increasing function on $\mathbb{R}_+$ and $\lambda_p^{(P)}$ as a positively increasing function on $\mathbb{R}_+$. Indeed, in this context, on the one hand, the variation in $(1)+(2)$ is negatively correlated with a volatility shock, and on the other hand, $(1)+(2)$ follows a positive relationship with the level of volatility predicted by an AR model due to component (2). A second example of an empirical pattern that the dual form of the risk premium can capture is the well-documented 'fly-to-quality' phenomenon ([11], [23], [7]). Indeed, let us suppose that y_q corresponds to a quality portfolio, and that the associated risk premiums take the form $\lambda_q^{(S)}(\sigma) = \bar{\lambda}_q^{(S)}\sigma^2$ and $\lambda_q^{(P)}(\sigma) = \bar{\lambda}_q^{(P)}\sigma^2$ with $0 < \bar{\lambda}_q^{(S)} = -\bar{\lambda}_q^{(P)}$. In this setup, during prolonged periods of stability in the volatility level, $(1)+(2)$ is close to zero. However, if the instantaneous volatility suddenly spikes, $(1)+(2)$ turns positive due to a positive delta between spot volatility and the historical volatility factor. Conversely, when volatility decreases, $(1)+(2)$ turns negative, spot volatility becoming lower than historical volatility.

Regarding (3), this feedback feature enables the modeling of momentum and mean-reversion phenomena, two important determinants of empirical price dynamics ([62], [36], [71]). The nature of the feedback generated by this component depends on the value of $\boldsymbol{\delta}_p$, as clearly evident from its mathematical expression provided by 4. Thus, when $\delta_p \in \mathbb{R}_+^{n_y}$ and if at least one coordinate of $\boldsymbol{\delta}_p > 0$, (3) may be viewed as a momentum factor. Conversely, if $\boldsymbol{\delta}_p \in \mathbb{R}_-^{n_y}$ and at least one coordinate of $\boldsymbol{\delta}_p < 0$, (3) can be seen as a reversal factor. Beyond these two polar cases, the form of (3) allows capturing more complex feedback structures, such as the coexistence of positive autocorrelation in returns over short horizons and negative autocorrelation over longer horizons highlighted by Poterba and Summers ([62]).

The component (4) differs from the first three components of 12 on several fronts. Firstly, it is not necessarily solely a function of endogenous information but may also depend on exogenous information contained in $\mathcal{I}_t$. Consequently, this component enables the consideration of exogenous determinants of factor drifts, allowing for the concrete incorporation of factors such as 'views' in the Black-Litterman sense

on portfolios. Moreover, (4) can serve to ensure certain relationships between different components of the model, as in the case of the CAPM specification of $\boldsymbol{\mu}$ discussed in appendix C.2.

3 The Factorial Path-Dependent Market Generator

3.1 The market generator framework

3.1.1 The considered RPDV model

The theoretical framework proposed in section 2, defining the RPDV model, is intentionally very general and encompasses a large number of possible specifications. This section aims to define the RPDV under consideration, which will be used, in a discretized version, as a market generator in the remainder of the article.

3.1.1.1 The set of factor portfolios $\mathcal{Y}$

Firstly, the set of factor portfolios $\mathcal{Y}$ is defined as:

$$\mathcal{Y} = \{\mathbf{e}_p\}_{p=1}^{m},$$

where $\mathbf{e}_p$ is an $m \times 1$ vector, with the value of the p-th row being 1 and 0 for all other rows. This is one of the simplest possible specifications of $\mathcal{Y}$, which implies that the features on which the factor drifts and volatilities depend are each a function of a single elementary factor. Furthermore, through equations 4 and 5, the drift vector of the elementary factors can be simplified as follows:

$$\boldsymbol{\mu}_t = \sum_{p=1}^{m} \boldsymbol{\beta}_{p,t} \cdot \Gamma_{p,t} = \sum_{p=1}^{m} \frac{\boldsymbol{\Omega}_t \mathbf{e}_p}{\mathbf{e}_p \boldsymbol{\Omega}_t \mathbf{e}_p} \cdot \Gamma_{p,t} = \sum_{p=1}^{m} \mathbf{e}_p \cdot \Gamma_{p,t} = \boldsymbol{\Gamma}_t.$$

Therefore, the drift of an elementary factor j depends solely on the performance factor $\Gamma_{j,t}$ associated with it (since $\boldsymbol{\beta}_{j,t} = \mathbf{e}_j$).

3.1.1.2 The kernels

The specification adopted for the kernels associated with the features on which the factor drifts and volatilities depend is also made with a concern for parsimony. Thus, the considered model includes only two kernels: the first shared by all convolved dynamics features and the second shared by all historical volatility features, respectively defined by:

$$g^{(\hat{\mu})}(s) = \sum_{k=1}^{n_\tau} \frac{\boldsymbol{\delta}_k}{\tau_k} e^{-\frac{s}{\tau_k}} \quad \text{and} \quad g^{(\hat{\sigma})}(s) = \sum_{k=1}^{n_\tau} \frac{\mathbf{w}_k}{\tau_k} e^{-\frac{s}{\tau_k}}, \tag{13}$$

where $\boldsymbol{\delta}, \boldsymbol{w} \in \mathbb{R}_+^{n_\tau}$ and with

$$\tau_k = \exp\left(\log(\tau_-) + \frac{\log(\tau_+) - \log(t_-)}{n_\tau - 1}(k-1)\right). \tag{14}$$

This method of defining the parameters $\{\tau_k\}_{k=1}^{n_\tau}$, introduced in [58], allows for a good approximation for the majority of positive decreasing kernels on $\mathbb{R}_+$ as long as one opts for a specification of n_τ, t_-, t_+. In

this case, we choose the following specification: $n_\tau = 10$ and $1/365$ and 5 (expressed in years).

3.1.1.3 The volatilities of elementary factors

The specification of the vector of elementary factors is based on the idea that one elementary factor, corresponding to the market factor, plays a specific role in the overall level of factorial volatilities. In this framework, the purely path-dependent component of the volatility of an elementary factor depends solely on its own past trajectory and the past trajectory of the market factor. Following this principle, by associating the market factor with the first (in terms of index) elementary factor, we define this purely path-dependent component for an elementary factor k as:

$$\left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)\right)_k = \mathrm{b}_{0,k} + \mathrm{b}_{1,k} \cdot \hat{\mu}_t\left(\mathbf{e}_k, \boldsymbol{\delta}\right) + \mathrm{b}_{2,k} \cdot \hat{\sigma}_t\left(\mathbf{e}_k, \mathbf{w}\right) + \mathrm{b}_{3,k} \cdot \left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)\right)_1, \tag{15}$$

where $\forall\ k : \mathrm{b}_{0,k}, \mathrm{b}_{2,k}, \mathrm{b}_{3,k} \in \mathbb{R}_+$. In a less compact form, equation 15 can be rewritten as follows:

$$\begin{aligned}\left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)\right)_k = \mathrm{b}_{0,k} + \mathrm{b}_{3,k}\mathrm{b}_{0,1} + \mathrm{b}_{1,k}\int_{-\infty}^{t} K^{(\hat{\mu})}(u)\mathrm{d}(\mathbf{F}_u)_k + \mathrm{b}_{3,k}\mathrm{b}_{1,1}\int_{-\infty}^{t} K^{(\hat{\mu})}(u)\mathrm{d}(\mathbf{F}_u)_1 \\ + \mathrm{b}_{2,k}\sqrt{\int_{-\infty}^{t} K^{(\hat{\sigma})}(u)(\boldsymbol{\Omega}_u)_{k,k}\mathrm{d}u} + \mathrm{b}_{3,k}\mathrm{b}_{2,1}\sqrt{\int_{-\infty}^{t} K^{(\hat{\sigma})}(u)(\boldsymbol{\Omega}_u)_{1,1}\mathrm{d}u}.\end{aligned}$$

Certainly, for the specific case of the market factor (i.e. $k = 1$), $b_{3,1} = 0$. Consequently, the purely path-dependent component of the market factor volatility takes a form analogous to the PDV model by Guyon and Lekeudfack:

$$\left(\boldsymbol{\mathcal{V}}\left(\mathcal{F}_t\right)\right)_1 = \mathrm{b}_{0,1} + \mathrm{b}_{1,1}\int_{-\infty}^{t} K^{(\hat{\mu})}(u)\mathrm{d}(\mathbf{F}_u)_1 + \mathrm{b}_{2,1}\sqrt{\int_{-\infty}^{t} K^{(\hat{\sigma})}(u)(\boldsymbol{\Omega}_u)_{1,1}\mathrm{d}u}.$$

Regarding $\mathbf{X}$, the process responsible for capturing the exogenous dynamics of factorial volatilities, we adopt here the form suggested in section 2.2, namely an exponential OU process. More precisely, reusing the form 9, we consider the case where both $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Psi}$ are diagonal matrices, such that the coordinates are independent. Furthermore, we assume that the mean-reversion parameters associated with this process are very high, corresponding to high or medium frequency phenomena. Thus, for the time step Δt considered in the market generator, the realizations of $\mathbf{X}_t$ and $\mathbf{X}_{t+\Delta t}$ are approximately independent. This modeling assumes that, for low-frequency observations, the autocorrelation structure of volatilities is entirely captured by $\boldsymbol{\mathcal{V}}$ and $\mathcal{S}$.

To model the dynamics of the market sensitivity operator $\mathcal{S}$, we assume that $\log(\mathcal{S})$ follows a CARMA process. In practice, since the market generator is a discrete model, the selected model will be a standard ARMA(1,1). This choice, which may appear arbitrary at first glance, is actually determined by the empirical dynamics of $\mathcal{S}$, which are fairly well-modeled by this type of process (see appendix G.3).

3.1.1.4 The factorial drifts

The specification of the factor drift vector follows a CAPM-like framework under the assumption of zero interest rates: only the market factor has a non-zero drift. Specifically, this vector is defined by the expression:

$$\boldsymbol{\mu}_t = \boldsymbol{\Gamma}_t = \mathbf{e}_1 \cdot \underbrace{\left(\bar{\mu} + \zeta \cdot \hat{\mu}_t\left(\mathbf{e}_1, \boldsymbol{\delta}\right) + \lambda \cdot \hat{\sigma}_t\left(\mathbf{e}_1, \mathbf{w}\right)\right)}_{\Gamma_t^\star}.$$

where $\Gamma^\star$ respresents the market premium. This comprises three components. The first, $\bar{\mu}$, represents the invariant part of the market premium. The second component, $\zeta \cdot \hat{\mu}_t\left(\mathbf{e}_1, \boldsymbol{\delta}\right)$, models the effect of past market factor dynamics on their current market premium. If ζ is positive, the market premium exhibits momentum: the market factor is subject to a positive feedback effect. Conversely, if ζ is negative, the market premium exhibits reversal: the market factor is subject to a negative feedback effect. The third component, $\lambda \cdot \hat{\sigma}_t\left(\mathbf{e}_1, \mathbf{w}\right)$, represents a premium for historical volatility.

Several additional remarks can be made. Firstly, the market risk premium, and consequently the drift, is purely path-dependent. Only the feature variables determining the path-dependent component of the volatility of the market factors make the drifts time-variable. Furthermore, the market risk premium can be rewritten as follows:

$$\Gamma_t^\star = \bar{\mu}' + \lambda^{\mathcal{V}} \cdot \left(\mathcal{V}_t\right)_1 + \zeta' \cdot \hat{\mu}_t\left(\mathbf{e}_1, \boldsymbol{\delta}\right) + \lambda' \cdot \hat{\sigma}_t\left(\mathbf{e}_1, \mathbf{w}\right),$$

$\bar{\mu}' = \bar{\mu} - \lambda^{\mathcal{V}} \cdot b_{0,1}, \zeta' = \zeta - \lambda^{\mathcal{V}} \cdot b_{1,1}$ and $\lambda' = \lambda - \lambda^{\mathcal{V}} \cdot b_{2,1}$. Therefore, it has as a special case $\Gamma_t^\star = \bar{\mu} + \lambda^{\mathcal{V}} \cdot \left(\mathcal{V}_t\right)_1$. Another important point is that the drift vector of the assets is defined as:

$$\mathbf{A}\boldsymbol{\mu}_t = \left(\mathbf{A}\right)_{[:,1]} \cdot \Gamma_t^\star.$$

Consequently, in this framework, changes in the asset drifts are solely caused by movements in the market risk premium.

3.1.2 The market generator

The considered market generator is a discrete version of the FPDM model discussed in dection 3.1.1. It is defined by algorithm 1. In addition to the time step Δt and the number of simulations n_s, it takes as input the set of parameters $\boldsymbol{\delta}, \mathbf{w}, \mathbf{A}, \boldsymbol{\tau}, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{S}, \bar{\mu}, \zeta, \lambda, a_0, a_1, a_2$, and the set of state variables $\mathbf{P}_0, \mathbf{M}_0^{(\tilde{\boldsymbol{\mu}})}, \mathbf{M}_0^{(\tilde{\boldsymbol{\Omega}})}, \epsilon_0$, whose determination is discussed in next section 3.2. These inputs are then used to generate a time series $\left\{\mathbf{P}_{u\Delta t}\right\}_{0 \leq t \leq u}$, which constitutes the output of the generator[3]. The objective of this section is to elucidate the various components of this market generator, including the discretization choices made and the rationale behind these choices.

Firstly, the SDEs associated with the dynamics of the price vector and elementary factors are discretized using a simple Euler scheme. Interestingly, the price is subject to the positive part operator $(.)_+$. This characteristic has two major advantages. Firstly, it ensures that prices remain positive or zero. Secondly, it allows for the possibility of bankruptcy or default: when the price reaches 0, its value remains null for all subsequent dates. In the case where the price vector corresponds to a set of stocks, this property allows,

[3]Of course, other elements used for simulating this time series can be added to the output, depending on the purpose of using the market generator. Typically, it may be beneficial to also retain the trajectories of elementary factors or volatility.

for example, to estimate by simulation the probability of default at a given time horizon for the different companies in the considered investment universe.

Algorithm 1 The Factorial Path-dependent Market Generator.

Input: $\mathbf{P}_0, \mathbf{M}_0^{(\tilde{\boldsymbol{\mu}})}, \mathbf{M}_0^{(\tilde{\boldsymbol{\Omega}})}, \epsilon_0, \boldsymbol{\delta}, \mathbf{w}, \mathbf{A}, \boldsymbol{\tau}, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{S}, \bar{\mu}, \zeta, \lambda, a_0, a_1, a_2, \Delta t, n_s$.

For $t = 1$ to n_s **do**

1. Sample $\left(\boldsymbol{W}_{t+\Delta t}^\top, \boldsymbol{B}_{t+\Delta t}^\top, \epsilon_{t+\Delta t}\right)^\top$ from $\mathcal{N}\left(\mathbf{0}_{2m+1}, \mathbf{I}_{2m+1}\right)$.

2. Update system equations:

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_{t+\Delta t} &\leftarrow \boldsymbol{\delta}^\top \mathbf{M}_t^{(\tilde{\boldsymbol{\mu}})} \\
\hat{\boldsymbol{\sigma}}_{t+\Delta t} &\leftarrow \left(\mathbf{w}^\top \mathbf{M}_t^{(\tilde{\boldsymbol{\Omega}})}\right)^{\circ 1/2} \\
\mathcal{V}_{t+\Delta t} &\leftarrow \mathbf{b}_0 + \mathbf{b}_1 \odot \hat{\boldsymbol{\mu}}_{t+\Delta t} + \mathbf{b}_2 \odot \hat{\boldsymbol{\sigma}}_{t+\Delta t} + \mathbf{b}_3 \cdot \left(\mathbf{b}_0 + \mathbf{b}_1 \odot \hat{\boldsymbol{\mu}}_{t+\Delta t} + \mathbf{b}_2 \odot \hat{\boldsymbol{\sigma}}_{t+\Delta t}\right)_1 \\
\mathbf{X}_{t+\Delta t} &\leftarrow \exp \circ \left(\mathbf{S} \odot \left(\boldsymbol{B}_{t+\Delta t} - \mathbf{S}\right)\right) \\
\log\left(\mathcal{S}_{t+\Delta t}\right) &\leftarrow a_0 + a_1 \log\left(\mathcal{S}_{t+\Delta t}\right) + a_2 \epsilon_t + \epsilon_{t+\Delta t} \\
\boldsymbol{\mu}_{t+\Delta t} &\leftarrow \mathbf{e}_1 \cdot \left(\bar{\mu} \cdot \mathbf{1}_m + \zeta \cdot \hat{\boldsymbol{\mu}}_{t+\Delta t} + \lambda \cdot \hat{\boldsymbol{\sigma}}_{t+\Delta t}\right) \\
\sqrt{\boldsymbol{\Omega}}_{t+\Delta t} &\leftarrow \operatorname{diag}\left(\mathcal{V}_{t+\Delta t} \odot \mathbf{X}_{t+\Delta t} \cdot \mathcal{S}_{t+\Delta t}\right) \\
\Delta \mathbf{F}_{t+\Delta t} &\leftarrow \boldsymbol{\mu}_t \cdot \Delta t + \sqrt{\boldsymbol{\Omega}_t} \boldsymbol{W}_{t+\Delta t} \cdot \sqrt{\Delta t} \\
\mathbf{P}_{t+\Delta t} &\leftarrow \left(\mathbf{P}_t + \mathbf{P}_t \odot \left(\mathbf{A}^\top \Delta \mathbf{F}_{t+\Delta t}\right)\right)_+ \\
\mathbf{M}_{t+\Delta t}^{(\tilde{\boldsymbol{\mu}})} &\leftarrow \left(\mathbf{1}_{n_\tau}^\top \otimes \exp \circ \left(\boldsymbol{\tau}^{-1} \cdot \Delta t\right)\right) \odot \mathbf{M}_t^{(\tilde{\boldsymbol{\mu}})} + \left(\boldsymbol{\tau}^{-1}\right)^\top \otimes \Delta \mathbf{F}_{t+\Delta t} \\
\mathbf{M}_{t+\Delta t}^{(\tilde{\boldsymbol{\Omega}})} &\leftarrow \left(\mathbf{1}_{n_\tau}^\top \otimes \exp \circ \left(\boldsymbol{\tau}^{-1} \cdot \Delta t\right)\right) \odot \mathbf{M}_t^{(\tilde{\boldsymbol{\Omega}})} + \left(\boldsymbol{\tau}^{-1}\right)^\top \otimes \left(\Delta \mathbf{F}_{t+\Delta t}\right)^{\circ 2}
\end{aligned}
$$

3. Update the set of price vector trajectories:

$$
\left\{\mathbf{P}_{z\Delta t}\right\}_{0 \le z \le u} \leftarrow \left\{\mathbf{P}_{u\Delta t}\right\}_{0 \le z \le u-1} \cup \left\{\mathbf{P}_{u\Delta t}\right\}
$$

end for.

Output: $\left\{\mathbf{P}_{k\Delta t}\right\}_{1 \le k \le n_s}$.

The EWMA of the variations and quadratic variations of the elementary factors are, in turn, aggregated respectively into matrices $\mathbf{M}^{(\tilde{\boldsymbol{\mu}})}$ and $\mathbf{M}^{(\tilde{\boldsymbol{\Omega}})}$ of dimension $m \times n_\tau$. Thus, the element at coordinate (j, p) of $\mathbf{M}^{(\tilde{\boldsymbol{\mu}})}$ (resp. $\mathbf{M}^{(\tilde{\boldsymbol{\Omega}})}$) corresponds to the EWMAs with parameter τ_p of the variations (resp. quadratic variations) of the elementary factor j. Furthermore, the dynamics of these matrices are not directly modeled by discretizing the SDEs associated with them, but rather by discretizing the integrals that define these state variables. This approach, adopted for example in [68], has the beneficial property of ensuring the stability of the model regardless of the time step considered, which is not the case for classical discretization

schemes of the SDEs for EWMAs ([67]).

Now, concerning the drifts and factorial volatilities, only the components $\mathbf{X}$ and $\mathcal{S}$ need to be discretized. To start with, the process $\mathbf{X}$ is not regarded in this discrete model as a stochastic process, in the sense that the different realizations of this random vector generated from this model form an i.i.d. sample. This stems from the assumption of very rapid mean reversion of the coordinates of $\mathbf{X}$ as posited in section 3.1.1.3, which makes it consistent to model $\mathbf{X}_{t+\Delta t}$ from the asymptotic distribution of $\mathbf{X}$. Thus, for each simulation period:

$$\mathbf{X}_{t+\Delta t} \sim \mathcal{LN}(-\mathbf{S} \odot \mathbf{S},\ \mathrm{diag}(\mathbf{S})^2).$$

In practice, this choice is equivalent to opting for a non-Gaussian innovation process whenever $\mathbf{S} \neq \mathbf{0}_m$. Indeed, the marginal distributions of $\mathbf{X}_t \odot \boldsymbol{W}_t$ correspond (independent of each other) to Normal Log-normal (NLN) mixture ([74]) of the form $W \cdot \exp(s(B-s))$, with $(W, B)^\top \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$, These marginal distributions have the important properties of sharing the first three centered moments regardless of the value taken by $\mathbf{S}$: zero mean, a standard deviation (variance) equal to 1, and zero skewness (see details in appendix E.1). On the contrary, the kurtosis of the marginal distributions of $\mathbf{X}_t \odot \boldsymbol{W}_t$ depends on the specification of $\mathbf{S}$, since:

$$\mathbb{E}\left[\left(W\mathrm{e}^{s(B-s)}\right)^4\right] = 3\mathrm{e}^{4s^2}.$$

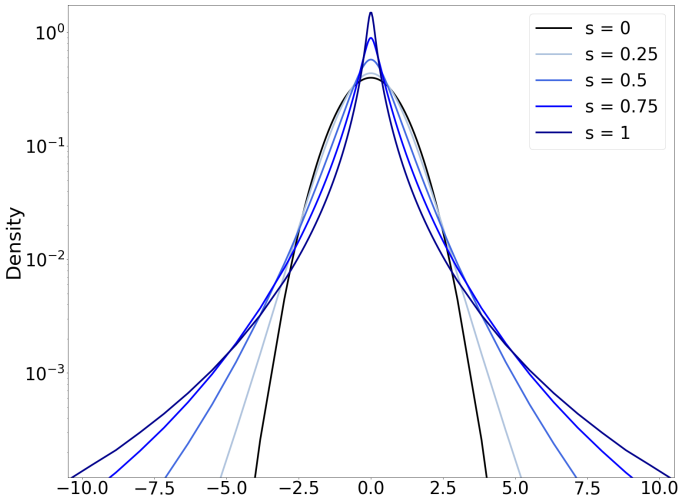


Figure 4: NLN mixture distributions of the form $W\mathrm{e}^{s(B-s)}$ with $(W, B)^\top \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$ for different values of s.

Thus, as illustrated in figure 4, the larger (the absolute value of) $(\mathbf{S})_j$, the thicker the tails of the distributions of $(\mathbf{X}_t \odot \boldsymbol{W}_t)_j$ exhibit. This characteristic enables capturing potential extreme variations of the elementary factors (and consequently, of the assets) occurring at short time scales. However, as mentioned earlier, the skewness of $(\mathbf{X}_t \odot \boldsymbol{W}_t)_j$ is necessarily zero due to the independence of $(\boldsymbol{W})_j$ and $(\boldsymbol{B})_j$, implying a symmetry of these factorial variations at the horizon Δt. Consequently, this modeling assumes that the potential asymmetry of the distributions of factorial increments is caused by the path-dependent component of volatility, and more specifically by the convolved dynamics features. For this reason, the choice of the simulation time step Δt is important. Indeed, in this framework, if the choice of a daily time step makes the model structurally unable to capture potential asymmetries in daily return distributions, opting for a smaller time step allows for a potential capture of these asymmetries through the impact of returns on the level of volatility.

Of course, alternative modeling choices are conceivable, either to capture a potential portion of the skewness unexplained by the dynamics of $\mathcal{V}_t$, or to maintain a relatively high simulation time step. One way is to relax the assumption of independence between $\boldsymbol{W}$ and $\boldsymbol{B}$, so that the correlation between $(\boldsymbol{W})_j$ and $(\boldsymbol{B})_j$ can be different from 0. In this framework, the innovation process is still distributed according to an NLN

mixture distribution but can exhibit a non-zero skew. However, the correlation between $(\boldsymbol{W})_j$ and $(\boldsymbol{B})_j$ results in a non-zero mean of the innovation process associated with factor j, which must be neutralized in order to maintain $\mathbb{E}[\Delta \mathbf{F}_{t+\Delta t}] = \boldsymbol{\mu}_t \cdot \Delta t$. Another possible approach to incorporate excess skewness is to replace $\mathbf{X}_t \odot \boldsymbol{W}_t$ with $\boldsymbol{Z}_t$, a vector whose marginals follow, for instance, skewed generalized t distributions or kernel density estimators (KDE). However, once again, this requires ensuring that $\boldsymbol{Z}_t$ remains centered so that the factorial drifts stay fully captured by $\boldsymbol{\mu}$.

Remark 1 *The algorithm 1 implicitly assumes that the simulation time step is equal to the observation time step of the model on which the ARMA model parameters were calibrated. When this condition is not met, it is necessary first to simulate the trajectory of S using the time step with which the model was calibrated, and then perform interpolation between the simulation periods $\Delta t, 2\Delta t, \ldots, n_s \Delta t$. In practice, linear interpolation will be used.*

3.2 Calibration of the market generator

3.2.1 Input information for model calibration

In the following sections, the calibration of 1 is performed based on a data matrix $\mathbf{D}$ of the form:

$$\mathbf{D} = \begin{pmatrix} P_{1,t_0} & \ldots & P_{n,t_0} \\ P_{1,t_1} & \ldots & P_{n,t_1} \\ \vdots & \ddots & \vdots \\ P_{1,t_N} & \ldots & P_{n,t_N} \end{pmatrix} \tag{16}$$

where $t_0 < t_1 < \ldots < t_N = T$. This matrix $\mathbf{D}$ will be used to compute the central matrix for model calibration, namely the matrix of returns denoted $\mathbf{R}$, defined as:

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & \ldots & r_{n,1} \\ \vdots & \ddots & \vdots \\ r_{1,N} & \ldots & r_{n,N} \end{pmatrix} \tag{17}$$

where:

$$r_{i,u} = \frac{P_{i,t_u} - P_{i,t_{u-1}}}{P_{i,t_{u-1}}}.$$

It is worth to note that if the formulation of $\mathbf{D}$ does not imply a strict constancy of the observation frequencies of the vector $\mathbf{P}$, the estimation method proposed in the following sections is based on the assumption of moderate heterogeneity in $\mathbf{P}$ frequencies. Thus, if moderate heterogeneity in daily observations caused by the presence of weekends or holidays is not problematic for the proposed estimation method, the same cannot be said for a mixture of daily and hourly observation frequencies, for instance.

3.2.2 The elementary factors decomposition

The price dynamics are driven in the FPDM model by the elementary factors. Accordingly, the first step in calibrating the market generator is to perform a decomposition of $\mathbf{R}$ in the form:

$$\mathbf{R} = \widehat{\mathbf{\Delta}}_{\mathbf{F}} \widehat{\mathbf{A}}^{\top} = \left(\widehat{\mathbf{\Delta}}_{\mathbf{F^c}}^{\top}, \widehat{\mathbf{\Delta}}_{\mathbf{F^i}}^{\top} \right)^{\top} \left((\widehat{\mathbf{A}}^{\mathbf{c}})^{\top}, \mathbf{I}_n \right) = \widehat{\mathbf{\Delta}}_{\mathbf{F^c}} \left(\widehat{\mathbf{A}}^{\mathbf{c}} \right)^{\top} + \widehat{\mathbf{\Delta}}_{\mathbf{F^i}},$$

where $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{A}}^{\mathbf{c}}$ are respectively estimators of the loading matrix $\mathbf{A}$ and $\mathbf{A}^{\mathbf{c}}$, $\widehat{\mathbf{\Delta}}\mathbf{F}$ is the history of estimated variations of elementary factors:

$$\widehat{\mathbf{\Delta}}_{\mathbf{F}} = \begin{pmatrix} \widehat{\mathbf{F}}_{\mathbf{t_1}} - \widehat{\mathbf{F}}_{\mathbf{t_0}} \\ \vdots \\ \widehat{\mathbf{F}}_{t_N} - \widehat{\mathbf{F}}_{t_{N-1}} \end{pmatrix}, \quad \widehat{\mathbf{\Delta}}_{\mathbf{F^c}} = \begin{pmatrix} \widehat{\mathbf{F}}^{\mathbf{c}}_{\mathbf{t_1}} - \widehat{\mathbf{F}}^{\mathbf{c}}_{\mathbf{t_0}} \\ \vdots \\ \widehat{\mathbf{F}}_{t_N} - \widehat{\mathbf{F}}_{t_{N-1}} \end{pmatrix}.$$

The method used for this purpose is defined by algorithm 2.

Algorithm 2 Decomposition of the returns matrix $\mathbf{R}$ into elementary factors.

Input: $\mathbf{R}, n_\delta$

1. Standardize the matrix $\mathbf{R}$ to obtain $\bar{\bar{\mathbf{R}}}$ (equation 18).
2. Perform a singular value decomposition of $\bar{\bar{\mathbf{R}}}$.
3. $v, q \leftarrow 1, \frac{n}{N}$
4. **Repeat until convergence**:
   (a) Compute

$$\phi_{\text{thr}} \leftarrow \underset{\phi_i}{\arg\min} \left| \phi_i - v \left(1 + \sqrt{q}\right)^2 \right|$$

$$\tilde{\phi}_{\min} \leftarrow \phi_{\text{thr}}(1 - \text{t})$$

$$\tilde{\phi}_{\max} \leftarrow \phi_{\text{thr}}(1 + \text{t})$$

   (b) Estimate f_{KDE}, the density of $\{\phi_i \mid \phi_i \leq \phi_+\}$ using a kernel density method.
   (c) Find the parameters v and q that minimize:

$$v, q \leftarrow \underset{v,q \in [0,1]^2}{\arg\min} \sum_{k=0}^{n_\delta} \left(\sqrt{f_{\text{KDE}}(\tilde{\phi}_k)} - \sqrt{f_{\text{MP}}(\tilde{\phi}_k | v, q)} \right)^2,$$

   where $\tilde{\phi}_k = \tilde{\phi}_{\min} + \frac{\tilde{\phi}_{\max} - \tilde{\phi}_{\min}}{n_\delta} k$.

5. Compute

$$\widehat{\mathbf{A}}^{\mathbf{c}} \leftarrow (\mathbf{U})_{[:,:m]}$$

$$\widehat{\mathbf{\Delta}}_{\mathbf{F^c}} \leftarrow (\mathbf{D})_{[:m,:m]} (\mathbf{V})_{[:,:m]}$$

$$\widehat{\mathbf{\Delta}}_{\mathbf{F^i}} \leftarrow (\mathbf{D})_{[m_c:,m_c:]} (\mathbf{V})_{[:,:m]} (\mathbf{V})^{\top}_{[:m_c,:]}$$

Output: $\widehat{\mathbf{A}}^{\mathbf{c}}, \widehat{\mathbf{\Delta}}_{\mathbf{F^c}}, \widehat{\mathbf{\Delta}}_{\mathbf{F^i}}$

To do this, we will use a matrix factorization commonly employed in various fields of data analysis: singular value decomposition. However, we will not perform this factorization on $\mathbf{R}$ but on $\bar{\bar{\mathbf{R}}}$ defined as:

$$\bar{\bar{\mathbf{R}}} = \mathbf{R}\mathbf{M_{st}} \quad \text{with} \quad \mathbf{M_{st}} = \text{diag}_{M\to D}\left(\frac{1}{N}\cdot\bar{\mathbf{R}}\bar{\mathbf{R}}^\top\right)^{-1/2}, \tag{18}$$
$$\text{and where} \quad \bar{\mathbf{R}} = \text{diag}_{M\to D}\left(\frac{1}{n}\cdot\mathbf{R}^\top\mathbf{R}\right)^{-1/2}\mathbf{R}.$$

Therefore, it involves double standardization: standardization per row then standardization per column. The row standardization helps reduce the impact of volatility level variability on the singular value decomposition. This standardization makes particular sense in the proposed model where the volatilities of elementary factors depend on the volatility level of the market factor and the sensitivity operator ($\mathcal{S}$). On the other hand, column standardization assigns equal importance to each asset, whereas a singular value decomposition performed directly on $\mathbf{R}$ implicitly weights more volatile assets. Moreover, it normalizes the variance of the matrix $\bar{\mathbf{R}}$, denoted as v, to 1, which will allow for certain simplifications in subsequent steps. By proceeding with a singular value decomposition of $\bar{\mathbf{R}}$, we obtain the following factorization:

$$\bar{\mathbf{R}} = \mathbf{U}\mathbf{D}\mathbf{V}^\top,$$

where $\mathbf{U}$ is a $N\times N$ matrix, $\mathbf{V}$ a $n\times n$ matrix, and $\mathbf{D}$ is a $N\times n$ matrix $\text{diag}_{M\to d}(\mathbf{D}) = (\sqrt{\phi_1},...,\sqrt{\phi_n})^\top$ such as $\phi_1 > ... > \phi_n$.

To build the estimators $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{\Delta}}_{\mathbf{F}}$ from the elements obtained through this decomposition, the next step involves distinguishing the dynamics of common elementary factors from those of idiosyncratic elementary factors. The m_c elements associated with the m_c largest singular values will define the component of $\bar{\mathbf{R}}$ generated by the common elementary factors, and the $n - m_c$ elements its component produced by the idiosyncratic elementary factors. More precisely, regarding the matrices associated with the common elementary factors:

$$\widehat{\mathbf{A}}^{\mathbf{c}} = \mathbf{M}_{\mathbf{st}}^{-1/2}(\mathbf{V})_{[:,:m_c]}, \quad \text{and} \quad \widehat{\mathbf{\Delta}}_{\mathbf{F^c}} = (\mathbf{U})_{[:,:m_c]}(\mathbf{D})_{[:m_c,:m_c]}. \tag{19}$$

As for the matrix of variations of idiosyncratic elementary factors:

$$\widehat{\mathbf{\Delta}}_{\mathbf{F^i}} = (\mathbf{U})_{[:,m_c+1:]}(\mathbf{D})_{[m_c+1:,m_c+1:]}\left(\mathbf{M}_{\mathbf{st}}^{-1/2}(\mathbf{V})_{[:,m_c+1:]}\right)^\top. \tag{20}$$

The challenge is therefore to determine the value taken by m_c. In practice, this involves finding a threshold value ϕ_{thr} for the eigenvalues of $\frac{1}{N}\bar{\mathbf{R}}\bar{\mathbf{R}}^\top$ to separate the noise from the signal, an issue that has been extensively addressed in the academic literature ([45], [17], [49]). In these works based on results from random matrix theory, ϕ_{thr} is determined using the upper bound ϕ_+ of a Marchenko-Pastur distribution, whose density is defined by:

$$f_{\text{MP}}(\phi|v,q) = \begin{cases} \dfrac{\sqrt{(\phi_+-\phi)(\phi-\phi_-)}}{2\pi q\phi} & \text{if } \phi\in[\phi_-:\phi_+], \\ 0 & \text{else,} \end{cases}$$

where $\phi_- = v(1-\sqrt{q})^2$ and $\phi_+ = v(1+\sqrt{q})^2$. However, if the values of the parameters $q = N/n$ and v correspond to the variance of $\frac{1}{N}\bar{\mathbf{R}}\bar{\mathbf{R}}^\top$ in the case where the entries of $\bar{\mathbf{R}}$ are independent identically distributed random variables, the choice of these values is no longer straightforward when this assumption

does not hold (which is a priori the case here). For this reason, several approaches coexist in the literature for determining these parameters ([45], [49]).

The method adopted here is largely inspired by the one proposed by Lopez de Prado ([49]). This involves, after initializing v and q to 1 and $\frac{n}{N}$, the iterative execution of three successive steps. The first one assigns a provisional value to ϕ_{thr} as the nearest eigenvalue to $v\left(1+\sqrt{q}\right)^2$. The second step consist in estimating the density of the distribution of eigenvalues that are less than or equal to ϕ_{thr} using a kernel density method, density denoted f_{KDE}. The third step updates the parameters v and q to the values that minimize (a proxy of) the Hellinger distance between $f_{\text{KDE}}(\cdot)$ and $f_{\text{MP}}(\cdot|v,q)$. These three steps are then repeated until the convergence of ϕ_{thr}.

After this iterative process, m_{c} is defined by the number of eigenvalues strictly greater than ϕ_{thr}. The matrices $\widehat{\mathbf{\Delta}}_{\mathbf{F}}^{\mathbf{c}}$, $\widehat{\mathbf{\Delta}}_{\mathbf{F}}^{\mathbf{i}}$, and $\widehat{\mathbf{A}}^{\mathbf{c}}$ are finally calculated using equations 19 and 20.

3.2.3 Estimation procedure

The objective now is to estimate all the parameters (other than the matrix $\mathbf{A}$) and state variables taken as input by the FPDM generator (algorithm 1), using the matrix of variations of the elementary factors $\widehat{\mathbf{\Delta}}_{\mathbf{F}}$ obtained via algorithm 2. The method chosen for this purpose is defined by algorithm 3.

This one starts by calculating the various EWMAs of the variations and quadratic variations of the elementary factors. These EWMA are initialized at period t_{rw} from the first rw periods (step 1.a. of algorithm 3) using a discretisation of

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{t_{rw}}^{(k)} &= \tilde{\boldsymbol{\mu}}_{t_0}^{(k)} \cdot \mathrm{e}^{\frac{t_0-t_{rw}}{\tau_k}} + \frac{1}{\tau_k}\int_{t_0}^{t_{rw}} \mathrm{e}^{\frac{u-t_{rw}}{\tau_k}}\,\mathrm{d}\mathbf{F}_u, \\
\text{and} \quad \tilde{\boldsymbol{\Omega}}_{t_{rw}}^{(k)} &= \tilde{\boldsymbol{\Omega}}_{t_0}^{(k)} \cdot \mathrm{e}^{\frac{t_0-t_{rw}}{\tau_k}} + \frac{1}{\tau_k}\int_{t_0}^{t_{rw}} \mathrm{e}^{\frac{t_0-t_{rw}}{\tau_k}}\left(\mathrm{d}\mathbf{F}_u \odot \mathrm{d}\mathbf{F}_u\right).
\end{aligned}$$

The values of $\tilde{\boldsymbol{\mu}}_{t_0}^{(k)}$ and $\tilde{\boldsymbol{\Omega}}_{t_0}^{(k)}$ are thus initialized using the means of the variations and quadratic variations of the elementary factors over the entire training period, which compensates for the lack of data for periods before t_0. In practice, when rw is large enough, the impact of this initialization choice is small, as only the EWMA associated with a high τ_k are affected. For periods beyond t_{rw}, updating the values of EWMA is done in the same way as adopted in the market generator.

This initial step is followed by a second phase that involves considering the log-likelihood function associated with the equation governing the dynamics of the vector of elementary factors, under the assumption that $\mathbf{X}_t = \mathbf{1}_m$, such as:

$$\Delta\mathbf{F}_{t_u} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{t_u}\cdot\Delta_u,\ \operatorname{diag}\left(\mathcal{S}_u\cdot\boldsymbol{\mathcal{V}}_u\cdot\Delta_u\right)^2\right),$$

where $\Delta_u = t_u - t_{u-1}$. This log-likelihood function is given by:

$$\begin{aligned}
\mathcal{L}\left(\mathbf{\Theta}|\widehat{\mathbf{\Delta}}_{\mathbf{F}}\right) = -\frac{1}{2}\sum_{u=rw+1}^{N}&\left((\widehat{\mathbf{\Delta}}_{\mathbf{F}})_u - (\boldsymbol{\mu})_u\cdot\Delta_u\right)^{\top}\left(\Delta_u\cdot\left(\mathcal{S}_u\cdot\boldsymbol{\mathcal{V}}_u\right)^{\circ 2}\mathbf{I}_m\right)^{-1}\left((\widehat{\mathbf{\Delta}}_{\mathbf{F}})_u - (\boldsymbol{\mu})_u\cdot\Delta_u\right) \\
&+\log\left|\Delta_u\cdot\left(\mathcal{S}_u\cdot\boldsymbol{\mathcal{V}}_u\right)^{\circ 2}\mathbf{I}_m\right| - m\log(2\pi), \qquad (21)
\end{aligned}$$

where $\boldsymbol{\Theta}$ represents the set of parameters on which $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{V}}$ depend. However, on the contrary to the classical maximum likelihood method, the proposed approach does not strictly maximize the likelihood of the model, due to its sequential structure. Indeed, step 2 of Algorithm 3 involves iteratively maximizing 21 only over subsets of the model's parameter space. This first subset (step 2a) is defined by:

$$\boldsymbol{\Theta}_1 = \left\{\boldsymbol{\delta}, \mathbf{w}, (\mathbf{b}_0)_1, (\mathbf{b}_1)_1, (\mathbf{b}_2)_1, (\bar{\boldsymbol{\mu}})_1, \zeta, \lambda\right\}.$$

Step 2a therefore consists in maximizing the likelihood of the univariate model of the dynamics of the first elementary factor, corresponding to the market factor. Since $\boldsymbol{\delta}$ and $\mathbf{w}$ are estimated in this step, the kernels $K^{(\hat{\mu})}$ and $K^{(\hat{\sigma})}$ are determined solely from the data for this factor. This separate and primary optimization of the parameters of the market factor is motivated by the singular role it plays in the model. Step 2b involves maximizing 21 with respect to the parameter set $\boldsymbol{\Theta}_2$, defined as:

$$\boldsymbol{\Theta}_2 = \left\{(\mathbf{b}_0)_j, (\mathbf{b}_1)_j, (\mathbf{b}_2)_j, (\mathbf{b}_3)_j\right\}_{2\leq j\leq m}.$$

Thus, this step determines the parameters that are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{V}}$ associated with all elementary factors, with the exception of those associated with the market factor (estimated in step 2a). Step 2c, which constitutes the final step of phase 2, involves maximizing the log-likelihood with respect to $\mathcal{S}_{t_u}$ for $rw+1 \leq u \leq N$. In this step, the values of $\mathcal{S}$ are directly obtained using the analytical solution of the problem presented in appendix F.

Once the value of the log-likelihood function has converged after iterating through steps 2a, 2b, 2c, the algorithm estimates a_1, a_2, a_3 by fitting an ARMA(1, 1) model to the series $\left\{\log\left(\hat{\mathcal{S}}_{t_u}\right)\right\}_{rw+1\leq u\leq N}$, using the standard maximum likelihood estimation approach ([18]).

The final phase involves estimating the vector $\mathbf{S}$ on which $\mathbf{X}$ depends, utilizing a method-of-moments approach whose rationale is explained in detail in appendix E.3. First, step 4a calculates the fourth-order moments of the m samples[4]:

$$\left\{\frac{\left(\Delta\hat{\mathbf{F}}_u - \hat{\boldsymbol{\mu}}_u \cdot \Delta_u\right)_j}{(\mathcal{S}_u \cdot \boldsymbol{\mathcal{V}}_u \cdot \sqrt{\Delta_u})_j}\right\}_{rw+1\leq u\leq N}.$$

The standardization of factor variations (i.e., $\hat{\mathbf{F}}_u$) serves two main purposes: firstly, it allows us to treat these samples as being i.i.d., and secondly, it enables us to isolate the impact of $\mathbf{X}$. Thus, in the specific case where the simulation time step Δt equals the average time step between observations, the computed empirical moments in step 4a directly serve as estimators of $\mathbb{E}\left[(\mathbf{X}\odot\boldsymbol{W})_j^4\right]$. In a broader context, these fourth-order moments enable us to estimate $\mathbf{S}$ using the formula derived in appendix E.3, which is precisely what step 4b accomplishes.

[4]The algorithm 3 uses the biased estimator of $\mathbb{E}\left[(\mathbf{X}\odot\boldsymbol{W})_j^4\right]$. A consistent alternative would be to use its unbiased estimator.

Algorithm 3 Estimation of the FPDM generator.

Input: $\widehat{\boldsymbol{\Delta}}_{\mathbf{F}}, \Delta t$

1. **For** $k = 1$ **to** n_τ **do**

   (a) $\left\{\tilde{\boldsymbol{\mu}}^{(k)}_{t_{rw}}\right\} \leftarrow \frac{\mathrm{e}^{\frac{t_0 - t_{rw}}{\tau_k}}}{N} \cdot \sum_{u=1}^{N} \frac{\left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u}{t_u - t_{u-1}} + \frac{1}{\tau_k} \sum_{u=1}^{rw} \mathrm{e}^{\frac{t_u - t_{rw}}{\tau_k}} \cdot \left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u$

   $\left\{\tilde{\boldsymbol{\Omega}}^{(k)}_{t_{rw}}\right\} \leftarrow \frac{\mathrm{e}^{\frac{t_0 - t_{rw}}{\tau_k}}}{N} \cdot \sum_{u=1}^{N} \frac{\left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}} \odot \widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u}{t_u - t_{u-1}} + \frac{1}{\tau_k} \sum_{u=1}^{rw} \mathrm{e}^{\frac{t_u - t_{rw}}{\tau_k}} \cdot \left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}} \odot \widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u$

   (b) **For** $u = rw + 1$ **to** N **do**

$$\left\{\tilde{\boldsymbol{\mu}}^{(k)}_{t_z}\right\}_{0 \le z \le u} \leftarrow \left\{\tilde{\boldsymbol{\Omega}}^{(k)}_{t_z}\right\}_{0 \le z \le u} \cup \left\{\mathrm{e}^{\frac{t_u - t_{u-1}}{\tau_k}} \cdot \tilde{\boldsymbol{\mu}}^{(k)}_{t_{u-1}} + \frac{1}{\tau_k} \cdot \left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u\right\}$$

$$\left\{\tilde{\boldsymbol{\Omega}}^{(k)}_{t_z}\right\}_{0 \le z \le u} \leftarrow \left\{\tilde{\boldsymbol{\Omega}}^{(k)}_{t_z}\right\}_{0 \le z \le u-1} \cup \left\{\mathrm{e}^{\frac{t_u - t_{u-1}}{\tau_k}} \cdot \tilde{\boldsymbol{\Omega}}^{(k)}_{t_u} + \frac{1}{\tau_k} \cdot \left(\widehat{\boldsymbol{\Delta}}_{\mathbf{F}} \odot \widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)_u\right\}$$

   End for.

   End for.

2. **Repeat until convergence**:

   (a) $\hat{\boldsymbol{\Theta}}^{\star}_1 \leftarrow \arg\min_{\boldsymbol{\Theta}_1 \in \mathcal{S}_\Theta} \mathcal{L}\left(\boldsymbol{\Theta} | \widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)$

   (b) $\hat{\boldsymbol{\Theta}}^{\star}_2 \leftarrow \arg\min_{\boldsymbol{\Theta}_2 \in \mathcal{S}_\Theta} \mathcal{L}\left(\boldsymbol{\Theta} | \widehat{\boldsymbol{\Delta}}_{\mathbf{F}}\right)$

   (c) **For** $u = rw + 1$ **to** N **do**

$$\hat{\mathcal{S}}^{\star}_u \leftarrow \sqrt{\sum_{j=1}^{m} \frac{\left(\left(\Delta \mathbf{F}_{t_u}\right)_j - \left(\hat{\boldsymbol{\mu}}_{t_u}\right)_j \cdot \Delta_{t_u}\right)^2}{\left(\boldsymbol{\mathcal{V}}_{t_u}\right)^2_j \cdot \Delta_{t_u}}}$$

3. Fit ARMA(1,1) based on the time series $\left(\log\left(\hat{\mathcal{S}}^{\star}_1\right), \ldots \log\left(\hat{\mathcal{S}}^{\star}_N\right)\right)$ to obtain $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{\epsilon}_N$.

4. **For** $j = 1$ **to** m **do**

   (a) $\hat{c}_{j,4} \leftarrow \frac{1}{N - rw - 1} \sum_{u=rw+1}^{N} \left(\frac{\left(\Delta \mathbf{F}_u - \hat{\boldsymbol{\mu}}_u \cdot \Delta_u\right)_j}{\left(\mathcal{S}_u \cdot \boldsymbol{\mathcal{V}}_u \cdot \sqrt{\Delta_u}\right)_j}\right)^4$

   (b) $(\hat{\mathbf{S}})_j \leftarrow 0.5 \sqrt{\log\left(\frac{\frac{1}{N-rw-1} \sum_{u=rw+1}^{N} (t_u - t_{u-1})}{\Delta t} \left(\frac{\hat{c}_{j,4}}{3} - 1\right)_+ + 1\right)}.$

   End for.

Output: $\left\{\tilde{\boldsymbol{\mu}}^{(k)}_{t_N}\right\}_{0 \le k \le n_\tau}, \left\{\tilde{\boldsymbol{\Omega}}^{(k)}_{t_N}\right\}_{0 \le k \le n_\tau}, \hat{\boldsymbol{\Theta}}^{\star}_1, \hat{\boldsymbol{\Theta}}^{\star}_2, \hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{\epsilon}_N$

4 Empirical assessment of the FPDM generator performance

4.1 Modalities of the conducted assessments

4.1.1 The considered market dataset

The objective of section 4 is to assess the capability of the FPDM generator defined in section 3.1.2 to produce realistic market scenarios, capturing various characteristics of the considered time series. We begin by defining the data under consideration.

Firstly, we consider an equity investment universe: the assets comprising the S&P500 index. More precisely, we focus on the 436 assets belonging to the S&P500 as of April 30, 2024, for which we have daily historical data from April 1, 2010, to April 30, 2024. The first 8 years of historical data, spanning from April 1, 2010, to April 30, 2018, constitute the training dataset used to calibrate the FPDM Generator. The remaining historical data, covering the period from May 1, 2018, to April 30, 2024, will be used as the test dataset for evaluating the model's performance. This division has the particularity that the training set does not include periods of very strong market turbulence similar to that of the 2020 stock market crash included in the test set, which enables to assess the model's ability to reproduce such events even without them appearing in the data used for its training.

4.1.2 The set of parameters of the generator

The various parameters of the market generator are determined by applying algorithms 2 and 3 successively on the training set.

To begin, the application of algorithm 2 results in obtaining 25 common elementary factors. Consequently, the FPDM Generator used comprises a total of 461 common elementary factors: 25 common factors and 436 idiosyncratic factors. If we focus more specifically on the first estimated common factor, the correlation between its daily variations and the returns of the S&P500 index over the same periods is 0.985 (see appendix G.2). Therefore, considering this factor as the market factor is entirely coherent. It is also interesting to examine the separation achieved by algorithm 2 between the eigenvalues associated with the common elementary factors and the remaining eigenvalues, as illustrated in figure 5. Firstly, the Marchenko Pastur distribution estimated by the algorithm fits the data quite well. Additionally, the separation threshold between the 25 largest and the smaller eigenvalues coincides with several changes in the behavior of the spectrum of $\frac{1}{N}\bar{\bar{\mathbf{R}}}\bar{\bar{\mathbf{R}}}^\top$. Indeed, as illustrated by the plot in the top left (figure 5), the density of eigenvalues as a function of their magnitude appears to follow a power-law relationship for the 25 largest eigenvalues, which is not the case for eigenvalues below this threshold. Relatedly, except for the first eigenvalue associated with the market factor, the eigenvalues associated with the other common risk factors follow the relationship: $\log(\phi_k) \approx a - b \cdot \log(\text{rank}(\phi_k))$. This is illustrated by the two plots at the bottom of figure 5, where the black line corresponds to the OLS regression line $a - b \cdot \log(\text{rank}(\phi_k))$ estimated from eigenvalues 2 to 25 (the parameters obtained here are $\hat{a} = 3.45$ and $\hat{b} = 0.92$). Even more remarkably, when performing the same regression on the oracle eigenvalues estimated using the approach proposed in [47], $\hat{b}$ is almost exactly equal to 1, which could suggest the existence of an underlying fundamental financial relationship.

The algorithm 3 is then employed, subsequent to the application of algorithm 2, to estimate the remaining parameters of the model. This entails specifying $\{\tau_k\}_{k=1}^{n_\tau}$, Δt, and rw, in addition to the matrix obtained from algorithm 2. To determine the values of $\{\tau_k\}_{k=1}^{n_\tau}$, we utilize equation 14 (section 3.1.1.2) with $n_\tau = 20$,

$\tau_- = 1/365$, and $\tau_+ = 5$, all expressed in years. The time step utilized for the simulations is $\Delta t = 1/3650$, also expressed in years. Moreover, we select a parameter $rw = 1008$, specified in the number of periods, which roughly corresponds to the initial 4 years of the training dataset used to initialize the various EWMA components of the model. Consequently, the estimation of the model parameters is practically conducted over a period of approximately 4 years, spanning from April 2014 to the end of April 2018.

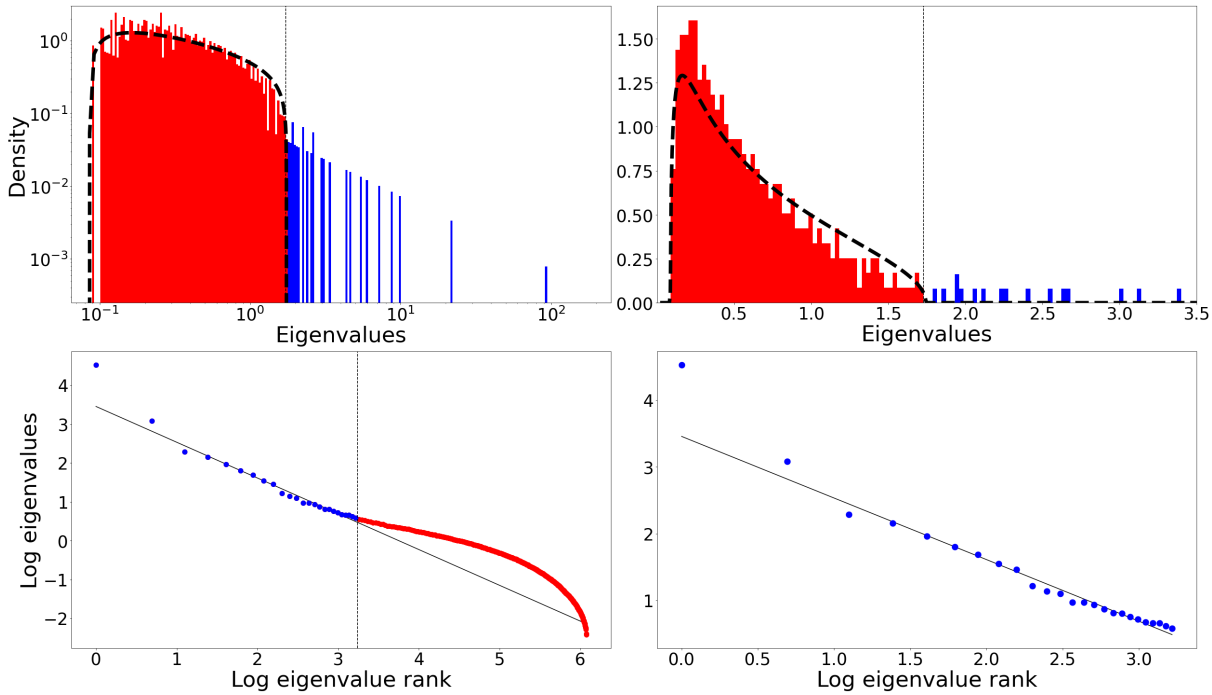


Figure 5: Histogram of the eigenvalue distribution of $\frac{1}{N}\bar{\bar{\mathbf{R}}}\bar{\bar{\mathbf{R}}}^\top$. The black dashed line that fits the red part of the histograms corresponds to the fitted Marchenko-Pastur distribution obtained from algorithm 2, used to separate common elementary factors from idiosyncratic elementary factors.

Several observations can be made following the application of algorithm 3. Firstly, as shown in figure 18, the obtained kernels are very close to time-shifted power law (TSPL) kernels, corroborating several recent findings in volatility literature ([39]). Regarding the parameters of the drift and the path-dependent component of the volatility associated with the market factor, they are respectively of the same sign: $\bar{\mu}, \zeta, \mathrm{b}_{0,1}, \lambda, \mathrm{b}_{2,1}$ are positive, and $\zeta, \mathrm{b}_{1,1}$ are negative. However, if the drift of the market factor is quite close to a positive affine relationship with the $(\boldsymbol{\mathcal{V}}_t)_1$, then $\bar{\mu}/\mathrm{b}_{0,1} > \zeta/\mathrm{b}_{1,1} > \lambda/\mathrm{b}_{2,1}$. Consequently, the constant part of the drift plays a more important role than in its approximation of the form $\Gamma_t^\star \approx \bar{\mu} + \lambda^{\mathcal{V}} \cdot (\boldsymbol{\mathcal{V}}_t)_1$. Similarly, the reversal effect is more significant than the historical volatility risk premium effect if the affine relationship were perfectly respected. Another notable observation is the strong correlation of 0.875 between the purely path-dependent volatility component associated with the market factor (i.e., $(\boldsymbol{\mathcal{V}})_1$) and the VIX, despite the latter not being included in the training set. This element suggests a high level of coherence between the volatility of the market factor estimated by the model results and the market's priced volatility. Additionally, significant differences in the fourth-order moments are observed between the common elementary factors and the idiosyncratic elementary factors. For the 25 common factors, this moment ranges from 2.75 to 4.1, with mean and median values of 3.26 and 3.18, respectively, close to the Gaussian assumption (where this moment equals 3). In contrast, for the idiosyncratic factors, this moment ranges from 2.98 to 78, with mean and median values of 12.8 and 9.64, respectively.

4.2 General properties of generated datasets

The objective of this initial series of evaluations is to compare the overall properties of synthetic data generated by the FPDM generator model with market data.

4.2.1 Evaluation based on the marginal distributions

To begin, we delve into the individual dynamics of assets using simulated data, aiming to assess the consistency of their marginal return distributions with their empirical counterparts. Our analysis focuses on the first four moments of the daily, weekly, and monthly returns distributions for this purpose.

The comparison method used is as follows: we start by computing the various empirical moments considered, for both the historical realization and for each market scenario generated by the market generator. Therefore, for each asset i/moment p/temporal horizon l combination (e.g. the empirical mean of daily returns for Apple stock), we obtain a sample of 1000 empirical moments from the corresponding 1000 simulations. From the empirical cumulative distribution function $\tilde{F}_{i,p,l}$ estimated from this sample, we compute the estimated cumulative probability of the empirical moment obtained from the actual market data sample:

$$p_{i,p,l} = \tilde{F}_{i,p,l}(\hat{m}_{i,p,l}),$$

where $\hat{m}_{i,p,l}$ is the empirical p-th moment of the returns with a temporal horizon l of asset i calculated on real data. Subsequently, for each moment/time horizon pair, we compute the mean of these cumulative probabilities across the 436 assets, along with the three quartiles and the proportion of cumulative probabilities between 0.05 and 0.95. The results are then report in the table 1.

		Mean $p_{i,p,l}$	Q1 $p_{i,p,l}$	Med. $p_{i,p,l}$	Q3 $p_{i,p,l}$	Prop. $p_{i,p,l} \in [0.05 : 0.95]$
Daily	Moment 1	0.535	0.349	0.589	0.749	0.913
	Moment 2	0.582	0.457	0.643	0.765	0.933
	Moment 3	0.462	0.203	0.43	0.718	0.966
	Moment 4	0.6	0.499	0.643	0.753	0.954
Weekly	Moment 1	0.426	0.233	0.412	0.574	0.954
	Moment 2	0.55	0.386	0.612	0.741	0.931
	Moment 3	0.488	0.265	0.462	0.698	0.961
	Moment 4	0.582	0.456	0.617	0.753	0.95
Monthly	Hist. data	0.427	0.236	0.41	0.571	0.954
	Sim. data	0.535	0.349	0.589	0.748	0.913
	Moment 3	0.398	0.205	0.315	0.561	0.97
	Moment 4	0.565	0.418	0.62	0.756	0.936

Table 1: Statistics related to the estimated cumulative probabilities set $\{p_{i,p,l}\}_{1 \leq i \leq 436}$. For example, the cell corresponding to the "Mean" column and the "Daily/Moment 3" row represents $\frac{1}{436} \sum_{i=1}^{436} p_{i,3,1} = 0.462$.

These obtained figures tend to demonstrate a very good fit between the distributions of the returns generated by the model and their empirical counterparts. Indeed, the various empirical moments of the real data are generally close to the mean and median levels of their counterparts obtained through simulations. Thus, except for the third-order moment of monthly returns, for all others pairs p and l, the mean and median of $\{\tilde{F}_{i,p,l}(\hat{m}_{i,p,l})\}_{i=1}^{436}$ fall between 0.4 and 0.6. Similarly, for each moment/time horizon pair, over 90% of the assets exhibit empirical moments within the 90% confidence interval of the model.

Moreover, the synthetic data generated by the FPDM Generator are not only coherent with market data in terms of asset return distributions, but also in terms of trajectories. Indeed, as illustrated in figure 6 with the example of Amazon stock[5], the various characteristics of the price and return trajectories of individual stocks are well reproduced. It is particularly striking regarding the joint dynamics of price and volatility: in line with empirical data, on the one hand, volatility spikes coincide in the vast majority of cases with price drops; on the other hand, volatility tends to decrease relatively slowly following these shocks. This accurate modeling enables the reproduction of another related feature of financial series: the volatility clustering effect, which denotes the coexistence of periods of low and high volatility.

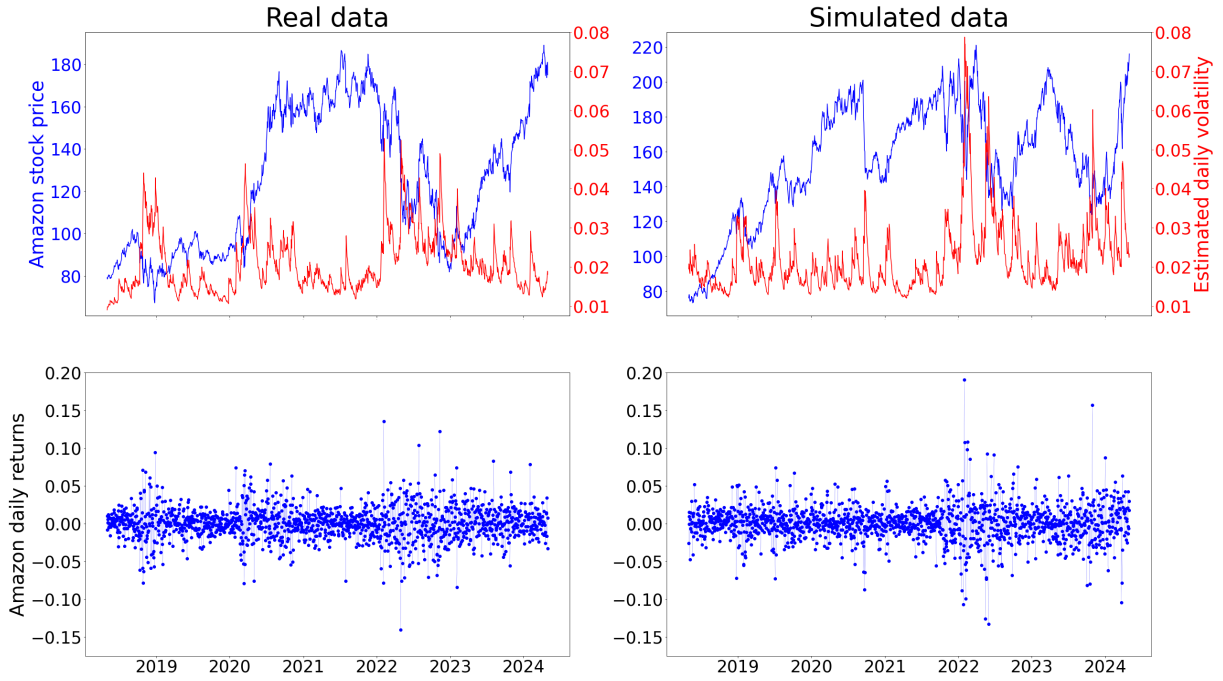


Figure 6: Price trajectory, volatility (estimated from a GARCH(1,1) model), and daily returns of Amazon stock: real data vs simulated data. The features characterizing the real data, such as the phenomenon of volatility spikes and volatility clustering, strong daily price swings, or the leverage effect, are well reproduced by the model.

4.2.2 Evaluation based on the joint asset prices dynamics

The objective of this section is to assess the model's ability to capture the correlation structure of price dynamics.

To this end, the first type of evaluation will focus on the covariance matrix and the correlation matrix of daily returns. More specifically, we will evaluate estimators of the correlation matrix and the covariance matrix of the daily returns of the considered universe obtained from the simulations generated by the FPDM generator, with the sample estimators of these matrices computed for the test period. For this purpose, we will use as benchmarks, estimators of these matrices calculated on historical data over the period on which the FPDM generator is trained (i.e. from April 2014 to the end of April 2018). For both

[5]We take the example of Amazon here, as it is one of the largest capitalizations in the S&P500. However, the observations in this paragraph hold for all simulated assets.

estimators calculated on historical and synthetic data, four types of estimators are considered[6]:

1. The unbiased sample estimator.
2. The Oracle approximating shrinkage (OAS) proposed in [24].
3. The first linear shrinkage estimator of Ledoit and Wolf introduced in [46].
4. The non-linear shrinkage estimator also proposed by Ledoit and Wolf, as detailed in [47].

To compare these estimators, two evaluation metrics are employed. The first one is the Frobenius norm of the difference between the sample test covariance (resp. correlation) matrix $\boldsymbol{C}$ and the estimator covariance (resp. correlation) matrix $\hat{\boldsymbol{C}}$:

$$L_F(\hat{\boldsymbol{C}}, \boldsymbol{C}) = \|\hat{\boldsymbol{C}} - \boldsymbol{C}\|_F.$$

This is equivalent to calculating the root sum squared error between the elements of the two matrices. The second one is the minimum loss function proposed in [30], defined by:

$$L_{ML}(\hat{\boldsymbol{C}}, \boldsymbol{C}) = \frac{\mathrm{Tr}\left(\hat{\boldsymbol{C}}^{-1}\boldsymbol{C}\hat{\boldsymbol{C}}^{-1}\right)/n}{\mathrm{Tr}\left(\hat{\boldsymbol{C}}^{-1}\right)/n} - \frac{1}{\mathrm{Tr}\left(\boldsymbol{C}^{-1}\right)/n}.$$

As Ledoit and Wolf specify ([30], [47]), this cost function is designed to measure the quality of an estimator of the covariance matrix for use cases "where variance minimization decisions must be taken". The table 2 reports the obtained results.

		Corr. Matrix		Cov. Matrix	
		L_F	L_{ML}	L_F	L_{ML}
Sample	Hist. data	5.98×10^{1}	4.67×10^{-1}	5.68×10^{-2}	2.04×10^{-4}
	Sim. data	6.17×10^{1}	2.78×10^{-1}	4.01×10^{-2}	1.30×10^{-4}
OAS	Hist. data	6.10×10^{1}	3.94×10^{-1}	5.71×10^{-2}	1.72×10^{-4}
	Sim. data	6.17×10^{1}	2.78×10^{-1}	4.01×10^{-2}	1.30×10^{-4}
LW linear	Hist. data	6.15×10^{1}	3.74×10^{-1}	5.73×10^{-2}	1.64×10^{-4}
	Sim. data	6.17×10^{1}	2.78×10^{-1}	4.01×10^{-2}	1.30×10^{-4}
LW non-linear	Hist. data	6.15×10^{1}	3.67×10^{-1}	5.68×10^{-2}	1.65×10^{-4}
	Sim. data	5.64×10^{1}	2.70×10^{-1}	4.00×10^{-2}	1.24×10^{-4}

Table 2: Comparison of different estimators of the correlation matrix and the covariance matrix of daily returns.

Firstly, the results concerning the correlation matrix are heterogeneous depending on the cost function considered. If we begin by focusing on the cost measured by the Frobenius loss function, the costs associated with correlation matrices calculated on historical data are generally close to their counterparts calculated on data simulated by the model. If we specifically center our analysis on the sample estimator, the estimator calculated on historical data slightly outperforms its counterpart calculated on synthetic data. While this result may initially appear somewhat disappointing regarding the model's ability to obtain better estimators of correlations, several factors must be considered to interpret it as accurately as possible. Firstly, the correlation structure is dynamic in the FPDM generator. However, outperforming the sample estimator on this metric and over this test period is much simpler with a constant correlation matrix model.

[6]For each type of estimator, we first calculate the covariance matrix estimator, then from this, we compute the associated correlation matrix.

In addition, due to this dynamic nature, the sample correlation matrix changes from one sample (generated by the model) to another. Thus, if we consider each synthetic sample separately, in 52.3% of cases, the sample correlation estimator obtained from these samples is associated with a Frobenius loss lower than that obtained with the sample correlation matrix calculated on the historical data. Furthermore, the most efficient estimator of the correlation matrix based on the criterion of Frobenius loss is the non-linear Ledoit-Wolf shrinkage estimator calculated on the simulated data. If we now focus on the comparison between the estimators of the correlation matrix based on the criterion of the minimum loss function, the results clearly favor the simulated data. Thus, on this criterion, the cost of the sample correlation matrix calculated from this simulated data is more than 40% lower than the cost associated with its counterpart calculated on historical data, and about 25% lower than the cost of the best estimator calculated on historical data (the non-linear Ledoit-Wolf shrinkage estimator). The results obtained for the covariance matrices are similar. Indeed, for both considered cost functions, the sample estimator of the covariance matrix calculated from simulated data significantly outperforms all estimators of this covariance matrix calculated on historical data. Thus, considering the Frobenius loss, incorporating the effect of the standard deviations of returns of different assets improves the relative performance of the sample estimator calculated on simulated data compared to its associated correlation matrix. Therefore, in this context of use (that of calculating covariance matrices from a set of simulated scenarios over a medium to long-term horizon), the strength of the model seems to lie less in its ability to better capture linear correlations between asset returns than in improving the estimation of their individual standard deviations while maintaining a realistic correlation structure.

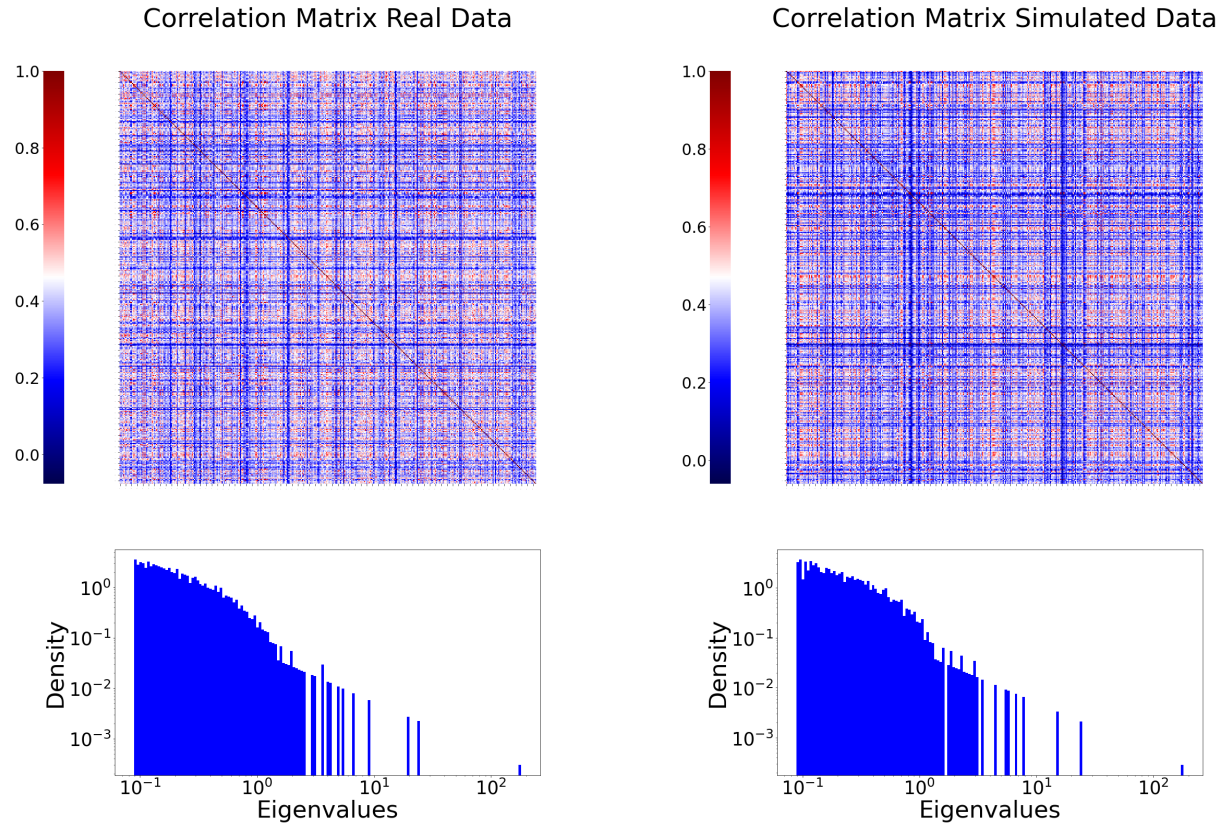


Figure 7: The sample correlation matrix and its spectrum: on the left obtained from the real test data, on the right obtained from a sample generated by the FPDM generator.

Other interesting findings emerge when considering, not the aggregation of the different generated scenarios, but each scenario separately. Specifically, while the sample correlation matrix varies significantly from one generated sample to another, its spectrum retains features characteristic of asset return correlation matrices. That's illustrated in figure 7. Thus, like its counterpart calculated on empirical data, the spectra of these matrices exhibit a similar structure, with, on one side, the largest eigenvalues, which are distributed following a power law distribution, and the other eigenvalues, which are distributed according to a Marchenko-Pastur law.

Due to the fact that stock return distributions are non-elliptical ([25]), it is interesting to complement this analysis of the data's correlation structure by focusing on correlation measures capable of capturing more complex dependencies and tail behaviors. To this end, we now focus on two rank correlation measures: Spearman's rho and Kendall's tau. Similarly to the evaluation conducted on linear correlations, we then compare for each pair of assets (i.e. 94830 pairs, $436 \times (436-1)/2 = 94830$) the empirical Spearman's rho and Kendall's tau of the test sample with the empirical Spearman's rho and Kendall's tau of the simulated data. In addition,two benchmarks for these measures are used to put the obtained results into perspective: the empirical estimators and the estimators associated with the Gaussian copula calculated from the training sample[7]. The table 3 reports the obtained results.

	Spearman's Rho		Kendall's Tau	
	RMSE	MAE	RMSE	MAE
Emp. hist. data	9.006×10^{-2}	7.191×10^{-2}	6.510×10^{-2}	5.208×10^{-2}
Gaussian Copula	10.672×10^{-2}	8.653×10^{-2}	7.909×10^{-2}	6.426×10^{-2}
Emp. sim. data	9.069×10^{-2}	7.186×10^{-2}	6.603×10^{-2}	5.229×10^{-2}

Table 3: Comparison of different estimators of the correlation matrix and the covariance matrix of daily returns.

As shown in table 3, the performance of the rank correlation measures estimated on simulated data is substantially identical to their empirical counterparts estimated on historical data. Furthermore, both of these outperform the estimators obtained under the assumption of a Gaussian correlation structure. Consequently, the model effectively captures the rank relationship that links the different pairs of returns present in the sample on which the generator is trained. While this may not be sufficient to obtain rank correlation estimators that are better than the empirical estimators, it does allow for more realistic joint distributions of returns compared to using a Gaussian copula, while employing a dynamic rather than a static approach.

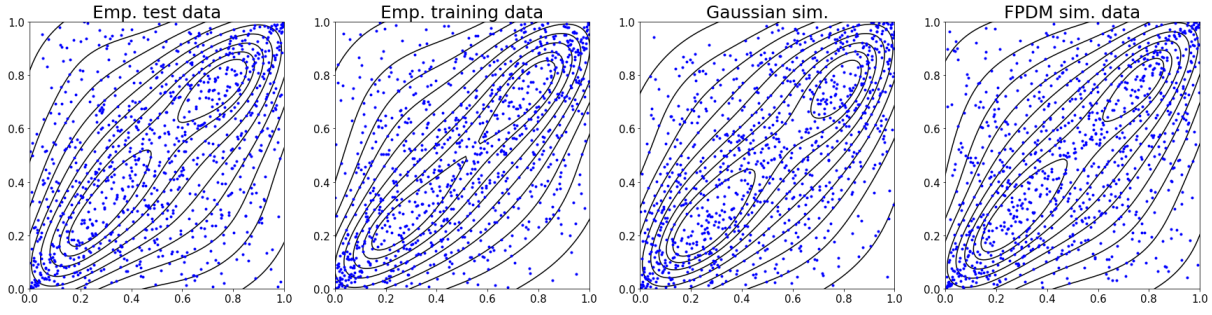


Figure 8: Dependograms of Johnson&Johnson and Pfizer stock daily returns.

[7]When two random variables are linked by a bivariate Gaussian copula with parameter ρ, the Spearman's rho and Kendall's tau are respectively equal to $\frac{6}{\pi}\arcsin(0.5\rho)$ and $\frac{2}{\pi}\arcsin(\rho)$ ([66]).

4.3 Evaluation of strategy features replication

4.3.1 Strategies considered for evaluating the generator

One of the major interest of a market generator is its utility as a tool for backtesting strategies ([48]). Therefore, it makes sense to evaluate the RPDM generator based on its ability to reproduce the various time-series features characterizing different strategies.

To this end, we consider the following 5 strategies:

1. **EW**: The equally weighted portfolio, one of the most well-known strategies, that consists of assigning equal weights to each asset in the portfolio such as:

$$\mathbf{x} = n^{-1} \cdot \mathbf{1}_n.$$

2. **MV**: A constrained minimum variance portfolio, whose composition is the solution of the following optimization program ([65]):

$$\mathbf{x} = \underset{\mathbf{x}}{\arg\min}\ \mathbf{x}^\top \hat{\mathbf{\Sigma}} \mathbf{x} \qquad \text{s.t.} \quad \begin{cases} \mathbf{1}_n^\top \mathbf{x} = 1 \\ \mathbf{x} > -n^{-1} \cdot \mathbf{1}_n, \end{cases}$$

   where $\hat{\mathbf{\Sigma}}$ corresponds to the estimator of the asset returns covariance matrix. In practice, the estimator $\hat{\mathbf{\Sigma}}$ used will be the daily returns covariance matrix over the last 4 years. Regarding the constraint of individual short selling on assets, this serves a dual purpose. First, it helps to stay within relatively realistic portfolios, as the sum of the absolute values of the weights is bounded by 3. Furthermore, it helps to restrict the range of the obtained metrics, as in the absence of constraints on the weights other than $\mathbf{1}_n^\top \mathbf{x}$, the behavior of the strategies can fluctuate extremely from one market scenario to another.

3. **TP**: A constrained tangency portfolio in the zero risk free-rate hypothesis, defined as the solution to ([65]):

$$\mathbf{x} = \underset{\mathbf{x}}{\arg\min}\ \frac{\mathbf{x}^\top \hat{\boldsymbol{\mu}}}{\sqrt{\mathbf{x}^\top \hat{\mathbf{\Sigma}} \mathbf{x}}} \quad \text{s.t.} \quad \begin{cases} \mathbf{1}_n^\top \mathbf{x} = 1 \\ \mathbf{x} > -n^{-1} \cdot \mathbf{1}_n, \end{cases}$$

   where $\hat{\boldsymbol{\mu}}$ is the estimated expected asset returns vector. Here, we use the vector of empirical means of the daily returns over the last 4 years as the estimator. Also, the constraint imposed on the weights serves the same purpose as that imposed on the minimum variance portfolio.

4. **TF**: A trend following strategy of cross-section momentum type ([42]) whose weight vector is defined by:

$$\mathbf{x} = \frac{\hat{\boldsymbol{\mu}}_t - Q_1(\hat{\boldsymbol{\mu}}_t)}{\mathbf{1}_n^\top \left(\hat{\boldsymbol{\mu}}_t - Q_1(\hat{\boldsymbol{\mu}}_t)\right)} \tag{22}$$

   where $Q_1(\hat{\boldsymbol{\mu}}_t)$ corresponds to the first quartile of $\hat{\boldsymbol{\mu}}_t$ itself defined by:

$$\hat{\boldsymbol{\mu}}_t = \mathbf{P}_t \oslash \mathbf{P}_{t-l} - \mathbf{1}_n.$$

   Here, the window length l of the moving average estimator will be fixed to 1 year.

5. **RV**: A reversal strategy of cross-section momentum type, whose weights are defined in the same manner as for the TF strategy via equation 22, but using:

$$\hat{\boldsymbol{\mu}}_t = -\mathbf{P}_t \oslash \mathbf{P}_{t-l} - \mathbf{1}_n,$$

where l is also set to 1 year.

Each of these strategies will be examined in three distinct modes: buy-and-hold, fixed-weight, and dynamic.

In the buy-and-hold approach, the portfolio is constructed at the close of the last trading day of April 2018, and no rebalancing is performed thereafter. In this framework, only the price evolution affects the asset weights in the composition of the various portfolios. However, this mechanism alone is enough to significantly impact the behavior of the portfolio value considered as a random variable. For instance, suppose the market factor drops and then does not experience a significant rebound. In this scenario, all else being equal, the weights of securities with a very high beta tend to decrease relative to other securities, and this event tends to reduce the portfolio's exposure to the market factor. The interest of backtesting buy-and-hold strategies lies in evaluating the model's ability to accurately reproduce this type of path dependence.

The fixed-weight approach, for its part, involves daily rebalancing of the strategies to maintain the same composition as initially set. In this way, the effect of portfolio weight changes induced by the price dynamics to which buy-and-hold strategies are subject is neutralized. It is therefore the model's ability to replicate various constant linear combinations of the random variables constituting the S&P500 prices that is evaluated through this approach.

Finally, the dynamic approach involves rebalancing the portfolios monthly by updating the various parameters on which the composition of the portfolios associated with each strategy depends. Consequently, depending on the different market scenarios, the composition of the portfolios can vary significantly between real and synthetic data. Therefore, it is the model's ability to reproduce the market, conceptualized as an ecosystem, that is evaluated through these strategies. In other words, the capacity to model the market not merely as a collection of assets with a simple correlation structure, but as a complex system whose mechanisms give rise to non-trivial statistical regularities.

Remark 2 *The parameters and estimators of the strategies considered are not necessarily optimal. Typically, to construct minimum variance and tangency portfolios, we use the sample covariance matrix, which constitutes a poor estimator of the covariance matrix from a portfolio optimization perspective. Furthermore, the "trend-following" and "reversal" portfolios with the buy-and-hold and fixed-weight approaches are not strictly speaking trend-following and reversal strategies since their composition is solely based on data as of April 30, 2018. However, these elements are not really problematic in this context since the goal is not to conduct a comparative study of the considered strategies, but rather to see if the RPDM generator is capable of reproducing the time-series features characterizing various types of investment strategies.*

4.3.2 Results of numerical experiments

We start by individually considering the 5 types of strategies across the 3 rebalancing modes presented in the previous section. Tables 4, 5 and 6 compare a set of financial metrics associated with these strategies between real data and simulated data: their volatility and Sharpe ratio calculated from daily returns and then annualized, their maximum drawdown, and their value-at-risk and expected shortfall at the 95% and 99% thresholds, empirically calculated on a daily basis. Figure 9 focuses on the moments of the returns distributions of these strategies at different frequencies (from daily returns to 6-month returns).

Overall, the obtained results tend to prove that the model reproduces very well the different characteristics of the various strategies considered. Thus, for almost all distributions of the different strategies considered, the first four empirical moments of the real data fall within the 90% interval of the model calculated from the simulated data. Moreover, in a significant number of cases, these empirical moments are close to their median levels calculated from the simulations. Additionally, the non-monotonic relationship of the skewness of the returns of different strategies as a function of the considered time horizon is noteworthy. Thus, for both market data and simulated data, the distribution of returns initially becomes increasingly negatively skewed as the time horizon lengthens. However, past an inflection point located around 10 trading days, the skewness of the return distribution decreases and tends towards zero in the long term. Similarly, the negatively convex relationship between the kurtosis value and the time horizon produced by the model for the different strategies is also consistent with market data. In the same way as the moments of the distributions, the set of empirical financial metrics - reported in tables 4, 5, and 6 - for almost each strategy falls within the model's 90% confidence intervals and, in most cases, is fairly close to the median level obtained through simulations.

		Vol.	SR	MD	$VaR_{95\%}$	$VaR_{99\%}$	$ES_{95\%}$	$ES_{99\%}$
EW	Real data	20.45%	76.58%	36.38%	1.85%	3.22%	3.06%	5.51%
	Sim. data med.	20.26%	95.57%	30.05%	1.71%	3.63%	3.05%	5.60%
	Sim. data D1	14.63%	36.60%	16.06%	1.27%	2.51%	2.13%	3.60%
	Sim. data D9	34.40%	152.58%	62.29%	2.65%	6.37%	5.23%	10.43%
MV	Real data	17.40%	79.11%	32.88%	1.48%	2.89%	2.57%	5.12%
	Sim. data med.	12.87%	60.38%	21.92%	1.16%	2.19%	1.86%	3.16%
	Sim. data D1	9.94%	6.51%	12.23%	0.92%	1.65%	1.40%	2.18%
	Sim. data D9	22.14%	113.97%	47.75%	1.74%	3.97%	3.29%	6.50%
TP	Real data	22.25%	106.88%	36.72%	1.96%	3.64%	3.11%	5.28%
	Sim. data med.	20.34%	65.51%	31.57%	1.83%	3.49%	2.94%	4.93%
	Sim. data D1	15.57%	9.94%	18.59%	1.44%	2.54%	2.17%	3.43%
	Sim. data D9	32.02%	116.96%	60.27%	2.64%	5.85%	4.90%	9.35%
TF	Real data	20.45%	81.59%	36.15%	1.83%	3.36%	3.01%	5.73%
	Sim. data med.	19.45%	89.81%	29.88%	1.66%	3.52%	2.93%	5.32%
	Sim. data D1	14.19%	32.26%	15.22%	1.24%	2.43%	2.05%	3.42%
	Sim. data D9	33.24%	143.55%	61.58%	2.60%	6.20%	5.11%	10.22%
RV	Real data	22.86%	65.53%	30.80%	2.10%	3.76%	3.40%	5.77%
	Sim. data med.	22.82%	90.26%	33.79%	1.96%	4.12%	3.40%	6.24%
	Sim. data D1	16.64%	33.87%	18.36%	1.45%	2.84%	2.41%	4.05%
	Sim. data D9	40.69%	145.51%	69.33%	3.08%	7.34%	6.08%	12.35%

Table 4: Financial metrics associated with different buy-and-hold strategies: real vs simulated data.

		Vol.	SR	MD	$VaR_{95\%}$	$VaR_{99\%}$	$ES_{95\%}$	$ES_{99\%}$
EW	Real data	21.01%	76.12%	37.92%	1.83%	3.32%	3.13%	5.71%
	Sim. data med.	19.53%	77.26%	29.28%	1.52%	3.38%	2.88%	5.77%
	Sim. data D1	16.21%	3.82%	15.16%	1.29%	2.74%	2.36%	4.21%
	Sim. data D9	32.57%	151.14%	61.73%	2.61%	5.79%	4.82%	9.55%
MV	Real data	16.97%	87.00%	31.80%	1.41%	2.73%	2.51%	4.94%
	Sim. data med.	11.39%	64.09%	18.43%	0.95%	1.93%	1.65%	3.14%
	Sim. data D1	9.78%	-2.97%	10.44%	0.83%	1.56%	1.39%	2.32%
	Sim. data D9	19.22%	134.70%	39.57%	1.59%	3.18%	2.81%	5.25%
TP	Real data	21.35%	101.64%	27.63%	1.82%	3.71%	3.03%	5.31%
	Sim. data med.	17.08%	62.58%	27.52%	1.44%	2.86%	2.46%	4.61%
	Sim. data D1	14.73%	-5.01%	16.28%	1.26%	2.37%	2.07%	3.43%
	Sim. data D9	28.68%	134.11%	54.66%	2.42%	4.75%	4.12%	7.66%
TF	Real data	20.26%	68.83%	38.80%	1.79%	3.19%	2.98%	5.70%
	Sim. data med.	18.13%	76.19%	27.71%	1.43%	3.14%	2.66%	5.31%
	Sim. data D1	15.12%	5.63%	14.26%	1.22%	2.52%	2.19%	3.87%
	Sim. data D9	30.41%	148.46%	58.75%	2.42%	5.24%	4.48%	8.67%
RV	Real data	23.03%	79.16%	37.11%	2.08%	3.84%	3.42%	5.93%
	Sim. data med.	21.89%	76.56%	32.19%	1.72%	3.79%	3.22%	6.40%
	Sim. data D1	18.21%	4.34%	17.22%	1.46%	3.04%	2.64%	4.69%
	Sim. data D9	36.40%	150.06%	65.93%	2.92%	6.39%	5.35%	10.53%

Table 5: Financial metrics associated with different fixed-weight strategies: real vs simulated data.

		Vol.	SR	MD	$VaR_{95\%}$	$VaR_{99\%}$	$ES_{95\%}$	$ES_{99\%}$
EW	Real data	21.01%	76.12%	37.92%	1.83%	3.32%	3.13%	5.71%
	Sim. data med.	19.5%	96.04%	29.28%	1.65%	3.53%	2.93%	5.39%
	Sim. data D1	13.95%	37.85%	15.16%	1.2%	2.42%	2.03%	3.46%
	Sim. data D9	34.73%	152.87%	61.73%	2.59%	6.39%	5.23%	10.45%
MV	Real data	15.6%	44.99%	33.55%	1.19%	2.54%	2.29%	4.82%
	Sim. data med.	11.29%	70.18%	17.91%	1.01%	1.94%	1.64%	2.79%
	Sim. data D1	9.05%	12.03%	10.25%	0.83%	1.49%	1.27%	2.02%
	Sim. data D9	16.58%	126.55%	36.8%	1.39%	2.9%	2.49%	4.78%
TP	Real data	23.75%	89.95%	30.05%	2.17%	3.73%	3.53%	6.41%
	Sim. data med.	17.38%	75.3%	26.49%	1.54%	3.01%	2.53%	4.32%
	Sim. data D1	13.45%	19.95%	15.19%	1.21%	2.2%	1.89%	3.02%
	Sim. data D9	26.56%	129.78%	52.43%	2.22%	4.7%	4.01%	7.64%
TF	Real data	21.23%	75.43%	31.23%	1.89%	3.63%	3.09%	5.42%
	Sim. data med.	21.36%	91.4%	31.77%	1.83%	3.83%	3.19%	5.79%
	Sim. data D1	15.74%	36.51%	17.23%	1.39%	2.74%	2.28%	3.78%
	Sim. data D9	34.45%	144.22%	63.48%	2.74%	6.42%	5.27%	10.61%
RV	Real data	29.64%	50.69%	46.69%	2.55%	4.82%	4.17%	7.44%
	Sim. data med.	20.41%	85.35%	31.17%	1.72%	3.64%	3.05%	5.54%
	Sim. data D1	14.19%	30.52%	16.02%	1.24%	2.4%	2.02%	3.37%
	Sim. data D9	41.3%	141.02%	68.33%	2.97%	7.49%	6.2%	12.7%

Table 6: Financial metrics associated with different dynamic strategies: real vs. simulated data.

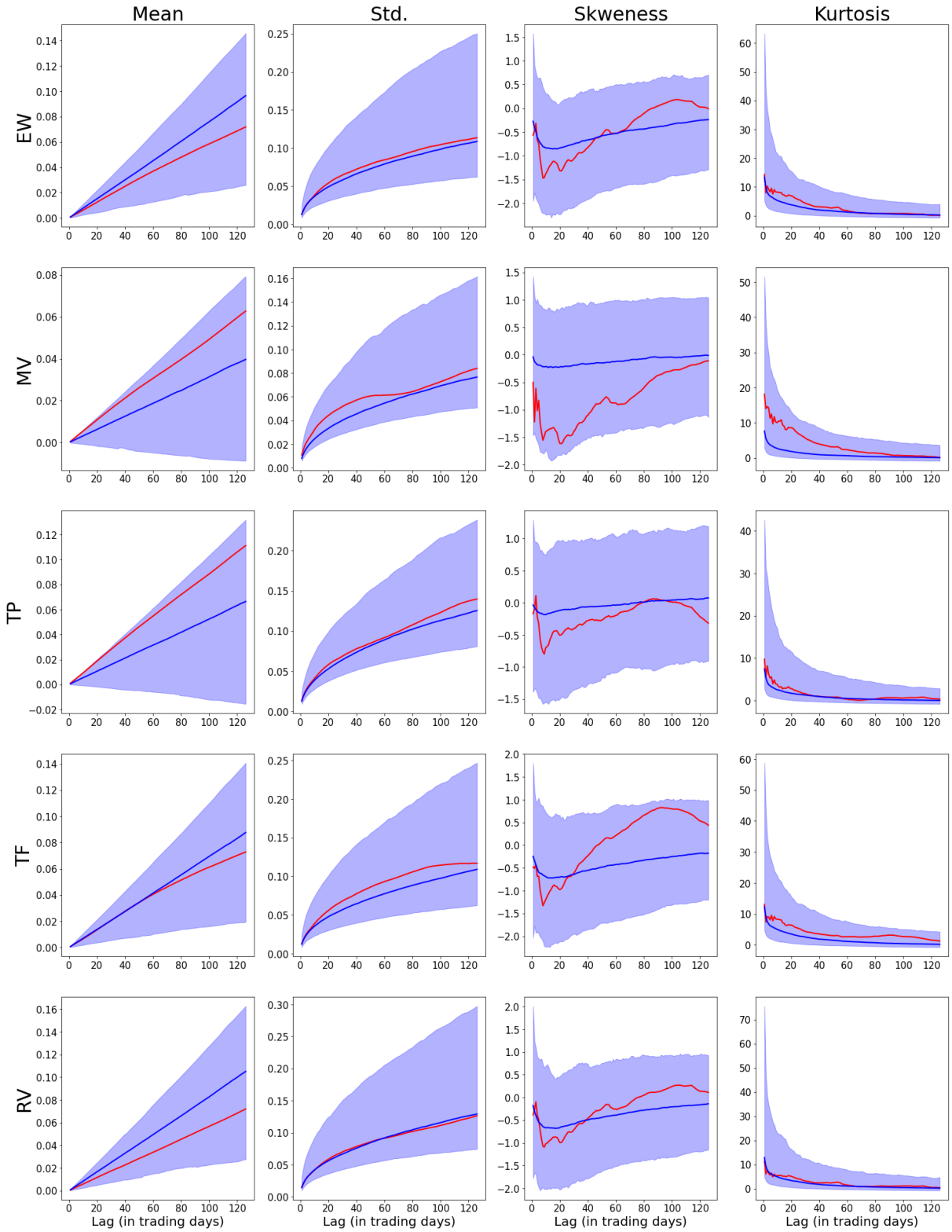


Figure 9: Evolution of the four first moments of the returns distributions for the different considered buy-and-hold strategies as a function of the time-horizon. The red curve corresponds to the real data, the blue curve to the median of the simulated data, and the blue area represents the 90% interval of the simulations.

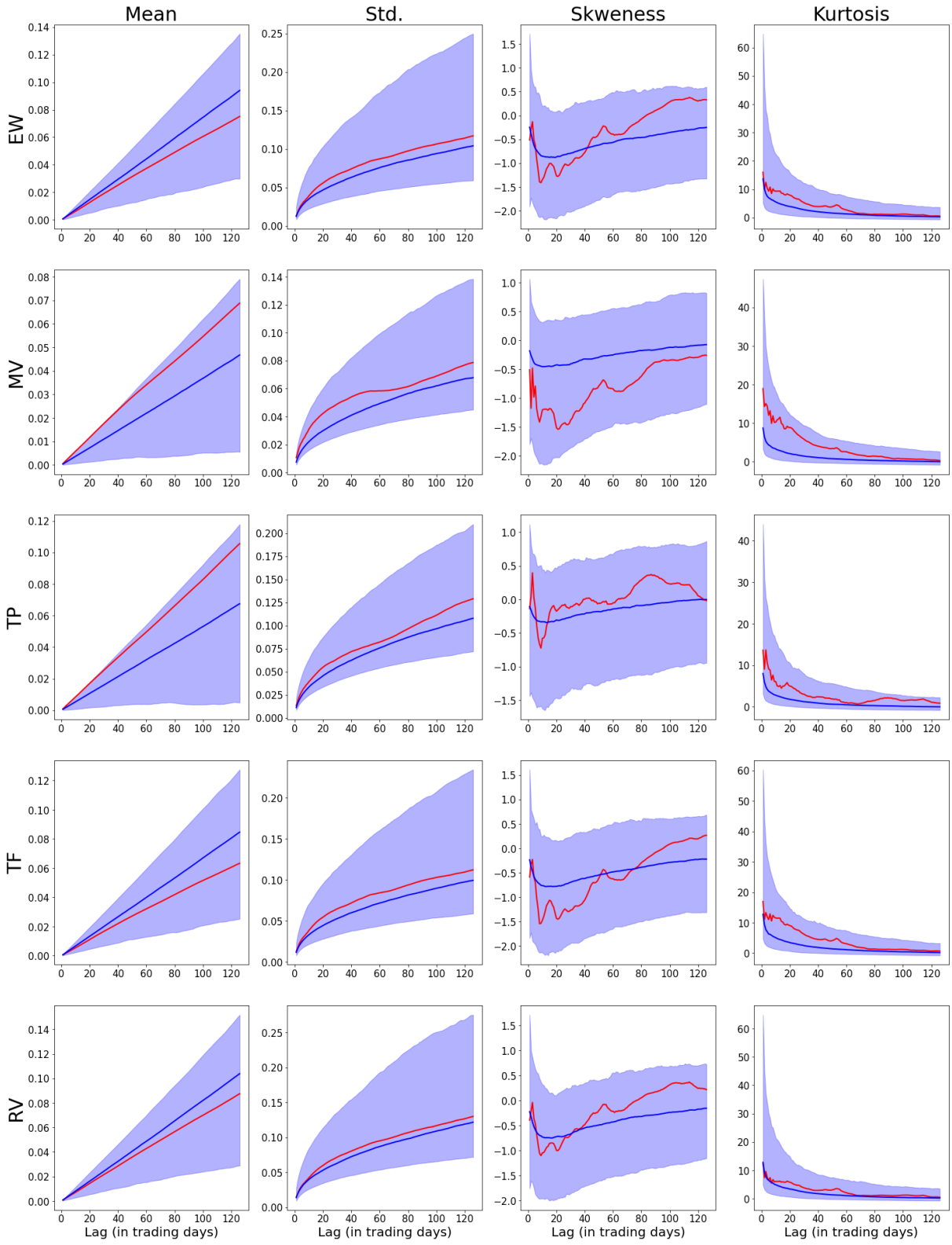


Figure 10: Evolution of the four first moments of the returns distributions for the different considered constant-weighted strategies as a function of the time-horizon. The red curve corresponds to the real data, the blue curve to the median of the simulated data, and the blue area represents the 90% interval of the simulations.

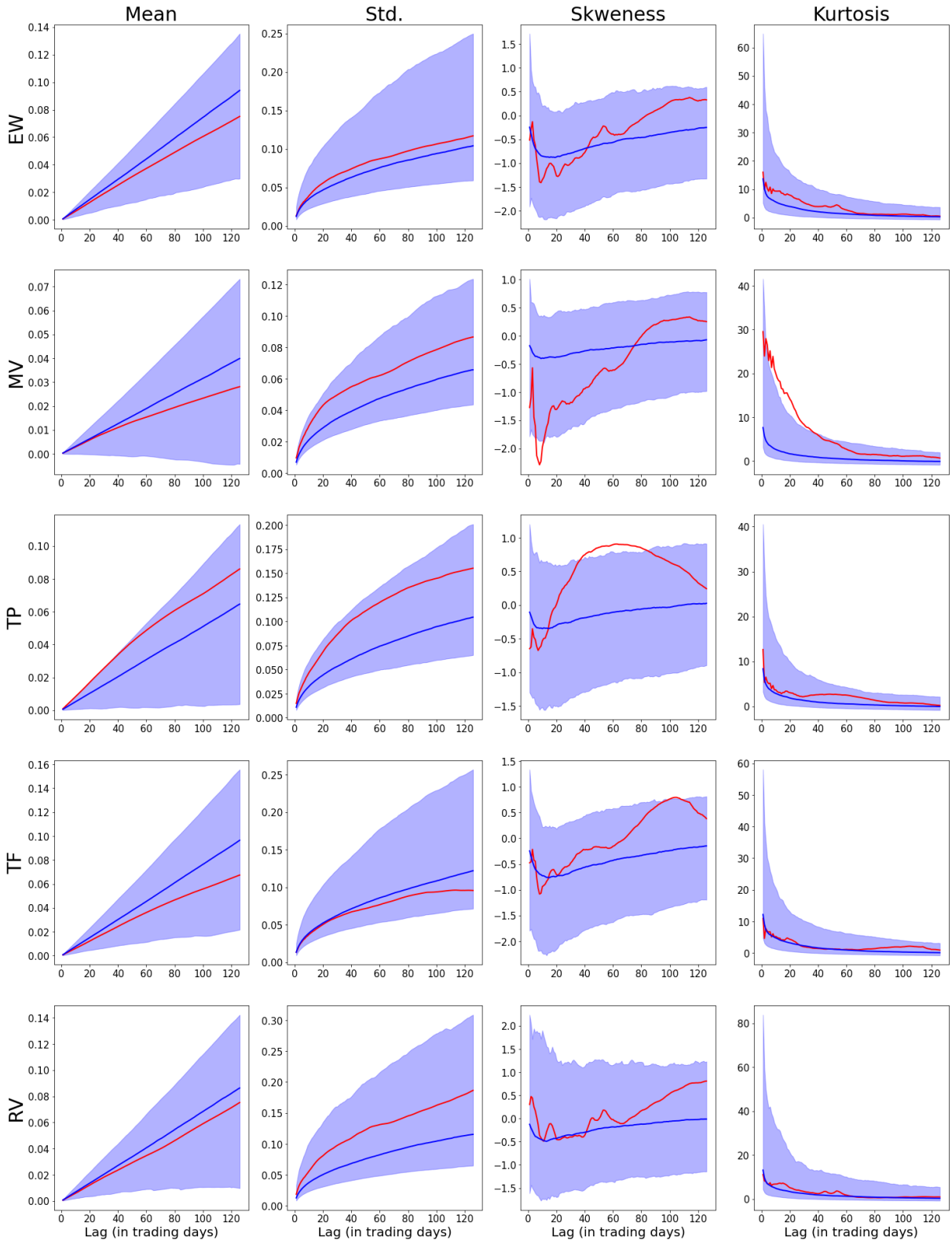


Figure 11: Evolution of the four first moments of the returns distributions for the different considered dynamic strategies as a function of the time-horizon. The red curve corresponds to the real data, the blue curve to the median of the simulated data, and the blue area represents the 90% interval of the simulations.

Beyond this generally very positive big picture regarding the FPDM generator's ability to generate realistic market scenarios, there is, however, a relative disparity in performance when it comes to reproducing certain time-series features depending on the strategies considered.

To begin, the data generated by the model seems, on average, to slightly underestimate the risk associated with the various MV portfolios in these different approaches (buy-and-hold, fixed-weighted, and dynamic) compared to real data. Thus, the different financial risk metrics of these strategies are significantly higher than their median model level. Similarly, the empirical distributions of returns of these MV strategies are more skewed than the median level of the return distribution of these strategies on simulated data. This potential underestimation of risk could be explained by the fact that this strategy is built from the empirical estimator of the covariance matrix. Indeed, in the case of buy-and-hold and fixed-weighted approaches, and even for the initialization of the dynamic approach, this matrix is calculated over the exact period on which the model is fitted. Consequently, this strategy tends to maximize the model risk. However, this potential underestimation of risk requires nuance. Firstly, the various empirical risk metrics associated with the MV strategies remain within the 1st and 9th deciles of the simulated data. As a result, the empirical features of these strategies are not outliers when viewed through the lens of the model. Furthermore, it is noteworthy that the proposed model seems to avoid, or at least partially mitigate, the phenomenon of overfitting. This shows up in the fact that, while the empirical estimator of the daily annualized volatility of the constant-weighted MV portfolio computed from the learning period is 7.89% (even when using a shrinkage estimator of the covariance matrix, this estimated volatility remains below 8%), the first decile of this volatility calculated from the simulations exceeds 9%.

Additionally, the rebalancing approach plays a somewhat significant role in the proximity between metrics on real data and their median levels on simulated data. Thus, while with the buy-and-hold approach, the empirical risk metrics of the strategies TP and RV are close to their median levels on simulated data, they are more off-center towards higher deciles when considering dynamic rebalancing. However, once again, like for the MV strategy, the financial metrics associated with real data remain within the 80% confidence interval of the simulated data. Therefore, it is difficult to conclusively assert based on these results that the model underestimates the risk of these strategies.

Beyond effectively capturing the features of the individual strategies, the model also reproduces the correlation between these strategies quite well overall, as shown in figures 13, 15, and 17. On this point, however, the proximity between real and simulated data depends significantly on the rebalancing approach adopted. For the buy-and-hold and constant-weighted approaches, the correlations between the different strategies in the simulated data are very close to those in the real data (figures 13 and 15). This correlation structure is however less similar to the real data for the dynamic rebalancing approach (figure 17). Furthermore, within this approach, the disparity between real correlations and correlations in simulated data depends on the pairs of strategies considered. Thus, the correlation of the pair of strategies EW/MV is well reproduced by the model. Conversely, the empirical correlations between pairs involving the TP, TF, and RV strategies differ significantly from the correlations obtained from simulated data. However, these three strategies share a common characteristic: their composition depends on an estimator of the drift vector, specifically a moving average estimator (over the last 4 years for the TP strategy and over the last year for the TF and RV strategies). This observation thus suggests that the disparity between the correlations in real and simulated data could be attributable to an imperfect modeling of drift dynamics by the generator. Therefore, the assumption of the market factor as the unique determinant of asset drift is likely too simplistic and warrants amendment.

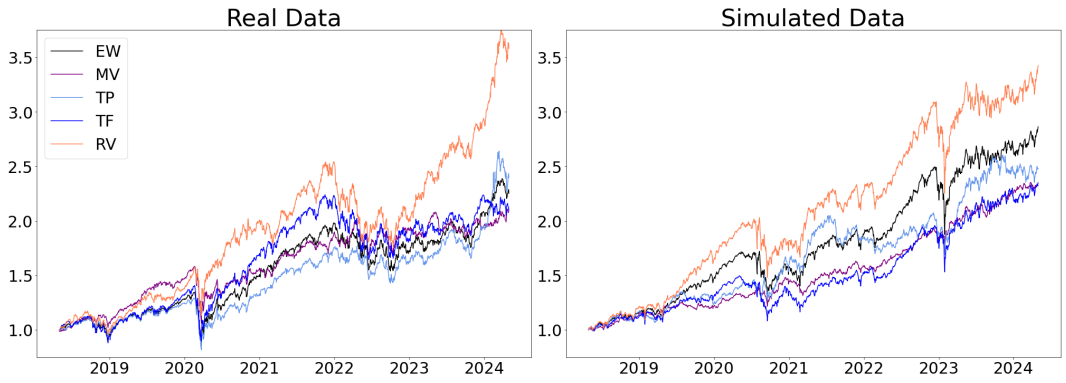


Figure 12: Cumulative returns of the considered buy-and-hold strategies: real vs simulated data.

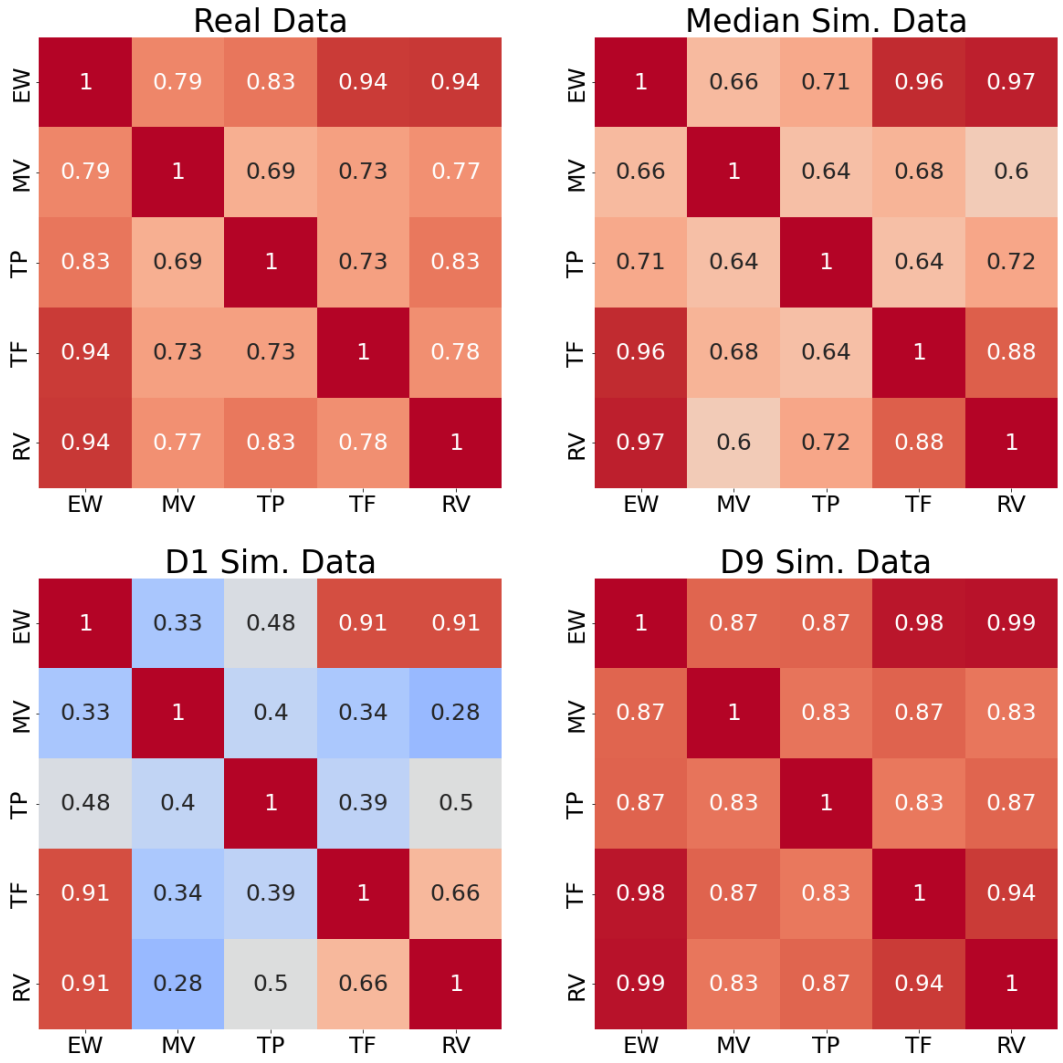


Figure 13: Correlation matrix of daily returns of the different buy-and-hold strategies: real vs. simulated data.

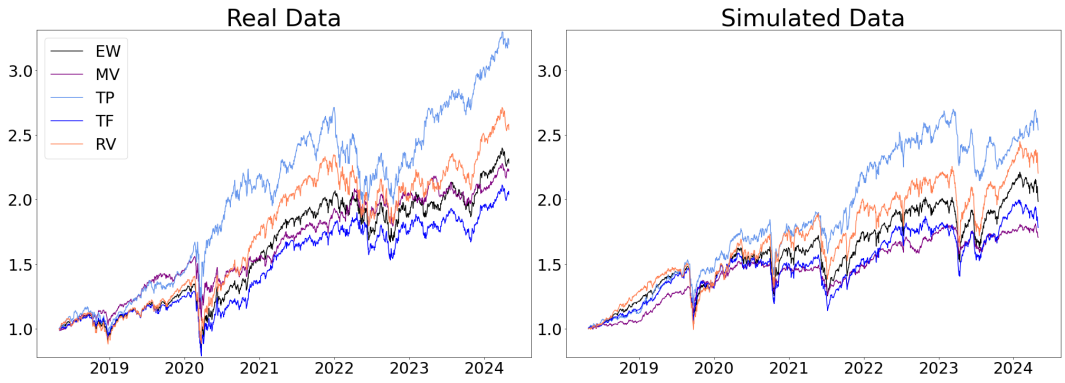


Figure 14: Cumulative returns of the considered fixed-weighted strategies: real vs simulated data.

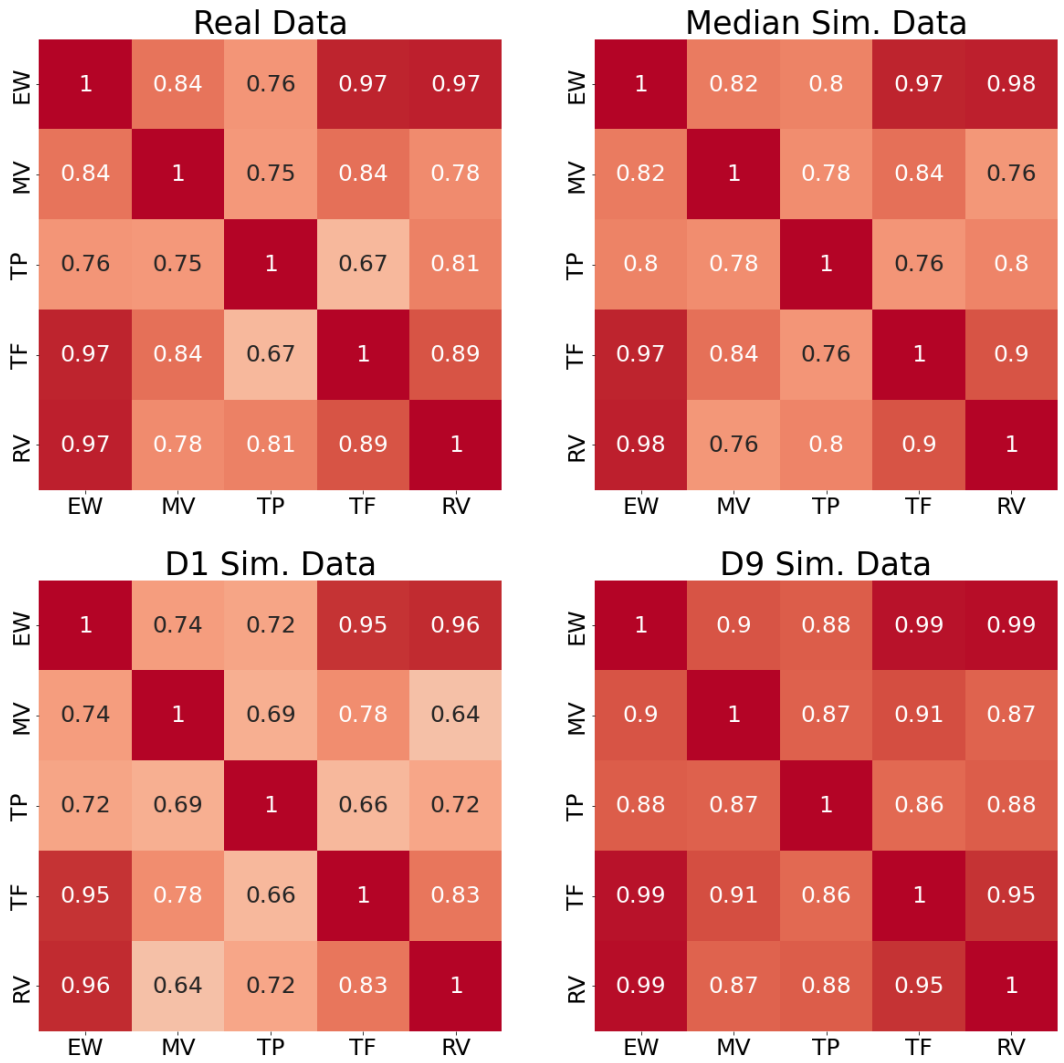


Figure 15: Correlation matrix of daily returns of the different fixed-weighted strategies: real vs simulated data.

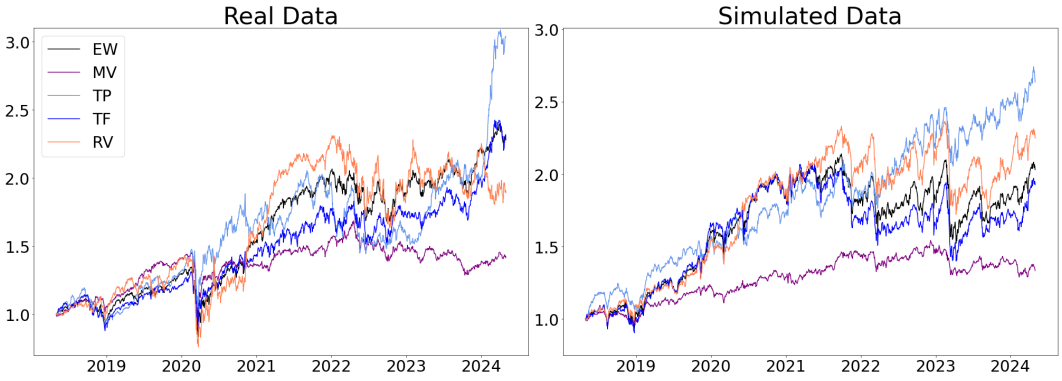


Figure 16: Cumulative returns of the considered dynamic strategies: real vs simulated data.

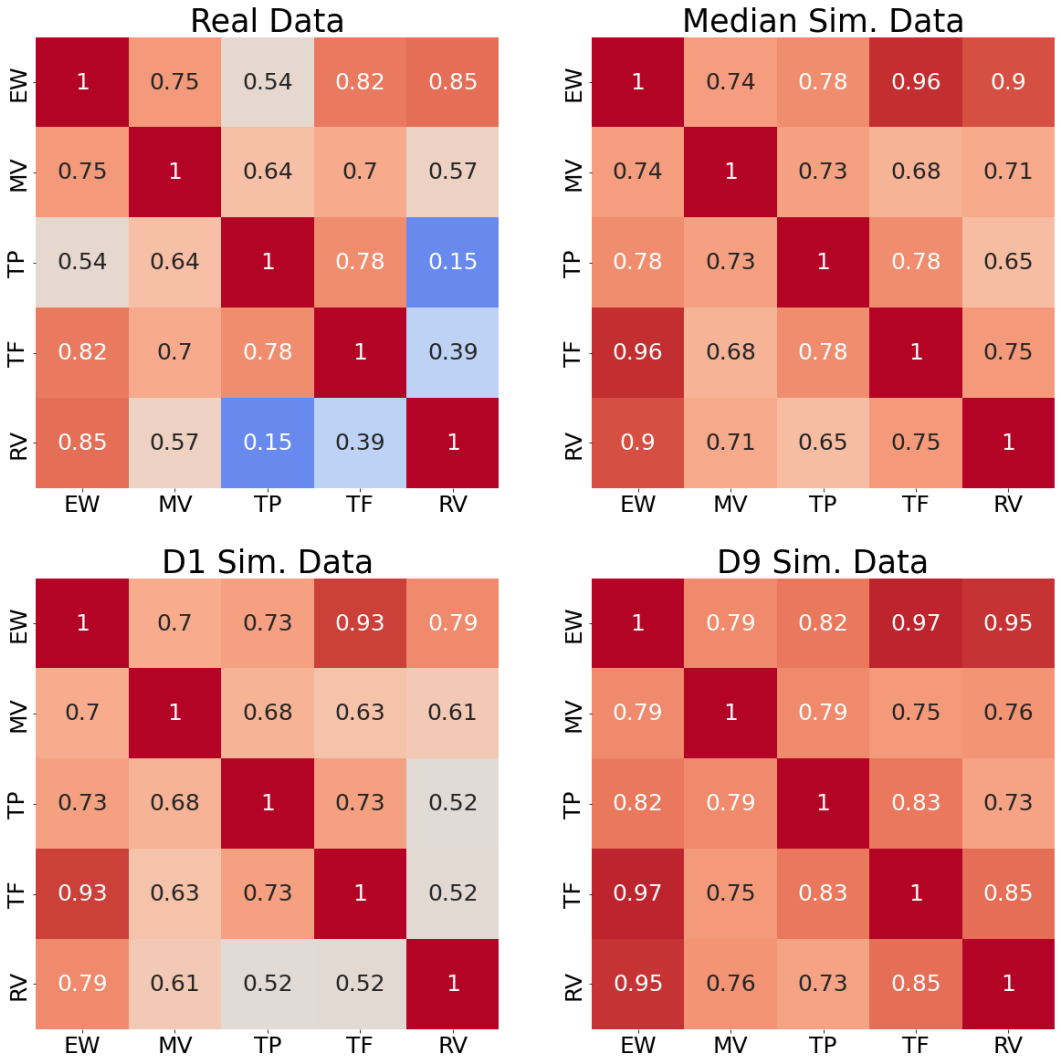


Figure 17: Correlation matrix of daily returns of the different dynamic strategies: real vs simulated data.

5 Conclusion

In this article, we have introduced the Factorial Path-Dependent Market (FPDM) model, a new theoretical framework for modeling multivariate asset price dynamics. In this framework, asset prices are driven by a set of elementary factors of two kinds: common elementary factors that link the dynamics of assets together and generate the correlation structure between individual price variations, and idiosyncratic elementary factors specific to each asset. The dynamics of the vector of these elementary factors are described by a multidimensional Itô SDE, whose drifts and volatility vectors depend either partially or completely on its own past trajectory. As a result, the proposed approach falls within the realm of path-dependent modelings. This choice aims to account for the fundamentally endogenous nature of financial markets, which manifests in various forms, ranging from the volatility formation process to the emergence of price trends caused by feedback effects from previous price dynamics. In practice, the path-dependence in the proposed model is implemented by modeling factor drifts and volatilities as functions of features encoding the past trajectory of factor portfolios.

After defining this theoretical framework in section 2, section 3 introduced a market generator derived from a specific version of the FPDM model. In this one, the path-dependent component of the volatility of each elementary factor depends on both the past trajectory of the elementary factor to which it is associated and the past trajectory of the elementary factor that constitutes the market factor. The market factor thus plays a singular role in this framework, firstly in the overall volatility of the investment universe under consideration, and secondly in shaping the level and dynamics of the correlation structure among the assets it comprises. Furthermore, in the considered specification, the market factor is the only elementary factor with a non-zero drift, thus producing a CAPM-like modeling.

Once this market generator has been defined, a calibration method for it has been proposed. This process consists of two main phases. The first involves factorizing a historical dataset of returns from the investment universe under consideration, such that each return is expressed as a linear combination of variations corresponding to what, in the model, are the elementary factors. The second phase takes as input the elements resulting from this factorization to estimate the various model parameters using an adapted form of maximum likelihood estimation.

Section 4 provided an extensive evaluation of this market generator using an investment universe of 436 assets from the S&P500. For this purpose, the data set was divided into two parts: the first, spanning from April 1, 2010, to April 30, 2018, was used to fit the model, while the second part, covering data from May 1, 2018, to April 30, 2024, was used to conduct the various tests. All evaluations of the market generator were thus conducted out-of-sample.

A first set of tests showed that the moments of the marginal distributions of the asset returns generated by the model were highly consistent with their empirical counterparts calculated from market data. A second type of evaluation, based on the joint dynamics of the assets, was then conducted. A first key result from these tests is that using synthetic data generated by the FPDM generator enables us to obtain an estimator of the covariance matrix of returns that is significantly better than standard or shrinkage-based estimators derived from historical data. Supplementary results from these tests suggest that this improvement seems to stem less from the model's ability to better capture linear correlations between asset returns and more from its enhancement in estimating their individual standard deviations while maintaining a realistic correlation structure. This, however, does not imply that the model's ability to

capture correlations is poor. For example, the estimated Kendall's tau and Spearman's rho calculated on simulated data exhibit performance on test data similar to the empirical estimators of these correlation measures calculated on historical data. Furthermore, these results must also be seen in the light of the fact that, on the one hand, the correlation structure in the model is dynamic, and on the other, the test horizon is relatively long (six years), which constitutes rather unfavorable evaluation conditions for the model.

The final set of tests aimed to evaluate the model's ability as a tool for backtesting investment strategies. For this purpose, five types of strategies were considered, each through three different approaches: buy-and-hold, constant-weight with daily rebalancing, and dynamic with monthly rebalancing. The results initially demonstrated that the various time series features characterizing the individual trajectories of these strategies were generally well replicated by the generator. This includes moments of return distributions calculated at different frequencies, as well as various risk metrics such as value-at-risk and expected shortfall at different confidence levels, along with maximum drawdown. In this respect, then, the FPDM generator seems to be a relevant and powerful tool for backtesting various strategies on synthetic data. The results regarding the model's capture of the correlation between these strategies are slightly more nuanced. Thus, while the correlations of returns among strategies under the buy-and-hold and constant-weighted modes are well reproduced by the model, the difference between the correlations among dynamic strategies from simulated data differs quite significantly from their counterparts obtained from empirical data. However, this latter observation only pertains to the strategies whose composition depends on an estimator of the drift vector, specifically a moving average estimator. This implies that the difference in correlations between real and simulated data may stem from an imperfect modeling of drift dynamics by the generator. Therefore, the assumption that the market factor is the sole determinant of asset drift is likely overly simplistic and may require adjustments, such as allowing other elementary factors to have a non-zero drift.

More generally, the market generator presented in this article could be improved in multiple ways. As an example, the model could incorporate "views" in the Black-Litterman sense ([15]) to integrate exogenous information into the model. Furthermore, the market generator proposed is just one possible (discrete) specification of the FPDM model introduced in section 2. Hence, for example, more sophisticated specifications that directly incorporate sectoral decomposition via elementary factor portfolios could further refine the proposed framework. Additionally, the model can be used to generate conditional scenarios, such as given a certain trajectory of the market factor, which also deserves further investigation.

Regardless, the results obtained from the market generator proposed in this article already demonstrate that the FPDM model constitutes an extremely rich framework that can be used as a white box to generate highly realistic simulations of price trajectories in a high-dimensional investment universe.

References

[1] Abi Jaber E., and El Euch O. (2019). Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 10(2), 309-349.

[2] Ali U., and Hirshleifer D. (2020). Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics*, 136(3), 649-675.

[3] Andersen T. G., Bollerslev T., Diebold F. X., and Labys P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579-625.

[4] Ang A., and Chen J. (2002). Asymmetric correlations of equity portfolios. *Journal of financial Economics*, 63(3), 443-494.

[5] Arratia A., Cabaña A., and Cabaña, E.M. (2018). Embedding in law of discrete time ARMA processes in continuous time stationary processes. *Journal of Statistical Planning and Inference*, 197, 156-167.

[6] Baele L. (2005). Volatility spillover effects in European equity markets. *Journal of Financial and Quantitative Analysis*, 40(2), 373-401.

[7] Baele L., Bekaert G., Inghelbrecht K., and Wei M. (2020). Flights to safety. *The Review of Financial Studies*, 33(2), 689-746.

[8] Bakshi G., and Kapadia N. (2003). Delta-hedged gains and the negative market volatility risk premium. *The Review of Financial Studies*, 16(2), 527-566.

[9] Babura M., Giannone D., and Lenza M. (2015). Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections. *International Journal of forecasting*, 31(3), 739-756.

[10] Bekaert G., Engstrom E. C., and Xu N. R. (2022). The time variation in risk appetite and uncertainty. *Management Science*, 68(6), 3975-4004.

[11] Bernanke B. S., Gertler M., and Gilchrist S. (1994). The financial accelerator and the flight to quality.

[12] Bianchi M. L., Hitaj A., and Tassinari G. L. (2020). Multivariate non-Gaussian models for financial applications. *arXiv* preprint arXiv:2005.06390.

[13] Bianchi M.L., Hitaj A., and Tassinari G.L. (2020). Multivariate non-Gaussian models for financial applications. *arXiv* preprint arXiv:2005.06390.

[14] Black F., Studies of stock price volatility changes. *Studies of Stock Price Volatility Changes. Proceedings of the 1976 Meetings of the American Statistical Association*, 171-181.

[15] Black F., and Litterman R. (1992). Global portfolio optimization. *Financial analysts journal*, 48(5), 28-43.

[16] Bouchaud J. P., Matacz A., Potters M. (2001). Leverage effect in financial markets: The retarded volatility model. *Physical review letters*, 87(22), 228701.

[17] Bouchaud J. P., and Potters M. (2009). Financial applications of random matrix theory: a short review. *arXiv* preprint arXiv:0910.1205.

[18] Box G.E., Jenkins G.M., Reinsel G.C., and Ljung G.M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[19] Brockwell P. J. (2001). Continuous-time ARMA processes. *Handbook of statistics*, 19, 249-276.

[20] Brockwell P. J. (2004). Representations of continuous-time ARMA processes. *Journal of Applied Probability*, 41(A), 375-382.

[21] Brooks C., and Persand G. (2001). Volatility forecasting for risk management. *Journal of Forecasting*, 20(5), 341-356.

[22] Chambers M.J., and Thornton M. A. (2012). Discrete time representation of continuous time ARMA processes. *Econometric Theory*, 28(1), 219-238.

[23] Ciner C., Gurdgiev C., and Lucey B. M. (2013). Hedges and safe havens: An examination of stocks, bonds, gold, oil and exchange rates. *International Review of Financial Analysis*, 29, 202-211.

[24] Chen Y., Wiesel A., Eldar Y. C., and Hero A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE transactions on signal processing*, 58(10), 5016-5029.

[25] Chicheportiche R., and Bouchaud J.P. (2012). The joint distribution of stock returns is not elliptical. *International Journal of Theoretical and Applied Finance*, 15(03), 1250019.

[26] Christiansen, C. (2007). Volatilityspillover effects in European bond markets. *European Financial Management*, 13(5), 923-948.

[27] Colacito R., Engle R.F., and Ghysels E. (2011). A component model for dynamic correlations. *Journal of Econometrics*, 164(1), 45-59.

[28] Cont R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2), 223.

[29] Cont R., Cucuringu M., Xu R., and Zhang, C. (2022). Tail-gan: Learning to simulate tail risk scenarios. *arXiv* preprint arXiv:2203.01664.

[30] Engle R.F., Ledoit O., and Wolf M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2), 363-375.

[31] Engle R.F., Lilien D. M., and Robins R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. . *Econometrica: journal of the Econometric Society*, 391-407.

[32] Flaig S., and Junike G. (2022). Scenario generation for market risk models using generative neural networks. *Risks*, 10(11), 199.

[33] Foschi P., and Pascucci A. (2008). Path dependent volatility. Decisions in Economics and Finance, 31, 13-32.

[34] French K. R., Schwert G. W., and Stambaugh R. F. (1987). Expected stock returns and volatility. *Journal of financial Economics*, 19(1), 3-29.

[35] Gatheral, J. Jusselin P., and Rosenbaum M. (2020). The quadratic rough Heston model and the joint S&P 500/VIX smile calibration problem. *arXiv* preprint arXiv:2001.01789.

[36] Gropp J. (2004). Mean reversion of industry stock returns in the US, 19261998. *Journal of Empirical Finance*, 11(4), 537-551.

[37] Gustafsson J., and Jonsson C. (2023). *Scenario Generation for Stress Testing Using Generative Adversarial Networks: Deep Learning Approach to Generate Extreme but Plausible Scenarios.*

[38] Guyon J. (2014). Path-dependent volatility. *Risk*.

[39] Guyon J., and Lekeufack J. (2023). Volatility is (mostly) path-dependent. *Quantitative Finance*, 23(9), 1221-1258.

[40] Han Y. (2011). On the relation between the market risk premium and market volatility. *Applied financial economics*, 21(22), 1711-1723.

[41] Hong Y. (2001). A test for volatility spillover with application to exchange rates. Journal of Econometrics, 103(1-2), 183-224.

[42] Jusselin P., Lezmi E., Malongo H., Masselin C., Roncalli T., and Dao T. L. (2017). Understanding the momentum risk premium: An in-depth journey through trend-following strategies. Available at *SSRN* 3042173.

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

[44] Kondratyev A., and Schwarz C. (2019). The market generator. Available at SSRN3384948.

[45] Laloux L., Cizeau P., Bouchaud J. P., and Potters M. (1999). Noise dressing of financial correlation matrices. *Physical review letters*, 83(7), 1467.

[46] Ledoit O., and Wolf M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2), 365-411.

[47] Ledoit O., and Wolf M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5), 3043-3065.

[48] Lezmi E., Roche J., Roncalli T., and Xu J. (2020). Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks. Available at SSRN 3645473.

[49] Lopez de Prado M. (2016). A robust estimator of the efficient frontier. Available at *SSRN* 3469961.

[50] Lopez de Prado M. (2019). Tactical investment algorithms. Available at *SSRN* 3459866.

[51] Mandelbrot B.B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4),394-419.

[52] Merton R.C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867-887.

[53] Meucci A. (2007). Risk contributions from generic user-defined factors. *Risk*, 84-88.

[54] Meucci A. (2009). Review of statistical arbitrage, cointegration, and multivariate Ornstein-Uhlenbeck.

[55] Moskowitz T.J., and Grinblatt M. (1999). Do industries explain momentum?. *The Journal of finance*, 54(4), 1249-1290.

[56] Ni H., Szpruch L., Wiese M., Liao S. and Xiao B. (2020). Conditional Sig-Wasserstein GANs for Time Series Generation. *arXiv* preprint arXiv:2006.05421.

[57] Oksendal B. (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

[58] Parent L. (2023). Deep Estimation for Volatility Forecasting. Available at *SSRN* 4470221.

[59] Parent L. (2023). Investigating Approaches to Modeling Rough Path-Dependent Volatility: Insights and Implications. Available at *SSRN* 4579759.

[60] Parent L. (2022). Rough Path-Dependent Volatility Models. Available at *SSRN* 4270481.

[61] Pelger M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics* 208.1 (2019): 23-42.

[62] Poterba J. M., and Summers L. H. (1988). Mean reversion in stock prices: Evidence and implications. Journal of financial economics, 22(1), 27-59.

[63] Potluru V. K., Borrajo D., Coletta A., *et al.* (2023). Synthetic Data Applications in Finance. *arXiv* preprint arXiv:2401.00081.

[64] Rizzato M., Wallart J., Geissler C., Morizet N., and Boumlaik N. (2023). Generative Adversarial Networks applied to synthetic financial scenarios generation. *Physica A: Statistical Mechanics and its Applications*, 623, 128899.

[65] Roncalli T. (2011). Understanding the impact of weights constraints in portfolio theory. *Available* at SSRN 1761625.

[66] Roncalli T. (2020). *Handbook of financial risk management*. Chapman and Hall/CRC.

[67] Rosenbaum M., and Zhang J. (2021). Deep calibration of the quadratic rough Heston model. *arXiv* preprint arXiv:2107.01611.

[68] Rosenbaum M., and Zhang J. (2022). On the universality of the volatility formation process: when machine learning and rough volatility agree. *arXiv* preprint arXiv:2206.14114.

[69] Ross S. (1976), The arbitrage theory of capital pricing, *Journal of Economic Theory* 13, 341-360.

[70] Scruggs J. T. (1998). Resolving the puzzling intertemporal relation between the market risk premium and conditional market variance: A twofactor approach. *The Journal of Finance*, 53(2), 575-603.

[71] Serletis A., Rosenberg A.A. (2009). Mean reversion in the US stock market. *Chaos, Solitons & Fractals*, 40(4), 2007-2015.

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

[73] Tasche D. (1999). Risk contributions and performance measurement. *Report of the Lehrstuhl für mathematische Statistik, TU München.*

[74] Yang M. (2004). Normal log-normal mixture: Leptokurtosis, skewness and applications. *Econometric Society.*

[75] Zeevi A., and Mashal R. (2002). Beyond correlation: Extreme co-movements between financial assets. Available at *SSRN* 317122.

[76] Zumbach G. (2010). Volatility conditional on price trends. *Quantitative Finance*, 10(4), 431-442.

Appendix A List of notations

The following table presents the matrix operators and standard matrices/vectors used in this article.

Symbol	Description
$\mathbf{1}_n$	$n \times 1$ vector of ones
$\mathbf{e}_i$	$n \times 1$ vector whose value is 1 for coordinate i and 0 elsewhere
$\mathbf{I}_n$	Identity matrix of dimension $n \times n$
$(\boldsymbol{V})_i$	Element i of the vector $\boldsymbol{V}$
$(\boldsymbol{M})_{[i,j]}$	Entry at row i and column j of the matrix $\boldsymbol{M}$
$(\boldsymbol{M})_{[i,:]}$	Row i of the matrix $\boldsymbol{M}$
$(\boldsymbol{M})_{[:,j]}$	Column j of the matrix $\boldsymbol{M}$
$\boldsymbol{M}^\top$	Transpose of the matrix $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	Inverse of the square matrix $\boldsymbol{M}$
$\boldsymbol{M}^{\circ p}$	Element-wise application of the power p to each coordinate of $\boldsymbol{M}$
$f \circ (\boldsymbol{M})$	Element-wise application of the function f to each coordinate of $\boldsymbol{M}$
$\odot$	Hadamard product
$\oslash$	Hadamard division
$\otimes$	Kronecker product
$\oplus$	Kronecker sum
$\mathrm{diag}(\boldsymbol{V})$	Converts the vector $\boldsymbol{V}$ into a diagonal matrix
$\mathrm{diag}_{M \to d}(\boldsymbol{M})$	Transformation of the square matrix $\boldsymbol{M}$ into a vector from the diagonal of $\boldsymbol{M}$
$\mathrm{diag}_{M \to D}(\boldsymbol{M})$	Transformation of the square matrix $\boldsymbol{M}$ into a diagonal matrix by retaining only the diagonal of $\boldsymbol{M}$
$\mathrm{Tr}(\boldsymbol{M})$	Trace of the square matrix $\boldsymbol{M}$
$\mathrm{vec}(\boldsymbol{M})$	The vec operator that transforms a matrix into a column vector
$\mathrm{vec}^{-1}_{m \times m}(\boldsymbol{V})$	The inverse vec operator that transforms the $m \times 1$ vector $\boldsymbol{V}$ into a square matrix $m \times m$
$\|\boldsymbol{M}\|_F$	The Frobenius norm of the matrix $\boldsymbol{M}$

Appendix B Stochastic differential equations involved in the FPDM model

B.1 Calculation of the asset price vector solution

The SDE describing the dynamics of the price vector $\mathbf{P}$ is given by:

$$\mathrm{d}\mathbf{P}_t = \mathbf{P}_t \odot (\mathbf{A}\boldsymbol{\mu}_t \mathrm{d}t) + \mathbf{P}_t \odot \left(\mathbf{A}\sqrt{\boldsymbol{\Omega}}_t \mathrm{d}\mathbf{W}_t\right).$$

However, if f is a C^2 function such that $\mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n$, using the general Itô formula ([57]), we have:

$$\mathrm{d}f_k(t, \mathbf{P}_t) = \frac{\partial f_k(t, \mathbf{P}_t)}{\partial t}\mathrm{d}t + \sum_i \frac{\partial f_k(t, \mathbf{P}_t)}{\partial P_i}\,\mathrm{d}(\mathbf{P}_t)_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f_k(t, \mathbf{P}_t)}{\partial P_i \partial P_j}\,\mathrm{d}(\mathbf{P}_t)_i\,\mathrm{d}(\mathbf{P}_t)_j.$$

Considering $f(t, \mathbf{P}_t) = \ln \circ \mathbf{P}_t$, we have:

$$\begin{aligned}
\frac{\partial f_k(t, \mathbf{P}_t)}{\partial t}\mathrm{d}t &= 0, \\
\sum_i \frac{\partial f_k(t, \mathbf{P}_t)}{\partial P_i}\,\mathrm{d}(\mathbf{P}_t)_i = \frac{\mathrm{d}(\mathbf{P}_t)_k}{(\mathbf{P}_t)_k} &= \left(\mathbf{A}\boldsymbol{\mu}_t \mathrm{d}t + \mathbf{A}\sqrt{\boldsymbol{\Omega}_t}\mathrm{d}\mathbf{W}_t\right)_k, \\
\frac{1}{2}\sum_{i,j} \frac{\partial^2 f_k(t, \mathbf{P}_t)}{\partial P_i \partial P_j}\,\mathrm{d}(\mathbf{P}_t)_i\,\mathrm{d}(\mathbf{P}_t)_j = -\frac{1}{2(\mathbf{P}_t)_k}\mathrm{d}(\mathbf{P}_t)_k \mathrm{d}(\mathbf{P}_t)_k &= -\frac{1}{2}\cdot\left(\mathbf{A}\boldsymbol{\Omega}_t\mathbf{A}^\top\right)_{k,k}\mathrm{d}t.
\end{aligned}$$

It follows that:

$$\mathrm{d}\ln \circ \mathbf{P}_t = \left(\mathbf{A}\boldsymbol{\mu}_t - \frac{1}{2}\cdot \mathrm{diag}\left(\mathbf{A}\boldsymbol{\Omega}_t\mathbf{A}^\top\right)\right)\mathrm{d}t + \mathbf{A}\sqrt{\boldsymbol{\Omega}}_t \mathrm{d}\mathbf{W}_t$$

Consequently, the solution to the asset price vector is given by:

$$\mathbf{P}_t = \mathbf{P}_0 \odot \exp \circ \left(\int_0^t \mathbf{A}\boldsymbol{\mu}_u - \frac{1}{2}\cdot \mathrm{diag}\left(\mathbf{A}\boldsymbol{\Omega}_u\mathbf{A}^\top\right)\mathrm{d}u + \int_0^t \mathbf{A}\sqrt{\boldsymbol{\Omega}}_u \mathrm{d}\mathbf{W}_u.\right).$$

B.2 EWMA estimators and their stochastic differential equations

Consider the following vector stochastic differential equation (SDE):

$$\mathrm{d}\tilde{\boldsymbol{\mu}}_t^{(j)} = \frac{1}{\tau_j}\cdot\left(\mathrm{d}\mathbf{F}_t - \tilde{\boldsymbol{\mu}}_t^{(j)}\mathrm{d}t\right).$$

By setting $f\left(t, \tilde{\boldsymbol{\mu}}_t^{(j)}\right) = e^{\frac{t}{\tau_j}}\tilde{\boldsymbol{\mu}}_t^{(j)}$, and applying Ito's formula, we obtain:

$$\mathrm{d}f\big(t, \tilde{\boldsymbol{\mu}}_t^{(j)}\big) = \frac{e^{\frac{t}{\tau_j}}\tilde{\boldsymbol{\mu}}_t^{(j)}}{\tau_j}\mathrm{d}t + e^{\frac{t}{\tau_j}}\mathrm{d}\tilde{\boldsymbol{\mu}}_t^{(j)} = \frac{e^{\frac{t}{\tau_j}}}{\tau_j}\cdot\mathrm{d}\mathbf{F}_t$$

It follows that the solution corresponds to the following EWMA:

$$\tilde{\boldsymbol{\mu}}_t^{(j)} = \tilde{\boldsymbol{\mu}}_0^{(j)} e^{\frac{-t}{\tau_j}} + \frac{1}{\tau_j} \int_0^t e^{\frac{u-t}{\tau_j}} \cdot \mathrm{d}\mathbf{F}_u.$$

In the same manner, when considering

$$\mathrm{d}\tilde{\boldsymbol{\Omega}}_t^{(j)} = \frac{1}{\tau_j} \cdot \left(\mathrm{d}\mathbf{F}_t \mathrm{d}\mathbf{F}_t^\top - \tilde{\boldsymbol{\Omega}}_t^{(j)} \mathrm{d}t \right) = \frac{1}{\tau_j} \cdot \left(\tilde{\boldsymbol{\Omega}}_t - \tilde{\boldsymbol{\Omega}}_t^{(j)} \right) \mathrm{d}t,$$

and setting $f\left(t, \tilde{\boldsymbol{\Omega}}_t^{(j)}\right) = e^{\frac{t}{\tau_j}} \tilde{\boldsymbol{\Omega}}_t^{(j)}$, we have:

$$\mathrm{d}f\left(t, \tilde{\boldsymbol{\Omega}}_t^{(j)}\right) = \frac{e^{\frac{t}{\tau_j}} \tilde{\boldsymbol{\Omega}}_t^{(j)}}{\tau_j} \mathrm{d}t + e^{\frac{t}{\tau_j}} \mathrm{d}\tilde{\boldsymbol{\Omega}}_t^{(j)} = \frac{e^{\frac{t}{\tau_j}}}{\tau_j} \cdot \tilde{\boldsymbol{\Omega}}_t^{(j)} \mathrm{d}t.$$

It follows that:

$$\tilde{\boldsymbol{\Omega}}_t^{(j)} = \tilde{\boldsymbol{\Omega}}_0^{(j)} e^{\frac{-t}{\tau_j}} + \frac{1}{\tau_j} \int_0^t e^{\frac{u-t}{\tau_j}} \cdot \tilde{\boldsymbol{\Omega}}_u^{(j)} \mathrm{d}u.$$

Appendix C Important specifications of the FPDM model

C.1 The 4-factor PDV as a specific case of the FPDM model

Suppose that $m = n = 1$, $\mathbf{A} = 1$, and $n_\tau = 2$. In this univariate single-factor framework, there exists a unique factor portfolio, which is $y = 1$. Therefore $\mathcal{Y}_t = 1 \; \forall \; t$. If we additionally assume that the drift is zero, we then have the following FPDM model:

$$\begin{cases} \dfrac{\mathrm{d}P_t}{P_t} & = \mathrm{d}F_t = \sigma_t \mathrm{d}\mathrm{W}_t, \\ \sigma_t & = \bar{\sigma} + \mathrm{b}_1 \cdot \hat{\mu}_t\left(1, \boldsymbol{\delta}\right) + \mathrm{b}_2 \cdot \hat{\sigma}_t\left(1, \mathbf{w}\right), \\ \hat{\mu}_t\left(1, \boldsymbol{\delta}\right) & = (\boldsymbol{\delta})_1 \cdot \tilde{\mu}_t^{(1)} + (\boldsymbol{\delta})_2 \cdot \tilde{\mu}_t^{(2)}, \\ \hat{\sigma}_t\left(1, \mathbf{w}\right) & = \sqrt{(\mathbf{w})_1 \cdot \tilde{V}_t^{(1)} + (\mathbf{w})_2 \cdot \tilde{V}_t^{(2)}}, \\ \mathrm{d}\tilde{\mu}_t^{(j)} & = \dfrac{1}{\tau_j} \cdot \left(\dfrac{\mathrm{d}P_t}{P_t} - \tilde{\mu}_t^{(j)} \mathrm{d}t \right), \\ \mathrm{d}\tilde{V}_t^{(j)} & = \dfrac{1}{\tau_j} \cdot \left(\left(\dfrac{\mathrm{d}P_t}{P_t} \right)^2 - \tilde{V}_t^{(j)} \mathrm{d}t \right), \end{cases}$$

which corresponds to the 4-factor PDV model by Guyon and Lekeufack.

C.2 Definition of the factorial drift vector in the context of the CAPM

Let π represent the value of the market portfolio, and assume that at any time, there exists a vector $\mathbf{y}_t^\star$ such that, $\forall\ t$:

$$\frac{\mathrm{d}\pi_t}{\pi_t} = \left(\mathbf{y}_t^\star\right)^\top \mathrm{d}\mathbf{F}_t.$$

Let us now suppose that the vector of drifts for elementary factors depends on the $n+1$ factor portfolios $y_t^\star, \mathbf{e}_1, ..., \mathbf{e}_n$ (where $\mathbf{e}_j$ represents the factor portfolio composed entirely of factor j), such that:

$$\mu_t = \boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot \Gamma_t^\star + \sum_{p=1}^{n} \boldsymbol{\beta}_t\left(\mathbf{e}_1\right) \cdot \Gamma_{p,t}^{(I)} = \boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot \Gamma_t^\star + \sum_{p=1}^{n} \Gamma_{p,t}^{(I)},$$

where

$$\Gamma_{p,t}^{(I)} = \mathcal{E}_{p,t}^{(I)} = \left(1 - \left(\mathbf{A}\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right)\right)_i\right) \cdot r_t + \left(\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right)\right)_i \cdot r_t.$$

From the expression of μ and since for $i > m, (\mathbf{y}_t^\star)_i = 0$, the drift of $\mathrm{y}^\star$ is given by:

$$\begin{aligned}
\mu(\mathbf{y}_t^\star) &= \left(\mathbf{y}_t^\star\right)^\top \boldsymbol{\mu}_t \\
&= \left(\mathbf{y}_t^\star\right)^\top \Gamma_t^{(I)} + \left(\mathbf{y}_t^\star\right)^\top \boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot \Gamma_t^\star \\
&= \left(\mathbf{y}_t^\star\right)^\top \frac{\boldsymbol{\Omega}_t \mathbf{y}_t^\star}{\left(\mathbf{y}_t^\star\right)^\top \boldsymbol{\Omega}_t \mathbf{y}_t^\star} \cdot r_t + \left(\mathbf{y}_t^\star\right)^\top \frac{\boldsymbol{\Omega}_t \mathbf{y}_p^\star}{\left(\mathbf{y}_p^\star\right)^\top \boldsymbol{\Omega}_t \mathbf{y}_p^\star} \cdot \Gamma_t^\star \\
&= r_t + \Gamma_t^\star
\end{aligned}$$

It follows that the vector of the asset drift is defined by:

$$\begin{aligned}
\boldsymbol{\mu}_t &= \mathbf{A}\boldsymbol{\mu}_t \\
&= \mathbf{A}\mathcal{E}_t + \mathbf{A}\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot \Gamma_t^\star \\
&= \left(\mathbf{1}_n - \mathbf{A}\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right)\right) \cdot r_t + \mathbf{A}\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot r_t + \mathbf{A}\boldsymbol{\beta}_t\left(\mathbf{y}_t^\star\right) \cdot \Gamma_t^\star \\
&= \mathbf{1}_n \cdot r_t + \frac{\mathrm{Cov}\left(\mathrm{d}\mathbf{P}_t \oslash \mathbf{P}_t, \mathrm{d}\pi_t/\pi_t\right)}{\mathrm{Var}\left(\mathrm{d}\pi_t/\pi_t\right)} \cdot \Gamma_t^\star \\
&= \mathbf{1}_n \cdot r_t + \boldsymbol{\beta}_t^{(A)} \cdot \left(\mu(\mathbf{y}_t^\star) - r_t\right).
\end{aligned}$$

This then results in a relationship analogous to that of the CAPM.

Appendix D Approximation of the Wasserstein distance for eigenvalues

If the support of the distributions f_{MP} and f_{KDE} is included in $[\phi_{\min} : \phi_{\max}]$, then

$$\int_{\mathbb{R}} \left(\sqrt{f_{\mathrm{KDE}}(\phi)} - \sqrt{f_{\mathrm{MP}}(\phi|v,q)}\right)^2 \mathrm{d}\phi = \int_{\phi_{\min}}^{\phi_{\max}} \left(\sqrt{f_{\mathrm{KDE}}(\phi)} - \sqrt{f_{\mathrm{MP}}(\phi|v,q)}\right)^2 \mathrm{d}\phi.$$

It follows that

$$\int_{\mathbb{R}} \left(\sqrt{f_{\mathrm{KDE}}(\phi)} - \sqrt{f_{\mathrm{MP}}(\phi|v,q)}\right)^2 \mathrm{d}\phi \approx \frac{1}{n_x} \sum_{i=1}^{n_x} \left(\sqrt{f_{\mathrm{KDE}}\left(\phi_{\min} + \frac{\phi_{\max} - \phi_{\min}}{n_\delta} i\right)} - \sqrt{f_{\mathrm{MP}}(\delta_x i|v,q)}\right)^2.$$

Appendix E Normal log-normal mixture

E.1 First four moments a normal log-normal mixture of the form $W\mathrm{e}^{s(B-s)}$

Using the results of [74], if $X = W\mathrm{e}^{sB}$ with $(W,B)^\top \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$: $\mathbb{E}[Z] = 0, \mathbb{E}\left[Z^2\right] = \mathrm{e}^{2s^2}, \mathbb{E}\left[Z^3\right] = 0$, and $\mathbb{E}\left[Z^4\right] = 3\mathrm{e}^{8s^2}$. Therefore, setting $X = W\mathrm{e}^{s(B-s)}$, we have:

$$\begin{aligned}
\mathbb{E}[Z] &= \mathrm{e}^{-s^2}\mathbb{E}[X] = 0,\\
\mathbb{E}\left[Z^2\right] &= \mathrm{e}^{-2s^2}\mathbb{E}\left[X^2\right] = 1,\\
\mathbb{E}\left[Z^3\right] &= \mathrm{e}^{-3s^2}\mathbb{E}\left[X^3\right] = 0,\\
\mathbb{E}\left[Z^4\right] &= \mathrm{e}^{-4s^2}\mathbb{E}\left[X^4\right] = 3\mathrm{e}^{4s^2}.
\end{aligned}$$

E.2 First four moments of an i.i.d. sample from a normal log-normal mixture

We consider the random variable Z defined by:

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

where $Y_1, ..., Y_n$ is an i.i.d sample such that $\mathbb{E}\left[Y\right] = \mathbb{E}\left[Y^3\right] = 0$, $\mathbb{E}\left[Y^2\right] = 1$, and $\mathbb{E}\left[Y^4\right] = c_4$. Using the multinomial theorem, we can express the moments of Z of order q as follows:

$$\mathbb{E}\left[Z^q\right] = n^{\frac{-q}{2}} \cdot \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^q\right] = n^{\frac{-q}{2}} \cdot \sum_{k_1+...+k_n=q} \mathbb{E}\left[\binom{q}{k_1, ..., k_n} Y_1^{k_1} Y_2^{k_2} ... Y_n^{k_n}\right]$$

From this expression, it is easy to verify that $\mathbb{E}[Z] = \mathbb{E}\left[Z^3\right] = 0$ and $\mathbb{E}\left[Z^2\right] = 1$. To determine the 4th moment of Z, we can start by noting that only two types of combinations, such that $k_1 + ... + k_n = 4$, are associated with a non-zero expectation due to $\mathbb{E}[Y] = 0$ and $\mathbb{E}\left[Y^3\right] = 0$. The first combination implies that one of the terms k_j is equal to 4, while the remaining $n-1$ terms are equal to 0. The second combination

involves two of the terms k_j being equal to 2, with the rest being equal to 0. Therefore:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right)^4\right] &= \frac{1}{n^2}\left(\frac{4!}{4!(0!)^{n-1}}\cdot\binom{n}{1}\cdot\mathbb{E}\left[Y^4\right] + \frac{4!}{(2!)^2(0!)^{n-2}}\cdot\binom{n}{2}\cdot\mathbb{E}\left[Y^2\right]\right) \\ &= \frac{1}{n^2}\left(n\cdot c_4 + 6\cdot\frac{n(n-1)}{2}\right) \\ &= \frac{c_4+3(n-1)}{n} \end{aligned} \tag{23}$$

Consequently, if $Y_i = W_i \mathrm{e}^{s(B_i - s)}$ where W_i and B_i are i.i.d. standard normal realizations, using the expression for the third moment of Y_i calculated in section E.1, we have:

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right)^4\right] = \frac{3e^{4s^2}+3(n-1)}{n}.$$

E.3 Estimation of the vector S using the method of moments

We aim to estimate the vector $\mathbf{S}$ on which $\mathbf{X}$ depends. To begin with, we assume that the time step between observations is constant, and that the ratio of the sample time step to the simulation time step is equal to n, a strictly positive natural number:

$$\frac{\frac{1}{N-rw-1}\sum_{u=rw+1}^{N}(t_u - t_{u-1})}{\Delta t} = \frac{\bar{\Delta}_u}{\Delta t} = n, \quad \text{with } n \in \mathbb{N}^*.$$

In this context, and assuming $\hat{\boldsymbol{\mu}}_u, \hat{\mathcal{S}}_u, \hat{\boldsymbol{\mathcal{V}}}_u$ are constant between t_{u-1} and t_u, the equation for the variation of elementary factors over this period is given by:

$$\Delta\hat{\mathbf{F}}_u = \sum_{i=1}^{n} \hat{\boldsymbol{\mu}}_u \cdot \Delta t + \hat{\mathcal{S}}_u \cdot \hat{\boldsymbol{\mathcal{V}}}_u \cdot \mathbf{X}_i \cdot \boldsymbol{W}_i \cdot \sqrt{\Delta t}$$

We then introduce the process $\hat{Z}_j$, such that:

$$\hat{Z}_{j,u} = \frac{\left(\Delta\mathbf{F}_u - \hat{\boldsymbol{\mu}}_u \cdot \Delta_u\right)_j}{(\mathcal{S}_u \cdot \boldsymbol{\mathcal{V}}_u \cdot \sqrt{\Delta_u})_j} = \frac{\sum_{i=1}^{n} \mathbf{X}_i \cdot \boldsymbol{W}_i \cdot \sqrt{\Delta t}}{\sqrt{\Delta_u}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\exp\circ\left(\mathbf{S}\odot(\boldsymbol{B}_i - \mathbf{S})\right)\odot\boldsymbol{W}_i\right)_j.$$

Therefore $\hat{Z}_{j,rw+1}, ..., \hat{Z}_{j,N}$ forms an i.i.d. sample drawn from an NLN mixture distribution of the form considered in section E.1. To estimate $(\mathbf{S})_j$, we use a method-of-moments approach by equating $\hat{c}_4$ as the empirical estimator of $\mathbb{E}\left[\hat{Z}_j^4\right]$ associated with the sample $\{\hat{Z}_{j,u}\}_{rw+1\geq u\geq N}$ with its theoretical counterpart

obtained in section E.2. This gives us:

$$\begin{aligned}
\hat{c}_4 &= \frac{3e^{4(\mathbf{S})_j^2} + 3(n-1)}{n} \\
\frac{n\hat{c}_4}{3} &= e^{4(\mathbf{S})_j^2} + n - 1 \\
e^{4(\mathbf{S})_j^2} &= n\left(\frac{\hat{c}_4}{3} - 1\right) + 1 \\
(\mathbf{S})_j &= 0.5\sqrt{\log\left(n\left(\frac{\hat{c}_4}{3} - 1\right) + 1\right)}
\end{aligned}$$

However, in practice, $\hat{c}_4$ may be less than 3, resulting in the absence of a real solution for $(\mathbf{S})_j$. Additionally, the choice of Δt may lead to the ratio of the sample time step to the simulation time step not being an integer. Therefore, to address these issues, it is appropriate to use the following estimator for $(\mathbf{S})_j$:

$$(\hat{\mathbf{S}})_j = 0.5\sqrt{\log\left(\frac{\frac{1}{N-rw-1}\sum_{u=rw+1}^{N}(t_u - t_{u-1})}{\Delta t}\left(\frac{\hat{c}_4}{3} - 1\right)_+ + 1\right)}.$$

Appendix F Estimation of the market sensitivity operator by maximum likelihood

The solution to this optimization program is reached when the partial derivatives of 21 with respect to $\mathcal{S}_{t_u}$ are equal to 0. These partial derivatives are given by:

$$\frac{\partial \mathcal{L}\left(\widehat{\mathbf{\Delta}}_{\mathbf{F}}; \mathbf{\Theta}_j\right)}{\partial \mathcal{S}_{t_u}} = \frac{1}{\mathcal{S}_{t_u}^3}\sum_{j=1}^{m}\left(\frac{\left((\Delta\mathbf{F}_{t_u})_j - \left(\hat{\boldsymbol{\mu}}_{t_u}\right)_j \cdot \Delta_{t_u}\right)^2}{\left(\boldsymbol{\mathcal{V}}_{t_u}\right)_j^2 \cdot \mathcal{S}_{t_u}} - \mathcal{S}_{t_u}^2\right).$$

We can then deduce the solution value of $\mathcal{S}_{t_u}$:

$$\begin{aligned}
0 &= \frac{1}{\mathcal{S}_{t_u}^3}\sum_{j=1}^{m}\left(\frac{\left((\Delta\mathbf{F}_{t_u})_j - \left(\hat{\boldsymbol{\mu}}_{t_u}\right)_j \cdot \Delta_{t_u}\right)^2}{\left(\boldsymbol{\mathcal{V}}_{t_u}\right)_j^2 \cdot \Delta_{t_u}} - \mathcal{S}_{t_u}^2\right) \\
\mathcal{S}_{t_u}^2 &= \sum_{j=1}^{m}\frac{\left((\Delta\mathbf{F}_{t_u})_j - \left(\hat{\boldsymbol{\mu}}_{t_u}\right)_j \cdot \Delta_{t_u}\right)^2}{\left(\boldsymbol{\mathcal{V}}_{t_u}\right)_j^2 \cdot \Delta_{t_u}} \\
\mathcal{S}_{t_u} &= \sqrt{\sum_{j=1}^{m}\frac{\left((\Delta\mathbf{F}_{t_u})_j - \left(\hat{\boldsymbol{\mu}}_{t_u}\right)_j \cdot \Delta_{t_u}\right)^2}{\left(\boldsymbol{\mathcal{V}}_{t_u}\right)_j^2 \cdot \Delta_{t_u}}}.
\end{aligned}$$

Appendix G Additional results of the numerical experiment

G.1 Fitted kernels

The table 7 shows the weights of the different exponential kernels in the composition of the kernels $K^{(\hat{\mu})}$ and $K^{(\hat{\sigma})}$ obtained after calibration by the algorithm 3[8]. Figure 18 shows the shape of these kernels and compares them to their respective approximations by a TSPL kernel.

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}
$(\boldsymbol{\delta})_k$	0	0.2%	39.7%	55.9%	0.3%	0	3.8%	0	0	0
$(\mathbf{w})_k$	23.2%	20.8%	28.7%	23.4%	0	3.2%	0.6%	0	0	0.1%

Table 7: Weight of the different exponential kernels in the composition of the kernels $K^{(\hat{\mu})}$ and $K^{(\hat{\sigma})}$.

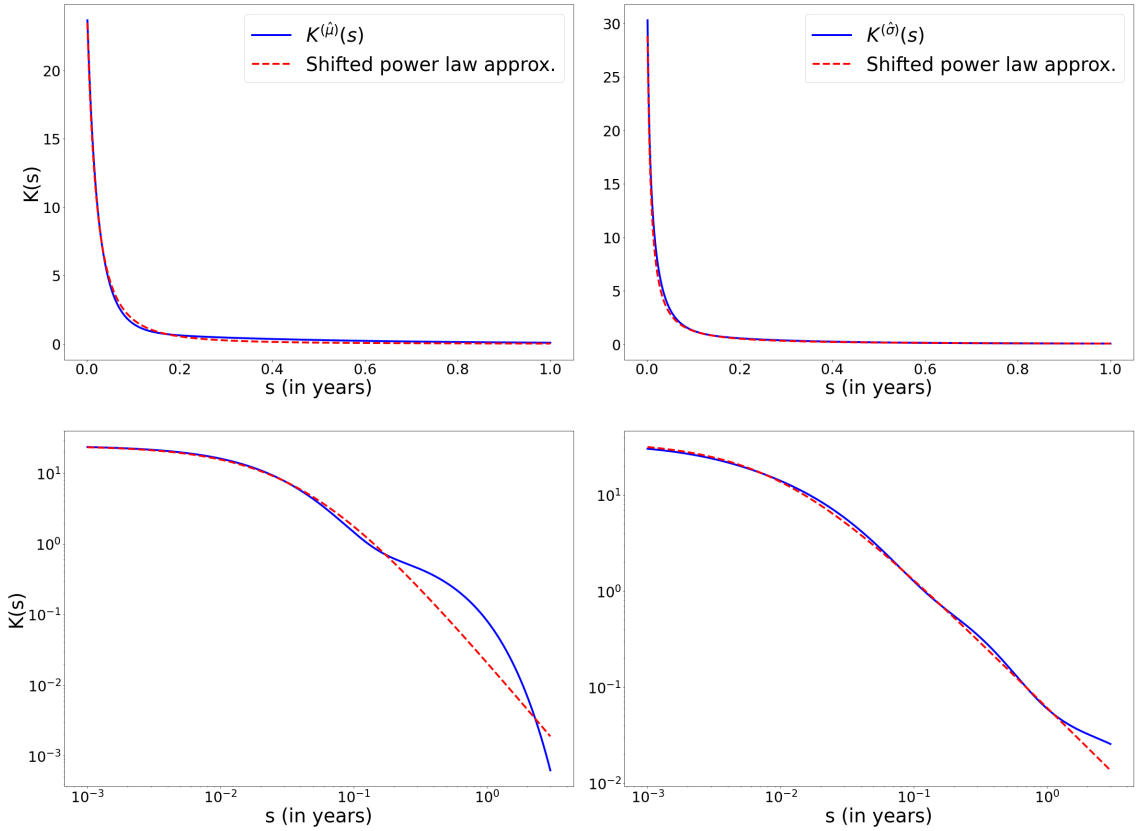


Figure 18: Kernels $K^{(\hat{\mu})}$ and $K^{(\hat{\sigma})}$ obtained as a result of the calibration performed by algorithm 2 and their respective approximations by a TSPL kernel.

[8] As exposed in section 3.1.1.2, the values of $\{\tau_k\}_{k=1}^{10}$ are defined by:

$$\tau_k = \exp\left(\log(1/365) + \frac{\log(5) - \log(1/365)}{9}(k-1)\right).$$

G.2 The market factor

The first common elementary factor obtained from the algorithm output 2 is hypothesized to be the market factor. Figure 19 allows for evaluating the coherence of this hypothesis by comparing the data related to this factor with the data related to the S&P500 index, which is a good proxy for the market portfolio. In addition, the path-dependent component estimated for this market factor, obtained after calibrating the market generator form algorithm 3, i.e. $(\boldsymbol{\mathcal{V}})_1$, is compared with the VIX over the same period.

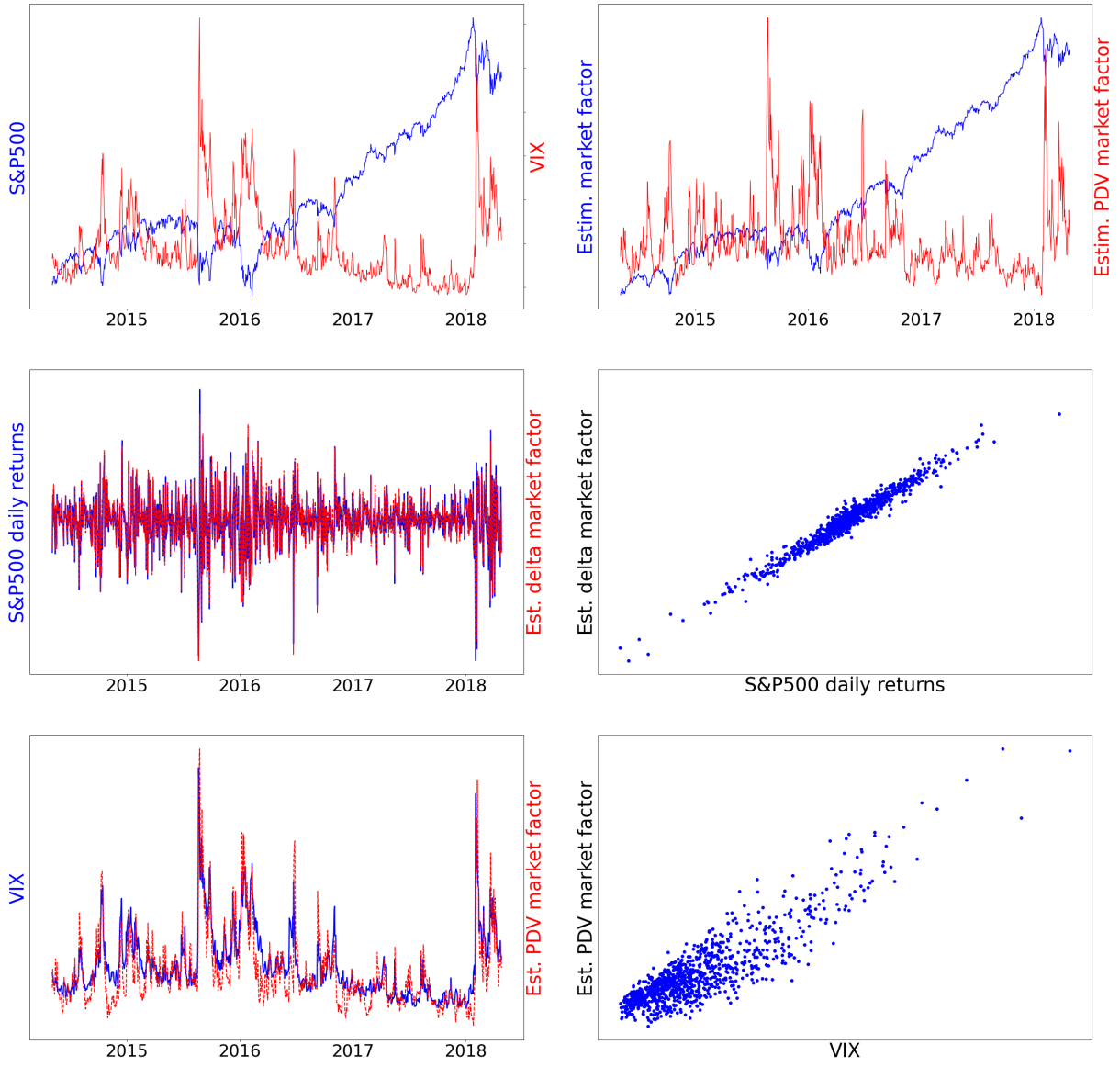


Figure 19: Comparison between the market factor estimated by the algorithm 2 and the S&P500 index, as well as between $(\boldsymbol{\mathcal{V}})_1$ and the VIX.

G.3 The market sensitivity operator

The figure 20 allows us to compare the estimated trajectory of the market sensitivity operator $\mathcal{S}$ obtained from the algorithm 3 with the trajectory simulated by the model over the same period. The figure 21 allows us to compare the empirical distribution and its autocorrelation of this market sensitivity operator with those generated by the model. Overall, the ARMA(1,1) model used to simulate the trajectory of $\log(\mathcal{S})$ reproduces the characteristics of the time series of this operator quite well. However, it can be noted that the empirical autocorrelation of $\mathcal{S}$ tends to become more negative compared to the simulated data. Additionally, the estimated empirical value of the market sensitivity operator appears to exhibit a seasonal component. Specifically, it seems to reach its lowest values during the latter half of December, then rise again in the first days of January. Consequently, taking this potential seasonality into account could improve the modeling quality of the proposed model.

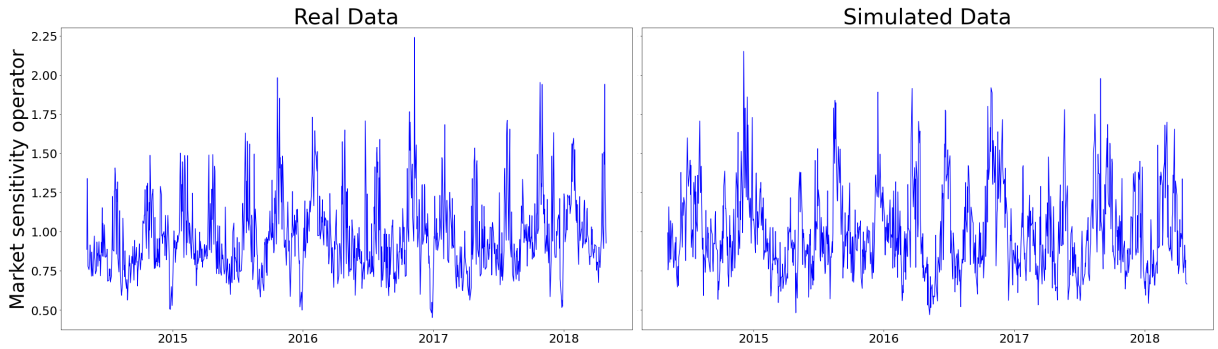


Figure 20: On the left, the estimated trajectory of the market sensitivity operator; on the right, a trajectory simulated from the model fitted on this path.

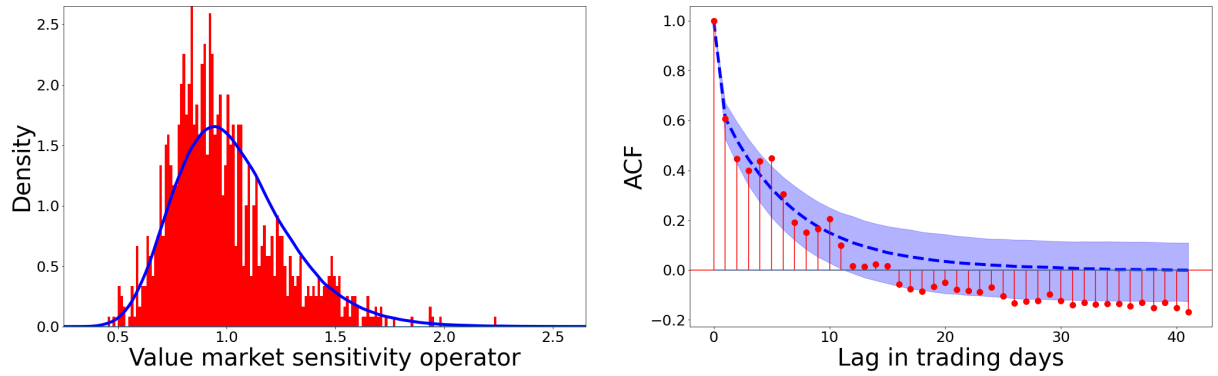


Figure 21: Distribution and autocorrelation of the market sensitivity operator: real vs. simulated data.

G.4 Some distributions of individual stock returns

Figures 22, 23, and 24 illustrate and compare, through three examples, the stock return distributions generated by the model with empirical data for daily, weekly, and monthly horizons.

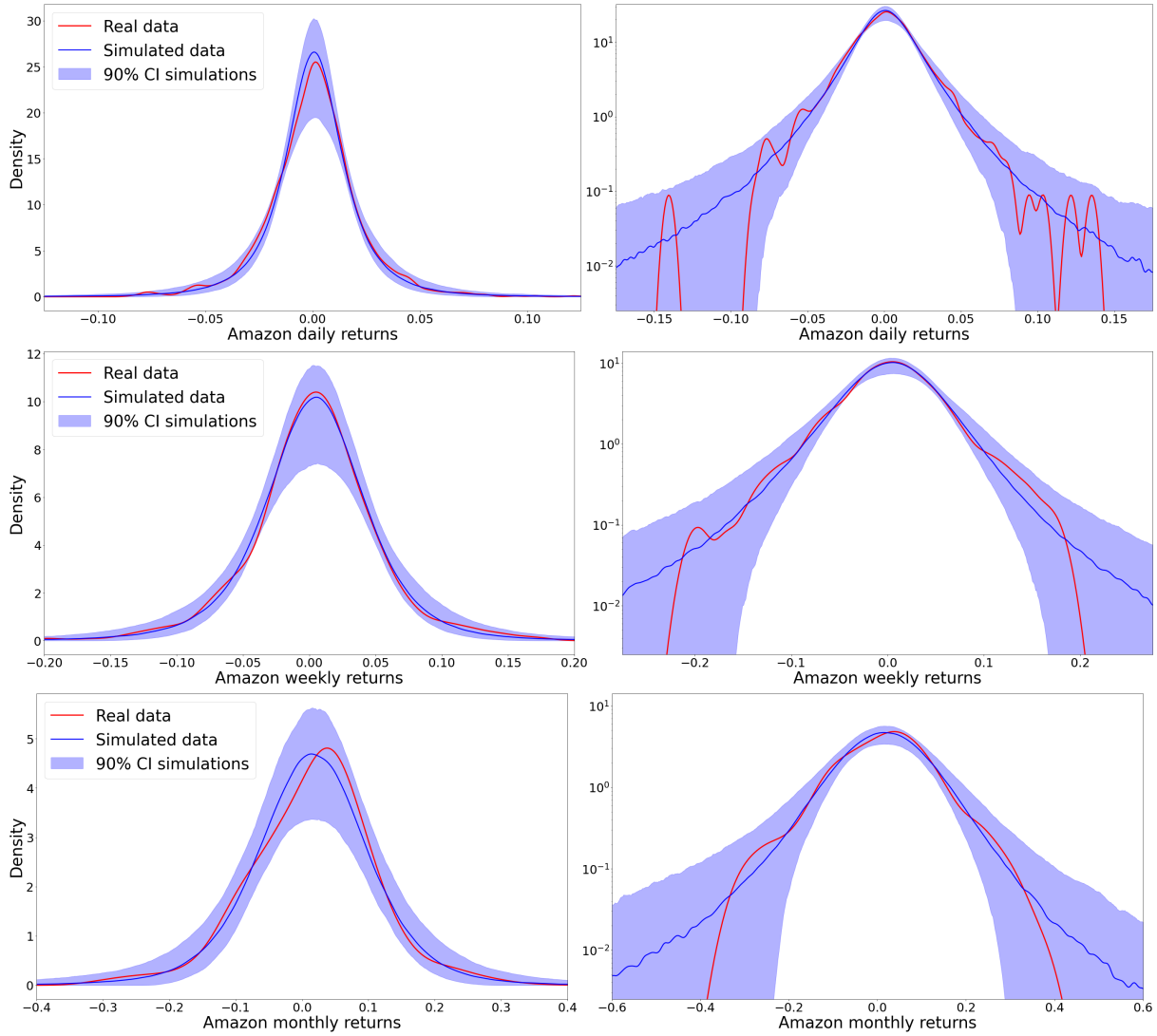


Figure 22: Comparison of the distribution of daily returns of Amazon stock between real data and data simulated using the FPDM generator. The KDE distribution of simulated data is estimated using the entire daily data set from the 1000 conducted simulations. The confidence interval is obtained by estimating the KDE distribution of daily returns for each simulation and then calculating the 5th and 95th percentiles of densities for each considered return level.

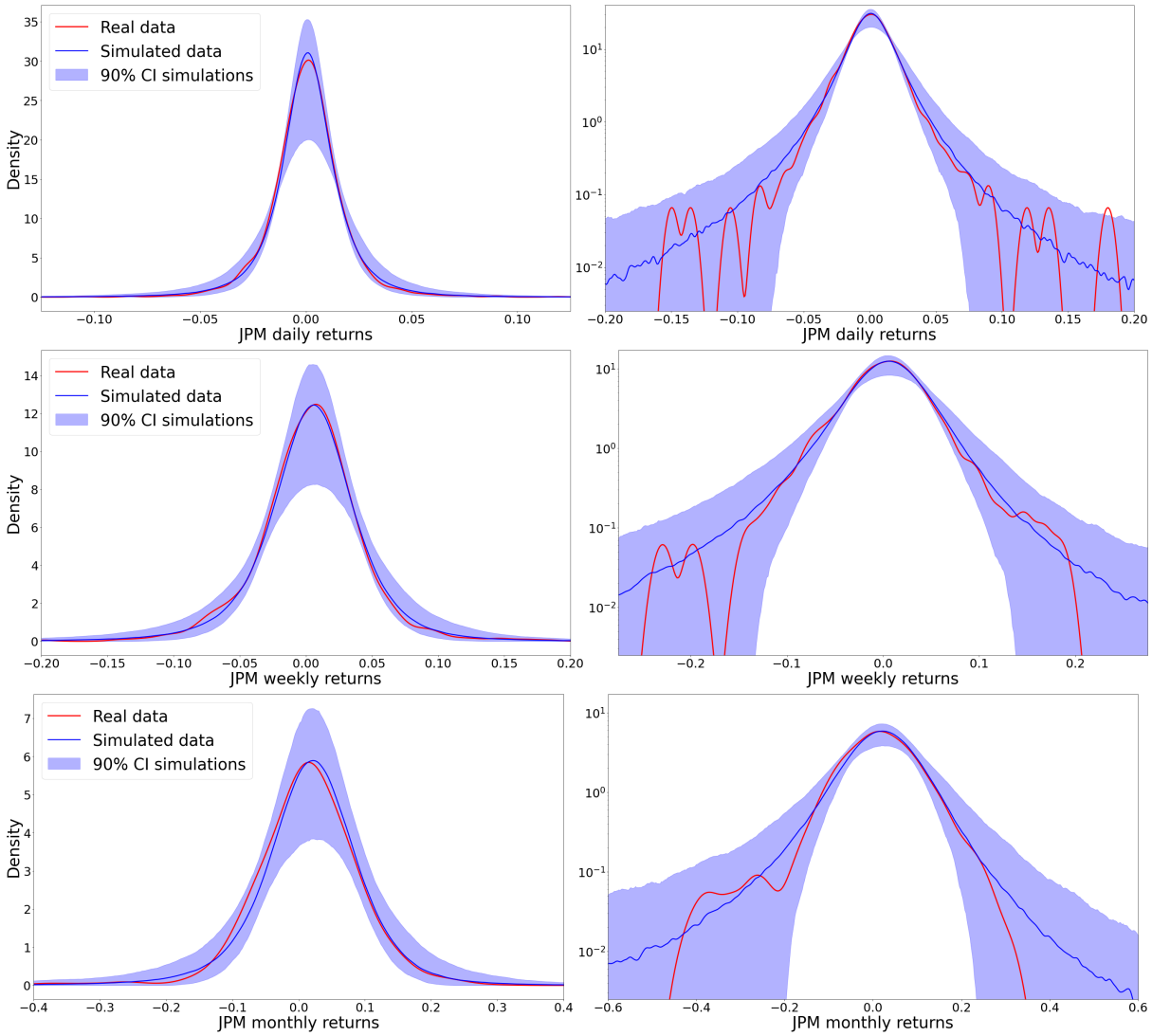


Figure 23: Comparison of the distribution of daily returns of JPMorgan Chase & Co. stock between real data and data simulated using the FPDM generator. The KDE distribution of simulated data is estimated using the entire daily data set from the 1000 conducted simulations. The confidence interval is obtained by estimating the KDE distribution of daily returns for each simulation and then calculating the 5th and 95th percentiles of densities for each considered return level.

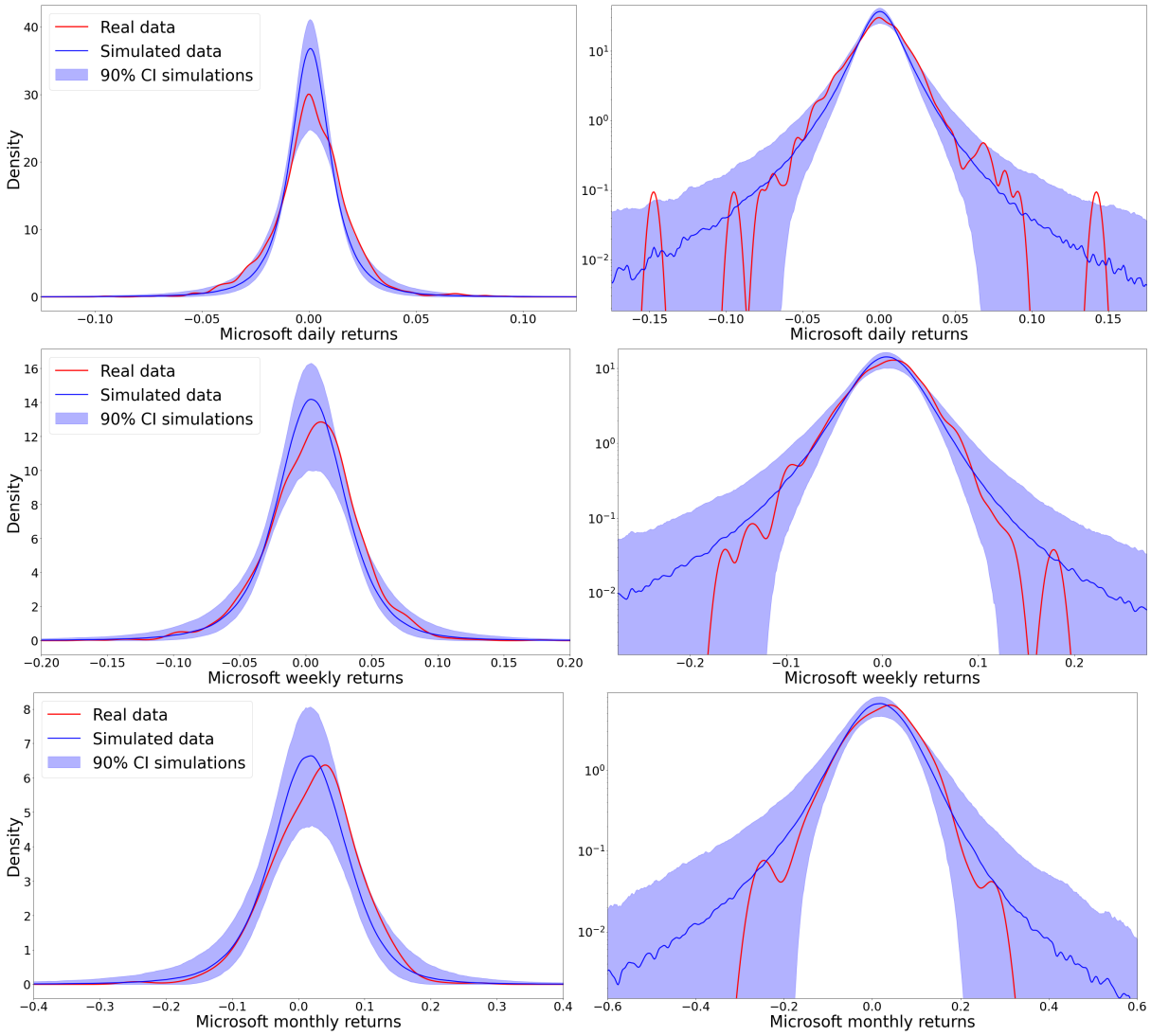
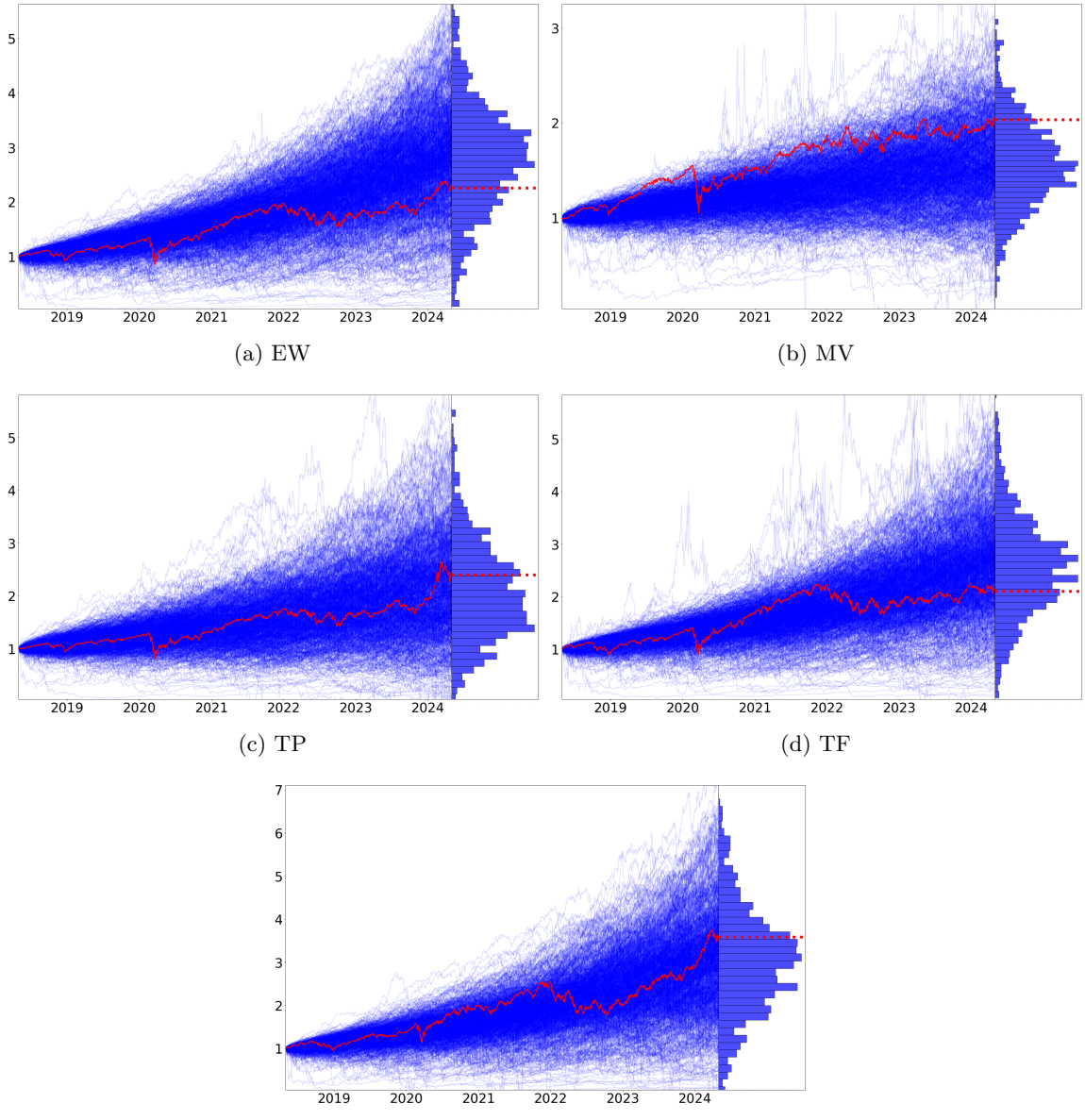


Figure 24: Comparison of the distribution of daily returns of Microsft stock between real data and data simulated using the FPDM generator. The KDE distribution of simulated data is estimated using the entire daily data set from the 1000 conducted simulations. The confidence interval is obtained by estimating the KDE distribution of daily returns for each simulation and then calculating the 5th and 95th percentiles of densities for each considered return level.

Appendix H Cumulative returns of the considered investment strategies

The following graphs compare the dynamics of different strategies between real data and data simulated from the FPDM generator.



(a) EW

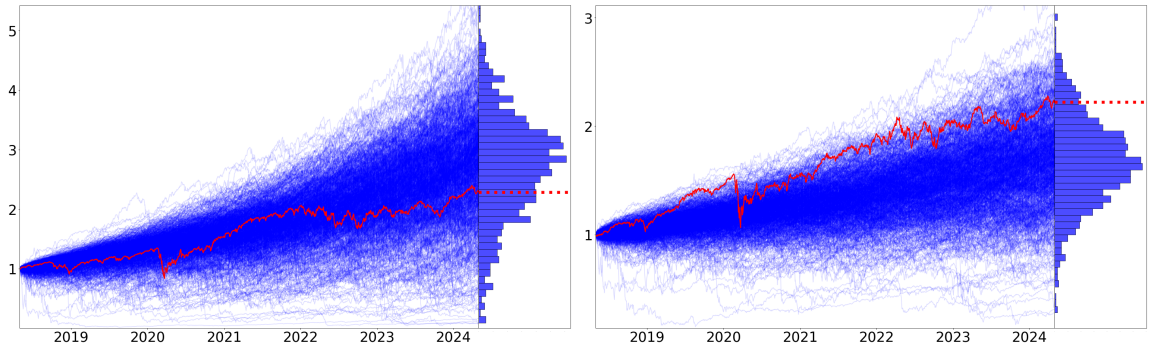
(b) MV

(c) TP

(d) TF

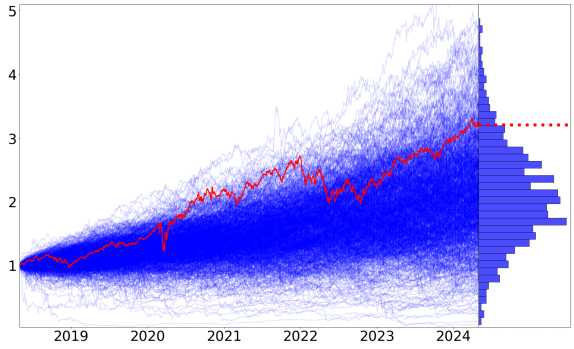
(e) RV

Figure 25: Cumulative returns of the different considered strategies with the buy-and-hold rebalancing approach: real data vs simulated data. In each plot, the red line corresponds to the cumulative returns of the strategy computed from the real data, while the blue lines represent those computed from the 1000 simulations considered in section 4.3.1.

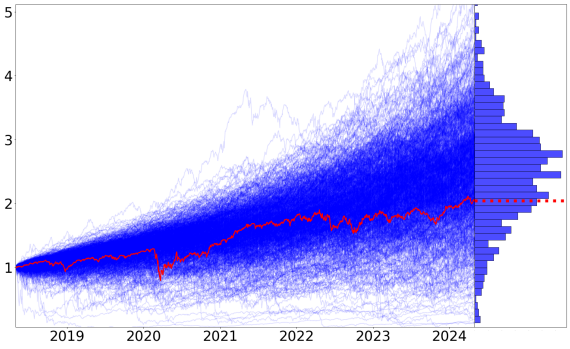


(a) EW

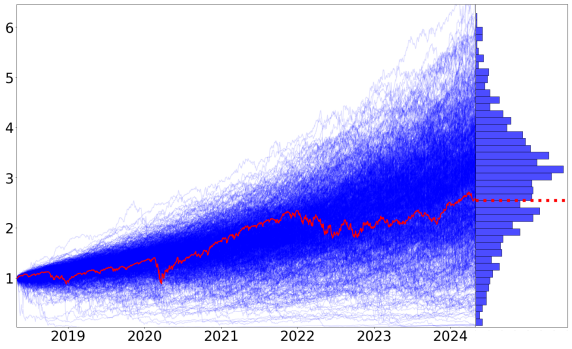
(b) MV



(c) TP

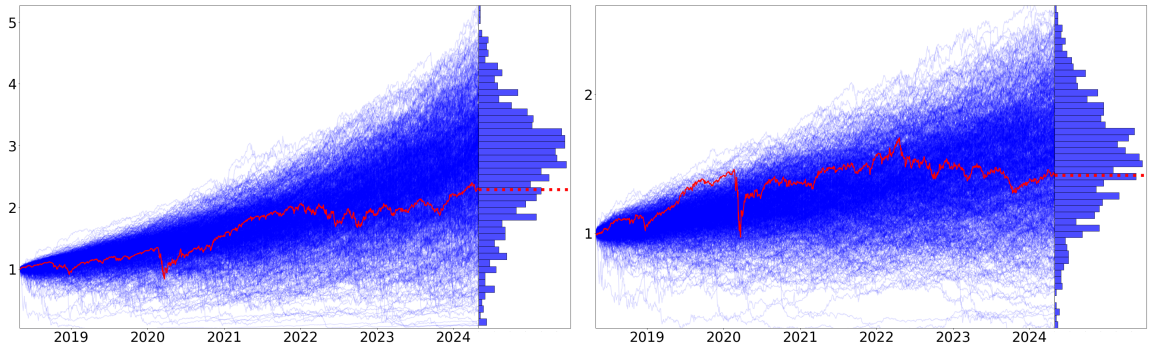


(d) TF



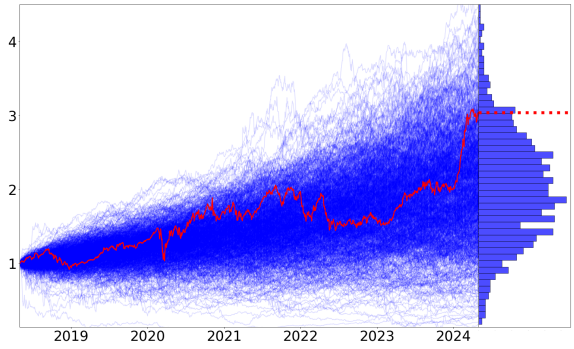
(e) RV

Figure 26: Cumulative returns of the different considered strategies with the constant-weighted rebalancing approach: real data vs simulated data. In each plot, the red line corresponds to the cumulative returns of the strategy computed from the real data, while the blue lines represent those computed from the 1000 simulations considered in section 4.3.1.

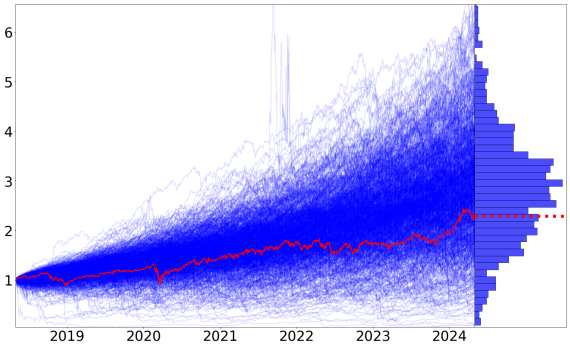


(a) EW

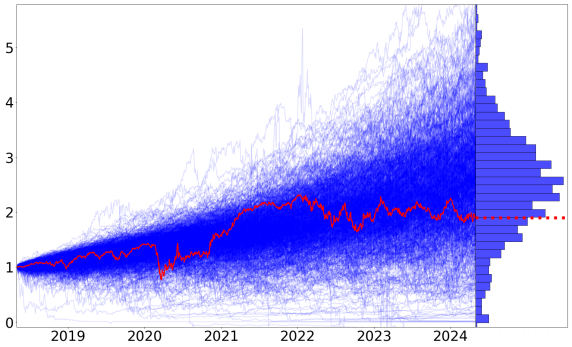
(b) MV



(c) TP



(d) TF



(e) RV

Figure 27: Cumulative returns of the different considered strategies with the dynamic rebalancing approach: real data vs simulated data. In each plot, the red line corresponds to the cumulative returns of the strategy computed from the real data, while the blue lines represent those computed from the 1000 simulations considered in section 4.3.1.