Factor Sufficiency in Asset Pricing: An Application for the Brazilian Market

Rafaela Dezidério
FEARP-USP

Márcio Laurini
FEARP-USP

Abstract

The asset pricing model derived from the Fama-French approach is extensively used in asset risk premium estimation procedures. Even including a considerable number of factors, it is still possible that omitted factors affect the estimation of this model. In this work, we compare estimators robust to the presence of omitted factors in estimating the risk premium in the Brazilian market. Initially, we applied the Mean Group and Common Correlated Effects panel data estimators to detect the presence of omitted factors. We then compare the results with those obtained by the estimator proposed by Giglio and Xiu (2021), which uses a principal components approach to correct the estimation in the case of omission of latent factors. We conclude that there is evidence of omitted factors and the best estimator is the Common Correlated Effects estimator.
Keywords: Robust Estimation, Risk Premia, Asset Pricing, Misspecification.

1 Introduction

In 1993 Fama and French proposed the three-factor pricing model [11]. In that paper they identified three systematic risk factors in stock returns, extending the usual estimation based on the CAPM estimation to a multi-factor structure constructing characteristics based risk factors. They built portfolios empirically to reproduce the market portfolio, size and book-to-market systematic risks. However, evidence from studies by Novy-Marx [23] and Titman, Wei, and Xie [29], show that the three-factor model can be insufficient, as they lose much of

the variation in average returns related to profitability and investment. With that in mind, in 2015, Fama and French added the profitability and investment factors [13] to the three-factor model, the so-called Five-Factor Model. They conclude that the Five Factor model explains 71% to 94% of the cross-section variance in expected returns for size, book-to-market, profitability and investment.

In 2017, Fama and French [14] tested the five-factor model for 4 regions - North America, Europe, Japan and Asia Pacific - and despite the global model not having a fully satisfactory result, the local models, that is, constructed with local data from each region, explained most of the variance in returns, as expected.

There is a relevant literature on new factors for asset pricing, and the large number of possible factors became known as the zoo factor, indicating that the estimation of traditional models of three, four and five factors derived from the Fama-French approach may be subject to problems caused by the absence of relevant risk factors in the estimation, leading to inconsistent estimators of the risk premium of the factors included in the model. One of the first to mention the zoo factor was Cochrane in 2011 [7]. Harvey, Liu and Campbell in 2015 [16], McLean and Pontiff in 2016 [21] and more recently Hou, Xue, and Zhang in 2017 [17] talk about the zoo factor and how these factors can influence the risk pricing procedures.

Recognizing the importance of omitted factors in risk premium estimation, Giglio and Xiu [15] propose a three-step method for estimating the risk premium of an observable factor that is valid even in the presence of omitted risk factors in the model, in addition to also controlling for a possible measurement error in the observable factors and also detecting whether if a factor is spurious or "useless", in the sense that it does not influence the estimation.

We study two ways of using the five-factor model to price stocks with Brazilian data, assuming that the relevance of the factors is constant over time. First, we focused our study on the possible variable omission, and thus the existence of a possible bias in the estimation. For this, we chose two estimators for panel data to estimate the risk premium of the factors: the Mean Group (MG) [24] and the Correlated Common Effects estimator (CCE) [25]. The MG estimator is defined as an average of OLS estimators, while the CCE estimator is an extension of the MG estimator that assumes an unobserved common factor structure for the errors. If there are factors omitted in the Fama French five-factor model, the CCE

estimator should capture them, and therefore, their coefficients would be different from the estimated coefficients for the MG estimator. The second way used to study the relevance of the five factors was to test whether they are sufficient to correctly price the assets, that is, whether these factors are able to estimate an approximately correct price for the set of assets in question. For this, we estimate the risk premium using the Giglio and Xiu [15] method, which theoretically corrects the estimation for possible omission of variables and the presence of measurement error, and we use this estimate to predict returns. Finally, we compare which model best fits the observed returns, the Fama-French five-factor model or the predictions made with the risk premium obtained by the Giglio and Xiu [15] method.

The results indicate that there is a strong indication that the coefficients of the CCE estimator are statistically different from the coefficients of the MG estimator and, therefore, there is the possibility of variable (factor) omission. In the second stage, we observed that, despite the estimator proposed by Giglio and Xiu supposedly correcting the estimation for omission of variables, the CCE estimator presents the best results in the asset pricing procedures.

This work has the following structure: the bibliographic reference is presented in Section 2; the methodology is reviewed in Section 3. In Section 4 we will present the data used. Section 5 presents the main results obtained. Final conclusions are presented in Section 6.

2 Bibliographic References

Asset pricing has been extensively studied, and in recent years, several factors and factor models have emerged to better understand how certain characteristics influence the prices of assets. These factors include size (market value), book equity (book value of equity), book-to-market equity (the ratio of book value to market value), leverage, earnings/price, and dividend/price, among others. Many models combine these different factors to more accurately price assets.

Some of the most significant works in asset pricing, such as those by Sharpe, Lintner, Black, and others, were based on Markowitz's [19] groundbreaking portfolio selection problem, which he studied in 1952. Markowitz divided the process of selecting an optimal portfolio into

two stages, with the second stage focused on the Mean-Variance rule and the construction of efficient mean-variance combinations. In 1959, he expanded on this work by developing an analysis based on the maximization of expected utility, offering a solution to the portfolio selection problem.

Building on Markowitz's studies, Sharpe [28] proposed a market equilibrium theory of asset prices under risky conditions, concluding that there is a linear relationship between expected returns and the standard deviation of returns for efficient combinations of risky assets in equilibrium. He also found a consistent relationship between expected returns and their "systematic risk," which can be measured by market beta, a metric that assesses a stock's volatility relative to the market.

Similar to Sharpe's work, Lintner [18] and Black [5] also studied the relationship between average return and risk, with the central prediction portfolio being the efficient portfolio proposed by Markowitz. Like the Sharpe model, the Lintner and Black models conclude that expected returns are positive linear functions of market beta. They also found that market betas absorb the effect of leverage on prices and are sufficient to describe the cross-section of expected returns.

In 1988, Fama and French [9] studied the relationship between dividend yields and expected returns on stocks, finding that dividend yields explain less than 5% of the variances of monthly or quarterly returns but generally explain more than 25% of the variance of returns over two to four years. This indicates that dividend yields have a greater influence on long-term returns.

In 1992, Fama and French [10] published another paper on expected stock returns, evaluating the relationship between expected returns and market beta, size (market value measured by the share price times the shares in circulation), leverage, book-to-market equity (BE/ME), and earning/price (E/P). Contrary to the predictions of Sharpe, Lintner, and Black, they did not find any reliable relationship between market betas and expected returns. Additionally, they concluded that leverage is well captured by book-to-market equity, and the combination of size and book-to-market equity absorbs the relationship between E/P and expected returns.

In their 1993 article "Common risk factors in the returns on stocks and bonds" [11],

Fama and French expanded on their previous research by using the time series regression approach of Black, Jensen, and Scholes (1972) [6] to construct two risk factors related to size and BE/ME for stocks, and two risk factors related to the term structure for bonds. The factors related to size and BE/ME are known as SMB and HML, respectively. To build these factors, they sorted the stocks by size (*Big and Small*) and BE/ME (*Low, Medium, and High*). This classification by BE/ME is based on dividing the stock population into three groups, with the lower 30% classified as *Low*, the middle 40% as *Medium*, and the upper 30% as *High*. From this classification, six portfolios are created based on the intersections between the size and BE/ME classifications: *Small/Low (S/L)*, *Small/Medium (S/M)*, *Small/High (S/H)*, *Big/Low (B/L)*, *Big/Medium (B/M)*, and *Big/High (B/H)*. These six portfolios provide returns on the large (B) and small (S) size portfolios.

$$R_B = \frac{1}{3}(R_{B/l} + R_{B/m} + R_{B/h})$$
$$R_S = \frac{1}{3}(R_{S/l} + R_{S/m} + R_{S/h})$$

From these two portfolio returns shown above, the returns of zero SMB net investment factors (*small minus big*, i.e. long position in low capitalization stocks and short position in high capitalization stocks) are constructed:

$$R_{SMB} = R_S - R_B$$

Similarly, the returns of the high (H) and low (L) portfolios are:

$$R_H = \frac{1}{2}(R_{S/h} + R_{B/h})$$
$$R_L = \frac{1}{2}(R_{S/l} + R_{B/l})$$

From these two portfolios, the zero HML net investment factor is created (*high minus*

low, that is, long position in high BE/ME and short position in low BE/ME):

$$R_{HML} = R_H - R_L.$$

They also created two portfolios to measure common risk related to unexpected changes in interest rates for bonds, called TERM and DEF. These five factors were found to explain well the common variation in bond and stock returns.

In a subsequent paper from 1995 [12], Fama and French attempted to find economic foundations for their empirical results and whether pricing is rational. They hypothesized that there must be common risk factors in returns associated with size and BE/ME, and that the size and BE/ME patterns in returns must be explained by earnings behavior. However, they could not find evidence that returns respond to the BE/ME factor in earnings. This paper left important questions open, such as the underlying economic state variables that produce variation in earnings and returns related to size and BE/ME.

Despite the popularity of the three-factor model, subsequent studies have shown that it is insufficient. In 2008, Titman, Wei, and Xie [29] studied the returns of stocks and capital investment based on the three-factor model, while in 2013, Novy-Marx [23] examined profitability. Both studies found that the three-factor model misses much of the variation in average returns related to profitability and investment.

After acknowledging the limitations of the three-factor model, Fama and French introduced the five-factor asset pricing model in 2015 [13]. This model extends the three-factor model by including the profitability and investment factors. Specifically, the RMW and CMA portfolios capture the differences in returns between firms with robust and weak profitability, and between conservative and aggressive companies, respectively. This new model has been shown to better explain average returns than the previous three-factor model. However, one challenge with using these models is the potential for omitted variable bias and measurement errors, which can lead to inconsistent estimates and less accurate asset pricing predictions.

To address these issues, researchers have explored new factors for asset pricing, leading to the emergence of a large number of potential factors, commonly referred to as the Factor

Zoo. Cochrane was among the first to highlight this phenomenon in 2011 [7], and since then, numerous articles have explored the impact of these additional factors on pricing models, including the work of Harvey, Liu, and Campbell [16], McLean and Pontiff [21], and Hou, Xue, and Zhang [17].

These works indicate the possible existence of a high number of possible risk factors, and although many are redundant or not significant for relevant periods of the sample, the omission of factors is a pervasive problem in estimating the risk premium. To address these challenges, Giglio and Xiu [15] propose a three-step methodology that utilizes a rotation invariance result for risk premium estimation in linear factor models, combined with Principal Component Analysis (PCA), to provide consistent risk premium estimates for any observed factor in the presence of omitted factors and misspecification in the model.

2.1 The consistent estimator for the risk premium in the presence of misspecification

We describe the fundamental elements of the three-step estimator proposed by Giglio and Xiu [15] in this section. The general idea of the method proposed by Giglio and Xiu [15] is to use a principal component estimation to recover the effects of the systematic factors omitted from the model, and thus carry out a consistent estimation of the risk premium associated with the factors included in the model.

To perform the first step, a consistent estimator of the number of factors is needed. The estimator used by them has the same idea as the factor estimators proposed by Bai and Ng [3] and Bai [4].

Bai and Ng [3] demonstrate that the penalty for overfitting should be a function of both N, the cross-section dimension, and T, the time dimension, to consistently estimate the number of factors. So, the usual AIC and BIC do not work well when both dimensions are large. So, considering the model

$$\underset{(T\times N)}{R} = \underset{(T\times p)}{v}\underset{(p\times N)}{\beta'} + \underset{(T\times N)}{e}$$

where $R = (\underline{R}_1, \ldots, \underline{R}_N)$, $\underline{R}_i = (R_{i1}, \ldots, R_{iT})$ for $i = 1, \ldots, N$, $v = (v_1, \ldots, v_T)$, $e = (\underline{e}_1, \ldots, \underline{e}_N)$, $\underline{e}_i = (e_{i1}, \ldots, e_{iT})$ for $i = 1, \ldots, N$, $\beta = (\beta_1, \ldots, \beta_N)$ and assume four hypotheses. The first hypothesis is related to the fourth moment of the factors, which converge to a definite positive matrix. The second hypothesis is that the norm of the vectors that constitute hypotheses regarding the properties of the factor loading matrix. third hypothesis refers to Cross-Section dependency, temporal dependency and heteroscedasticity. and the fourth and last hypothesis refers to the weak dependence between the factors and idiosyncratic errors.

Bai and Ng [3] also assume that the p factors are estimated by principal components, they show that the estimator

$$\hat{p} = arg \min_{0 \leq p \leq pmax} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(R_{it} - \bar{\beta^p}' \hat{v}^{p_t} + p\phi\left(N, T\right) \right)$$

$\bar{\beta^p}$ is constructed as $\sqrt{N}$ time the eigenvectors corresponding to the p largest eigenvalues of the matrix $N \times N$ $R'R$, $\bar{v^p} = R\bar{\beta^{p/N}}$ and $\hat{v}^p = \bar{v^p}(\bar{v^p}'\bar{v^{p/T}})^{1/2}$, has the following property:

$$\lim_{N,T \to \infty} Prob[\hat{p} = p] = 1$$

if (i) $\phi(N,T) \to 0$ and (ii) $\left(\min\left\{\sqrt{N}, \sqrt{T}\right\}\right)^2 \cdot \phi(N,T) \to \infty$ when $N, T \to \infty$.

Based on studies by Bai and Ng [3], Giglio and Xiu [15] assumed the following assumptions both for the development of the estimator in three steps and for building a consistent estimator for p:

I. f_t is a vector of asset pricing factors, where R_t denotes a vector of excess returns $N \times 1$ of test assets. The pricing model satisfies:

$$R_t = \beta\gamma + \beta v_t + u_t, \tag{1}$$

$$f_t = f + v_t, \tag{2}$$

$$E(v_t) = E(u_t) = 0, \ and \tag{3}$$

$$Cov(u_t, v_t) = 0, \tag{4}$$

where v_t is a vector $p \times 1$ of innovations of f_t, u_t is a vector $N \times 1$ of idiosyncratic components, β is a matrix $N \times 1$ of factor loadings, and γ is the risk premium $p \times 1$.

II. There is an observable vector $d \times 1$, g_t, of factors, which satisfies:

$$g_t = \delta + \eta v_t + z_t, \tag{5}$$

$$E(z_t) = 0, \ e \tag{6}$$

$$Cov(z_t, v_t) = 0, \tag{7}$$

where the g load in v, η is a matrix $d \times p$, δ is a constant $d \times 1$, and z_t is a measurement error vector $d \times 1$.

III. There is a positive constant K, such that for all N and T,

$$(i) \quad T^{-1} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left| E\left(N^{-1} \sum_{i=1}^{N} u_{it} u_{it'} \right) \right| \leq K, \qquad \max_{1 \leq t \leq T} E\left(N^{-1} \sum_{i=1}^{N} u_{it}^2 \right) \leq K.$$

$$(ii) \quad T^{-2} \sum_{s=1}^{T} \sum_{t=1}^{T} E\left(\sum_{j=1}^{N} (u_{js} u_{jt} - E(u_{js} u_{jt})) \right)^2 \leq KN.$$

IV. The factor innovations V obeys:

$$\|\bar{V}\|_{MAX} = O_p(T^{-1/2}),$$
$$\|T^{-1}VV' - \Sigma^v\|_{MAX} = O_p(T^{-1/2}),$$

where Σ^v is a positive definite matrix $p \times p$ and $0 < K_1 < \lambda_{min}(\Sigma^v) \leq \lambda_{max}(\Sigma^v) < K_2 < \infty$.

V. The factor loading matrix β satisfies

$$\left\|N^{-1}\beta'\beta - \Sigma^\beta\right\| = o_p(1), \text{ quando } N \to \infty,$$

Σ^β is a positive definite matrix $p \times p$ and $0 < K_1 < \lambda_{min}(\Sigma^\beta) \leq \lambda_{max}(\Sigma^\beta) < K_2 < \infty$.

VI. The factor loading matrix β and the idiosyncratic errors u_t satisfy the following moment conditions, for all $1 \leq j \leq p$ and for all N and T:

$$(i) \quad E\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\beta_{ij}u_{it}\right)^2 \leq KNT.$$
$$(ii) \quad E\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\beta_{ij}u_{it}\right)^2 \leq KNT.$$

The estimator proposed by Giglio and Xiu [15] is

$$\hat{p} = \arg\min_{1 \leq j \leq p_{max}} \left(N^{-1}T^{-1}\lambda_j(\bar{R}'\bar{R}) + j \times \phi(N,T)\right) - 1 \tag{8}$$

where p_{max} is some upper bound of p, $\phi(N,T)$ is a penalty function, and $\lambda_j(\bar{R}'\bar{R})$ is the jth largest eigenvalue of matrix $\bar{R}'\bar{R}$. They show that if $\phi(N,T) \to 0$ when $N,T \to \infty$, then we have $Prob(\hat{p} \geq p) \to 1$. And if, in addition, $\phi(N,T)/(N^{-1/2} + T^{-1/2}) \to \infty$, then $\hat{p} \xrightarrow{P} p$.

This estimator is used to build the estimator of factors and factor loadings in the first stage by conducting the PCA of the matrix $N^{-1}T^{-1}\bar{R}'\bar{R}$, defining the following estimators for the factors and for the factor loadings:

$$\hat{V} = T^{1/2}(\xi_1 : \xi_2 : \cdots : \xi_{\hat{p}})', \text{ e} \tag{9}$$

$$\hat{\beta} = T^{-1}\bar{R}\hat{V}' \tag{10}$$

where $\xi_1, \ldots, \xi_{\hat{p}}$ are the eigenvectors corresponding to the $\hat{p}$ largest eigenvalues of the PCA of the matrix $N^{-1}T^{-1}\bar{R}'\bar{R}$, and $(\xi_1 : \xi_2 : \cdots : \xi_{\hat{p}})$ is the horizontal concatenation of matrices, column by column, where the columns are equivalent to vectors ξ_i, for $i \in \{1, \ldots, \hat{p}\}$.

The second step is to perform a cross-sectional ordinary least squares (OLS) regression of the mean returns against the estimated factor loadings $\hat{\beta}$ to obtain the risk premium for the estimated latent factors

$$\hat{\gamma} = \left(\hat{\beta}'\hat{\beta}\right)^{-1}\hat{\beta}'\bar{R}. \tag{11}$$

The last step consists of performing a regression of g_t on the factors extracted by the PCA, $\hat{V}$, to obtain the $\hat{\eta}$ estimator and the corrected value of the observed factor:

$$\hat{\eta} = \bar{G}\hat{V}'\left(\hat{V}\hat{V}'\right)^{-1}, \tag{12}$$

$$\hat{G} = \hat{\eta}\hat{V} \tag{13}$$

where $\bar{G}$ is the mean of the matrix $G = (g_1, g_2, \ldots, g_T)$.

Finally, the g_t risk premium estimator is obtained by

$$\hat{\gamma}_g = \hat{\eta}\hat{\gamma} \tag{14}$$

$$= \bar{G}\hat{V}' \left(\hat{V}\hat{V}'\right)^{-1} \left(\hat{\beta}'\hat{\beta}\right)^{-1} \hat{\beta}'\bar{R} \tag{15}$$

3 Risk Premium Estimation Methodology

In this work we will assume that the evolution in the cross-section of assets and in time can be summarized through a panel structure, with the general specification:

$$R_{it} = \beta_i' d_t + e_{it}, \qquad i = 1, \ldots, N \tag{16}$$

and specifically we use the structure:

$$\beta_i = \begin{bmatrix} \alpha_i \\ \beta_{iM} \\ \beta_{iSMB} \\ \beta_{iHML} \\ \beta_{iIML} \\ \beta_{iWML} \end{bmatrix}, \qquad d_t = \begin{bmatrix} 1 \\ R_{Mt} - R_{ft} \\ SMB_t \\ HML_t \\ IML_t \\ WML_t \end{bmatrix} \tag{17}$$

- R_{Mt} is the market return in period t,
- R_{ft} is the risk free in period t,
- SMB_t is the factor related to size in period t.
- HML_t is the factor related to BE/ME in period t.
- IML_t is the factor related to liquidity in period t.
- WML_t is the factor related to past returns in period t.

This model is based on the *Fama-French five-factor model* [13]. What differs from the Fama-French model are the liquidity (IML) and past returns (WML) factors that replace the profitability (RMW) and investment (CMA) factors. This substitution of factors was necessary because our objective was to carry out the study with the factors available through NEFIN - Brazilian Center for Research in Financial Economics of the University of São Paulo, and the main source of risk factors used in the Brazilian financial market.

Our work studied this model in two stages. We first study whether these five factors are significant, and therefore useful to explain portfolio returns. So finally, we made predictions with the factors in the usual way and the predictions with the method of Giglio and Xiu [15], and compared which of the predictions gave us a better result.

3.1 Sufficiency of Factors

Our objective is to try to identify the possibility of omitting factors in the model (16), using a simple diagnosis comparing the estimation of a non-robust panel data model to the presence of omitted factors (Mean Group estimator - MG) with a robust estimator for panel data in the presence of latent factors, given by the Common Correlated Effects (CCE) estimator. Note that the use of a panel in risk premium estimation is a common procedure in factor risk premium estimation, and can be thought of as an alternative estimation method in relation to the Fama-Macbeth procedure [8], and the use of panel models for estimating multifactor models is discussed in Petersen (2009) [27].

The MG estimator is a simple average of the OLS estimators of each group, while the CCE estimator is an extension of the MG estimator assuming unobserved common correlated factors in the errors. So, the idea behind the CCE estimator is the same as the one we want to test. For this reason, we chose to compare the Mean Group with the CCE. In the next subsections we will detail these estimators in more detail.

3.1.1 Mean Group Estimator

To obtain the Mean Group estimator for the heterogeneous panel data model (16) we consider the following matrices:

$$D = \begin{bmatrix} d_1 & d_2 & \dots & d_T \end{bmatrix}' \tag{18}$$

$$R_{i.} = \begin{bmatrix} R_{i1} & R_{i2} & \dots & R_{iT} \end{bmatrix}' \tag{19}$$

The first step was to calculate the OLS estimators of each β_i, according to the equation below:

$$\hat{\beta}_i = (D'D)^{-1}D'R_{i.} \tag{20}$$

Finally, we obtain the MG estimator according to the equation below:

$$\hat{\beta}^F = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i \tag{21}$$

3.1.2 Correlated Common Effects Estimator

To calculate the correlated common effects estimator, we consider the heterogeneous panel data model (16) and assume that the error e_{it} has the following common factorial structure

$$e_{it} = \sum_{j=1}^{m}\gamma_{ij}f_{jt} + \varepsilon_{it} = \boldsymbol{\gamma}_i'\mathbf{f}_t + \varepsilon_{it} \tag{22}$$

where $\boldsymbol{f_t} = (f_{1t}, ..., f_{mt})'$ is a vector of unobserved common factors and $\boldsymbol{\gamma_i} = (\gamma_{i1}, ..., \gamma_{im})'$ is the factor loading vector. We assume that the number of factors, m, is fixed and $m << N$, where N is the number of assets.

So, substituting (22) into (16), our model has the following form:

$$R_{it} = \beta_i'd_t + \boldsymbol{\gamma}_i'\mathbf{f}_t + \varepsilon_{it} \tag{23}$$

The correlated common effects (CCE) estimator consists of approximating the linear combination of unobserved factors by means of the cross-section of the dependent and explanatory variables, and then calculating the regression for the augmented standard panel

with the means of the cross-section .

To calculate the averages we consider a non-stochastic vector of weights $\mathbf{w_t} = (w_{1t}, w_{2t}, ..., w_{Nt})'$, for $t \in \mathcal{T} \subset \mathbb{Z}$, where $\mathcal{T}$ is our time horizon. The vector $\mathbf{w_t}$ was chosen to satisfy the two hypotheses below:

$$|\mathbf{w}_t| = (\mathbf{w}_t'\mathbf{w}_t)^{\frac{1}{2}} = O\left(N^{-\frac{1}{2}}\right), \tag{24}$$

$$\frac{w_{jt}}{|\mathbf{w}_t|} = O\left(N^{-\frac{1}{2}}\right) \text{ uniformly in j} \in \mathbb{N}. \tag{25}$$

Thus, the averages were calculated as follows

$$\overline{R}_{wt} = \overline{\beta}_w' d_t + \overline{\boldsymbol{\gamma}}_w f_t + \overline{\varepsilon}_{wt} \tag{26}$$

where

$$\overline{R}_{wt} = \sum_{i=1}^{N} w_i R_{it},\ \overline{\beta}_w = \sum_{i=1}^{N} w_i \beta_i, \tag{27}$$

$$\overline{\boldsymbol{\gamma}}_w = \sum_{i=1}^{N} w_i \boldsymbol{\gamma}_i,\ \overline{\varepsilon}_{wt} = \sum_{i=1}^{N} w_i \varepsilon_{it} \tag{28}$$

And from the model's regression (26) we calculate f_t and $\hat{\beta}_i^{\,P}$.

3.1.3 Wald test

The idea of comparing the MG and CCE estimators is to identify the possible presence of omitted factors in the estimation, since the CCE estimator assumes a structure of unobserved common factors for the errors. By estimating the MG and the CCE we obtained their coefficients and the covariances of their parameters for the model (16). We chose to perform the Wald test, since it consists of evaluating the restrictions on the statistical parameters based on the weighted distance between the unconstrained estimate and its hypothetical value un-

der the null hypothesis.

The first test performed was a traditional Wald test assuming the model proposed by the MG estimator and, therefore, without common factors in the errors. We assume as a null hypothesis that the coefficients are equal to the coefficients estimated by the MG estimator, and we test whether the CCE estimator is equal.

Similarly, in the second test, we assume that the estimated model is the model proposed by the CCE estimator and therefore has common factors in the errors. Our null hypothesis was that the coefficients are equal to the coefficients estimated by the CCE estimator, and we tested whether the MG estimator is equal.

Finally, the last test boiled down to using the covariance estimates of the parameters of the two models in the Wald test. This test is similar to the F test that is done in analysis of variance (ANOVA).

3.2 Giglio and Xiu Method

Our model (16) only has observed factors. We applied the Giglio and Xiu method to calculate the risk premium for these five factors, controlling for the presence of possible omitted factors and measurement errors.

To apply this method we define $R_t = (R_{1t}, \ldots, R_{Nt})'$, and we assume equations 1 to 3.

We define the vector $g_t = (R_{Mt} - R_{ft}, SMB_t, HML_t, IML_t, WML_t)'$ (5×1). Note that $d_t = (1, g_t')'$.

Our objective is to estimate the risk premium of g_t corrected for the latent factors and use this risk premium to obtain the model parameters (16). For that we also assume equation 4.

We denote by R, V, G, U and Z the following matrices

$$\underset{(N\times T)}{R} = \begin{bmatrix} R_1 & R_2 & \dots & R_T \end{bmatrix}, \tag{29}$$

$$\underset{(p\times T)}{V} = \begin{bmatrix} v_1 & v_2 & \dots & v_T \end{bmatrix}, \tag{30}$$

$$\underset{(5\times T)}{G} = \begin{bmatrix} R_{M1} - R_{f1} & R_{M2} - R_{f2} & \dots & R_{MT} - R_{fT} \\ SMB_1 & SMB_2 & \dots & SMMB_T \\ HML_1 & HML_2 & \dots & HML_T \\ IML_1 & IML_2 & \dots & IML_T \\ WML_1 & WML_2 & \dots & WML_T \end{bmatrix}, \tag{31}$$

$$\underset{(N\times T)}{U} = \begin{bmatrix} e_1 & e_2 & \dots & e_T \end{bmatrix}, \tag{32}$$

$$\underset{(5\times T)}{Z} = \begin{bmatrix} z_1 & z_2 & \dots & z_T \end{bmatrix}. \tag{33}$$

And with these matrices we rewrite the model used by Giglio and Xiu as follows

$$R = \beta\gamma + \beta V + U \tag{34}$$

$$G = \xi + \eta V + Z \tag{35}$$

We denote by $(\overline{R}, \overline{V}, \overline{G}, \overline{U}, \overline{Z})$ the matrices of the means of the respective variables. And therefore, we have that the above equations become

$$\overline{R} = \beta\overline{V} + \overline{U}, \tag{36}$$

$$\overline{G} = \eta\overline{V} + \overline{Z}. \tag{37}$$

According to Bai and Ng [3], the number of factors estimated by the asymptotic principal component method is $\min\{N, T\}$. As we use principal components in future steps, we adopt $p_{max} = \min\{N, T\}$. We analyze two estimators $\hat{p}_j$, $j = 1, 2$:

$$\hat{p}_1 = \arg \min_{1 \leq j \leq p_{max}} \left(N^{-1}T^{-1}\lambda_j(\bar{R}'\bar{R}) + j \times \phi(N,T) \right) - 1 \tag{38}$$

$$\hat{p}_2 = \arg \min_{1 \leq j \leq p_{max}} \left(N^{-1}T^{-1}\lambda_j(\bar{R}'\bar{R}) + j \times \phi(N,T) \right) \tag{39}$$

.

The $\hat{p}_1$ estimator is the same estimator proposed by Giglio and Xiu, and they show that the penalty function can be sufficiently small when it is dominated by the large eigenvalues, so they add -1 to cover this case. While the $\hat{p}_2$ is based on the estimator proposed by Bai and NG.

For each $\hat{p}_i$, $i = 1, 2$, we test 4 different functions $\phi_k(N,T)$, $k = 1, 2, 3, 4$:

$$\phi_1 = \left(\log \left(\left(N^{-1/4} + T^{-1/4} \right)^{-1} \right) \right) \times \left(N^{-1/4} + T^{-1/4} \right) \tag{40}$$

$$\phi_2 = \left(\log \left(\frac{N \times T}{N + T} \right) \right) \times \left(\frac{N + T}{N \times T} \right) \tag{41}$$

$$\phi_3 = \left(\log \left(\min \left\{ N, T \right\}^2 \right) \right) \times \left(\frac{N + T}{N \times T} \right) \tag{42}$$

$$\phi_4 = \frac{\log \left(\min \left\{ N, T \right\}^2 \right)}{\min \left\{ N, T \right\}^2} \tag{43}$$

and we choose the estimator that obtained the best result. In all, we tested 8 estimators defined by the following equation:

$$\hat{p}_j^k = \begin{cases} \arg\min_{1 \leq l \leq p_{max}} \left((NT)^{-1}\lambda_l(\bar{R}'\bar{R}) + l \times \phi_k(N,T) \right) - 1 & \text{, se } j = 1 \\ \arg\min_{1 \leq l \leq p_{max}} \left((NT)^{-1}\lambda_j(\bar{R}'\bar{R}) + l \times \phi_k(N,T) \right) & \text{, se } j = 2 \end{cases} \tag{44}$$

We have $\phi_k(N,T) \to 0$ when $N, T \to \infty$, for $k \in \{1, 2, 3, 4\}$. However, only the function ϕ_1 has the following property: $\phi_1(N,T)/(N^{-1/2} + T^{-1/2}) \to \infty$, when $N, T \to \infty$.

Upon obtaining the estimate $\hat{p}$ of the number of factors, we perform the first step of the Giglio and Xiu method, calculating the factor estimator $\hat{V}$ and the factor loading estimator $\hat{\beta}$ was calculated as equations 9 and 10.

In the second stage of the method, we calculate, through an OLS on the average of the returns $\bar{R}$, the estimator of the risk premium of the latent factors $\hat{\gamma}$ according to 11.

Finally, with the last step we obtained the $\hat{\eta}$ and $\hat{G}$ estimators of the factor loadings of g in v and the corrected value of the factors observed after removing errors of measurement, respectively. The $\hat{\eta}$ estimator and the $\hat{G}$ estimator were obtained as 12 and 13.

Then, using the previous estimators to estimate the risk premium of g_t, which are the five observed factors, as 15.

3.3 Predictions

The last part of our work was the comparison of the usual predictions of the Fama-French model with the one corrected by the method of Giglio and Xiu.

We use the risk premium vector $\hat{\gamma}_g$ to recover the factor loadings of g_t thus obtaining an estimator $\hat{\beta}^G$. Then we apply this estimator to the model (16) to make forecasts of the returns of N assets compared to the forecasts $\hat{\beta}^F$ obtained by the usual regression of the model of (16).

4 Database

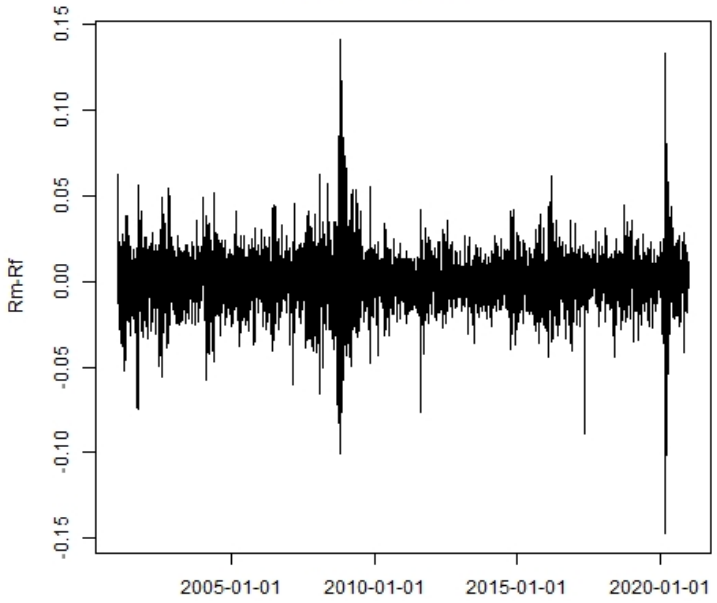
All data from this work were constructed by NEFIN from USP and its period is from January 2001 to December 2020, to data on a daily basis. Below is a description of the construction carried out by NEFIN of the factors and the 12 portfolios that we selected.

The one-year risk-free factor (R_f - risk free) was calculated from the 360-day DI Swap, deflated by expected inflation measured by the IPCA index (data available on the website of the Central Bank of Brazil).

The Market Factor ($R_M - R_f$) is the difference between the daily value-weighted return of the market portfolio and the daily risk-free rate, which is calculated from the 30-day DI-Swap. Figure 1 shows the Market factor returns.

The size factor SMB (*Small Minus Big*) is the return of a portfolio long on stocks with low market capitalization ("*Small*") and short stocks with high market capitalization ("*Big*"). Every January of the year t the shares are classified as eligible according to the market

Figure 1: Market Factor Returns

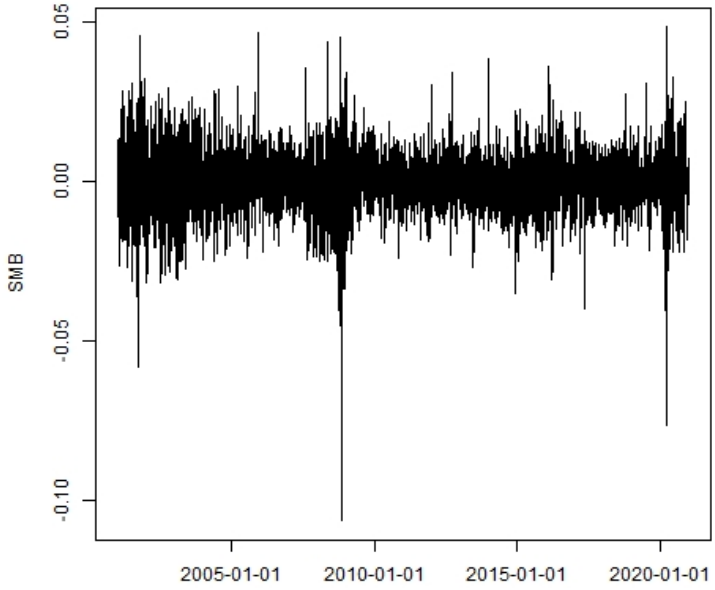


capitalization of December of the year $t-1$, and are separated into 3 quantiles (portfolios). Then, we calculate with equal weight the returns of the first portfolio ("*Small*") and the third portfolio ("*Big*"). The SMB factor is the return of the "*Small*" portfolio minus the return of the "*Big*" portfolio. Figure 2 shows the Size factor returns.

The factor related to BE/ME is the HML factor (*High Minus Low*). This return is the return of a portfolio long on stocks with a high book-to-market ratio ("*High*") and short on a low book-to-market ratio ("*Low*"). Every January of the year t, the shares are classified as eligible (increasingly) and divided into 3 quantiles (portfolios) according to the firm's book-to-market ratio in June of the year $t-1$. Then we calculate with equal weight the returns of the "*High*" portfolio minus the returns of the "*Low*" portfolio. Figure 2 presents the Book-to-Market factor returns.

The WML factor (*Winners Minus Losers*) is the return of a portfolio long on stocks with high past returns ("*Winners*") and short on low past returns ("*Losers* "). Every month t shares are classified as eligible (increasingly) and divided into 3 quantiles (portfolios) according to their cumulative returns between months $t-12$ and $t-2$. Then we calculate with equal weight the returns of the first portfolio ("*Losers*") and the third portfolio ("*Winners*"). The WML factor is the return of the "*Winners*" portfolio minus the return of the "*Losers*"

Figure 2: Size Factor Returns



portfolio. The returns of WML factor are shown in Figure 4.

The IML factor ("*Illiquid Minus Liquid*") is the return of a portfolio long on highly illiquid stocks ("*Illiquid*") and short on low illiquid ("*Liquid* "). Every t month, we sort eligible stocks (in ascending order) into 3 quantiles (portfolios) according to the moving average of illiquidity over the previous twelve months (stock illiquidity is calculated according to Acharya and Pedersen [1]) . Then we calculate with equal weight the returns of the first portfolio ("*Liquid*") and the third portfolio ("*Illiquid*"). The factor IML is the return on the "*Illiquid*" portfolio minus the return on the "*Liquid*" portfolio. Figure 5 presents the Book-to-Market factor returns.

The 12 portfolios are divided into four groups:

(i) 3 portfolios sorted by size.

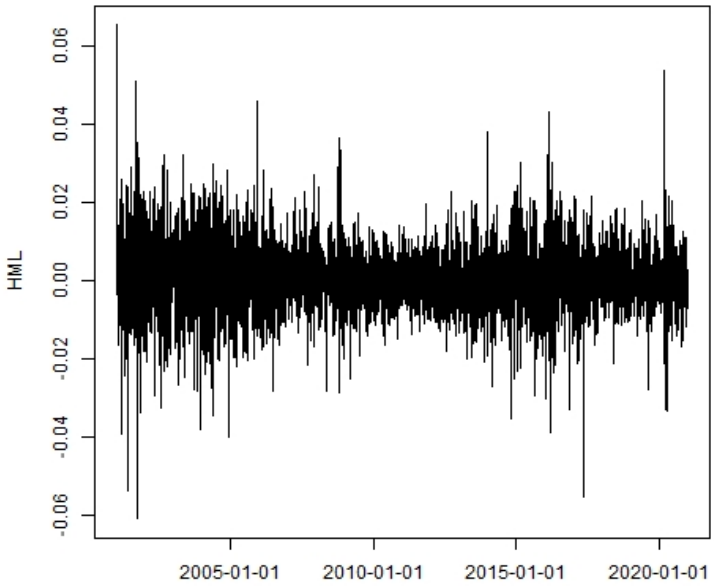
(ii) 3 portfolios classified by book-to-market.

(iii) 3 portfolios sorted by momentum.

(iv) 3 portfolios classified by illiquidity.

Portfolios sorted by size are obtained as follows: Every January of year t, eligible stocks are sorted in ascending order into tertiles according to their market capitalization in December

Figure 3: Book-to-Market Factor Returns



of year $t-1$. Then the portfolios are held for the year t.

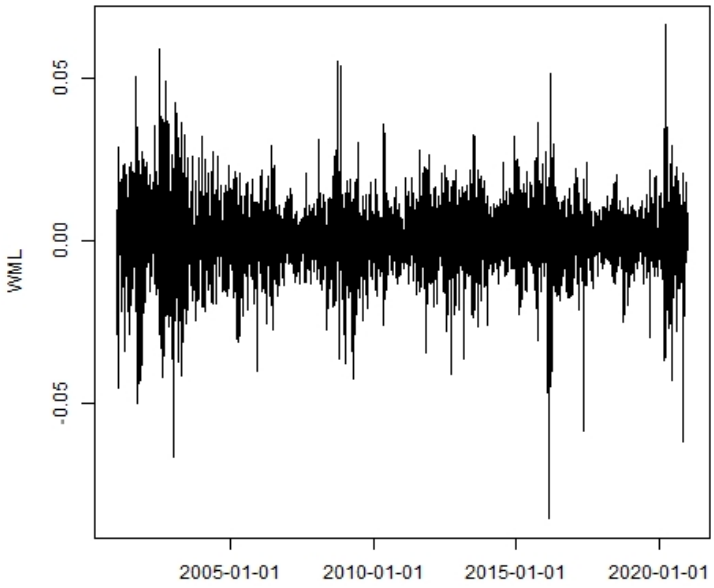
Portfolios sorted by book-to-market are obtained as follows: Every January of the year t, eligible stocks are sorted in ascending order in terciles, according to the ratio between book value and market value in June of the year $t-1$. Then the portfolios are held for the year t.

Momentum sorted portfolios are obtained as follows: Every month t, eligible stocks are sorted in ascending terciles according to their cumulative returns for month $t-12$ and month $t-2$. Then the portfolios are held for the month t.

The portfolios sorted by illiquidity are obtained as follows: every month t, eligible stocks are sorted in ascending tertiles according to the moving average of illiquidity of the previous twelve months, as in Amihud (2002) [2]. We then hold the portfolios for the month t.

The stock shares traded on BOVESPA considered eligible meet three criteria: The share is the company's most traded share (that is, the one with the highest volume traded during the last year); The shares were traded in more than 80% of the days of the year $t-1$, with a volume greater than R$500,000.00 per day, and if the share was listed in the year $t-1$, the period considered runs from the day of listing to the last day of the year; The shares were initially listed before December of the year $t-1$.

Figure 4: Momentum Factor Returns



5 Results

5.1 Wald test for MG and CCE estimators

The first step to perform the Wald tests was to calculate the coefficients of each of the estimators for the model (16). In Table 1 we present the results obtained for the Mean Group estimator for the model (16). We observe that for this estimator the t-statistic of the Market factor and the size factor is greater than two. In addition, the estimated R-square for the model was approximately 0.91133, which leads us to believe that it explains well the observed variation in returns and the factors that most influence are market and size.

Table 1: MG Estimator (16)

	$R_m - R_f$	SMB	HML	IML	WML
Estimative	0.9548	0.2987	0.0121	0.0653	−0.0511
Std. Dev.	0.0041	0.0644	0.0623	0.0641	0.0617
t Stat	236.5998	4.6361	0.1934	1.0192	−0.8270

Nota: The estimated $R - Square$ for the model was approximately 0.91133. The above results were obtained by calculating the MG estimator for a panel with nxT, where $n = 12$ and $T = 4950$, which results in a total of 59400 data.

In Table 2 we present the results of the CCE estimator for the model (16). In Table 3 we present the results of the Wald test with the null hypothesis that the coefficients are

Figure 5: Illiquidity Factor Returns

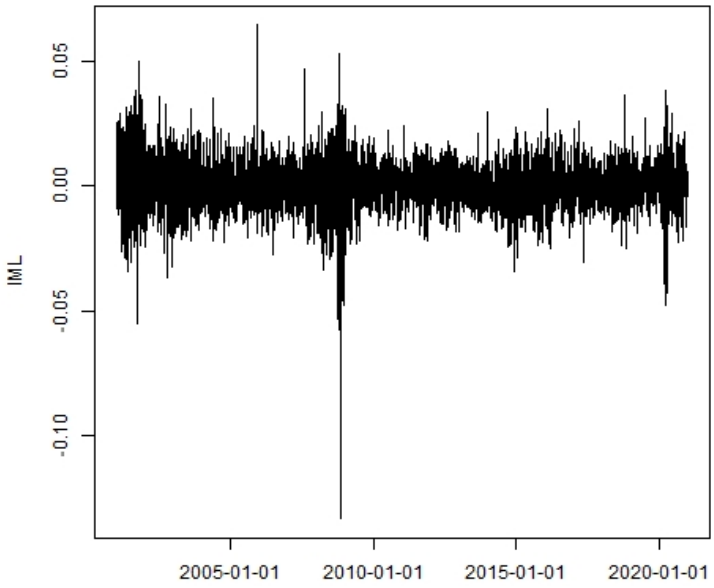
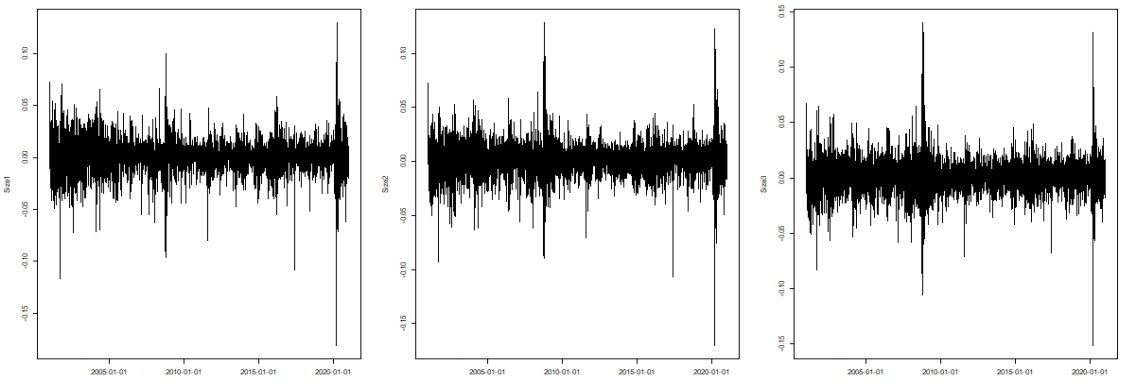


Figure 6: Portfolio Returns Sorted by Size



equal to the coefficients obtained by the Mean Group estimator $\hat{\beta}^F$. In Table 4 we present the results of the Wald test that the coefficients are equal to the coefficients obtained by the CCE estimator $\hat{\beta}^P$.

As explained in section 3.1.3, we performed three Wald tests. In **Test 1** we performed a simple Wald test, assuming the model (16) and that its coefficients were equal to $\hat{\beta}^F$ and the null hypothesis was that $\beta = \hat{\beta}^P$. We can see that **Test 1** rejects the null hypothesis. For **Test 2**, we assumed that the model (16) has the error structure of the CCE estimator and the coefficients were equal to $\hat{\beta}^P$ and the null hypothesis was that $\beta = \hat{\beta}^F$. We note in Table 3 that **Test 2** rejects the null hypothesis. **Test 3** is the Wald test comparing the covariance

Figure 7: Portfolio Returns Sorted by Book-to-Market

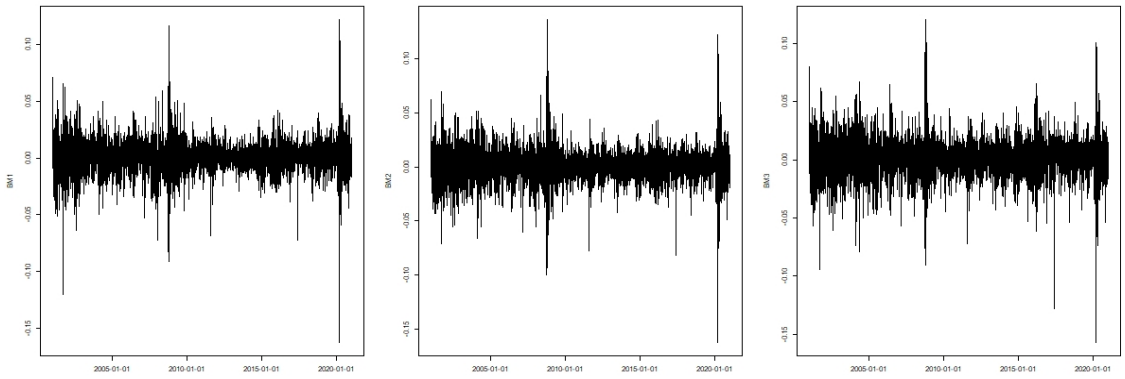
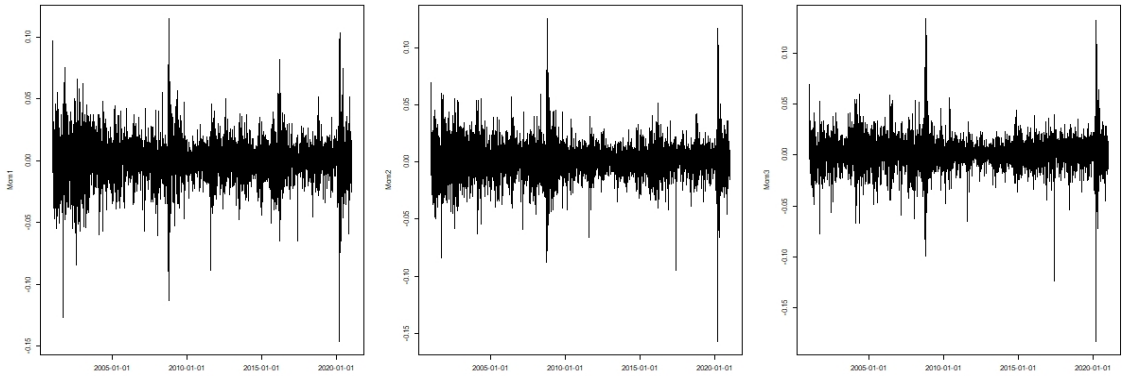


Figure 8: Portfolio Returns Sorted by Momentum



estimates of the two models. In this test we test our null hypothesis is that the covariances are equal. It also rejects the null hypothesis. With this, we conclude that there are strong indications that the model has omitted factors.

5.2 Simulation Results for $\hat{p}$

The Giglio and Xiu estimator is performed in three steps. To complete the first step, an estimator of the number of factors is required. In the article on which we base ourselves, the estimator of equation (45) was proposed. However, it is necessary to find a penalty function $\phi(n,t)$ that has the necessary properties for convergence and gives good estimation results. We chose four penalty functions and based ourselves on Bai and Ng [3]'s paper to carry out the simulations. Our idea was to carry out a test similar to the simulations of the homocedastic model that they adopted.

For each estimator $\hat{p}_j^k$, where $k \in \{1,\ldots,4\}$ and $j \in \{1,2\}$, defined by

Figure 9: Returns on Portfolios Sorted by Illiquidity

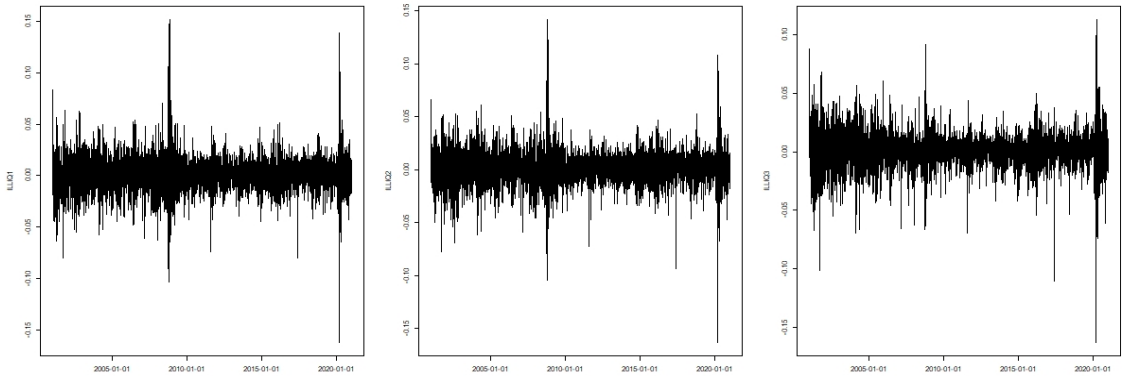


Table 2: CCE Estimator(16)

	$R_m - R_f$	SMB	HML	IML	WML
Estimative	0.0918969	1.2449408	0.1621465	0.0074829	−0.1936320
Std. dev.	0.1189371	0.8526953	0.1179169	0.3221000	0.1493925
t Stat	0.7727	1.4600	1.3751	0.0232	−1.2961

Note: The model's $R - Squared$ was approximately 0.97151. The above results were obtained by calculating the CCE-MG estimator for a panel with nxT, where $n = 12$ and $T = 4950$, which results in a total of 59400 data.

$$\hat{p}_j^k = \begin{cases} \arg\min_{1 \le l \le p_{max}} \left((NT)^{-1} \lambda_l(\bar{R}'\bar{R}) + l \times \phi_k(N,T) \right) - 1 & \text{, if } j = 1 \\ \arg\min_{1 \le l \le p_{max}} \left((NT)^{-1} \lambda_j(\bar{R}'\bar{R}) + l \times \phi_k(N,T) \right) & \text{, if } j = 2 \end{cases} \tag{45}$$

We chose 19 pairs (N, T). For each pair (N, T), we will generate data X that depends on a f amount of factors, $f \in \{1, 3, 4\}$. That is, X will be generated from one factor, or from three factors, or from 5 factors. Below is the equation representing the process:

$$\underset{(N\times T)}{X} = \underset{(N\times f)}{C}\underset{(f\times T)}{F} + \underset{(T\times N)}{E}' \tag{46}$$

All of our matrices were generated from a normal multivariate process: C is the charge matrix $(N \times f)$ generated by a random variable that follows a $\mathcal{N}(\mu_f, \Sigma_f)$ of size N, F is the matrix of factors $(f \times T)$ generated by a random variable that follows a $\mathcal{N}(\mu_T, \Sigma_T)$ of size f and E is the error matrix $(N \times T)$ generated by a random variable that follows $\mathcal{N}(\mu_N, \Sigma_N)$ of size T, where

Table 3: Wald tests

	Chisq	Pr(>Chisq)
Test 1	434.53	0.00
Test 2	49066.00	0.00
Test 3	601.00	0.00

Note: This test consists of a linear hypothesis test that calculates an F statistic comparing the model and the results obtained by the CCE-MG estimator with the coefficients obtained by the MG model.

$$\underset{(r\times 1)}{\mu_r} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \text{ for } r \in \{f, t_i, n_i\} \tag{47}$$

$$\underset{(r\times r)}{\Sigma_r} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \text{ for } r \in \{f, t_i, n_i\} \tag{48}$$

1000 simulations were performed and in each simulation the number of factors of X was estimated using each of the estimators $\hat{p}_j^k$. Finally, an estimator $\bar{p}_{j,(c_i,f)}^k$ was obtained, which is the average of the estimators obtained in the 1000 simulations.

In tables 4 to 6, we show the results obtained by the estimators p_j^k for each pair (N, T). We also report the mean squared error of each estimator across the 1000 simulations.

Table 4: Average and MSE of the estimators for the number of factors $f = 1$.

N	T	$\bar{p}_1^1$	$\bar{p}_1^2$	$\bar{p}_1^3$	$\bar{p}_1^4$	$\bar{p}_2^1$	$\bar{p}_2^2$	$\bar{p}_2^3$	$\bar{p}_2^4$
40	100	1.000	1.000	1.000	1.911	2.000	2.000	2.000	2.919
60	100	1.000	1.000	1.000	3.358	2.000	2.000	2.000	4.307
60	200	1.000	1.000	1.000	1.568	2.000	2.000	2.000	2.590
60	500	1.000	1.000	1.000	1.057	2.000	2.000	2.000	2.041
60	2000	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000
100	40	1.000	1.000	1.000	1.917	2.000	2.000	2.000	2.890
100	60	1.000	1.000	1.000	3.335	2.000	2.000	2.000	4.297
100	100	1.000	1.000	1.000	8.841	2.000	2.000	2.000	9.902
200	60	1.000	1.000	1.000	1.554	2.000	2.000	2.000	2.554
200	100	1.000	1.000	1.000	2.956	2.000	2.000	2.000	4.049
500	60	1.000	1.000	1.000	1.059	2.000	2.000	2.000	2.059
500	100	1.000	1.000	1.000	1.253	2.000	2.000	2.000	2.268
1000	60	1.000	1.000	1.000	1.001	2.000	2.000	2.000	2.000
1000	100	1.000	1.000	1.000	1.037	2.000	2.000	2.000	2.036
2000	60	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000
2000	100	1.000	1.000	1.000	1.001	2.000	2.000	2.000	2.000
4000	60	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000
4000	100	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000
12	4950	0.991	0.998	0.946	1.000	1.989	1.996	1.959	2.000
	MSE	0.0000043	0.0000002	0.0002	4.1345	0.9988	0.9996	0.9958	7.0754

Table 5: Average of estimators for the number of factors $f = 3$

N	T	$\bar{p}_1^1$	$\bar{p}_1^2$	$\bar{p}_1^3$	$\bar{p}_1^4$	$\bar{p}_2^1$	$\bar{p}_2^2$	$\bar{p}_2^3$	$\bar{p}_2^4$
40	100	2.970	2.999	2.950	14.914	3.971	4.000	3.956	15.886
60	100	2.971	3.000	2.997	24.844	3.981	4.000	4.000	26.062
60	200	2.941	3.000	3.000	13.514	3.943	4.000	4.000	14.497
60	500	2.907	3.000	3.000	5.266	3.891	4.000	4.000	6.286
60	2000	2.742	3.000	3.000	3.162	3.747	4.000	4.000	4.168
100	40	2.959	3.000	2.945	14.801	3.983	4.000	3.965	16.094
100	60	2.956	3.000	2.997	24.933	3.985	4.000	3.998	25.895
100	100	2.971	3.000	3.000	46.294	3.963	4.000	4.000	47.287
200	60	2.944	3.000	3.000	13.548	3.952	4.000	4.000	14.544
200	100	2.879	3.000	3.000	29.873	3.923	4.000	4.000	30.779
500	60	2.880	3.000	3.000	5.240	3.895	4.000	4.000	6.319
500	100	2.782	3.000	3.000	10.012	3.790	4.000	4.000	11.066
1000	60	2.835	3.000	3.000	3.607	3.826	4.000	4.000	4.630
1000	100	2.658	3.000	3.000	4.898	3.644	4.000	4.000	5.955
2000	60	2.776	3.000	3.000	3.156	3.788	4.000	4.000	4.141
2000	100	2.502	3.000	3.000	3.510	3.564	4.000	4.000	4.526
4000	60	2.675	3.000	3.000	3.017	3.720	4.000	4.000	4.018
4000	100	2.436	3.000	3.000	3.113	3.54	4.000	4.000	4.146
12	4950	2.662	2.887	1.836	2.999	3.651	3.877	2.815	4.000
	MSE	0.0603	0.0007	0.0716	216.9143	0.7073	0.9878	0.9408	236.7685

Table 6: Average of estimators for the number of factors $f = 5$

N	T	$\bar{p}_1^1$	$\bar{p}_1^2$	$\bar{p}_1^3$	$\bar{p}_1^4$	$\bar{p}_2^1$	$\bar{p}_2^2$	$\bar{p}_2^3$	$\bar{p}_2^4$
40	100	3.456	4.993	2.824	40.000	4.461	5.99	3.945	41.000
60	100	2.680	5.000	4.467	41.944	3.587	6.000	5.429	42.8100
60	200	1.513	5.000	4.987	60.000	2.494	6.000	6.000	61.000
60	500	0.709	5.000	5.000	60.000	1.678	6.000	6.000	61.000
60	2000	0.370	5.000	5.000	5.840	1.416	6.000	6.000	6.823
100	40	3.485	4.995	2.883	30.275	4.545	5.991	4.052	31.391
100	60	2.578	5.000	4.448	41.699	3.649	6.000	5.505	42.716
100	100	1.446	5.000	4.981	64.981	2.387	6.000	5.979	66.002
200	60	1.426	5.000	4.996	33.890	2.457	6.000	6.000	34.917
200	100	0.547	5.000	5.000	59.760	1.508	6.000	6.000	60.943
500	60	0.693	5.000	5.000	14.502	1.736	6.000	6.000	15.647
500	100	0.246	5.000	5.000	30.310	1.222	6.000	6.000	31.619
1000	60	0.501	5.000	5.000	7.901	1.461	6.000	6.000	8.913
1000	100	0.146	5.000	5.000	13.515	1.157	6.000	6.000	14.562
2000	60	0.395	5.000	5.000	5.848	1.388	6.000	6.000	6.842
2000	100	0.131	5.000	5.000	7.480	1.114	6.000	6.000	8.474
4000	60	0.365	5.000	5.000	5.236	1.347	6.000	6.000	6.251
4000	100	0.101	5.000	5.000	5.719	1.087	6.000	6.000	6.693
12	4950	3.084	3.952	1.744	12.000	4.016	4.925	2.723	13.000
	MSE	15.3767	0.0578	1.1267	996.0771	8.7872	0.8930	1.1365	1046.2293

With the results presented in Tables 4-6, we observe that the estimator

$$\hat{p}_1^2 = \arg \min_{1 \leq l \leq p_{max}} \left[(NT)^{-1} \lambda_j(\bar{R}'\bar{R}) + l \times \left(\log \left(\frac{N \times T}{N + T} \right) \right) \times \left(\frac{N + T}{N \times T} \right) \right] - 1 \tag{49}$$

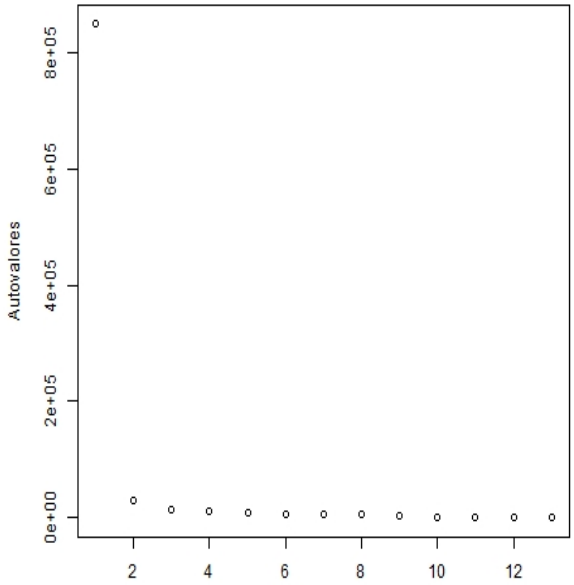
presents the smallest mean squared error for all factors and, therefore, we conclude that it is the best estimator among the chosen estimators. For this reason it will be used to estimate the number of factors in the next step.

We also believe that the fact that the function ϕ_4 converges more slowly than the previous ones in the limit, may have caused this erratic behavior of the $\hat{p}^4$, mainly for low N and T and larger numbers of factors.

5.3 Applying Giglio and Xiu method for Nefin portfolios

The results of the previous section helped us choose the $\hat{p}_1^2$ estimator. After this choice, the first step was to calculate the PCA of the matrix $(NT)^{-1}\bar{R}'\bar{R}$. With the eigenvalues obtained by PCA, we calculate the $\hat{p}_1^2$ estimator. For our portfolios $\hat{p}_1^2 = 2$, defined using the criteria discussed in the previous section, that is, two omitted latent factors influence our estimation.

Figure 10: Twelve first eigenvalues obtained by PCA

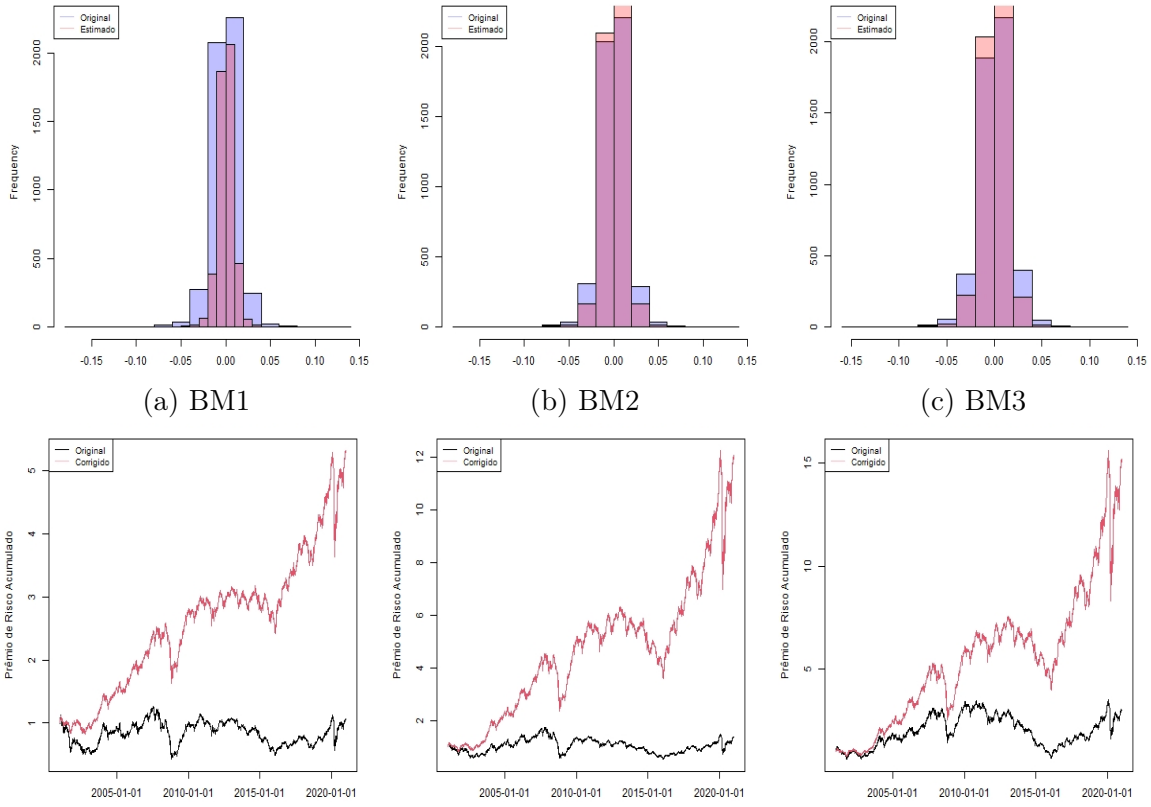


With $\hat{p}_1^2$, we estimate $\hat{\beta}$ and $\hat{V}$. Both are used to obtain the $\hat{\gamma}$ risk premium. The risk premium $\hat{\gamma}$ was obtained for each of the twelve portfolios and compared with the original values.

We can observe in Figure 11 that for portfolios classified by book-to-market, the estimated histograms have a more concentrated distribution than the original ones. When we look at the accumulated risk premium, we observe that, despite having the same movement, the estimated accumulated risk premium is far from the original series.

For portfolios 1 and 2 classified by illiquidity, the results shown in Figure 12 are similar to those we obtained with portfolios classified by book-to-market. However, for portfolio 3, the estimated risk premiums were very concentrated at zero, causing a very large difference between the histogram distribution of the original risk premiums and the histogram distribution of the estimated risk premiums.

Figure 11: Comparison of risk premiums for portfolio classified by book-to-market



(a) BM1 (b) BM2 (c) BM3

(d) Cummulated BM1 (e) Cummulated BM2 (f) Cummulated BM3

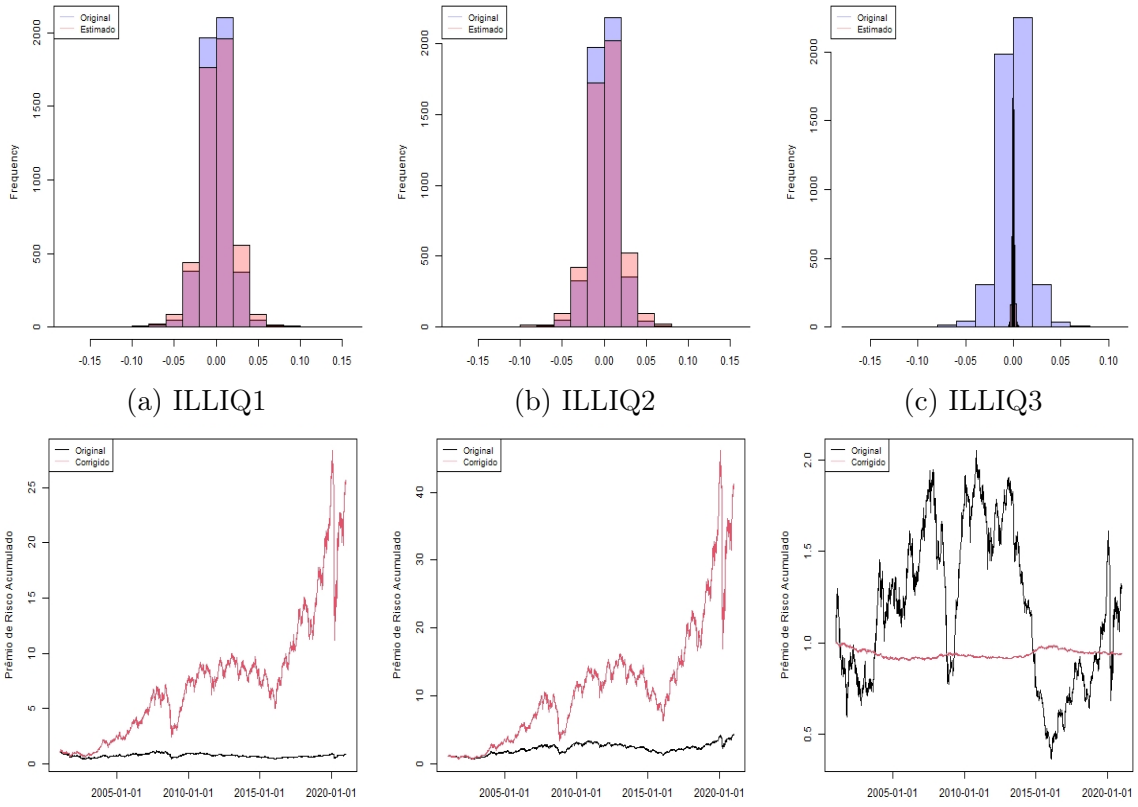
Both for portfolios classified by momentum and for portfolios classified by size, we observe in Figures 13 and 14 that the estimated histograms have a more concentrated distribution than the original ones. When we look at the cumulative risk premium, we can see that, despite having the same movement, the estimated cumulative risk premium is far from the original series.

5.4 Comparison of the three models

The comparison of the pricing of the models was performed using the residuals. Our idea was for each of the three models to calculate the residual series defined by

$$\hat{e}_{it} = R_{it} - \hat{R}_{it} \tag{50}$$

Figure 12: Comparison of risk premiums for portfolio classified by illiquidity



(a) ILLIQ1 (b) ILLIQ2 (c) ILLIQ3

(d) Cummulated ILLIQ1 (e) Cummulated ILLIQ2 (f) Cummulated ILLIQ3

where R_{it} is the return on portfolio i for period t and $\hat{R}_{it}$ is the estimated return on portfolio i for period t.

The residual series were used to calculate the following metrics:

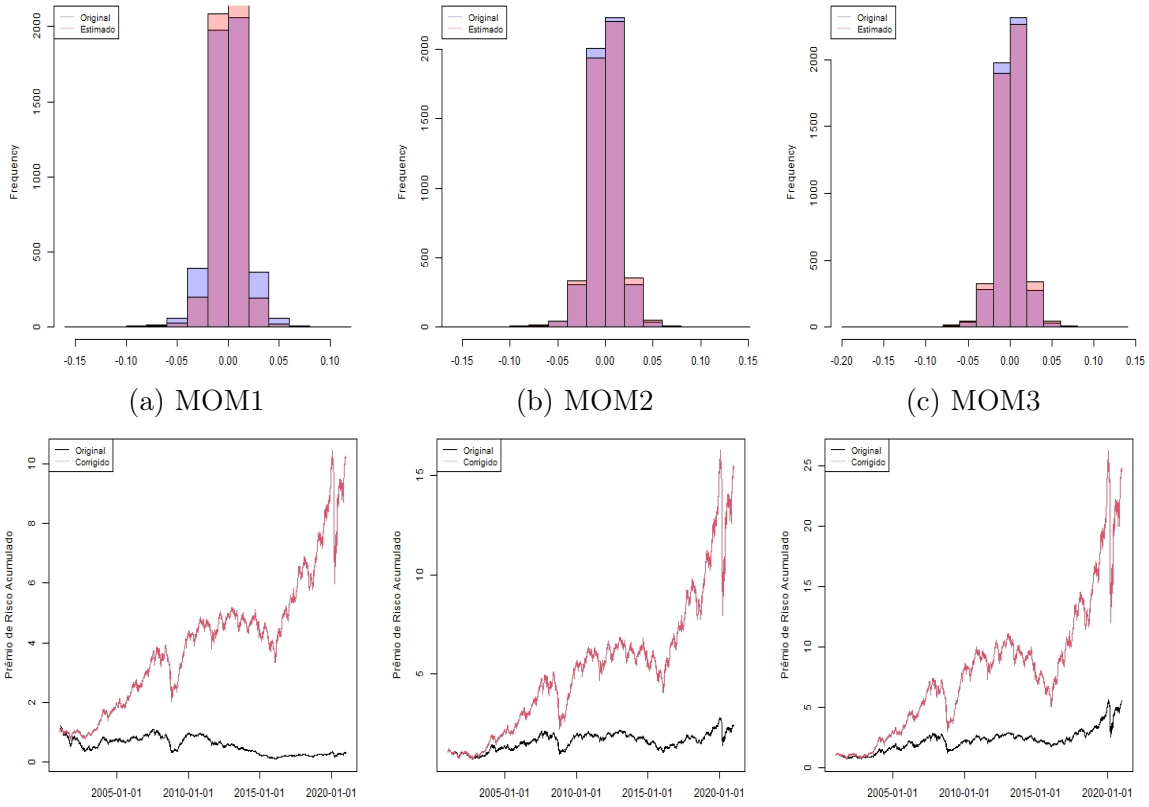
- Mean:

$$\bar{e}_i = \frac{1}{N} \sum_{t=1}^{T} \hat{e}_{it} \tag{51}$$

- Standard Deviation:

$$\sigma_i = \left(\sum_{t=1}^{T} \frac{(\hat{e} - \bar{e}_{it})^2}{N} \right)^{1/2} \tag{52}$$

Figure 13: Comparison of risk premiums for portfolio classified by Momentum



(a) MOM1 (b) MOM2 (c) MOM3

(d) Cummulated MOM1 (e) Cummulated MOM2 (f) Cummulated MOM3

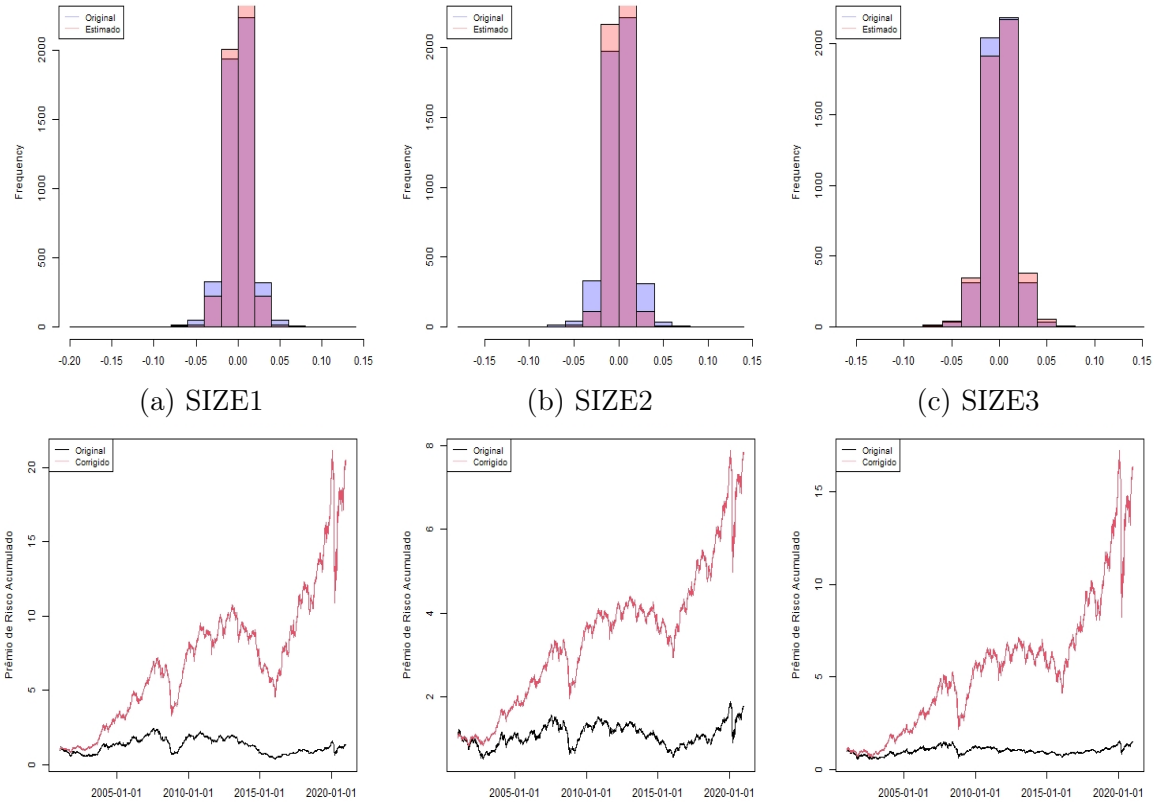
- Mean Squared Error (MSE):

$$MSE_i = \frac{1}{N} \sum_{t=1}^{T} \hat{e}_{it}^2 \tag{53}$$

where, N = 12, that is, the number of portfolios and T = 4950, which is the number of periods.

In Table 7 we present the results obtained for the portfolios classified by the book-to-market method. We note that in all cases the MG and CCE estimators have a mean of zero residuals, as expected. Therefore, we can conclude that the best one is the one with the lowest mean squared error, which is the CCE. The GX estimator has a mean different from zero, which contradicts hypothesis (3) of the Giglio and Xiu model. Either way, it features the highest MSE values for at least two of the three portfolios ranked by book-to-market.

Figure 14: Comparison of risk premiums for portfolio classified by Size



(a) SIZE1 (b) SIZE2 (c) SIZE3

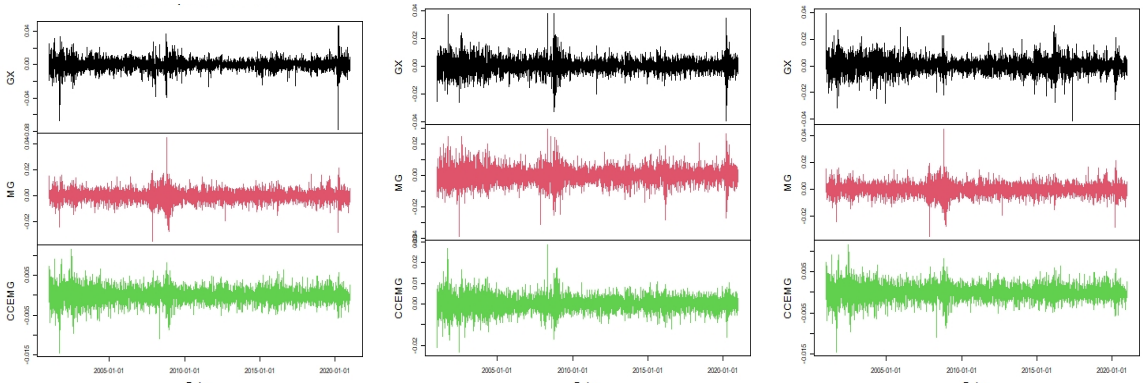
(d) Cummulated SIZE1 (e) Cummulated SIZE2 (f) Cummulated SIZE3

Table 7: Residuals for NEFIN BM portfolios

Portfolio	Model	Min	Max	Mean	Std. Dev.	MSE
BM1	GX	-0.0778	0.0472	0.0002	0.0068	0.000047
BM1	MG	-0.0346	0.0448	0.0000	0.0044	0.000019
BM1	CCE	-0.0145	0.0116	0.0000	0.0019	0.000004
BM2	GX	-0.0392	0.0381	0.0001	0.0051	0.000026
BM2	MG	-0.0385	0.0294	0.0000	0.0053	0.000028
BM2	CCE	-0.0232	0.0286	0.0000	0.0037	0.000014
BM3	GX	-0.0410	0.0396	0.0002	0.0055	0.000030
BM3	MG	-0.0346	0.0448	0.0000	0.0044	0.000019
BM3	CCE	-0.0145	0.0116	0.0000	0.0019	0.000004

In Table 8 we present the results obtained for the portfolios classified by illiquidity. We note that in all cases the MG and CCE estimators have a mean of zero residuals, as expected. Therefore, we can conclude that the best one is the one with the lowest mean squared error, which is the CCE. The GX estimator has a mean different from zero, which contradicts

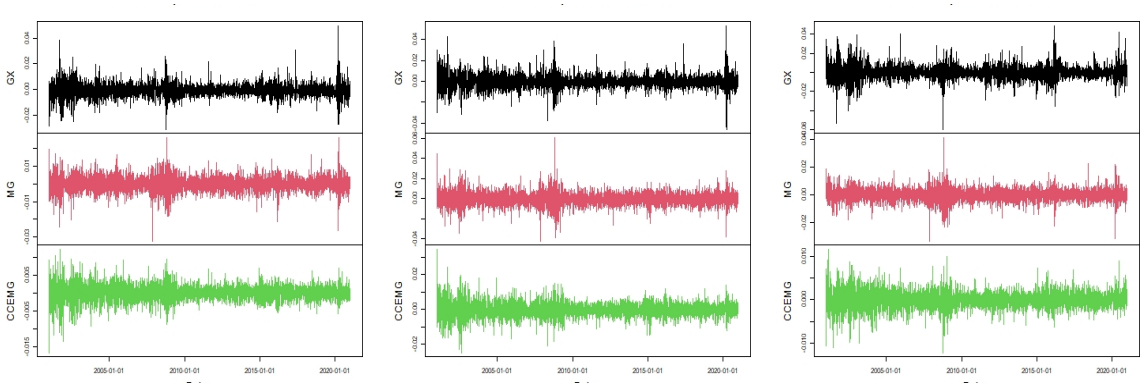
Figure 15: Portfolio Residuals for Book-to-Market portfolios



(a) BM1 (b) BM2 (c) BM3

hypothesis (3) of the Giglio and Xiu model. In any case, it has the highest MSE values for at least two of the three portfolios ranked by illiquidity.

Figure 16: Portfolio Residuals for Illiquidity portfolios



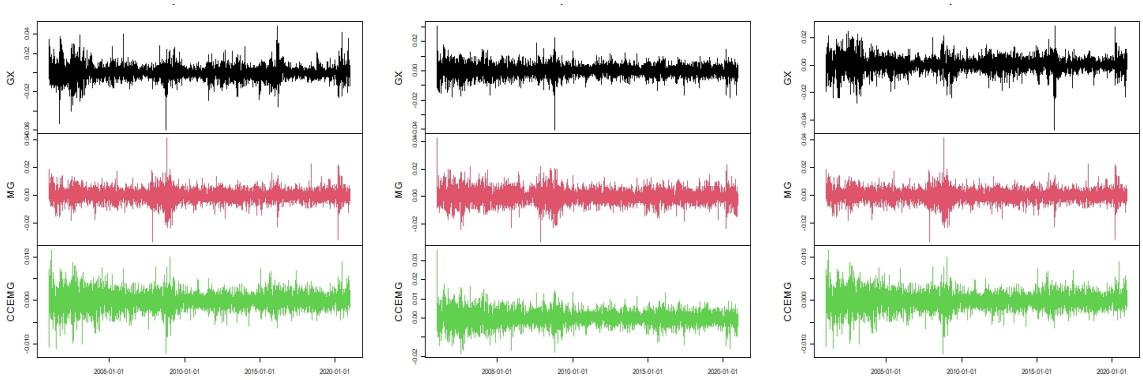
(a) ILLIQ1 (b) ILLIQ2 (c) ILLIQ3

Table 8: Residuals for NEFIN ILLIQ portfolios

Portfolio	Model	Min	Max	Mean	Std. Dev.	MSE
ILLIQ1	GX	-0.0305	0.0496	-0.0003	0.0051	0.000026
ILLIQ1	MG	-0.0322	0.0261	0.0000	0.0042	0.000018
ILLIQ1	CCE	-0.0168	0.0122	0.0000	0.0022	0.000005
ILLIQ2	GX	-0.0457	0.0533	-0.0001	0.0064	0.000041
ILLIQ2	MG	-0.0420	0.0609	0.0000	0.0064	0.000042
ILLIQ2	CCE	-0.0248	0.0346	0.0000	0.0041	0.000017
ILLIQ3	GX	-0.1681	0.1139	0.0007	0.0156	0.000244
ILLIQ3	MG	-0.0322	0.0261	0.0000	0.0042	0.000018
ILLIQ3	CCE	-0.0168	0.0122	0.0000	0.0022	0.000005

We observe the results presented in Table 9 for portfolio 1 classified by momentum, we observe that the GX estimator has a mean different from zero, which contradicts the hypothesis of the models. The MG and CCE estimators, on the other hand, have zero mean, however the CCE estimator is the one with the lowest standard deviation and MSE and, therefore, the one that presents better results for this portfolio. For portfolio 2 classified by momentum, we observed that the MG estimator has a mean different from zero, which contradicts the model's assumptions. The CCE and GX estimators have zero mean, however the CCE estimator continues to show better results with lower standard deviation and MSE. The results of portfolio 3 sorted by momentum are analogous to the results of portfolio 1 sorted by momentum.

Figure 17: Portfolio Residuals for Momentum portfolios



(a) MOM1 (b) MOM2 (c) MOM3

Table 9: Residuals for NEFIN MOM portfolios

Portfolio	Model	Min	Max	Mean	Std. Dev.	MSE
MOM1	GX	-0.0597	0.0491	-0.0002	0.0073	0.000053
MOM1	MG	-0.0332	0.0412	0.0000	0.0045	0.000021
MOM1	CCE	-0.0123	0.0117	0.0000	0.0020	0.000004
MOM2	GX	-0.0330	0.0427	0.0000	0.0053	0.000028
MOM2	MG	-0.0400	0.0308	0.0001	0.0040	0.000016
MOM2	CCE	-0.0186	0.0355	0.0000	0.0037	0.000014
MOM3	GX	-0.0471	0.0287	0.0001	0.0053	0.000028
MOM3	MG	-0.0332	0.0412	0.0000	0.0045	0.000021
MOM3	CCE	-0.0123	0.0117	0.0000	0.0020	0.000004

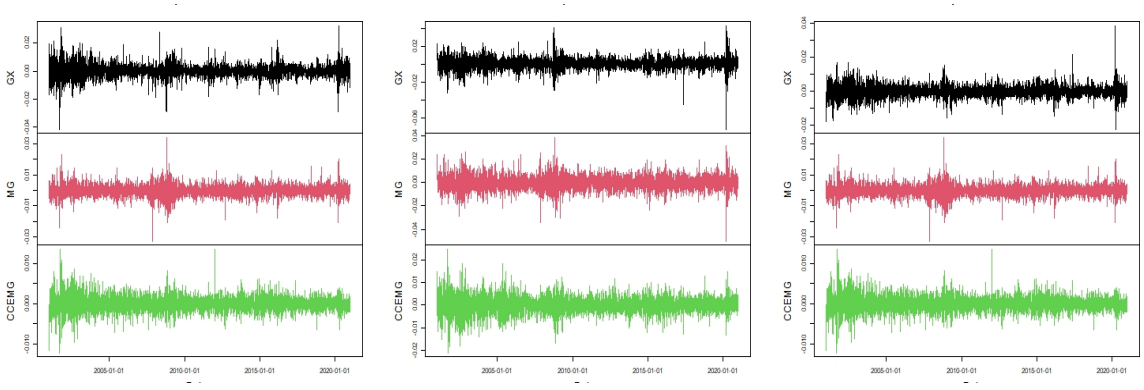
In Table 10 we present the results obtained for the portfolios classified by size. We note

that in all cases the MG and CCE estimators have a mean of zero residuals, as expected. Therefore, we can conclude that the best one is the one with the lowest mean squared error, which is the CCE. The GX estimator has a mean different from zero, which contradicts hypothesis (3) of the Giglio and Xiu model. In any case, it has the highest MSE values for at least two of the three portfolios sorted by size.

Table 10: Residuals for NEFIN SIZE portfolios

Portfolio	Model	Min	Max	Mean	Std. Dev.	MSE
SIZE1	GX	-0.0416	0.0325	0.0000	0.0047	0.000023
SIZE1	MG	-0.0327	0.0342	0.0000	0.0038	0.000015
SIZE1	CCE	-0.0122	0.0135	0.0000	0.0019	0.000004
SIZE2	GX	-0.0726	0.0429	0.0002	0.0065	0.000042
SIZE2	MG	-0.0501	0.0388	0.0000	0.0062	0.000039
SIZE2	CCE	-0.0212	0.0247	0.0000	0.0037	0.000014
SIZE3	GX	-0.0225	0.0387	-0.0001	0.0036	0.000013
SIZE3	MG	-0.0327	0.0342	0.0000	0.0038	0.000015
SIZE3	CCE	-0.0122	0.0135	0.0000	0.0019	0.000004

Figure 18: Residuals for NEFIN SIZE portfolios



(a) SIZE1 (b) SIZE2 (c) SIZE3

6 Conclusion

In this study, we explored the applicability of the Fama-French five-factor model to Brazilian data, utilizing two different estimation methods: the Mean Group (MG) estimator and the Correlated Common Effects (CCE) estimator. The primary objective was to compare these

two estimators and identify any potential omitted factors. Additionally, we compared the predictions of the three models - MG, CCE, and the model proposed by Giglio and Xiu [15] - by calculating estimated returns and analyzing the residuals generated by each model.

In addition to our analysis of the Fama-French model, we also evaluated two new estimators proposed by Giglio and Xiu [15] - the number of factors estimator and the three-step estimator. Our findings suggest that these new estimators may offer improved accuracy over traditional methods.

In the first part of our study, we observed notable differences in the estimated coefficients for the model (16) between the MG and CCE estimators, suggesting the potential omission of factors in the model.

We also evaluated the number of factors estimator proposed by Giglio and Xiu using four penalty functions. Although all penalty functions approach zero as N and T increase, we were unable to identify a penalty function that satisfies the second condition and accurately estimates the factors in our simulations. Despite this limitation, we found that the Giglio and Xiu estimator performed well with three of the penalty functions, particularly in simulations with one or three factors.

Regarding the three-step estimator proposed by Giglio and Xiu [15], we estimated the risk premium of the NEFIN portfolios in the second step and obtained the risk premium of the observed factors corrected for possible omission of factors and measurement error in the third step. While the estimated series generally exhibited similar movements to the original series, we noted significant divergence in the last year of the sample for most portfolios. However, we observed an inverse characteristic in the estimated series for portfolio 3, which was classified by illiquidity, as well as for the factor constructed based on illiquidity, in comparison to the other portfolios and factors.

We compared the residuals generated by the Fama-French model estimated by the MG estimator, the CCE estimator, and the model proposed by Giglio and Xiu [15]. While we expected the three-step estimator to yield better results, we found that it did not perform as well as we had hoped. We believe that this may be due to the limited number of portfolios and our inability to identify a penalty function that guarantees convergence in probability. Therefore, we conclude that the CCE estimator is the most effective approach for estimating

the returns of NEFIN portfolios.

References

[1] ACHARYA, V. V.; PEDERSEN, L. H. Asset pricing with liquidity risk. *Journal of Financial Markets*, 77(2), p. 31-56, 2002.

[2] AMIHUD, Y. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1), p. 31-56, 2002.

[3] BAI, J.; NG, S. Determining the Number of Factors in Approximate Factors Models. *Econometrica*, 70(1), p. 191-221, 2002.

[4] BAI, J. Inferential Theory for Factors Models of Large Dimensions. *Econometrica*, 71(1), p. 135-171, 2003.

[5] BLACK, F. Capital Market Equilibrium with Restricted Borrowing. *Journal of Business*, 45(3), p. 444-455, 1972.

[6] BLACK, F.; JENSEN, M. C.; SCHOLES, M. The Capital Asset Pricing Model: Some Empirical Tests. *Studies in the Theory of Capital Markets* Praeger Publishers Inc, 1972.

[7] COCHRANE. J. H., Presidential Address: Discount Rates. *Journal of Finance* 66, 1047-110, 2011.

[8] FAMA, E. F,. MACBETH, J. D. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3), p. 607-636, 1973.

[9] FAMA, E. F.; FRENCH, K. R. Dividend Yields and Expected Stocks Returns. *Journal of Financial Economics*, 22(1), p. 3-25, 1988.

[10] FAMA, E. F.; FRENCH, K. R. The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2), p. 427-465, 1992.

[11] FAMA, E. F.; FRENCH, K. R. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), p. 3-56, 1993.

[12] FAMA, E. F.; FRENCH, K. R. Size and Book-to-Market Factors in Earnings and Returns. *The Journal of Finance*, 50(1), p. 131-155, 1995.

[13] FAMA, E. F.; FRENCH, K. R. A Five-Factors Asset Pricing Model. *The Journal of Financial Economics*, 116(1), p. 1-22, 2015.

[14] FAMA, E. F.; FRENCH, K. R. International Tests of a Five-Factor Assets Pricing Model. *The Journal of Financial Economics*, 123(3), p. 441-463, 2017.

[15] GIGLIO, S.; XIU, D. Asset Pricing with Omitted Factor. *Journal of Political Economy*, Accepted February, 2021.

[16] HARVEY, C. R.; LIU, Y.; ZHU, H. ...And The Cross-Section of Expected Returns, *Review of Financial Studies*, 29(1), p. 5-68, 2015.

[17] HOU, K.; XUE, C.; ZHANG, L., Replicating Anomalies: An investment Approach. *Review of Financial Studies*, 28, p. 650-705, 2017.

[18] LINTNER, J. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics*, 47(1), p. 13–37, 1965.

[19] MARKOWITZ, H. Portfolio Selection. *The Journal of Finance*, 7(1), p. 77-91, 1952.

[20] MARKOWITZ, H. Portfolio Selection: Efficient Diversification of Investments. *Yale University Press*, 1959.

[21] MCLEAN, R. D.; PONTIFF, J. Does Academic Research Destroy Stock Return Predictability? *Journal of Finance*, 71(1), 5-32, 2016.

[22] MORGENSTERN, O.; VON NEUMANN, J. Theory of Games and Economic Behavior. Princeton University Press, third edition, 1953.

[23] NOVY-MARX, R. The Other Side of Value: The Gross Profitability Premium, Journal of Financial Economics 108(1), p. 1-28, 2013

[24] PESARAN, M. H.; SMITH, R. Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics*, 68(1), p. 79-113, 1995.

[25] PESARAN, M. H. Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica*, 74(4), p. 967-1012, 2006.

[26] PESARAN, M. H., Time Series and Panel Data Econometrics. Oxford University Press, first edition, 2015.

[27] PETERSEN, M. A. Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, 22(1) p. 435-480, 209.

[28] SHARPE, W. F. Capital Asset Prices: a Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance*, 19(3), p. 425-442, 1964.

[29] TITMAN, S.; WEI, K.; XIE, F. Capital Investments and Stocks Returns, Journal of Financial and Quantitative Analysis 39(4), 677-700, 2004.