



UNIVERSITÀ
DI PAVIA

Dipartimento di Scienze Economiche e Aziendali

Corso di Laurea magistrale in Finance

Estimating Risk Premia with Omitted Factors: Evidence from European Markets

Relatore:

Chiar.mo Prof. Lorenzo Trapani

**Tesi di Laurea
di Elena Galioto**

Matr. n.558044

Anno Accademico 2025-2026

ABSTRACT

Italian version

Il titolo “*Estimating Risk Premia with Omitted Factors: evidence from European markets*” può essere tradotto in “Stima dei premi per il rischio in presenza di fattori omessi: evidenza dai mercati azionari europei”.

L’obiettivo del presente elaborato consiste nell’indagare i *bias*, dovuti alla presenza di variabili omesse, nelle stimazioni dei premi per i rischi nei modelli di *asset pricing*. Sussistono molteplici metodi per stimare i premi quali, ad esempio, il modello Fama-MacBeth o il modello dei *mimicking portfolios*, le cui variabili omesse, però, comportano dei *bias*. Gli economisti Giglio e Xiu propongono una soluzione a tale problema, utilizzando una procedura a tre passaggi considerabile un’estensione dei due modelli citati.

Lo studio di Giglio e Xiu è interamente condotto su dati nord americani, e di conseguenza considera unicamente il contesto economico dell’area americana, rimanendo esclusi gli altri territori. Per tale ragione, la nostra indagine intende estendere l’analisi in un contesto europeo considerando 25 portafogli e tre fattori Fama-French (Mkt-Rf, SMB and HML) che coprono un periodo di 428 mesi da luglio del 1990 a febbraio del 2026.

L’analisi empirica utilizza sette fattori latenti estratti tramite PCA (*Principal Component Analysis*). La presente analisi applica la *three-pass procedure* e consente altresì di stimare i premi per il rischio. Le risultanze ottenute sono poi state confrontate con i risultati ottenuti da Giglio e Xiu, nonché applicando al nostro *dataset* il metodo di Fama-MacBeth.

In conclusione, l’analisi ha consentito di stabilire quale sia il fattore significativo nello stabilire i *risk premia* in un contesto economico europeo.

English version

The title of the thesis is “Estimating Risk Premia with Omitted Factors: evidence from European markets”. The objective is to investigate the omitted variables bias which affects the risk premia estimators in asset pricing. There are several methods to estimate the risk premia such as the Fama-MacBeth regression and the mimicking portfolio, but they are affected by the issue of potential omitted variables bias. Giglio and Xiu proposed a solution using the three-pass procedure that extends the two-pass procedure and the mimicking portfolio approach with a PC-augmented cross-sectional regression.

The study proposed by Giglio and Xiu is entirely conducted on data from North America and of consequently it considers only the American economic context. With this thesis, we extended the analysis to a European context by considering 25 portfolios sorted on size and book-to-market, and three Fama-French factors (Mkt-Rf, SMB and HML), and we covered a period of 428 months from July 1990 to February 2026.

The empirical analysis used seven latent factors ($\hat{p} = 7$) extracted with the Principal Component Analysis (PCA). The analysis applies the three-pass procedure and leads to the estimation of the risk premia of the factors.

The results of the analysis were then compared to those obtained by Giglio and Xiu and with those obtained with the Fama-MacBeth method.

In conclusion, the analysis allowed us to identify the significant factor in the risk premia estimation with omitted variables in the European context.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION	5
1.1 Motivation	5
1.2 European context.....	6
1.4 Structure of the thesis.....	7

CHAPTER II

LITERATURE REVIEW	9
2.1 Classical Asset Pricing Models	9
2.1.1 The Capital Asset Pricing Model	9
2.1.2 Critique of the CAPM.....	14
2.1.3 The Fama-French Three-Factor Model (1993).....	14
2.2 The Fama-MacBeth Two-Pass Procedure	17
2.2.1 Step 1: time-series beta estimation	19
2.2.2 Step 2: cross-sectional regression.....	21
2.2.3 Shanken correction for standard errors.....	22
2.2.4 Omitted variable bias	23
2.3 The Mimicking Portfolio Approach	24
2.3.1 The bias in mimicking portfolios.....	25

CHAPTER III

THE THREE-PASS PROCEDURE	27
3.1 The Linear Factor Model	27
3.1.1 Setup and notation	27
3.1.2 Rotational Invariance	28
3.1.3 Tradable vs. non-tradable factors.....	30
3.2 The Three-Pass Estimator	30
3.2.1 Step 1: Principal Component Analysis	32
3.2.2 Step 2: Cross-sectional Regression.....	34

3.2.3 Step 3: Time Series Regression	34
3.3 Empirical tests	35

CHAPTER IV

DATA DESCRIPTION AND PRELIMINARY ANALYSIS	38
4.1 Dataset Description	38
4.2 Descriptive statistics.....	39
4.3 Volatility Analysis	43
4.4 Outlier Analysis.....	47

CHAPTER V

EMPIRICAL ANALYSIS.....	51
5.1 Three-pass with $\hat{p}=7$	51
5.1.1 Step 1: Principal Component Analysis	51
5.1.2 Step 2: Cross-Sectional Regression	54
5.1.3 Step 3: Time-series regression	55
5.1.4 Bootstrap standard errors	57
5.2 Comparisons of three-pass procedure	58
5.2.1 Comparison with Giglio and Xiu results	58
5.2.2 Comparison with Fama MacBeth results.....	59
5.3 Weak factor test	62
5.4 Measurement Error Analysis.....	64
5.5 Robustness of Risk Premia	67
5.6 Subsample Analysis	69

CHAPTER VI

CONCLUSIONS	72
6.1 Summary of the results	72
6.2 Limitations	74
6.3 Future research directions	74
REFERENCES	76

CHAPTER I

INTRODUCTION

1.1 Motivation

A common issue in factor models is given by the estimator for risk premia. There are different estimators proposed with different approaches, such as the Fama-MacBeth regressions and the mimicking-portfolio projections. These methods are affected by a potential issue: omitted variables bias.

We can have an omitted variable bias in standard risk premia estimators when the model used does not fully include all the sources of risks that price the market. The solution to this problem is typically *ad hoc* and it changes from case to case. But there is no guarantee that these solutions work and that they are adequate to solve the problem.

Giglio and Xiu propose an alternative methodology to try to solve the omitted variables bias. They use the three-pass procedure that can be seen as a PC-augmented cross-sectional regression or as an extension of the mimicking portfolio estimator of risk premia.

The purpose of this thesis is to analyze the omitted variables bias.

We will focus on the three-pass procedure by Giglio and Xiu and we will see the application of the approach explained in the “Asset Pricing with Omitted Factors” paper of 2021 in a European context.

We will conduct an empirical analysis on a smaller dataset compared to that of Giglio and Xiu, composed of 25 European portfolios and of three European Fama-French factors. All the data are taken from the *Kenneth-French Data Library*, and they cover a period of 428 months from July 1990 to February 2026. The details of the dataset will be explained in Chapter IV.

The empirical analysis of the original paper was based on data from North America and therefore the results obtained refer to an American context. With this thesis we want to verify if the results found by Giglio and Xiu are still applicable to a European context and if there are some differences to consider.

1.2 European context

For this thesis, we chose to focus on the European context to investigate what the results are obtained applying the three-pass procedure to European data and what the main differences are from those of the US.

The main empirical asset pricing literature and the main methodologies applied to factors are developed around US data, while European research is really scarce if compared to the American one and it rarely addresses the omitted-variable problem, which is the central issue of this thesis.

In fact, the paper by Giglio and Xiu on which we base our work and from which we use the methodology of the three-pass procedure is entirely based on US conditions and it uses only US data. The dataset is composed of 647 test asset portfolios built from US equities, corporate bonds, currency portfolios sorted on interest rate differentials and currency portfolios sorted on currency momentum.

Also, the majority of the literature is composed of American economists like Fama and MacBeth (1973), who created the two-pass cross-sectional regression that is still used nowadays and, like the majority of the Capital Asset Pricing Model (CAPM), and like Fama and French, who extended the CAPM with the three and the five-factor models.

The asymmetry between US and European research makes it difficult to understand whether some results come from the US markets and from some other characteristics related to the US background and whether these results are still valid in the perspective of other countries like the Europeans.

With this thesis, we will see some differences between the US and the European results that are visible even with the relatively small dataset we chose.

1.3 My contributions

With this thesis and in particular with the empirical analysis we make some contributions to asset pricing.

In particular, we applied the three-pass estimator to a European cross-section. We provide an application of the three-pass procedure of Giglio and Xiu outside the usual US context extending the method to European countries that are not frequently considered.

With this thesis, we do not consider only theoretical evidence of the fact that the three-pass methodology is superior in terms of accuracy and bias compared to the Fama-MacBeth two-pass methodology. We also analyze empirically the divergence between the two procedures, and we give a demonstration of the omitted-variable bias in Fama-MacBeth.

We also proved that the three-pass method works with a smaller cross-section dataset. With a smaller dataset we found that the key theoretical properties are still applicable and give defensible empirical results. We checked the robustness of the number of latent factors, the weak factor test and the measurement error in the observable factors.

We analyzed the stability of the factors over time with a subsample analysis highlighting the periods where there are historical crises which influence the estimation of the risk premia.

1.4 Structure of the thesis

In the second chapter, we will explain the literature review that leads to the creation of the three-pass procedure. We will illustrate the Capital Asset Pricing Model and

its extensions with the three Fama-French factors. Then, we will focus on the two-pass regression by Fama and MacBeth and on the use of the mimicking portfolios. These two methods are used to find risk premia in asset pricing models, but they lead to an omitted factor bias.

In the third chapter, we will illustrate the three-pass procedure, and we will focus on the methodology proposed by Giglio and Xiu.

After the theoretical components, we will explore the empirical results found by employing the three-pass estimator to a European dataset.

In the fourth chapter, we will explore the construction of the dataset, and what its main characteristics are, such as the descriptive statistics, the factor volatility, the outlier analysis and the correlation between the observable factors.

In the fifth chapter, we will illustrate the empirical results of our analysis; we will analyze the comparison with the Giglio and Xiu results and we will add additional analysis such as the weak factor test, the measurement error in observable factors, the robustness to the number of latent factors and the subsample analysis.

CHAPTER II

LITERATURE REVIEW

In this chapter, we show the literature review that leads to the creation of the three-pass procedure by Giglio and Xiu. We start with the Capital Asset Pricing Model and its extension using the Fama-French Three-Factors Model. Then we will focus on two methods used to estimate the risk premia: the Fama-MacBeth two-pass procedure and the mimicking portfolio approach.

2.1 Classical Asset Pricing Models

2.1.1 The Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM) was developed in the mid-1960s, and it offered a new logic to describe, under uncertainty, the preferences of investors and how to find an equilibrium in asset pricing such that the demand for a risky asset matches the supply in the market.

Before the CAPM, the traditional models were focused on the prediction of the capital markets under certainty, while under uncertainty, they focused on some assertions about the market risk premium.

The Capital Asset Pricing Model is based on some premises. The first hypothesis is that all agents are risk-averse and they make their choices on the basis of mean and variance. They have the same beliefs about future returns and on risk/reward tradeoffs in the market.

Another assumption is that there exists a risk-free asset, which pays a safe return r_f and investors can borrow and lend at a riskless rate. In the CAPM, there exists only a single risk factor which represents the market return.

To reach equilibrium, investors should have the same identical sets of means, variances and covariances of the return distribution.

Investors would prefer higher expected wealth, but they will accept higher risk only if they are compensated by a higher expected return. Hence, the riskier investments should have a higher expected return; this is the reason why there are upward-sloping indifference curves.

To evaluate a potential investment, agents base their choice on the Mean-Variance Criterion. This is a criterion used to select investments with the same magnitude. The suggestion for investments with the same expected return is to choose the one with the lowest standard deviation, while, for investments with the same standard deviation, the suggestion is to choose the one with the greatest expected return. Therefore, the two parameters to consider, and that investors care about, are expected return (E) and standard deviation (σ). In this chapter we will indicate the returns with r and we refer to the market using the letter M and to the portfolios using the letter P .

The Separation Theorem is used in CAPM to simplify the problem of portfolio selection. There are two phases. The first phase is called the technical phase, and, in this, we identify an optimal combination of risky assets that leads to the maximization of returns. In the second phase, investors decide, based on their preferences, how to allocate the resources between the optimal risky portfolio and a risk-free asset.

All the optimal portfolios are allocated on the capital market line (CML). It is a line that starts at point $(0, r_f)$ and goes through $(E\tilde{r}_M, \sigma_M)$.

The slope of CML is $\frac{E\tilde{r}_M - r_f}{\sigma_M}$ and it represents the “market price of risk”. It tells us that “*an investor considering a marginally riskier efficient portfolio would obtain, in exchange, an increase in expected return of $\frac{E\tilde{r}_M - r_f}{\sigma_M}$* ” (Intermediate Financial Theory, Danthine and Donaldson, 2015).

Therefore, the expectation of the market return is $E\tilde{r}_M = r_f + \frac{E\tilde{r}_M - r_f}{\sigma_M} \sigma_p$.

Where $E\tilde{r}_M - r_f$ is the risk premium. The concept of risk premium is at the base of this thesis, and it is important to understand the basic concept that then can be applied to the factor models.

The risk premium is a way to compensate agents for their exposure to some sources of risk.

In the case of portfolios, the risk premium can be computed as the average excess return of the factor; this is typically used for the capital asset pricing model.

In other theoretical models instead, investors focus on risks which are not portfolios such as consumption, liquidity and so on. In this case, to estimate the risk premium we should create a tradable portfolio which is able to isolate a specific risk holding the others constant. This point is important to know especially for better understanding the Giglio and Xiu work; they, in fact, as we will see in chapter III, use also the non-tradable factors.

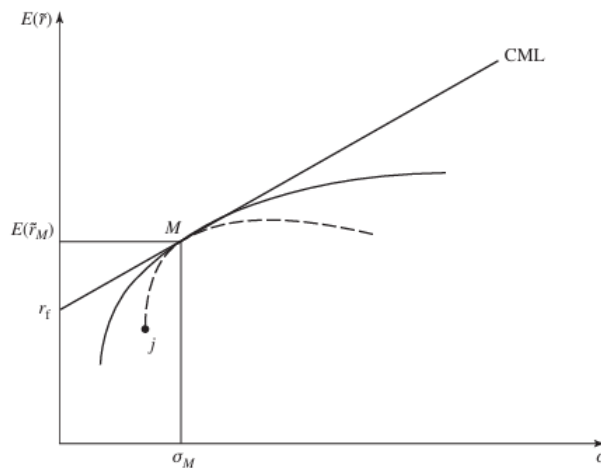


Figure 2.1 The CML by Danthine and Donaldson

From the graph in Figure 2.1, we can note the capital market line and the efficient frontier; the other curve represents the arbitrary asset j that does not belong to the efficient frontier. A condition that must be respected is that the curve which describes the combination of any asset with the efficient portfolio must be tangent to the CML. This is the tangency condition and in the figure the tangency point is M .

Investors can reduce the volatility of the portfolio through diversification without incurring significant costs. Considering some stocks, their returns will not be perfectly correlated, and it means that they will not experience extreme movements simultaneously. The volatility in this way can be reduced; in particular, we can observe that the standard deviation decreases as the number of securities in the portfolio increases.

Through diversification we can eliminate the unsystematic risk that is a “noise” that does not covary with the market, but we cannot eliminate the systematic risk because it is sensitive to the economic or market swings.

In an efficient portfolio all the diversifiable risks have been eliminated and therefore, the correlation between the asset j and the market is 1 ($\rho_{jM} = 1$).

As a measure of systematic risk, we can use a parameter called beta that measures “*the risk contribution of an individual security to a well diversified portfolio*” (Tuck School of Business at Dartmouth, Case 03-111).

Beta can also be seen as the ratio of the covariance between the return of the asset and the return of the market and the variance of the market return. The formula of beta is:

$$\beta_j = \frac{Cov(r_j, r_M)}{\sigma_M^2} = \frac{\sigma_{jM}}{\sigma_M^2}$$

Where:

r_j is the return of the asset

r_M is the return of the market

σ_M^2 is the variance of the return of the market

$Cov(r_j, r_M)$ is the covariance between the return of the asset and the return of the market.

The literature says that the expected excess return of any asset is proportional to β_j and defines the security market line (SML) as:

$$E\tilde{r}_j - r_f = (E\tilde{r}_M - r_f)\beta_j$$

That can be written also as

$$E(\tilde{R}_j) = r_f + \beta_j(E\tilde{R}_M - r_f)$$

The security market line can be seen graphically in Figure 2.2.

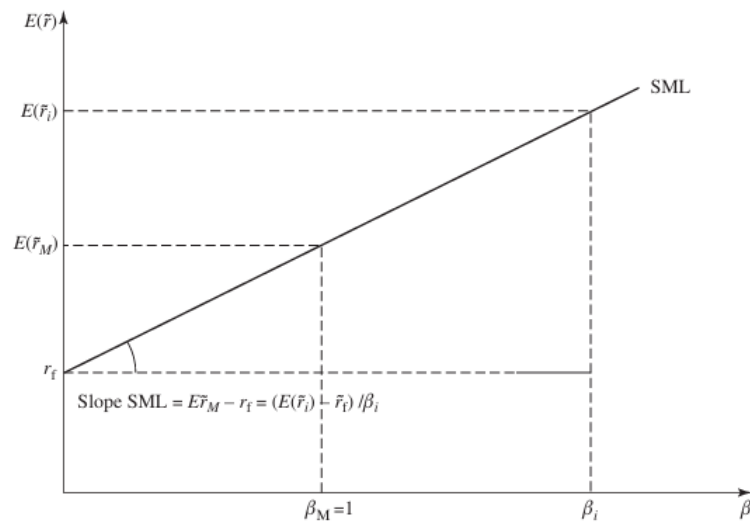


Figure 2.2 The SML by Danthine and Donaldson

The main properties of the security market line are:

1. There is a single risk factor whose variation is explained by beta.
2. Linearity: there is a linear relationship between the expected return and beta.
3. Positive risk-return trade-off: $E\tilde{r}_M - r_f$ which is the slope of the SML should be positive and equal across all assets.

2.1.2 Critique of the CAPM

Despite the fact that the CAPM is a useful tool, there are some empirical failures that have been found. In fact, the Capital Asset Pricing Model explains only 85% of the variations in returns observed ($R^2 = 85\%$), hence it is limited and the 15% of the variations are still unexplained.

A failure of CAPM is given by the assumption that all investors should have the same identical expectations because this does not happen in reality.

Another assumption is that the investors' wealth comes only from investments and risk-free assets but, in reality, the main income comes from the wage that is not perfectly correlated with the market return.

Another limitation is given by the assumption of only a single systematic risk factor that is not sufficient to explain the expected return of an asset. Many researchers believe that there are other factors which can have a significant impact on the expected returns in the market.

This last limitation has started more recent academic work on security price analysis.

2.1.3 The Fama-French Three-Factor Model (1993)

Eugene Fama and Kenneth French expanded the Capital Asset Pricing Model by identifying and incorporating in their model two other factors: SMB and HML.

The SMB factor accounts for the size risk while the HML factor accounts for the value risk.

Mathematically the Fama-French three-factor model can be written as:

$$r = r_f + \beta_1(r_M - r_f) + \beta_2(SMB) + \beta_3(HML) + \varepsilon$$

Where:

r is the expected rate of return

β indicates the coefficient of the factors and it is a measure of sensitivity

$r_M - r_f$ is the market risk premium

ε is the idiosyncratic risk

Therefore, the first factor in the equation is the market risk factor explained in the previous chapter with the Capital Asset Pricing Model.

The second factor is SMB, which stands for Small Minus Big, and it reflects the tendency of small-cap stocks to outperform big-cap stocks, hence, to have higher returns. It can be computed as the average return of the smallest 30% of stocks minus the average return of the largest 30% of stocks in a month.

Small companies have an undiversified nature, and they have a reduced capacity to absorb negative financial events compared to large companies. Due to these characteristics, small companies should be more sensitive than large companies, to various risk factors.

The beta coefficient can take positive or negative values as well. A positive SMB indicates that the small-cap stocks outperform the large-cap stocks while a negative SMB indicates that the large-cap stocks outperform.

HML stands for High Minus Low and it is a measure of the spread in terms of return between companies that have high book-to-market value ratio and companies with a low book-to-market value ratio.

In particular, the spread is between the average return of the 50% of stocks with the highest B/M ratio (these are value stocks) and the average return of the 50% of stocks with the lowest B/M ratio (these are growth stocks). The HML factor has a higher risk exposure for value stocks than for growth stocks.

As with the SMB factor, HML's beta can be positive or negative, and it is found by a regression. When HML is positive, it means that value stocks outperformed growth stocks. If it is negative, then growth stocks outperformed.

Thanks to the introduction of two new factors, Fama and French with their model can explain more than 90% of portfolios' returns.

With the three-factor model, investors can choose how much to be exposed to each specific risk factor by weighting their portfolios.

In 2014, the three-factor model was expanded by Fama and French by including two new factors: profitability (RMW) and investment (CMA).

The fourth factor highlights the fact that companies that have higher future earnings tend to have higher stock returns.

The fifth factor suggests that companies with aggressive investments in growth projects are likely to underperform in the future.

In our empirical work, we chose to employ only the three Fama-French factors for various reasons.

The first reason is that the three-factor model is still the canonical benchmark in the asset pricing literature despite the introduction of the five-factor model. Several studies use only three factors, and our choice makes it easier to make some comparisons with them.

Second, some studies have found that the Fama-French five-factor model improves upon the Fama-French three-factor model especially in North America while in other countries like Europe it has some limitations. The results of R^2 using the five factors seem lower for some countries compared to the results using the three factors and in Europe the CMA factor seems redundant as affirmed by Lin.

Giglio and Xiu, in their study, use the five Fama-French factors especially for the Monte Carlo simulation which we do not replicate because it is not the aim of this thesis. Therefore, this was another reason to use only the three Fama French factors.

Therefore, using only three factors simplified our analysis and it allowed us to focus only on the evaluation of the three-pass estimator without the complexity of two more factors.

2.2 The Fama-MacBeth Two-Pass Procedure

Fama and MacBeth in 1973 introduced a two-pass regression procedure for estimating the risk premia and for testing the beta of the asset pricing models. This test is based on the two-pass cross-sectional regression (CSR) of Black, Jensen and Scholes (BJS).

The main difference is that Fama and MacBeth employ regression with rolling betas. They avoided the problem of spurious cross-sectional relations, and they made the mean returns uncorrelated with the estimated betas.

It is composed of two steps: the time-series beta estimation and the cross-sectional regression.

The starting equation is given by:

$$E(\tilde{R}_i) = E(\tilde{R}_0) + [E(\tilde{R}_m) - E(\tilde{R}_0)]\beta_i$$

Where:

$E(\tilde{R}_i)$ is the expected return on security i

$E(\tilde{R}_0)$ is the expected return on a security which, in the portfolio m , is riskless

$E(\tilde{R}_m)$ is the expected return of the efficient portfolio m

β_i is a risk premium and it indicates a measure of the risk of security i in the efficient portfolio m .

From this equation, we can observe that there is a linear relationship between the expected return on a security and its risk in any portfolio. Another observation is that there is only one measure of risk in the equation and it is given by m . If an

investor is higher risk-averse, he should expect a higher expected return and therefore,

$$E(\tilde{R}_m) - E(\tilde{R}_0) > 0$$

The equation seen above is in terms of expected return but, to test the securities and the portfolio return with some data, Fama and MacBeth suggested using the stochastic generalization. Therefore, the main formula used is:

$$\tilde{R}_{it} = \tilde{\gamma}_{0t} + \tilde{\gamma}_{1t}\beta_i + \tilde{\gamma}_{2t}\beta_i^2 + \tilde{\gamma}_{3t}s_i + \tilde{\eta}_{it}$$

This equation allows the risk premia $\tilde{\gamma}_{0t}$ and $\tilde{\gamma}_{1t}$ to vary from period to period. $\tilde{\gamma}_{1t}$ would be the slope $E(\tilde{R}_m) - E(\tilde{R}_0) > 0$, hence, we must have

$$E(\tilde{\gamma}_{1t}) = E(\tilde{R}_{mt}) - E(\tilde{R}_{0t}) > 0.$$

The term beta squared was included to test linearity; in the previous equation the hypothesis was that $E(\tilde{\gamma}_{2t}) = 0$ but it varies stochastically. The same thing happens with the term s_i because the other formulation assumed $E(\tilde{\gamma}_{3t}) = 0$ but it can vary over time.

$\tilde{\eta}_{it}$ indicates a disturbance and it is independent from all the other variables. The disturbances are identically distributed over time with mean zero.

The hypotheses summarized by Fama and MacBeth are:

C1) linearity: $E(\tilde{\gamma}_{2t}) = 0$

C2) no systematic effects of non- β risk: $E(\tilde{\gamma}_{3t}) = 0$

C3) positive expected return-risk tradeoff: $E(\tilde{\gamma}_{1t}) = E(\tilde{R}_{mt}) - E(\tilde{R}_{0t}) > 0$

Sharpe-Lintner (S-L) Hypothesis: $E(\tilde{\gamma}_{0t}) = E(\tilde{R}_{ft})$

It is also required to have market efficiency (ME) and that the coefficients and the disturbances are fair games. In this way $E(\tilde{\gamma}_{2t}) = E(\tilde{\gamma}_{3t}) = 0$.

To estimate the risk premia, Fama and MacBeth studied common stock traded on the New York Stock Exchange and they chose monthly percentage returns from January 1926 to June 1968. These data are from the Center for Research in Security Prices of the University of Chicago.

Trying to test the two-pass model, Fama and MacBeth found that there is an “errors-in-the-variables” (EIV) problem. In fact, the equation to use is in terms of true values of β_i but only an estimate of β_i is available.

Fama and MacBeth used $\hat{\beta}_i = \frac{\widehat{cov}(\bar{R}_i, \bar{R}_m)}{\sigma^2(\bar{R}_m)}$.

2.2.1 Step 1: time-series beta estimation

In the first step there is the estimation of the rolling betas with a time-series regression. Fama and MacBeth formed portfolios from ranked $\hat{\beta}_i$. In particular, using the monthly returns of the first 4 years which range from 1926 to 1929, Fama and MacBeth formed 20 portfolios for individual securities.

They did the same thing for the following five years and so they recomputed the $\hat{\beta}_i$ as averages across securities to obtain 20 initial portfolios $\hat{\beta}_{pt}$ used in the risk-return tests. The subscript t indicates the month t of the following four years, used to recompute these $\hat{\beta}_{pt}$ as averages of individual securities.

The component $\hat{\beta}_i$ is computed and updated yearly from monthly returns from 1930.

All these passages and the continuous estimate of betas are motivated by the studies of Blume (1970). He showed that if the errors in the $\hat{\beta}_i$ are “less than perfectly positively correlated” the betas of the portfolios are more accurate in estimating the true value of beta rather than using the betas of the individual securities. This is the reason why we do not use directly the betas of individual securities, but we take a first step with the estimation of the rolling betas.

A bias to consider is that, in a cross section of $\hat{\beta}_i$, high values of $\hat{\beta}_i$ tend to be above the true value β_i and hence, the corresponding portfolio will overstate the true β_p .

Instead, if we consider low values of $\hat{\beta}_i$, we have the opposite situation. They tend to be lower than the true values and hence, the corresponding portfolio will underestimate the true one.

Therefore, to avoid this problem that Fama and MacBeth refer to as the “regression phenomenon”, they decide to use periods that are non-overlapping and to use beta re-estimation to create the portfolios.

Fama and MacBeth consider also the measure of the non- β risk of security i using the standard deviation $s(\hat{\epsilon}_i)$. This standard deviation is computed from the least-squares residual of $\hat{\epsilon}_{it}$ which comes from the following formula of the so-called market model:

$$\tilde{R}_{it} = \alpha_i + \beta^i \tilde{R}_{mt} + \tilde{\epsilon}_{it}$$

$s(\hat{\epsilon}_i)$ is computed as an estimate of that part of the return which is not directly related to β_i . This becomes the term s_i in the general equation and it is used along with beta. In particular, these values are updated annually, and their average is used for securities in portfolio p .

For the next step we use the independent variable $\hat{\beta}_{p,t-1}$ which is the average of $\hat{\beta}_i$, we use $\hat{\beta}_{p,t-1}^2$ which is the average of the squared values of β_i and we use $\bar{s}_{p,t-1}(\hat{\epsilon}_i)$ which is the average of the $s(\hat{\epsilon}_i)$.

The entire construction is predictive because the variables $\hat{\beta}_{p,t-1}$ and $\bar{s}_{p,t-1}(\hat{\epsilon}_i)$ are computed from data that were available at time t ; they were available a period prior to the month of the returns; the returns R_{pt} are the dependent variable of the regression.

Under the assumption that the disturbances are independently identically distributed and independent of $\tilde{R}_{mt}$, the standard error of $\hat{\beta}_i$ is

$$\sigma(\tilde{\beta}_i) = \frac{\sigma(\tilde{\epsilon}_i)}{\sqrt{n} \sigma(\tilde{R}_m)}$$

Where n indicates the number of periods.

And

$$\sigma(\tilde{\beta}_{p,t-1}) = \frac{\sigma(\tilde{\epsilon}_p)}{\sqrt{n} \sigma(\tilde{R}_m)}$$

2.2.2 Step 2: cross-sectional regression

In the second step we run a cross-sectional regression of the time-series average at each time period. We take the average of the return after having assumed that there are independent and identically distributed errors.

The cross-sectional equation is given by:

$$R_{pt} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\hat{\beta}_{p,t-1} + \hat{\gamma}_{2t}\hat{\beta}_{p,t-1}^2 + \hat{\gamma}_{3t}\bar{s}_{p,t-1}(\hat{\epsilon}_i) + \hat{\eta}_{pt}$$

$$p = 1, 2, \dots, 20$$

This regression produces a time series of the monthly coefficient estimates $\hat{\gamma}_{0t}$, $\hat{\gamma}_{1t}$, $\hat{\gamma}_{2t}$, and $\hat{\gamma}_{3t}$.

Those are the inputs of the two-parameter model. The procedure is then repeated for all the periods considered.

Fama and MacBeth present the t-statistics for testing the hypothesis $\bar{\gamma} = 0$; this is given by

$$t(\bar{\gamma}_j) = \frac{\bar{\gamma}_j}{s(\bar{\gamma}_j)/\sqrt{n}}$$

Where n is the number of months in the period and it is therefore equal to the number of estimates.

The authors of the two-pass procedure underline that the distributions of the stocks are commonly “thick-tailed” compared to the normal distribution and that the t-statistics work better with nonnormal symmetric distributions.

This leads to the fact that under the assumption of normality, the probability levels or the significance levels of the t-statistics are likely to be overestimated. Fama and MacBeth also note that there are upward-biased probability levels that create a bias toward the rejection of the hypothesis.

For this reason, economist Officer in 1971 estimated that the characteristic exponent for this kind of distribution is about 1.8 instead of the value 2.0 for normal distribution.

Fama and Roll found that the nonnormal distribution does not differ so much from the normal distribution except for the values in the extreme tails. But they also added that the interpretation of the t-statistic in the usual way does not lead to “serious errors”.

2.2.3 Shanken correction for standard errors

In the Fama-MacBeth two-pass regressions there are some errors-in-variables (EIV) problems that make the estimators biased in small samples.

Shanken (1992) finds that the bias and the measurement error in beta, however, tend to decrease as the sample size of the time-series T increases. For this reason, Shanken affirms that the second-pass estimator is T -consistent.

The time-series procedure of Fama and MacBeth to compute the standard errors fails because it “*reflect measurement error in the betas and overstate the precision of the gamma estimates*”.

To solve this problem, Black, Jensen and Scholes (1972), Blume and Friend (1973) and Fama and MacBeth (1973) employed portfolio grouping procedures to try to

minimize the measurement error and to still maintain a considerable cross-sectional variation in the betas of the portfolios.

The grouping leads to the construction of a finite set of portfolios that are well-diversified as $N \rightarrow \infty$. The assumption that makes this possible is the independence of the market factor model disturbances. In this way the residual variation is eliminated, and the betas have no more estimation errors.

Shanken presents a correction for the second-pass estimator considering the asymptotic distribution. Its main theorem affirms that the asymptotic distribution of the T-consistent estimator given by $\sqrt{T}(\hat{\Gamma} - \Gamma)$ converges in distribution, as $T \rightarrow \infty$, to a normal distribution. It has mean zero and its covariance is given by the matrix $(1 + c)\Omega + \Sigma_F^*$, where $c = \Gamma_{12}'\Sigma_F^{-1}\Gamma_{12}$ and Σ_F^* is the bordered version of Σ_F that is the top row and left column of zeros.

Shanken found that to account for errors-in-variables the standard error of the estimate should be multiplied by $\hat{c}$, which the author, after some empirical works, suggests should be equal to 0.36. Another correction is about the t-statistic that should be divided by $\sqrt{1.36}$, equivalent to being divided by $\sqrt{1 + \hat{c}}$.

Correcting is important because parameter c is related in a direct way to the magnitude of each price of risk and it is related inversely to the factor variance. For this reason, it does not surprise that $\hat{c}$ is greater in the multifactor model than in the single index model.

2.2.4 Omitted variable bias

An important issue to consider about the Fama-MacBeth two-pass regression is the omitted variable bias.

If for example, when estimating a risk premium we omit a factor, this can lead to a bias in the two steps of the model.

In the time-series beta estimation step, we have a bias in the estimate of the beta when the omitted factor is correlated with the other factors.

Giglio and Xiu affirm that, in particular, the time-series correlation of the factors influences the magnitude of the bias.

In the cross-sectional step, there is a second bias that occurs. In fact, in the second step, the regression of average returns is not onto the entire matrix of risk exposure with respect to all the beta factors because one factor has been omitted. Hence, we will have a second bias, the magnitude of which depends on the correlation of risk exposures (between the betas).

Another bias in the estimator of the risk premium is given by the measurement error. The measurement error, which is often referred to as z_t , is orthogonal to the factors but it is correlated with $\tilde{\eta}_{it}$. In fact, it can capture idiosyncratic risks or exposure to unpriced risks.

Giglio and Xiu say that a measurement error in the regression leads to an “attenuation bias” when we estimate the betas in the time-series regression step. This bias will affect the cross-sectional step, and it leads to a biased estimate of the risk premia.

2.3 The Mimicking Portfolio Approach

In this section we will see the mimicking portfolio approach, a methodology used in asset pricing. In particular, we can say that mimicking portfolios are important not only for the theory of asset pricing but also for empirical work.

Mimicking portfolios are employed to estimate the risk premia in factor models analysis when there are no traded assets. In general, in asset pricing models, they are used when there are stochastic changes in the opportunity set of consumption and investment.

In literature there are various examples of the use of mimicking portfolios, and they typically use methodologies where the weights of portfolios are fixed over time or vary over time in ad hoc ways, for example with a rolling regression.

Huberman, Kandel and Stambaugh in 1987 provided a mimicking portfolio which can replace the risk factors in asset pricing models with multiple betas.

Fama in 1998 interprets the three-factor model of Fama and French as a mimicking portfolio for the state variables.

In Giglio and Xiu (2021), the mimicking portfolio approach is used to estimate the risk premium of the factor g_t . The factor is projected onto a set of tradable assets returns and a portfolio is constructed that is called “maximally correlated mimicking portfolio” because it has the maximal correlation with g_t . The risk premium is obtained from the average excess return of the mimicking portfolio.

2.3.1 The bias in mimicking portfolios

Using the mimicking portfolio we can incur two kinds of bias: the omitted variable bias and the measurement error bias.

An omitted variable bias can arise by omitting assets onto which g_t is projected.

In the Giglio and Xiu paper, the expected excess return of the mimicking portfolio is given by the formula $r_t^g = (w^g)^T \tilde{r}_t$

Where w^g are the coefficients yielded by the projection and $\tilde{r}_t$ is the asset return of a set of test assets that is chosen.

Hence, Giglio and Xiu affirm that test assets follow a pricing model which can be written as $\tilde{r}_t = \tilde{\beta}\gamma + \tilde{\beta}v_t + \tilde{u}_t$

They find that the risk premium of the first factor, given by the mimicking portfolio, is $\gamma_g^{MP} = \{(\tilde{\beta}\Sigma^v\tilde{\beta}^T + \tilde{\Sigma}^u)^{-1}(\tilde{\beta}\Sigma^v e_1)\}^T \tilde{\beta}\gamma$.

Where e_1 is a column vector, Σ^v is the factor covariance matrix, and $\tilde{\Sigma}^u$ is the covariance matrix of the idiosyncratic risk of the asset chosen for the projection.

From this formula, we can note that not all choices of assets can give a consistent estimator of the risk premium. Therefore, it is not guaranteed that

$$\gamma_g^{MP} = \gamma_1.$$

This equation is valid when “*the assets are chosen to be p portfolios that are well diversified and fully span the true factor*” (Giglio and Xiu, 2021).

The mimicking portfolios will be biased if these conditions are not satisfied.

We can have a measurement error z_t which makes the mimicking portfolios biased. In particular, with a measurement error, the formula seen earlier for the risk premium of the first factor has an additional term $\tilde{\Sigma}^{z,u}$.

The risk premium formula becomes:

$$\gamma_g^{MP} = \{(\tilde{\beta}\Sigma^v\tilde{\beta}^T + \tilde{\Sigma}^u)^{-1}(\tilde{\beta}\Sigma^v e_1 + \tilde{\Sigma}^{z,u})\}^T \tilde{\beta}\gamma$$

Where $\tilde{\Sigma}^{z,u}$ is the covariance matrix between the measurement error z_t and the idiosyncratic component.

There is a bias unless the idiosyncratic errors in the spanning assets are uncorrelated with the measurement error.

CHAPTER III

THE THREE-PASS PROCEDURE

In this chapter we introduce the three-pass estimator methodology used by Giglio and Xiu to estimate the risk premia of the factors without incurring an omitted variable problem.

First, we will see in detail the notation used by Giglio and Xiu for the Linear Factor Model and the main property of Rotational Invariance. Then, we illustrate the three passages and the main goals achieved using the three-pass procedure.

3.1 The Linear Factor Model

3.1.1 Setup and notation

We start with the general linear factor model with p factors; this is given by:

$$r_t = \beta\gamma + \beta v_t + u_t \quad E(v_t) = E(u_t) = 0, \text{ and } \text{Cov}(u_t, v_t) = 0$$

Where:

r_t are excess returns on n assets

β are factor loadings

γ is the vector of risk premia for the p factors

v_t are the innovations of the p factors

u_t are idiosyncratic errors

The mean of v_t and u_t are equal to zero and this means that the factors and the errors, on average, revert to their mean over time. This makes sure that the shocks are temporary.

Another assumption is that the two components are uncorrelated; in fact, the covariance between the idiosyncratic error and the innovations of the factors is zero.

The objective of the thesis is to estimate the risk premia of one or more factors g_t , which are a proxy for the first factor v_{1t} , without necessarily observing all true factors v_t .

In general, we can define g_t as a set of d observable factors about which we want to estimate the risk premia. We can also say that we want to find the expected excess return of a portfolio with beta of 1 with respect to g_t and beta of 0 with respect to all other factors. This second definition can be expressed by the formula: $\gamma_g = \eta\gamma$ where η is the loading of g_t on the unobservable fundamental factors and γ is the risk premium of unobservable factors.

From the next formula we can observe the relationship between g_t and v_t :

$$g_t = \delta + \eta v_t + z_t \text{ with } E(z_t) = 0, \text{ and } \text{Cov}(z_t, v_t) = 0$$

Where z_t is a measurement error in g_t and η is the component that captures the relation between g_t and v_t .

It is important to underline that we cannot observe from the equations reported either η or γ without the possibility of observing v_t . For this reason, we will use the property of rotation invariance that we explain in detail in the next section.

3.1.2 Rotational Invariance

Rotational invariance is a basic result of the three-pass procedure; in fact, to estimate the risk premia in the linear factor model, Giglio and Xiu use this method.

With the term rotational invariance, we consider the fact that the observations of the risk premia g_t are invariant and do not depend on how the control factors $p-1$ are rotated.

This means that any $p-1$ linear combination of the factors can be seen as a control factor, provided that the rotated factor model spans the same risks as the original model.

Giglio and Xiu also want to emphasize that due to the rotational invariance, it is not necessary to know the identities of the true factors to estimate the corresponding risk premia.

Hence, the product $\eta\gamma$ can be identified also when we do not know their true values, but we can observe only an arbitrary full-rank rotation of the factors. For example, thanks to rotational invariance, we can observe $\hat{v}_t = Hv_t$ without observing either v_t or H .

The matrix H is a squared matrix that is unobservable.

We can rewrite the model in this way:

$$r_t = \beta H^{-1}H\gamma + \beta H^{-1}Hv_t + u_t$$

$$g_t = \delta + \eta H^{-1}Hv_t + z_t$$

We can define $\hat{\eta} = \eta H^{-1}$, $\hat{\gamma} = H\gamma$ and $\hat{\beta} = \beta H^{-1}$, and write the model in terms of rotated factors as:

$$r_t = \hat{\beta}\hat{\gamma} + \hat{\beta}\hat{v}_t + u_t$$

$$g_t = \delta + \hat{\eta}\hat{v}_t + z_t$$

If we can observe $\hat{v}_t$, in this way we can also identify $\hat{\eta}$ and $\hat{\gamma}$ and therefore we can compute the risk premia of g_t because

$$\hat{\eta}\hat{\gamma} = \eta H^{-1}H\gamma = \eta\gamma = \gamma_g.$$

Giglio and Xiu underline the fact that the invariance property is valid for the estimation of the risk premia of g_t (γ_g) but it is not valid to find other quantities of the model such as η , γ or β .

3.1.3 Tradable vs. non-tradable factors

We must distinguish between tradable factors and non-tradable factors to better understand the linear factor model.

A tradable factor is itself a portfolio and its risk premium can be found as the average excess return of the factor.

Non-tradable factors are not themselves portfolios and so the estimation of the risk premia requires the construction of a tradable portfolio; in this way we can isolate a specific risk while maintaining the others constant.

Giglio and Xiu in their analysis used both tradable and non-tradable factors. In particular, the tradable factors are the five Fama-French factors market, size, value, profitability and investment and other factors such as momentum (MOM), betting against beta (BAB) and quality minus junk (QMJ).

The non-tradable factors are AR (1) innovations in industrial production growth (IP), VAR (1) innovations in the first PCs of microfinance variables, a liquidity factor, two intermediary capital factors, four factors from Novy-Marx about temperature and two consumption-based factors which include the aggregate consumption series and the stockholders' consumption series.

3.2 The Three-Pass Estimator

In this section we illustrate the three-pass estimator, which is composed of three steps:

- 1) Principal Component Analysis
- 2) Cross-sectional regression
- 3) Time series regression

The three-pass procedure combines the two-pass cross-sectional regressions, explained in the previous chapter, with the principal component analysis (PCA) and it provides an estimator of the risk premia of the factors.

Before introducing the passages, we will cover some assumptions.

Giglio and Xiu impose that there are constant loadings and risk premia, an assumption that is valid for both stocks and sorted portfolios like those used in the empirical analysis.

They also make weak assumptions on the structure of the errors. This is because the results of the authors of the paper “*hold for nonstationary processes with heteroskedasticity and dependence in both the time series and the cross-sectional dimensions*” (Giglio and Xiu, 2021).

The third assumption is about the zero-beta rate, which they impose equal to the observed treasury bill rate. This assumption will not be useful for the empirical analysis of chapter V; in fact, we will not introduce the zero-beta rate.

Giglio and Xiu also underline an important assumption made by Bai and Ng (2002) and Bai (2003) to make it possible to employ the PCA. They say that to find β and $\bar{V}$, n and T must go to infinity. This is the asymptotic theory, and it assumes that a large n and a large T ensure that the eigenvalues that are associated with the factor components of the return covariance matrix increase rapidly at a rate n as n grows. This allows these eigenvalues to be distinguished from the idiosyncratic component, the eigenvalues of which grow at a lower rate. This assumption is the so-called “pervasive condition” for a factor model, and it requires that the factor is “*sufficiently strong that most assets have nonnegligible exposure*”¹.

As we observed above, the three-pass procedure can be seen as an extension of the two-pass procedure and for this reason it can also be called a PC-augmented two-pass cross-sectional regression.

The main advantage of this model is that, due to the rotational invariance property and the use of principal components as control factors, it is possible to find the risk

¹ Giglio and Xiu

premia of the factors without an omitted variable bias and without a measurement error.

The three-pass procedure can also surprisingly be viewed as a version of the mimicking portfolio approach.

Therefore, the three-pass procedure can be interpreted as an extension of both methods even if they have different settings and they differ in samples.

To start the analysis, Giglio and Xiu write the model in matrix form and denote R as the matrix of excess returns, V as the matrix of factors, G as the matrix of observable factors, U as the matrix of idiosyncratic errors and Z as the matrix of measurement error.

Knowing the equations $r_t = \beta\gamma + \beta v_t + u_t$ and $g_t = \delta + \eta v_t + z_t$, that we saw before, we can write the model as:

$$R = \beta\gamma l_T^T + \beta V + U.$$

Using the matrices of the demeaned variables, this becomes:

$$\bar{R} = \beta\bar{V} + \bar{U}$$

The matrix for g_t in matrix form becomes:

$$\bar{G} = \eta\bar{V} + \bar{Z}$$

Given all this information, we can compute the three-pass procedure starting with the PCA.

3.2.1 Step 1: Principal Component Analysis

The first step of the three-pass procedure is the Principal Component Analysis (PCA). It is usually used to reduce the observable variables into a smaller set of components with the highest possible variation of the observed variables. This mechanism is called data reduction, and it is employed when researchers want to

explore the information in the observed variables without including all of them in the analysis. PCA is used to estimate the Principal Components, and it is based on the observed variables.

In this case, PCA is employed to extract from a panel of test asset returns the factors and their loadings.

The first component PC1 has the highest variance explained because it extracts the largest amount of information from the data. The second component PC2 will be orthogonal to the first and, considering the subspace orthogonal to the first component, it will have the highest variance. And so on for all the principal components that we extract.

Giglio and Xiu extract the principal components from the $T \times T$ matrix $n^{-1}T^{-1}\bar{R}^T\bar{R}$ and they estimate the matrix of factors and their loadings as

$$\hat{V} = T^{1/2}(\xi_1: \xi_2: \dots: \xi_{\hat{p}})^T \text{ and } \hat{\beta} = T^{-1}\bar{R}\hat{V}^T$$

Where $\xi_1, \xi_2, \dots, \xi_{\hat{p}}$ denote the normalized eigenvectors, and they correspond to the $\hat{p}$ largest eigenvalues of $n^{-1}T^{-1}\bar{R}^T\bar{R}$.

Then, Giglio and Xiu say to select the first $\hat{p}$ returns. The value of $\hat{p}$ can be found by the following formula:

$$\hat{p} = \arg \min_{1 \leq j \leq p_{max}} ((nT^{-1}) \lambda_j(\bar{R}^T\bar{R}) + j \times \varphi(n, T)) - 1$$

Where $\lambda_j(\bar{R}^T\bar{R})$ indicates the j -th eigenvalue of the matrix $\bar{R}^T\bar{R}$ and $\varphi(n, T)$ is a penalty term that depends on the number of assets and the period that we consider.

In the first step, hence, we can find the factors v using the rotational invariance with $H\bar{V}$ and we can normalize the estimated factor so that $\hat{V}\hat{V}^T = \mathbb{I}_{\hat{p}}$.

If instead, we extract the principal components from the $n \times n$ matrix $n^{-1}T^{-1}\bar{R}^T\bar{R}$, normalization appears as $\hat{\beta}^T\hat{\beta} = \mathbb{I}_{\hat{p}}$.

The two normalization methods give the same risk premia estimates.

3.2.2 Step 2: Cross-sectional Regression

The second step of the three-pass procedure is the cross-sectional regression. Employing this, we want to find the risk premia using only the principal components.

There are several methods to estimate the risk premia such as OLS (Ordinary Least Squares), GLS (Generalized Least Squares) and WLS (Weighted Least Squares). All methods are equivalent and give feasible results. Giglio and Xiu suggested using the OLS method because it is the easiest to implement and it works best in a scenario with a large n . GLS and WLS instead require a large number of parameters, and they do not improve the efficiency of OLS.

In the second step, we estimate the risk premia of v_t which are given by $\hat{\gamma} = H\gamma$.

We regress the average returns $\bar{R}$ onto the factor loadings $\hat{\beta}$ estimated in the first step. Therefore, we obtain:

$$\hat{\gamma} = (\hat{\beta}^T \hat{\beta})^{-1} \hat{\beta}^T \bar{r}$$

Where $\bar{r} = [\bar{r}_1 \ \bar{r}_2 \ \dots \ \bar{r}_n]'$ indicates the returns of the n testing assets.

The authors of the original paper also underline that in this paper the typical assumptions of large T and fixed n do not exist because we know that both T and n go to infinity.

The errors, and so the covariance matrix of u_t , in the standard case, as in Fama-MacBeth, matter at the order of $O_p(n^{-1} + T^{-1})$; instead, the term $\hat{\gamma}_g$ is of $O_p(T^{-1/2})$.

3.2.3 Step 3: Time Series Regression

The third step is given by the estimate of a time-series regression of g_t onto the principal components. This step, using the PCs, allows us to uncover the relation

between g_t and the latent factor and to eliminate the potential measurement error which standard approaches cannot do.

In this step we can obtain the estimator $\hat{\eta}$ and the fitted value of the observable factor $\hat{G}$:

$$\hat{\eta} = \bar{G}\hat{V}^T(\hat{V}\hat{V}^T)^{-1} \text{ and } \hat{G} = \hat{\eta}\hat{V}.$$

We can hence estimate the risk premium of g_t by combining the second and the third step because:

$$\hat{\gamma}_g = \hat{\eta}\hat{\gamma}.$$

The three-pass estimator can also be written in a compact form as:

$$\hat{\gamma}_g = \bar{G}\hat{V}^T(\hat{V}\hat{V}^T)^{-1}(\hat{\beta}^T\hat{\beta})^{-1}\hat{\beta}^T\bar{r}.$$

An important characteristic to consider is that even if g_t can be multidimensional, the estimation $\hat{\gamma}_g$ for each observable factor is separate and the estimation of a risk premium for a factor does not influence the estimation of the others.

3.3 Empirical tests

Giglio and Xiu in their paper showed some significant results and theoretical concepts, some of which are useful for the empirical work we computed. We are going to explore in this section how they computed the weak factor tests and provide some information about the Zero-Beta Rate and the Sign of the Market Risk Premium.

Giglio and Xiu, in their paper, report the analysis on the R^2_g which are the R^2 of the time-series regression of each factor g_t onto the latent factors. They found that the R^2_g is lower than 100% whenever there is a measurement error in the factor g_t . For the SMB and the market factors, R^2_g is close to 1, meaning that there is essentially no error in these factors. For other factors, particularly the non-tradable

ones, R^2_g is considerably lower, indicating that these are dominated by noise. For example, the R^2_g of IP is around 2%. The authors test the p-value to discover which factors are weak. A rejection of the null hypothesis of the test indicates that the factor is strong; they found that for several non-tradable factors they fail to reject the null.

Giglio and Xiu in their paper also applied the procedure of three-pass in the case where the zero-beta rate is estimated instead of being fixed at the treasury bill rate.

The zero-beta rate “*serves as a substitute for the unobservable risk-free rate, representing the expected return on a portfolio that bears no systematic risk*”²

They found that the results are similar to those of the baseline case where we have a zero-beta rate restriction. For this reason, we can affirm that the results obtained with the three-pass procedure are robust.

There is however a case that the authors of the original paper wanted to remark. In fact, they document a phenomenon which occurs in the empirical asset pricing literature under conditions of free zero-beta rate. The factor models such as the Fama-French three factor model and the Fama-MacBeth regressions produce a negative estimate of the market risk premium in the cross section of equity portfolios.

These results are also confirmed with an empirical study that Giglio and Xiu did where they obtained a zero-beta rate estimate of 129 bps per month and a market risk premium of -24 bps using the Fama-MacBeth estimator.

This puzzle is due to the omission of some relevant controls; in fact, the measurement error is unlikely because of the large R^2_g . The three-pass estimator solves this puzzle, and the estimator yields a positive market risk premium in all cases.

² Resolving the Zero-Beta Rate Puzzle by Wang Rong (2026)

Giglio and Xiu in their paper also did two other analyses.

The first analysis is about the mixed-frequency data which results are not significant because they do not improve the asymptotic efficiency of the three-pass procedure.

The second analysis is about the ridge regression approach used to regularize the mimicking-portfolio estimator. The test was consistent, but the authors found that the PCA baseline approach is more efficient; this confirms the good theoretical properties of the three-pass procedure.

Because of these limited properties of the two analyses, we chose not to apply these approaches to our dataset.

CHAPTER IV

DATA DESCRIPTION AND PRELIMINARY ANALYSIS

4.1 Dataset Description

For the empirical studies, I constructed a dataset with 25 European portfolios sorted on size (market equity ME) and book-to-market ratio (BE/ME) and the European Three Fama-French factors.

All the data were found in the *Kenneth-French Data Library*, and they cover a sample period from July 1990 to February 2026. Hence, the dataset contains 428 monthly observations ($T = 428$).

The choice of the 25 European portfolios instead of a larger panel as the 647 portfolios used by Giglio and Xiu is motivated by several factors.

First of all, the purpose was to implement and replicate the three-pass estimator method using PCA and it does not require many portfolios to give consistent results.

From the Giglio and Xiu analysis we can note that also with a small $n=50$, the three-pass estimator outperforms the two-pass Fama-Macbeth, provided that T is sufficiently large.

In our case, we have $T = 428$ observations, a large number that compensates for the small dimension of the cross-section.

Another motivation of the portfolio choice is given by the fact that the corresponding American dataset is one of the most used in asset pricing and factor model applications.

We wanted to find significant results using a simple and well-understood dataset without introducing additional sources with a more difficult interpretation.

Despite the smaller n , we found remarkable results with our empirical analysis. It confirms that the cross-sectional information was sufficient to satisfy the PCA, the bootstrap analysis, the weak factor test and the robustness check with different $\hat{p}$. We will see this in the next chapter.

The Three Fama-French factors, as explained in the previous chapter, give us information about the market risk-free (Mkt-RF), the size effect (small minus big SMB) and the value effect (high minus low HML) in stock returns.

The Kenneth-French factors dataset also contained the risk-free rate factor (RF); this variable was removed because it was redundant and it could add unnecessary collinearity; in fact, the dataset already includes the market risk-free factor (Mkt-RF).

The portfolios and the factors are merged by date and therefore our dataset has a structure of (428, 28): there are 428 rows as the number of monthly observations and 28 columns ['SMALL LoBM', 'ME1 BM2', 'ME1 BM3', 'ME1 BM4', 'SMALL HiBM', 'ME2 BM1', 'ME2 BM2', 'ME2 BM3', 'ME2 BM4', 'ME2 BM5', 'ME3 BM1', 'ME3 BM2', 'ME3 BM3', 'ME3 BM4', 'ME3 BM5', 'ME4 BM1', 'ME4 BM2', 'ME4 BM3', 'ME4 BM4', 'ME4 BM5', 'BIG LoBM', 'ME5 BM2', 'ME5 BM3', 'ME5 BM4', 'BIG HiBM', 'Mkt-RF', 'SMB', 'HML'].

In particular, the return matrix R has a dimension (428,25) and the factor matrix G has a dimension (428,3).

4.2 Descriptive statistics

The main descriptive statistics that we wanted to analyze are Mean, Standard deviation (std), Minimum (min) and Maximum (max). In the factor analysis, also the Skewness and the Excess Kurtosis turned out to be interesting.

Starting with the 25 size BM portfolios, we can observe from the following table the main characteristics. We know that the count of each portfolio is 428 which corresponds to the number of monthly observations.

	SMALL LoBM	ME1 BM2	ME1 BM3	ME1 BM4	SMALL HiBM
mean	0.182827	0.530841	0.638318	0.752547	0.970911
min	-24.47	-25.99	-26.85	-25.25	-26.73
max	18.76	16.97	19.19	17.78	18.56
std	5.568077	5.327742	5.108624	4.897400	4.890601

	ME2 BM1	ME2 BM2	ME2 BM3	ME2 BM4	ME2 BM5
mean	0.441706	0.684533	0.751916	0.860000	0.991425
min	-26.73	-25.92	-24.97	-27.38	-26.48
max	17.57	17.62	20.97	20.46	19.79
std	5.793505	5.376586	5.189321	5.200215	5.373276

	ME3 BM1	ME3 BM2	ME3 BM3	ME3 BM4	ME3 BM5
mean	0.573318	0.780187	0.730140	0.769673	1.001145
min	-27.19	-26.76	-27.03	-26.25	-27.35
max	18.52	15.23	20.47	19.79	22.87
std	5.877273	5.345220	5.291189	5.267613	5.628075

	ME4 BM1	ME4 BM2	ME4 BM3	ME4 BM4	ME4 BM5
mean	0.725047	0.770070	0.743925	0.799393	0.907477
min	-25.13	-24.28	-23.32	-25.27	-27.59

max	15.49	16.15	17.36	21.14	24.21
std	5.422897	5.124015	5.058515	5.462661	5.905149

	BIG LoBM	ME5 BM2	ME5 BM3	ME5 BM4	BIG HiBM
mean	0.626285	0.794416	0.770584	0.871565	0.873037
min	-20.15	-17.21	-18.47	-19.01	-31.28
max	18.06	14.64	17.66	20.19	26.61
std	4.833950	4.735991	5.204419	5.295988	6.430866

Table 4.1

Observing the mean, we can note that all the portfolios document positive monthly average excess returns over the sample period; it aligns with the trend in European equity markets. The values go from 0.183 in SMALL LoBM to 1.001% in ME3 BM5.

The standard deviation range was between 4.7% and 6.4% per month, generally higher standard deviation and volatility are associated with greater risk exposure.

The minimum values are extremely negative returns associated with periods of market stress like the Dot-com bubble, the Global Financial Crisis and the COVID-19 market crash.

Considering the factors, we have the following results:

Factor	Mean	Min	Max
Mkt-RF	0.5326%	-22.0200%	16.6200%
SMB	-0.0534%	-6.9600%	9.3900%
HML	0.3442%	-11.3000%	12.0900%

Factor	Std	Skewness	Kurtosis
Mkt-RF	4.9053%	-0.5094	1.4778

SMB	2.1192%	0.0054	1.0533
HML	2.5958%	0.2947	3.4262

Table 4.2

Looking at the factor means, we observe that the market risk-free factor and the HML have a positive mean while the SMB has a slightly negative mean. The negative average return suggests that the size premium may be absent in the European market during the sample period. This preliminary evidence will be confirmed with the risk premia estimation.

The risk-free market factor is the most volatile factor, in fact, it has the highest standard deviation (4.9053%) while the others are around 2%. The volatility analysis will be examined in greater detail in the next section.

Skewness is a measure of the degree of asymmetry in distributions, in particular, a value between -0.5 and 0.5 is commonly interpreted as an indicator of a symmetric distribution.

In our case, SMB has a skewness of 0.0054 which is practically zero; this means that the distribution is symmetric. Even the HML skewness is low (0.2947), and it falls in the range of symmetric distributions despite suggesting a slightly positive skewness.

In contrast, Mkt-Rf has a moderate negative skewness (-0.5094). It indicates a tendency toward a higher probability of having extreme negative returns rather than extreme positive returns.

Kurtosis compares the dataset distributions to the normal distribution, and it gives information about the thickness of the distribution tails. A positive value, as in this case, indicates a leptokurtic distribution, suggesting that the factors tend to have heavy tails. Therefore, it's more probable to have extreme observations than would be expected under normality. A positive kurtosis is typical in financial markets. Among the factors, HML has the greatest level of kurtosis (3.4262). Mkt-Rf and

SMB have excess kurtosis (respectively 1.4778 and 1.0533), but they are less pronounced than the HML one.

Another analysis that we computed about the factors is the correlation among the three factors. We observe this in Figure 4.1.

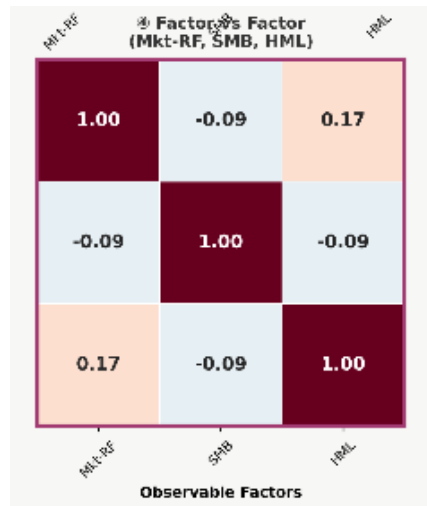


Figure 4.3

The three factors are not correlated with each other; in fact, all the correlations are low in absolute value (-0.09, -0.09, 0.17).

4.3 Volatility Analysis

To complement the statistics about the three Fama-French factors, we analyze in detail the risk profile along several dimensions: time-varying volatility, tail risk measures and regime-dependent return behavior.

In figure 4.2, we can observe the plot of the 12-month rolling annualized volatility for each factor over the sample period 1990-2026. The shaded regions underline the three major market stress episodes: the dot-com bubble and its consequences, the Global Financial Crisis (GFC) and the COVID-19 shock.

The market risk-free factor, consistent with its standard deviation, appears to be the most volatile factor with an average annual volatility of 15.41%. The time

series oscillates between 10% and 12% during a low-volatility regime, but it spikes during the crisis periods. The highest spike happens during the Global Financial Crisis in 2008-2009. It exhibits a peak of about 40%, which is much higher than that observed during any other crisis period. During the COVID shock in March 2020 the spike observed reaches nearly 30%. The increase in volatility during the dot-com correction of 2000-2002 is almost negligible if compared with the other crises. During the period between 2012 and 2020, we can observe greater stability in volatility at a low level, primarily due to the European Central Bank's monetary policy.

Considering the SMB factor, it exhibits an average annualized volatility of 6.91%. The size factor's volatility peaks during the dot-com bubble reaching nearly 16% while during the GFC and the COVID crisis, SMB volatility rises but remains well below the shocks observed for the Mkt-Rf.

The average annualized volatility of 7.41% of HML illustrates the propensity of the value strategy to exhibit large shifts in return patterns upon larger market changes. The peak is reached during the dot-com correction of 2000-2002 and it is about 20%. This peak is close to the peak observed during the COVID era while, during the Global Financial Crisis, we observed another peak but not so high compared to the others.

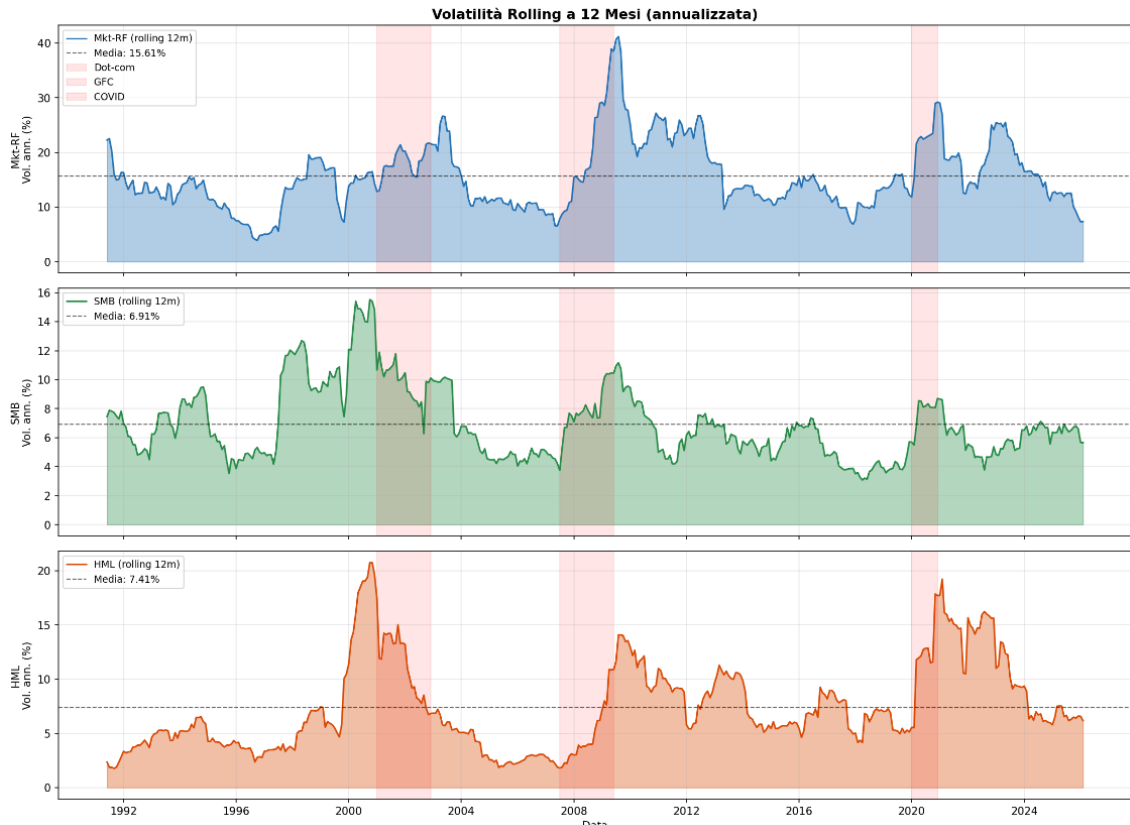


Figure 4.2

The empirical return distributions are used as a basis to estimate the risk measures, in order to calculate how much risk is associated with each source of return in the empirical data. We analyzed the risk measures VaR and CVaR for the three factors, as shown in Figure 4.3:

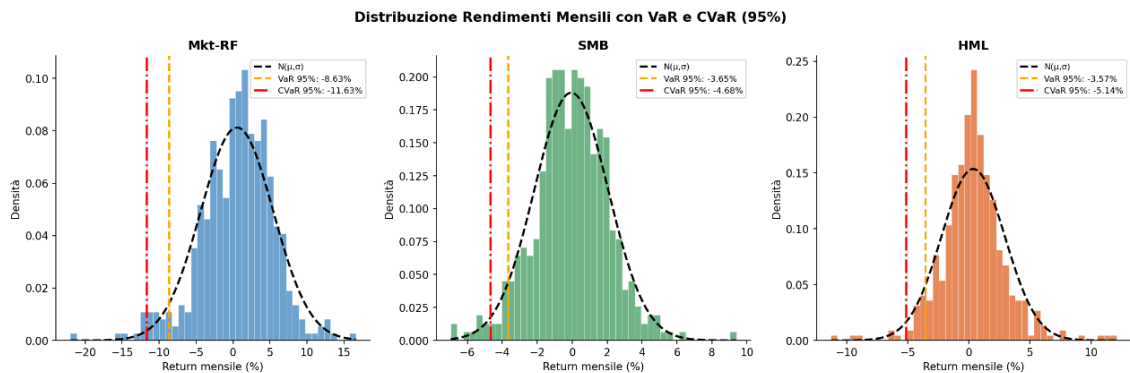


Figure 4.3

Value at Risk “*can be characterized as a maximum expected loss, given some time horizon and within a given confidence interval*” (Springer, 2017)

Conditional Value at Risk is a “*the expected value of the economic loss that is greater than the Value at Risk*” (ScienceDirect). It provides a more comprehensive view of risk compared to VaR.

The risks estimated at the 95% level for the VaR are -8.63% for Mkt-Rf, -3.65% for SMB and -3.57% for HML.

The values for the Conditional Value-at-Risk (CVaR) with 95% monthly are -11.63% for Mkt-Rf, -4.68% for SMB and -5.14% for HML. The ratio of CVaR to VaR is 1.35 for the market factor, 1.28 for the size factor and 1.44 for the value factor. This indicates that the tails are thick, especially the HML ones; these results are consistent with excess kurtosis seen in the descriptive statistics.

Figure 4.3 illustrates the return distribution of the three observable factors with VaR and CvaR at 95% threshold highlighted.

As expected, the return distribution of factor SMB is left-skewed which is reflected in the risk measures 95% VaR (-3.65%) and 95% CVaR (-4.68%). The Mkt-Rf distribution is left-skewed, and the worst realization reaches -22%. HML has a leptokurtic distribution, and it has heavy tails on both sides. SMB has the most normal shape.

The next figure illustrates the return behavior across volatility regimes; it classifies each month as a high-volatility month or a low-volatility month.

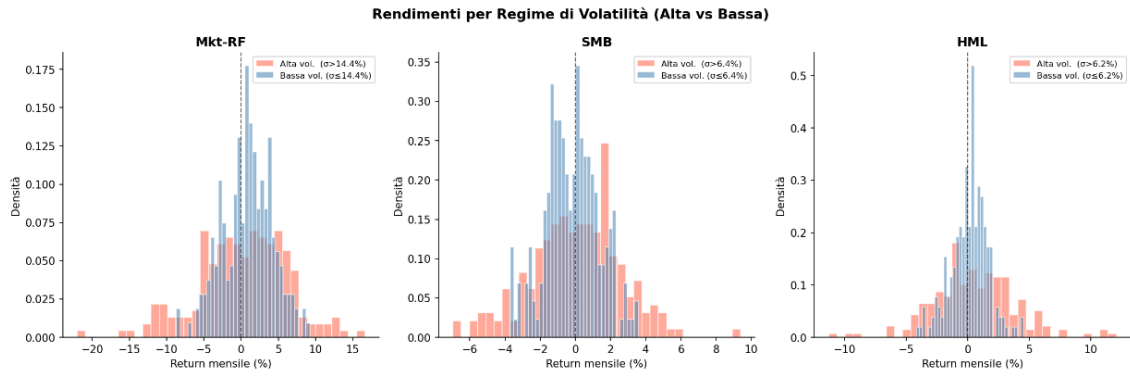


Figure 4.4

With the red distribution, we underline the high volatility while with a blue distribution we refer to the low volatility.

The thresholds are 14.4% for Mkt-Rf, 6.4% for SMB and 6.2% for HML. Below these values we have low volatility while above we have high volatility.

For Mkt-Rf, we can observe a more symmetric return distribution for high volatility while the low volatility is positive centered.

The SMB high volatility has a heavier left tail while the low volatility is more centered.

HML volatility has a bimodal distribution with heavy tails on both the tails.

4.4 Outlier Analysis

To analyze and detect the outliers of the factors, we considered three different and complementary methods: Z-score, IQR and Isolation Forest.

The Z-score method is a classic statistical method that works best when data have a bell-shaped curve. It considers the standard deviation of the mean and it treats as an outlier all the values lying beyond the ± 3 standard deviations. This is because in a normal distribution usually the 99.7% of the values lie within 3 standard deviations.

The Interquartile Range (IQR) method is non-parametric, and it does not need assumptions about the normal distribution, hence, it can also be used for skew distributions. The IQR focuses on the “core” of the distribution and not on the tails because it captures the middle 50%: $IQR = Q3 - Q1$ ³. It identifies the outliers via the IQR proximity rule, and it considers as outlier all the values below the lower bound⁴ and the values above the upper bound⁵.

Isolation Forest is an algorithm used especially in high-dimensional datasets. The method is based on the idea that anomalies can be separated, with few random splits, from the rest of the data.

Employing the three methods we identified 12 outliers with the Z-Score, 34 with the IQR and 22 with the Isolation Forest.

The observations are classified as consensus when outliers are found by at least two of three methods. The elements with consensus are 17.

In the following graph we can observe the boxplots of the *three factors with outliers* where the vertical axis represents the returns in percentage.

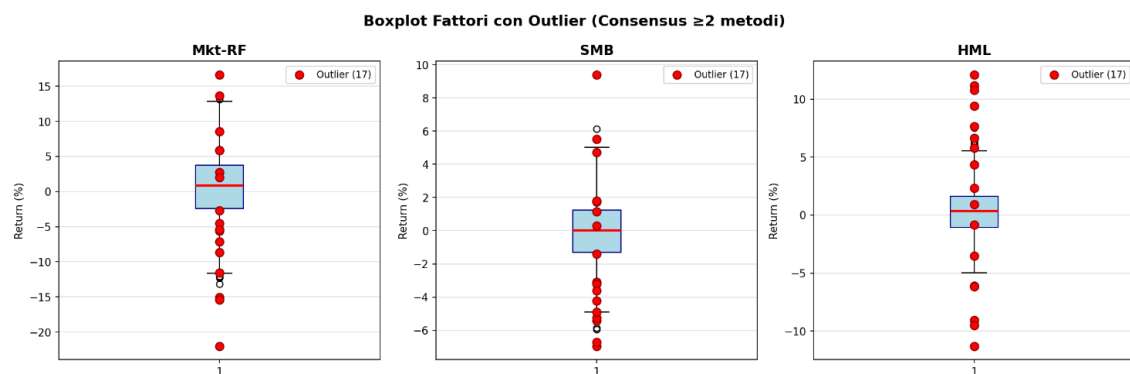


Figure 4.5

There are some important features of the factors' distribution. For Mkt-Rf, the downside level reaches -22.08% in October 2008 while the maximum does not

³ Q1 and Q3 are respectively the first (25th percentile) and third (75th percentile) quartile

⁴ Lower bound = $Q1 - (1.5 \cdot IQR)$

⁵ Upper bound = $Q3 + (1.5 \cdot IQR)$

exceed +16.62%. This reflects the negative skewness tendency, and it confirms that the negative market realizations are more likely than the positive ones.

Considering the SMB plot, we have a symmetric distribution around the box, consistent with the near-zero skewness. The interquartile range is between -6.96% and +9.39%.

In the HML plot, we can see that the extreme positive value is +12.09% and the most negative observation is -11.30% in March 2020. This is consistent with the skewness analysis.

In the following figure we can observe the time series of outliers:

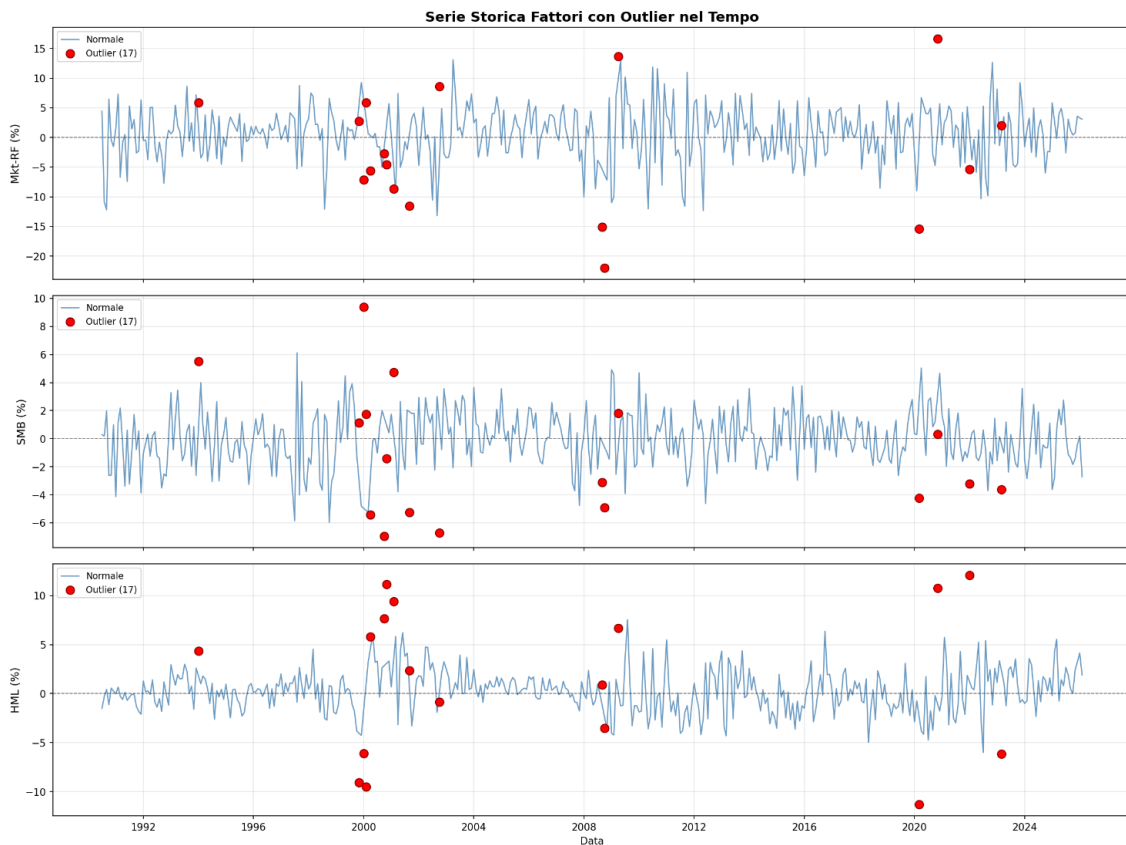


Figure 4.6

The time series shows the outlier distribution around the observed interval (1996-2026). We identified outliers especially in October 2008, between 2000 and 2003 and in the 2020-2022 period.

These observations are explained by economic phenomena. In the 1999-2003 window we have multiple values due to the dot-com bubble and its consequent collapse and due to the September 11 shock.

In 2008-2009, we observe a cluster corresponding to the Global Financial Crisis.

Between 2020 and 2022, there are some extreme observations because of Covid-19.

CHAPTER V

EMPIRICAL ANALYSIS

In this chapter we illustrate the main empirical results about the application of the Three-pass procedure to the dataset explained in the previous chapter. We will analyze the comparison with the Giglio and Xiu results and we will add additional analysis such as the weak factor test, the measurement error in observable factors, the robustness to the number of latent factors and the subsample analysis.

We started our analysis with a baseline of $\hat{p} = 7$; it means that we extracted 7 latent factors. This is the same baseline used by Giglio and Xiu.

5.1 Three-pass with $\hat{p}=7$

5.1.1 Step 1: Principal Component Analysis

The first step of the three-pass procedure is given by the principal component analysis. In the following figure, we can see the scree plot of variance explained by each PC.

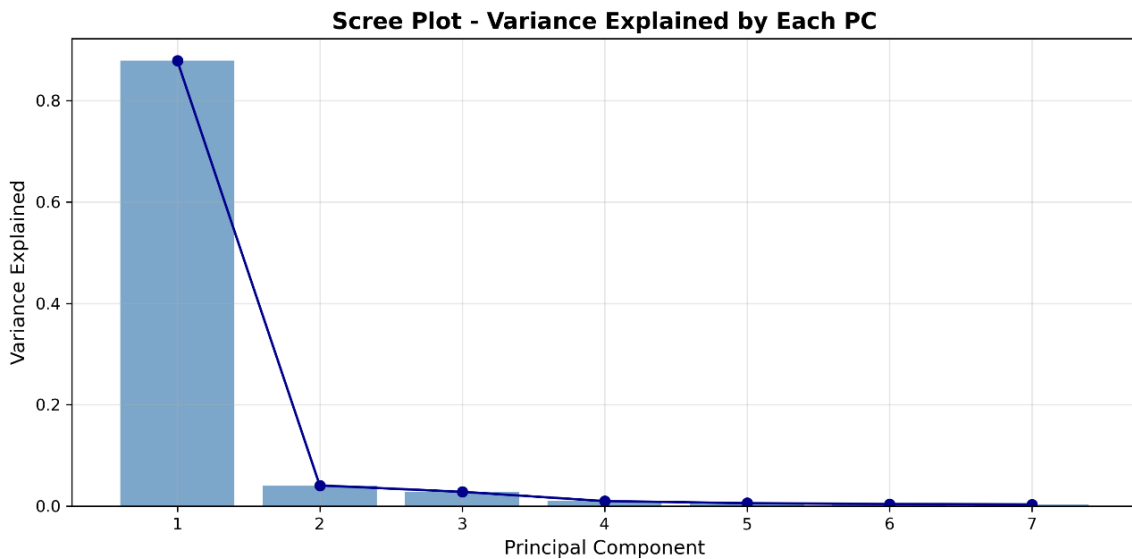


Figure 5.4

Using the PCA we find that the cumulative explained variance is 97.12%; it means that the seven components explain nearly all the variance.

The first principal component PC1 explains 87.87%, PC2 explains 4.09%, PC3 explains 2.82%, PC4 explains 0.99%, PC5 explains 0.59%, PC6 explains 0.41%, PC7 explains 0.34%.

PC1 alone explains 87.87% of the total variance which aligns with the canonical feature of equity returns panel: the first principal component has the highest variance explained and it extracts the highest amount of information from the data.

Another important observation is that each principal component from 4 to 7 explains less than 1% of the total variance. Their contribution is small, but they may still capture other sources of cross-sectional related to industry, country or liquidity effects.

Now, we can observe the correlation between the observed three factors and the PC factors.

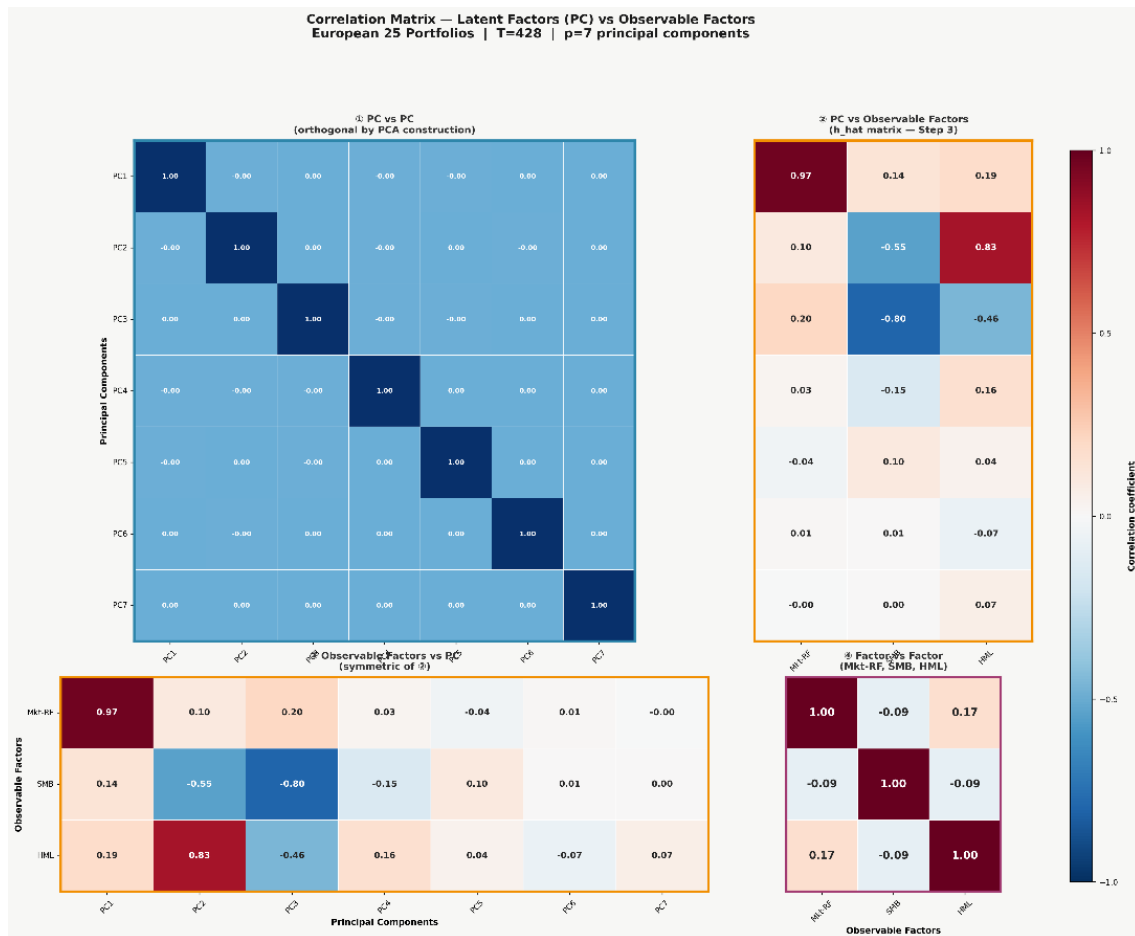


Figure 5. 5

In the first quadrant, we can observe the correlation between the seven principal components. PC factors have no correlation with each other; this is consistent with the theory of PCA. In fact, the extracted components are orthogonal: all the diagonal elements equal unity and all the off-diagonal elements are zero or negligible values. One of the prerequisites for the non-singularity of matrices which are involved in every step of the PCA implementation is satisfied in our case. This is a prerequisite for such steps of the three-pass procedure as Steps 2 and 3.

The second and third quadrants are symmetrical and they analyze the correlations between each PC and each observable factor; it's the core of our section. The highest correlation is +0.97 and it is between PC1 and Mkt-Rf. They correlate

almost perfectly. The correlations between PC1 and the other observable factors SMB and HML are negligible, being only +0.14 and +0.19.

PC2 is strongly correlated with HML (+0.83), and it is negatively correlated with SMB (-0.55) while the correlation with Mkt-Rf is negligible (+0.10). Therefore, PC2 identifies the value dimension of the European panel; it captures the spread between high book-to-market and low book-to-market.

PC3 is strongly negatively correlated with SMB (-0.80), and it has a negative correlation also with HML (-0.46) while its correlation with Mkt-Rf is not so high, as it's only +0.20. PC3 is the principal component that better defines the size factor, and we can say that it is the size factor of the return panel.

Looking at the other PCs, we can observe that there aren't significant correlations between PCs from 4 to 7 and the three factors. The positive correlations are below 0.16 and the negative ones are all above -0.15. This confirms that these PCs capture idiosyncratic cross-sectional variations which are orthogonal to the Three Fama-French risk dimension.

5.1.2 Step 2: Cross-Sectional Regression

The second step of the three-pass estimator is the cross-sectional regression. In this phase the vector of average excess returns contains the dependent variables that must be regressed against the matrix of PC betas.

The cross-sectional R^2 we found is 78.28% and it indicates that the seven principal components explain around three quarters of the dispersion in average portfolio returns.

We computed the risk premia of PC $\hat{\gamma}$ and found the following results:

PC1: 3.7023%

PC2: 0.6669%

PC3: -0.0198%

PC4: -0.3448%

PC5: -0.2010%

PC6: 0.1843%

PC7: 0.0961%

We can observe that the first PC is the one with the highest risk premium associated (3.7023% per month). Consistent with the literature, the first PC extracted from cross-section corresponds to the market risk.

The second PC gives a much smaller risk premium of 0.6669% and it reflects a size dimension or a value dimension hence it can be related to the SMB or HML Fama-French factor.

PCs 3 through 7 give negligible results, in fact, the values in absolute value are all nearly zero. The negative results can imply that these components carry a negative compensation or that maybe their correlation with priced risk is negative or this is due to their magnitude.

These results seem to be in line with the Giglio and Xiu analysis based on US data.

5.1.3 Step 3: Time-series regression

We found the Time-series R^2 (signal-to-noise ratio) for each factor R^2_g :

Mkt-RF: $R^2_g = 99.72\%$

SMB: $R^2_g = 98.78\%$

HML: $R^2_g = 96.94\%$

They measure the fraction of each factor's variance captured by the seven principal components. The results are particularly high; in fact, they are above the threshold

of 70%, typically used to indicate strong factors. Therefore, the observable factors are very well spanned by the PC structure, and they leave a minimal residual idiosyncratic noise.

We can make some hypotheses about the fact that the three factors are all strong factors; this will be confirmed in the next sections where we will see in detail the weak factor test.

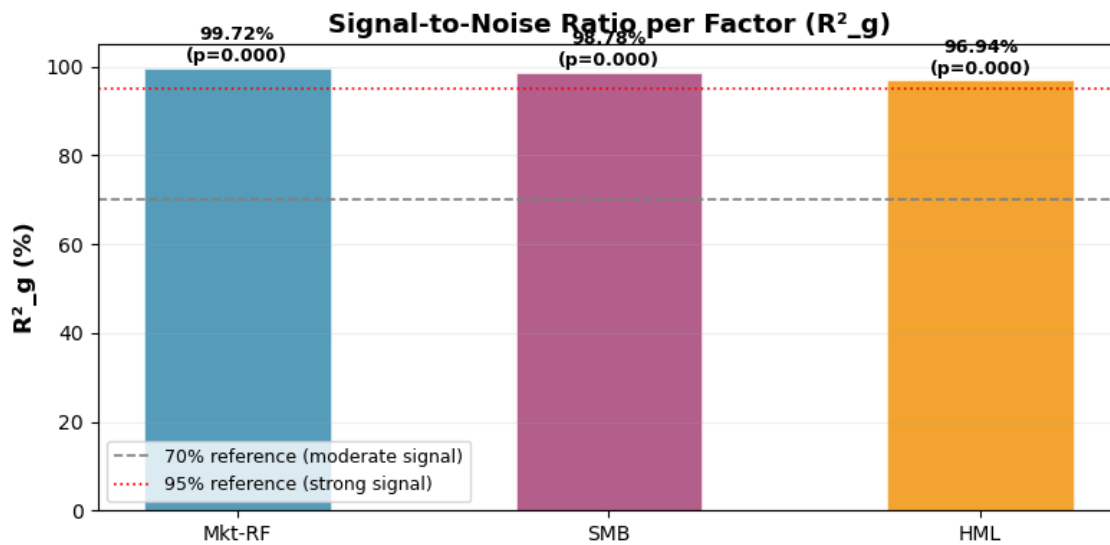


Figure 5.6

Then we found the three-pass risk premia estimates γ_g . We can see the results, reported in percentage per month, in the following table. We have also analyzed and compared the risk premia found employing the three-pass procedure to the time-series average.

Factor	Three-Pass	Time-Series Avg	Difference
Mkt-RF	0.7653%	0.5326%	0.2327%
SMB	-0.0699%	-0.0534%	-0.0165%
HML	0.2689%	0.3442%	-0.0753%

Table 5.3

The estimated market risk premium is 0.7653% per month and it exceeds the time-series average by 0.2327%.

The results for the size and the value factors are close using the two methods. The differences are only -0.0165% and -0.0753% respectively.

The discrepancy of the results is explained by the size of the dataset used ($n=25$ portfolios) and for $n \rightarrow \infty$ this vanishes asymptotically as explained by Giglio and Xiu in Proposition 1.

5.1.4 Bootstrap standard errors

From table 5.4, we can observe the three-pass risk premia estimates with bootstrap standard errors, t-statistics, p-value and the confidence intervals obtained from 1000 block bootstrap replication.

Factor	Three-Pass	SE	t-stat	p-value	CI Lower	CI Upper	Sig
Mkt-RF	0.7653%	0.2538%	3.016	0.0026	0.2680%	1.2627%	***
SMB	-0.069%	0.1152%	-0.607	0.5441	-0.296%	0.1559%	
HML	0.2689%	0.1893%	1.420	0.1555	-0.102%	0.6400%	

Table 5.4

SE = std of bootstrap replications | CI = estimate $\pm$ 1.96 $\times$ SE

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Considering the market risk premium, the bootstrap standard error is 0.2538% and it yields a t-stat of 3.016 which is significant for $p < 0.01$.

Mkt-Rf is the only factor to be significant. In fact, SMB and HML have a not so high t-stat and a high p-value hence, they are not significant.

5.2 Comparisons of three-pass procedure

5.2.1 Comparison with Giglio and Xiu results

In this section, we analyze the differences between the results obtained by the empirical analysis in section 5.1, which is computed, as we said, with European data, and the results found by Giglio and Xiu implementing the three-pass procedure.

	My empirical analysis (baseline $\hat{p}=7$)		Giglio and Xiu (baseline $\hat{p}=7$)	
Factors	γ	stderr	γ	stderr
Mkt-Rf	0.77***	(0.24)	0.51**	(0.23)
SMB	-0.07	(0.10)	0.20	(0.16)
HML	0.27**	(0.12)	0.20	(0.15)

Table 5.5

Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Considering the market premium, we can observe that there is a substantial difference between the results obtained from the two analyses. The European three-pass procedure estimates a risk premium of 0.77% per month while the US data give a much smaller result (0.51%). Therefore, the difference is approximately 0.26 percentage points per month. The estimates are both significant and the three *** indicate that they are statistically significant at the 1% level. The bootstrap standard errors are similar (0.24 and 0.23) and they are comparable despite the different sample sizes. These results align with the study of Dimson, Marsh and Staunton (2003), which analyzed the risk premia of 16 countries over the period 1950-2002. They found that the US risk premium is not representative of the global situation. There are some countries which have higher risk premia like France and Italy, even though the US risk premium is above the world average.

Observing the values of the SMB, we can note that there is a huge difference between the two regions. Giglio and Xiu found that the risk premium for the size

factor is positive (0.20%) while the analysis conducted on European data reveals a slightly negative result (-0.07%). It underlies that the size factor is not representative of the European context. This is confirmed also by Fama and French (2017) who show the risk premia of different countries over the period 1990-2015. These studies showed that in Europe, Japan and Asia Pacific the SMB risk premium was around 0, instead in North America it was slightly higher (0.17%).

For the HML factor, we can note that the results are similar in the American and European contexts, they are respectively 0.20% and 0.27% per month. Also their standard errors are similar (0.12 for European context and 0.15 for US context). These results are in line with those of other studies such as Fama and French and Ossola-Trifan.

5.2.2 Comparison with Fama MacBeth results

In this table we present the results found using the three-pass procedure (seen in the previous section), and the results of employing the Fama-MacBeth two-pass regression.

In the first step we estimated the betas and then in the second, we found the risk premia from the Fama-MacBeth regression: [-0.44176926; -0.11126159; 0.31283987].

	Three-Pass		Fama-MacBeth	
	(baseline $\hat{p}=7$)		(Two-Pass)	
Factors	γ	stderr	γ	stderr
Mkt-Rf	0.77***	(0.24)	-0.44	(0.41)
SMB	-0.07	(0.10)	-0.11	(0.10)
HML	0.27**	(0.12)	0.31**	(0.13)

Table 5.6

Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

Bootstrap SE: B=1000 iterations, seed=42

The results confirm the Giglio and Xiu study which affirms that the Fama MacBeth two-pass regression is biased in the presence of omitted factors. We can note, in fact, that the estimator of the market risk premium is -0.442% per month, economically absurd and it does not align with the results in the literature.

The results for SMB and HML seem similar using the two methods and also the standard deviations of the estimates are nearly identical.

In the following table we find the corrected results using the three-pass regression.

Factors	Three-Pass	Fama-MacBeth	Difference	Sign
Mkt-Rf	0.77***	-0.44	1.2071%	↑ (upward correction)
SMB	-0.07	-0.11	0.0414%	↑ (upward correction)
HML	0.27**	0.31**	-0.0440%	↓ (downward correction)

Table 5.7

We have a mixed pattern of correction: the market risk-free factor has an upward correction of 1.2071%, the SMB has an upward correction of 0.0414% and the HML has a downward correction of 0.0440% per month.

We also graphically computed the distribution of the three observable factors using both the three-pass estimator and the two-pass estimator. The distributions are visible in figure 5.4, where we have the histograms of the standardized estimates in simulations.

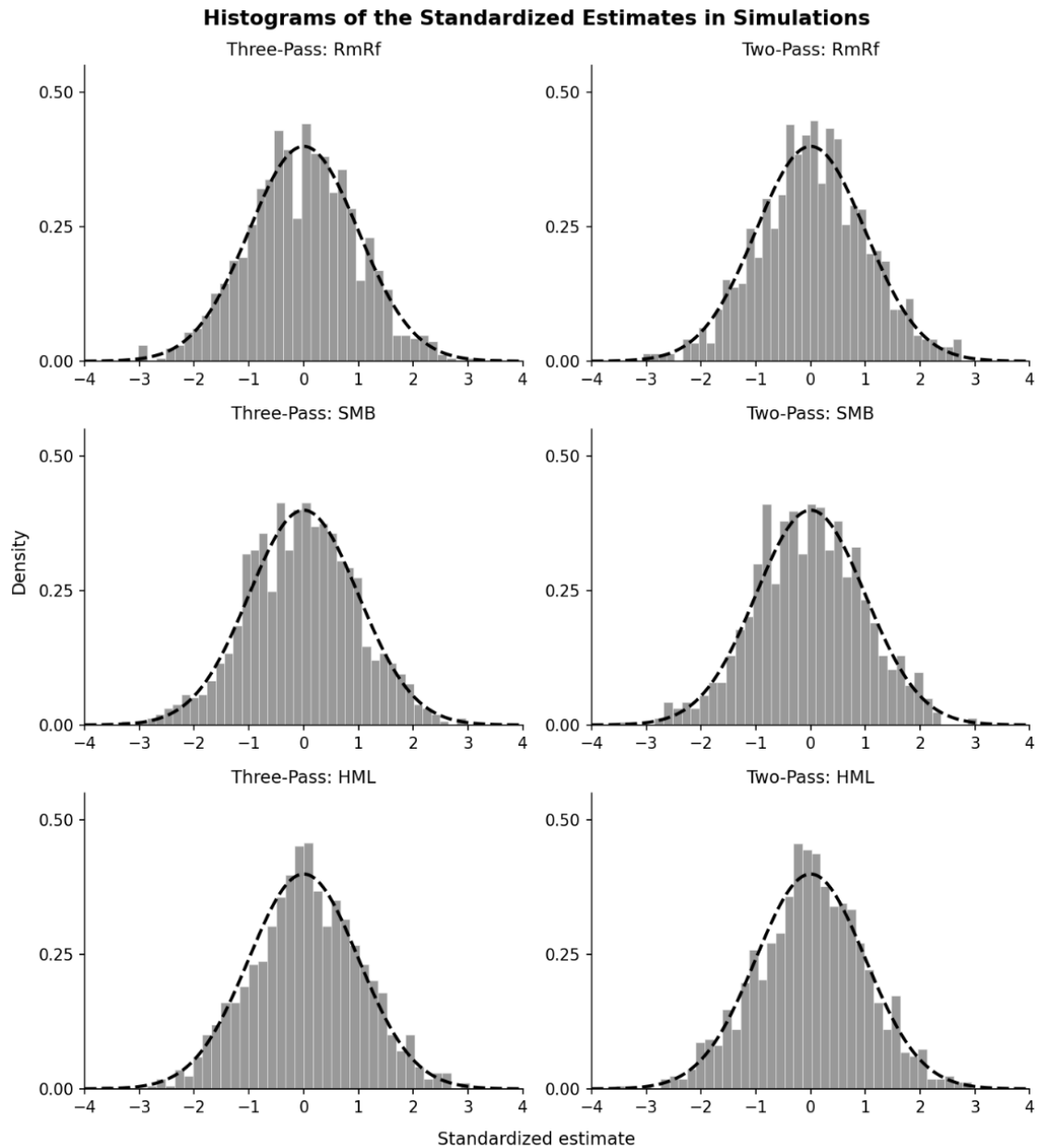


Figure 5.7

The difference between the distributions is not so evident. This is because the dataset has only $n=25$, although the distributions are well standardized using both methods.

In the Giglio and Xiu empirical analysis, instead, the difference between the two methods is more pronounced. We can observe from figure 5.5 that the three-pass estimator is the method that better fits the standardized distribution.

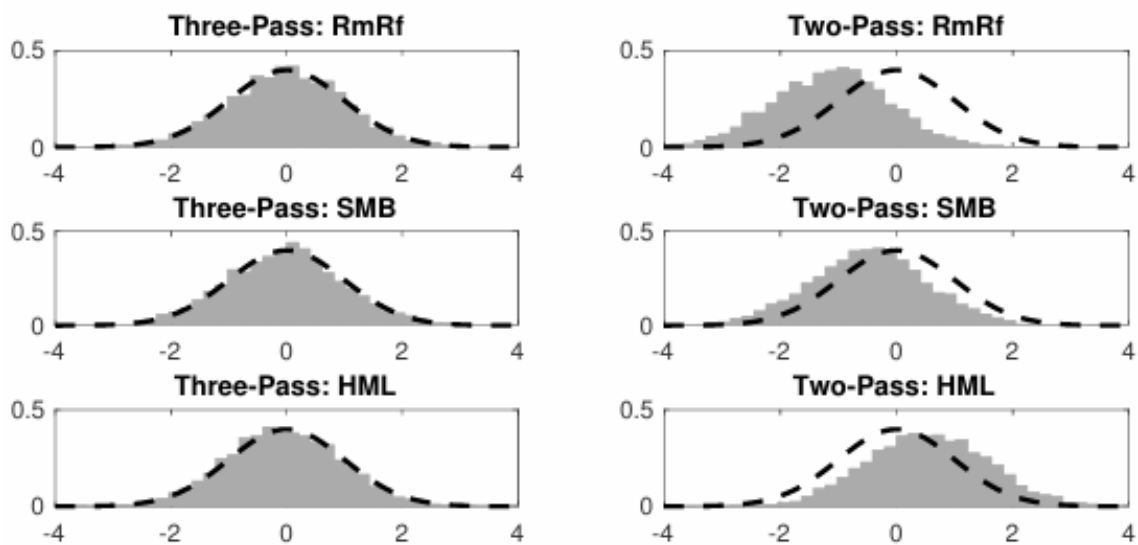


Figure 5.8 by Giglio and Xiu

5.3 Weak factor test

In this section, we analyze the weaknesses of the three observable factors. We considered the signal-to-noise ratio, the Wald statistic, which uses a χ^2 test with 7 degrees of freedom (because $\hat{p}=7$), and a p-value.

The hypothesis formulated for the Wald statistics is $H_0: h = 0$.

We obtained the following results:

Mkt-RF	
R^2_g (signal-to-noise ratio)	99.72%
Wald statistic ($\chi^2(7)$)	152386.6920
p-value	0.0000 ***
Conclusion	Strong factor

Table 5.8

This table suggests rejecting H_0 : Mkt-RF is NOT a weak factor.

SMB	
R^2_g (signal-to-noise ratio)	98.78%
Wald statistic (chi ² (7))	33983.9602
p-value	0.0000 ***
Conclusion	Strong factor

Table 5.9

This table suggests rejecting H_0 : SMB is NOT a weak factor.

HML	
R^2_g (signal-to-noise ratio)	96.94%
Wald statistic (chi ² (7))	13329.7893
p-value	0.0000 ***
Conclusion	Strong factor

Table 5.10

This table suggests rejecting H_0 : HML is NOT a weak factor.

In this table we can summarize the results obtained from the three observable factors.

Factor	R^2_g	Wald Statistic	df	p-value	Significant	Interpretation
Mkt-RF	0.9972	152386.6920	7	0.0000	Yes	Strong factor
SMB	0.9878	33983.9602	7	0.0000	Yes	Strong factor
HML	0.9694	13329.7893	7	0.0000	Yes	Strong factor

Table 5.11

Note: Wald $\sim \chi^2(\hat{p})$ under H_0 : $h = 0$ | *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

We observed that all the factors are strong because the null hypothesis is rejected for all the factors at the 1% level.

5.4 Measurement Error Analysis

Another analysis we wanted to compute is the measurement error analysis. The three-pass procedure allows for a decomposition of the observable factors into a systematic component, which is captured by the latent factor space, and a measurement error component.

To make visible the measurement error of the factors we created the following graphs. The red line indicates the cleaned factor, found after having denoised it, while the blue line indicates the original factor.

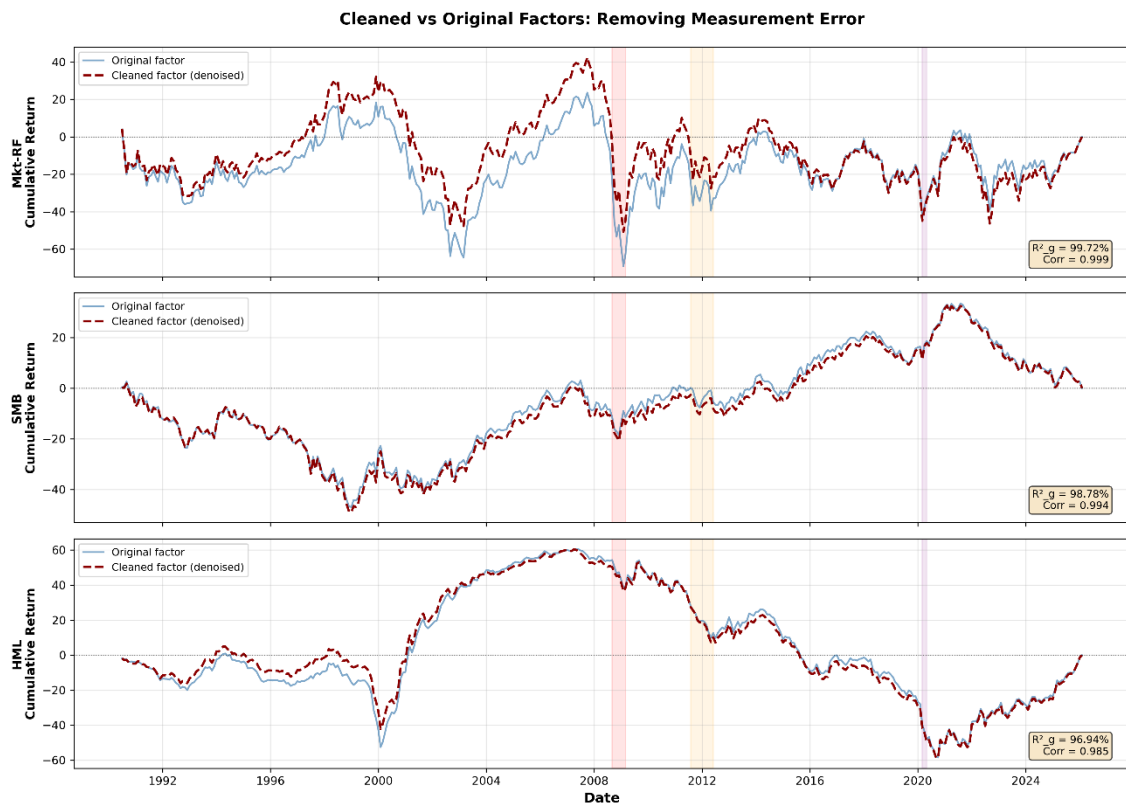


Figure 5.9

We know that when the blue line is above the red line, the original factor has a higher cumulative return than the systematic component alone. Therefore, the excess return is given by the idiosyncratic component and not the systematic component.

Instead, when the blue line is below the red line, we have the opposite situation: the higher cumulative return is coming from the systematic component.

In the graphs are also highlighted the three big periods of crisis from 1990 onward seen in the previous chapter: the dot-com bubble of 2000, the global financial crisis of 2008 and the COVID-19 crisis.

From the following tables, we can note in detail the data necessary for a measurement error analysis relative to the three factors like the original variance, the cleaned variance, the noise variance, the signal-to-noise ratio, the variance explained and the correlation between the original and cleaned factors.

Mkt-RF	
Original variance	24.062399
Cleaned variance	23.996104
Noise variance	0.066294
Signal-to-noise ratio	361.9636
R^2_g (variance explained)	0.9972
Correlation (orig-clean)	0.9986

Table 5.12

SMB	
Original variance	4.490933
Cleaned variance	4.435979
Noise variance	0.054954
Signal-to-noise ratio	80.7220
R^2_g (variance explained)	0.9878
Correlation (orig-clean)	0.9939

Table 5.13

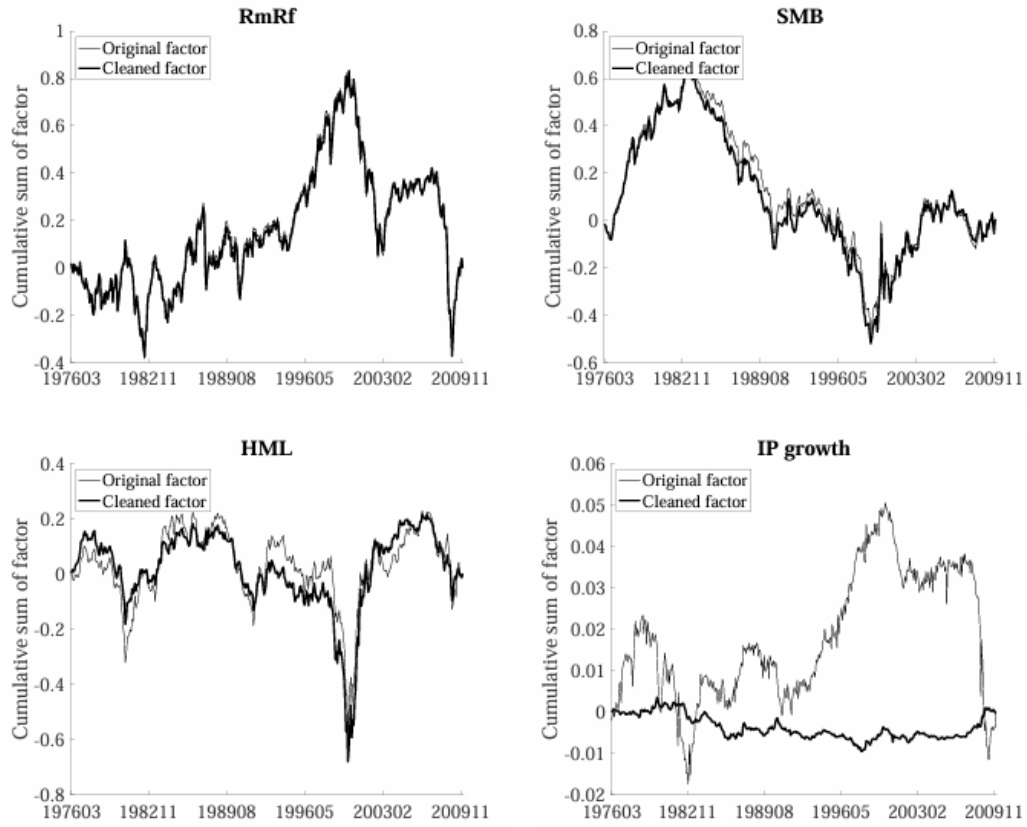
HML	
Original variance	6.738226
Cleaned variance	6.531925
Noise variance	0.206300
Signal-to-noise ratio	31.6622
R^2_g (variance explained)	0.9694
Correlation (orig-clean)	0.9846

Table 5.14

The high values of R^2_g and of correlation make clear that the two lines are close and that the three observable factors are almost entirely systematic, they do not have high idiosyncratic noise.

In the next figure, we report the time series about the measurement error for the Fama-French factors found by Giglio and Xiu. Our results seem consistent with the Giglio and Xiu's results. In fact, we can observe that the differences between the line of original and cleaned factors are low.

Another common feature between our study and that of Giglio and Xiu is given by the value of R^2_g . The variance explained, found by Giglio and Xiu, is high for all three factors: 99.57% for Mkt-Rf, 97.24% for SMB, 83.03% for HML.



Note: This figure reports the time series of cumulative factor innovations for RmRf, SMB, HML, and IP (thin line) together with the time series obtained from removing measurement error from the factor (thick line).

Figure 5.10 by Giglio and Xiu

5.5 Robustness of Risk Premia

In this section we analyze the robustness to the number of latent factors. We can observe from table 5.15 how the risk premia and the R^2_g change with the number of $\hat{p}$.

In the previous sections we chose $\hat{p}=7$, exactly as Giglio and Xiu did in their studies but we wanted to underline this important concept that Giglio and Xiu state: “*As long as our selected number of factors, denoted by $\hat{p}$, is greater than or equal to p , yet is not too large relative to n and T , then the risk premia estimator based on $\hat{p}$, denoted by $\hat{\gamma}_g$, remains consistent*” (Giglio and Xiu (2021)).

Risk Premia across $\hat{p}$:

Factor \ $\hat{p}$	3	5	7	10
Mkt-RF	0.7574%	0.7613%	0.7653%	0.7631%
SMB	-0.0924%	-0.0719%	-0.0699%	-0.0621%
HML	0.3427%	0.2772%	0.2689%	0.3071%

Table 5.15

Given the results from the table, we can observe that the market return is very stable. In fact, its risk premium range is [0.7574%, 0.7653%] and the coefficient of variation (CV) is 0.004; it is essentially invariant to the number of latent factors.

The SMB range is [-0.0924%, -0.0621%] and its CV=0.151. It maintains the negative sign through all the choices of $\hat{p}$, and it is reasonably stable.

The HML range goes from 0.2689% to 0.3427% and it is very stable with a CV=0.097.

The stability of the three-pass risk premia to the choice of $\hat{p}$ can be observed graphically in figure 5.8.

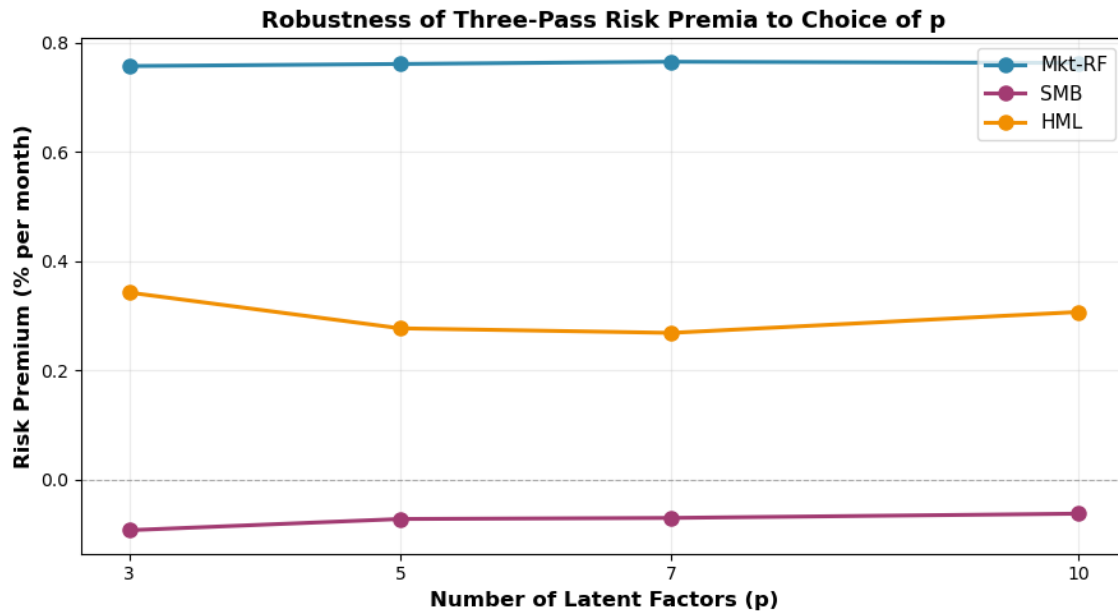


Figure 5.11

R^2_g across $\hat{p}$:

Factor \ p	3	5	7	10
Mkt-RF	99.45%	99.72%	99.72%	99.76%
SMB	95.51%	98.77%	98.78%	98.88%
HML	93.20%	96.00%	96.94%	97.98%

Table 5.16

From table 5.16, we can note that the R^2_g is higher and it grows with the increase of $\hat{p}$. This is valid for all three observable factors.

5.6 Subsample Analysis

In this section we analyze the risk premia across periods; in particular, we focus on two economically relevant events: the Global Financial Crisis of 2008 and the resolution of the European Sovereign Debt Crisis of 2012.

In table 5.17 we consider the risk premia of the three Fama-French factors for the full sample (T=428), for the period Pre-2008 Crisis (T=210), Post-2008 Crisis (T=218), Pre-Euro Debt Crisis (T=258) and Post-Euro Debt Crisis (T=170).

Factor	Full Sample	Pre-2008 Crisis	Post-2008 Crisis	Pre-Euro Debt Crisis	Post-Euro Debt Crisis
Mkt-RF	+0.7653	+0.9059	+0.6075	+0.6736	+0.9142
SMB	-0.0699	-0.1189	-0.0115	-0.1116	-0.0111
HML	+0.2689	+0.5116	+0.0234	+0.3549	+0.1890

Table 5.17

In table 5.18 we focus on the comparison Pre and Post 2008.

Factor	Pre-2008	Post-2008	Change	% Change
Mkt-RF	0.9059%	0.6075%	-0.2984%	-32.9%
SMB	-0.1189%	-0.0115%	0.1074%	90.4%

HML	0.5116%	0.0234%	-0.4882%	-95.4%
-----	---------	---------	----------	--------

Table 5.18

The market risk premium is positive, and it ranges from 0.6075% in Post-2008 Crisis to 0.9142% in Post-Euro Debt Crisis. We note a decline of 32.9% from pre-2008 to post-2008 crisis. This is consistent with the compressed equity premium which characterizes the post Global Financial Crisis.

The size factor is negative throughout all periods, and it confirms that the size premium is not relevant in Europe.

The value factor shows a decline of 95.4%, it goes from 0.5116% in the pre-2008 period to 0.0234% per month in the post-2008 period.

Figure 5.9 shows the four subsample comparisons for each factor, and it indicates with a line the full sample level. For Mkt-Rf, the bars are all positive and close to the full sample level. For SMB, the bars are negative and close to zero, especially post-2008 and post-euro debt crisis. For HML the samples are positive and we want to highlight the difference between the value pre-2008 and the value post-2008.

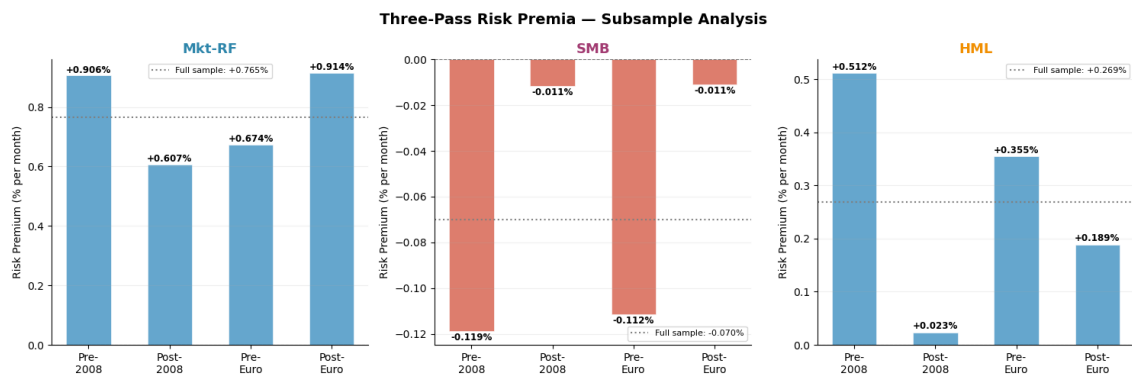


Figure 5.12

Figure 5.10 is a heatmap which allows us to visualize the subsamples of the three factors across 5 different periods. For Mkt-Rf, we can observe that it is uniformly red and it confirms the significance of the market risk-free factor. SMB is

uniformly blue across the periods, and it is negative and near zero. For HML, we can note the crisis periods and the recovery by the gradient from red to white.

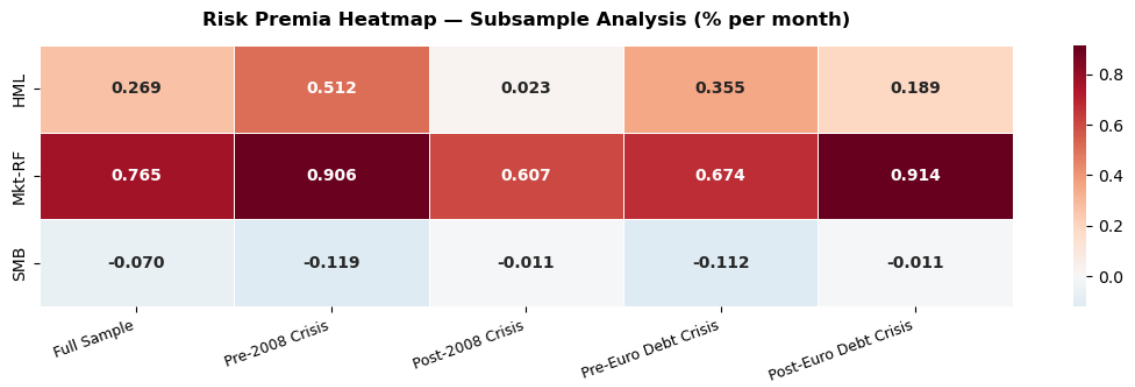


Figure 5.13

CHAPTER VI

CONCLUSIONS

6.1 Summary of the results

The objective of the thesis is to implement the three-pass procedure proposed by Giglio and Xiu in the paper “Asset pricing with omitted factors” of 2021 to find the risk premia of the factors on a European dataset.

We evaluated the results obtained by applying the three-pass procedure and we compared them with the results obtained by Giglio and Xiu and with those that we would have obtained using the two-pass procedure.

The empirical analysis was conducted on 25 European portfolios sorted on size and on book-to-market ratio and three European Fama-French factors and it covers a period of 428 monthly observations from July 1990 to February 2026.

The first step to obtain the three-pass estimator is Principal Component Analysis (PCA), and it was initially conducted using $\hat{p} = 7$, exactly as the original paper did. We found that the cumulative explained variance is 97.12%, so the seven components explain nearly all the variance. In particular, the first principal component explains the highest portion of the variance: 87.87% and it is highly correlated with the Mkt-Rf factor (+0.97). The correlation between PC1 and the other observable factors, instead, we found, was negligible. PC2 identified the value dimension of the European panel with a correlation with HML of +0.83. PC3, instead, is the best indicator of the size factor; in fact, the correlation between PC3 and SMB is very high even though negative (-0.80). The other PCs do not give significant results.

The second step is the cross-sectional regression, and, in this phase, we found a cross-sectional $R^2 = 78.28\%$, indicating that the seven principal components explain more than three quarters of the dispersion in average portfolio returns. The

highest risk premium $\hat{\gamma}$ is 3.7023% and it is associated with the first principal component; the other risk premia from PCs 3 to 7 are negligible.

The third step of time-series regression gives high signal-to-noise ratio results ($R^2_g = 99.72\%$ for Mkt-Rf, $R^2_g = 98.78\%$ for SMB, $R^2_g = 96.94\%$ for HML). The three factors are all above the threshold of 70% used to indicate strong factors. This means that the observable factors are very well spanned by the PC structure.

The risk premia found with the three steps are 0.7653% per month for the market factor, -0.0699% per month for the SMB factor and 0.2689% per month for the HML factor. Analyzing the bootstrap standard errors, we found that the market risk premium is the only significant factor; it is statistically significant at the 1% level. SMB and HML are not significant; in fact, the t-stat and the p-value are lower than the threshold.

These results were compared with those of Giglio and Xiu estimated with US data (0.51% for the market factor and 0.20% for the SMB factor and the HML factor). The comparison between the US and the European risk premia highlighted that the market factor in Europe is higher than in the US, the size factor gives a negative premium in Europe, and the HML premium is in line with the US one. These differences are in line with the international asset pricing literature such as that of Dimson, Marsh and Staunton (2003), which analyzes the heterogeneity of equity risk premium across countries, and that of Fama and French which documented the near-zero size premium outside North America.

The results were also compared with those obtained employing the Fama-MacBeth two-pass regression and highlighted the bias in the presence of omitted factors. In particular, the result for the market factor looked absurd with Fama-MacBeth while SMB and HML estimates were very similar with both methods.

Analyzing the robustness of risk premia using different $\hat{p}$ (3, 5, 7 and 10 latent factors) we found that R^2_g increases with the number of $\hat{p}$ and the coefficients of

variation were 0.004 for the market return, 0.151 for the SMB and 0.097 for the HML.

The last analysis we computed is the subsample analysis, where we focused on two economically relevant events: the Global Financial Crisis of 2008 and the resolution of the European Sovereign Debt Crisis of 2012. We showed that the market premium and the HML premium have positive values with some declines in the crisis periods. The size factor, instead, is negative throughout all periods; therefore this analysis reinforced the conclusion that the size factor is not relevant in European markets over the sample period.

6.2 Limitations

Despite the results obtained by applying the three-pass procedure by Giglio and Xiu were consistent, this study has some limitations.

The first limitation is the cross-sectional dimension. For the analysis, in fact, we used a relatively small cross-section because $n = 25$ portfolios and compared to the 647 portfolios used by Giglio and Xiu, it is much smaller. The large $T = 428$ compensates for the small n used and the risk premia are consistent, but a large cross-section would give more precise results because it can estimate the latent factors better and identify weak or omitted factors.

Another limitation is given by the choice of the observable factors. We decided to restrict the analysis using only the three Fama-French factors. We could instead extend the analysis to the five Fama-French factors and so add also RMW and CMA or maybe add some macroeconomic factors like Giglio and Xiu did.

6.3 Future research directions

The future research directions should solve the limitations of the study.

For this reason, the analysis should be extended using a larger European panel dataset. This would expand the cross-section, and we could also test how some specific additive factors react in European economies.

With a higher n , we could also try to find a greater number of latent factors $\hat{p}$ and observe what would happen. In this way, we could also replicate all the robustness tests employed by Giglio and Xiu with European data and also add some analysis done by the authors of the three-pass procedure that we chose not to replicate. For example, we could add to our work the zero-beta rate and the ridge regression and use mixed-frequency data.

Another extension of this study would be the incorporation of the Fama-French five-factor model or the addition of other factors such as momentum, profitability and investment.

We could also explore other methods of analysis for weak factors and bootstrap or other tests to validate the strength of the factors.

We could also expand the analysis by observing in detail the European countries and the differences between them, and by examining how the strength of the three factors differs across individual European markets.

REFERENCES

- Allee, K. D., Do, C., & Raymundo, F. G. (2022). *Principal component analysis and factor analysis in accounting research* (Working Paper). University of Arkansas, Walton College of Business.
- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1), 135–171.
- Black, F., Jensen, M. C., & Scholes, M. (1972). The Capital Asset Pricing Model: Some empirical tests. In M. C. Jensen (Ed.), *Studies in the Theory of Capital Markets* (pp. 79–121). Praeger.
- Corporate Finance Institute. *Fama-French three factor model*.
<https://corporatefinanceinstitute.com/resources/valuation/fama-french-three-factor-model/>
- Danthine, J.-P., & Donaldson, J. B. (2015). *Intermediate financial theory* (3a ed.). Academic Press/Elsevier.
- Dimson, E., Marsh, P., & Staunton, M. (2003). Global evidence on the equity risk premium. *Journal of Applied Corporate Finance*, 15(4), 27–38.
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.
- Fama, E. F., & French, K. R. (2017). International tests of a five-factor asset pricing model. *Journal of Financial Economics*, 123(3), 441–463.
- Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3), 607–636.
https://www.econ.sdu.edu.cn/__local/4/6F/F6/4D025D9B2F79E97BF52BE884B93_94C78037_2DAEA4.pdf

Ferson, W. E., Siegel, A. F., & Xu, P. (2006). Mimicking portfolios with conditioning information. *Journal of Financial and Quantitative Analysis*, 41(3), 607–635.

French, K. R. (n.d.). *Kenneth R. French – Data library*. Tuck School of Business at Dartmouth.
https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Giglio, S., & Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, 129(7), 1947–1990.

Hatem, G., Zeidan, J., Goossens, M., & Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. *BAU Journal – Science and Technology*, 3(2), 1–7.

Investopedia. *High minus low (HML)*.
https://www.investopedia.com/terms/h/high_minus_low.asp

Jolliffe, I. T. (2002). *Principal component analysis* (2a ed.). Springer-Verlag.

Kim, J. (2023). Signal-to-noise ratio in the big data era. *Medium*.
<https://medium.com/@jaekim8080/importance-of-checking-the-signal-to-noise-ratio-in-the-big-data-era-9fbdea99ad01>

Kolenikov, S., & Angeles, G. (2004). *The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices* (Working Paper WP-04-85). MEASURE Evaluation, Carolina Population Center, University of North Carolina at Chapel Hill.

Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2), 175–194.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1), 13–37.

Marghani, M. (2026). Isolation forest explained: The fastest way to spot outliers. *Medium*. <https://medium.com/@muskanmarghani13/isolation-forest-explained-the-fastest-way-to-spot-outliers-4a1285c91a3f>

More, P. (2025a). Outlier detection with the IQR method: A complete guide. *Medium*. <https://medium.com/@morepravin1989/outlier-detection-with-the-iqr-method-a-complete-guide-c0199bbc10bd>

More, P. (2025b). Outlier detection using the z-score method in Python. *Medium*. <https://medium.com/@morepravin1989/outlier-detection-using-the-z-score-method-in-python-ccfba52cb868>

Olson, D. L., & Wu, D. (2017). Value at risk models. In *Enterprise Risk Management Models* (2a ed., cap. 6, pp. 75–87). Springer.

Ossola, E., & Trifan, I. (2026). The risk premia from the European equity market: An application of the three-pass estimation methodology. *International Journal of Financial Studies*, 14(4), Articolo 96.

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

ScienceDirect Topics. *Conditional* *value-at-risk*.
<https://www.sciencedirect.com/topics/mathematics/conditional-value-at-risk>

Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial Studies*, 5(1), 1–33.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3), 425–442.

Wang, R. (2026). Resolving the Zero-Beta Rate Puzzle. *Job Market Paper* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4942202

Womack, K. L., & Zhang, Y. (2003). *Understanding risk and return, the CAPM, and the Fama-French three-factor model* (Tuck School of Business Working Paper n. 03-111).