Linear Factor Models
Theory, Applications and Pitfalls[1]

Attilio Meucci
attilio.meucci@symmys.com

this revision: December 07 2014
latest revision of code and article at symmys.com/node/336

Abstract

We clarify the rationale and differences between the two main categories of linear factor models, namely dominant-residual and systematic-idiosyncratic.

We discuss the five different, yet interconnected areas of quantitative finance where linear factor models play an essential role: multivariate estimation theory, asset pricing theory, systematic strategies, portfolio optimization, and risk attribution.

We present a comprehensive list of common pitfalls and misunderstandings on linear factor models.

An appendix details all the calculations. Supporting code is available for download.

JEL Classification: C1, G11

Keywords: generalized r-square, dimension reduction fundamental factor models, macroeconomic factor models, factor analysis, regression, random matrix theory, GICS industry classification, cross-sectional models, time-series models, statistical models

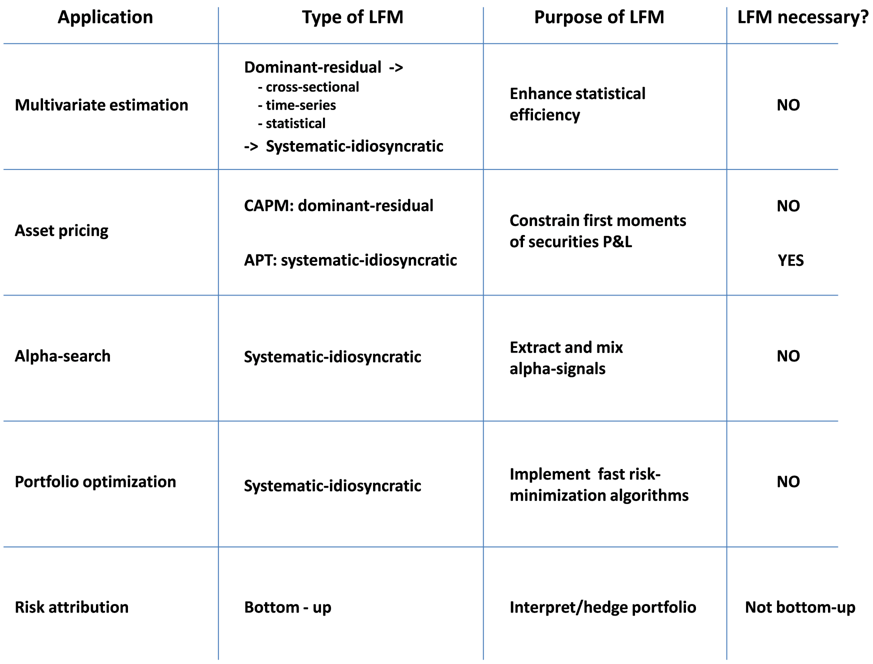
[1] The author is grateful to Eva Chan, Marcello Colasante, Erkko Etula, Sridhar Gollamudi, John Hall, Mutin Ren, Tai-Ho Wang.

Contents

1 Introduction

Linear factor models (LFM's) play a key role in five broad areas of finance: multivariate estimation, asset pricing theory, alpha search, portfolio optimization, and risk attribution/hedging, see Figure 1.

In this article we present an in-depth critical review of both the theory and the applications of LFM's. Then we present the numerous pitfalls that lurk behind these apparently simple models. As a result of our discussion, we also question the very necessity of LFM's and point in the direction of better modeling alternatives.



Application	Type of LFM	Purpose of LFM	LFM necessary?
Multivariate estimation	Dominant-residual -> - cross-sectional - time-series - statistical -> Systematic-idiosyncratic	Enhance statistical efficiency	NO
Asset pricing	CAPM: dominant-residual APT: systematic-idiosyncratic	Constrain first moments of securities P&L	NO YES
Alpha-search	Systematic-idiosyncratic	Extract and mix alpha-signals	NO
Portfolio optimization	Systematic-idiosyncratic	Implement fast risk-minimization algorithms	NO
Risk attribution	Bottom - up	Interpret/hedge portfolio	Not bottom-up

Figure 1: Overview of Linear Factor Models.

In Section 2 we provide a detailed account of the theory of LFM's. In particular, we discuss two distinct classes of LFM's that are often incorrectly identified as one single class. The first class are dominant-residual LFM's, which include time-series, cross-sectional, and statistical LFM's as special cases. As we shall see, these three sub-classes follow naturally from the same common generalized maximization principle. The second class, distinct from dominant-residual LFM's, are systematic-idiosyncratic LFM's.

In Section 3 we focus on the five applications of LFM's in finance. In multivariate estimation theory, LFM's are dimension-reduction techniques to obtain statistically efficient estimates in large markets. In the theory of asset pricing, CAPM and APT are constraints on the first moments of global LFM's for all the securities in the market. In the design of quantitative strategies, LFM's can be used to extract from the market and mix predictive signals that yield "alpha". In portfolio optimization, LFM's induce a structure on the securities covariances that allows for the implementation of fast risk minimization algorithms. In risk attribution, LFM's allow a manager to best interpret and analyze his projected P&L, or hedge the exposure to given risk factors. In each application we point out that, to the surprise of many, almost none of the above applications need LFM's and actually better implementation can be achieved without LFM's.

Often times confusion arises on the variables to which LFM's apply, such as past returns or log-changes in implied volatility, or future projected P&L's. Confusion also arises on the assumptions on which LFM's rely, such as the independence of the variables across time, or among the simultaneous residuals. This leads to suboptimal implementations of quantitative financial models. In Section 4 we analyze these pitfalls.

In the appendix we prove all the technical results. The code for the empirical case studies in this paper and additional case studies is available for download at symmys.com/node/336.

2 Theory of LFM's

In this section we define LFM's and we discuss the most important features of different classes of LFM's. To better grasp the theory, we illustrate the main points and caveats by means of easy-to-interpret low-dimensional simplified examples.

There exist three broad categories of LFM's, see Figure 2.

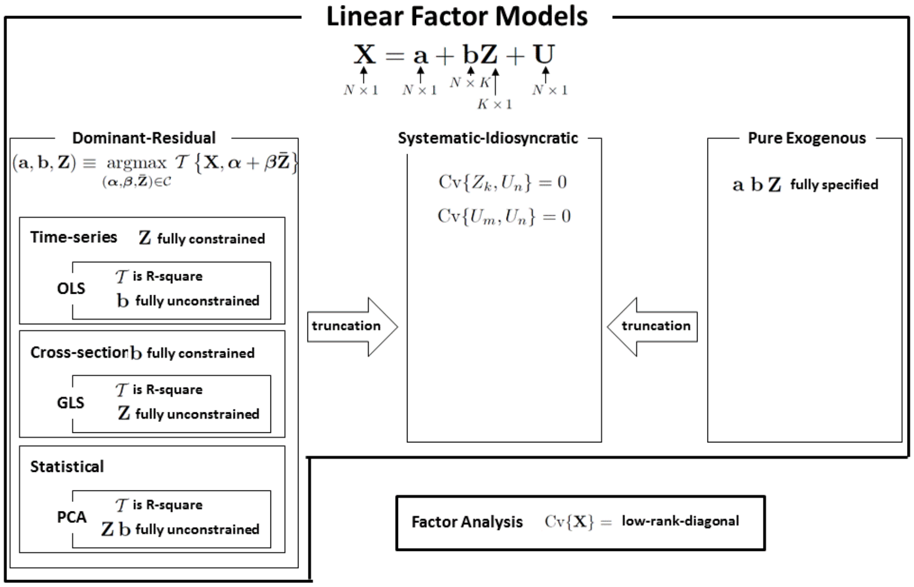

Linear Factor Models
$\mathbf{X} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$
$N \times 1$
$N \times 1$
$N \times K$
$K \times 1$
$N \times 1$
Dominant-Residual
$(\mathbf{a}, \mathbf{b}, \mathbf{Z}) \equiv \operatorname{argmax}_{(\boldsymbol{\alpha}, \boldsymbol{\beta}, \bar{\mathbf{Z}}) \in \mathcal{C}} \mathcal{T}\{\mathbf{X}, \boldsymbol{\alpha} + \boldsymbol{\beta}\bar{\mathbf{Z}}\}$
Time-series $\mathbf{Z}$ fully constrained
OLS
$\mathcal{T}$ is R-square
$\mathbf{b}$ fully unconstrained
Cross-section $\mathbf{b}$ fully constrained
GLS
$\mathcal{T}$ is R-square
$\mathbf{Z}$ fully unconstrained
Statistical
PCA
$\mathcal{T}$ is R-square
$\mathbf{Z}\,\mathbf{b}$ fully unconstrained
truncation
Systematic-Idiosyncratic
$\mathrm{Cv}\{Z_k, U_n\} = 0$
$\mathrm{Cv}\{U_m, U_n\} = 0$
truncation
Pure Exogenous
$\mathbf{a}\ \mathbf{b}\ \mathbf{Z}$ fully specified
Factor Analysis $\mathrm{Cv}\{\mathbf{X}\} =$ low-rank-diagonal

Figure 2: Types of Linear Factor Models.

In Section 2.1 we cover the first category, namely dominant-residual LFM's, which include time-series, cross-sectional, and statistical LFM's as special cases. As we shall see, surprisingly all these three sub-categories of LFM follow naturally from the same optimization, under different constraints.

In Section 2.2 we discuss the second category, namely systematic-idiosyncratic LFM's, which at times are incorrectly assumed to overlap with dominant-residual LFM's.

In Section 2.3 we discuss the third category, namely pure exogenous LFM's.

In Section 2.4 we discuss factor analysis, which is sometimes considered similar to principal component analysis, but which in reality is fundamentally different because it is not a LFM.

LFM's can be used to decompose different types of variables, namely the i.i.d. invariants that drive the dynamics of the market, such as changes in yield spreads for bonds, or log-changes in implied volatility for options; or the projected P&L's over a future period of a set of securities or portfolios; or the returns of the same securities over an arbitrary period. In order to cover LFM's in full generality in this theoretical section, we denote these random variables by $\mathbf{X}$, which we call "the market". In the applications in Section 3 we will specify in detail the variables $\mathbf{X}$.

Key definition. A LFM is a decomposition of the market $\mathbf{X} \equiv (X_1, \ldots, X_N)'$ as a linear combination of factors plus residuals, as follows

$$\begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix} \equiv \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} + \begin{pmatrix} b_{1,1} & \cdots & b_{1,K} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,K} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_K \end{pmatrix} + \begin{pmatrix} U_1 \\ \vdots \\ U_N \end{pmatrix}, \tag{1}$$

or in matrix notation

$$\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}. \tag{2}$$

In this expression $\mathbf{a} \equiv (a_1, \ldots, a_N)'$ are N constants; $\mathbf{Z} \equiv (Z_1, \ldots, Z_K)'$ are K factors, i.e. random variables correlated with the market $\mathbf{X}$; $\mathbf{b} \equiv \{b_{n,k}\}_{n=1,\ldots,N}^{k=1,\ldots,K}$ is a $N \times K$ matrix of factor loadings, i.e. coefficients that transfer the randomness of the factors $\mathbf{Z}$ into the market $\mathbf{X}$; and $\mathbf{U} \equiv (U_1, \ldots, U_N)'$ are N random residuals.

The randomness recovered by the LFM is

$$\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}. \tag{3}$$

The different components of a LFM $\mathbf{X}, \mathbf{a}, \mathbf{b}, \mathbf{Z}, \mathbf{Y}, \mathbf{U}$ can be specified in one of two ways. First, with the joint distribution $f_{\mathbf{X},\mathbf{Z}}$ of market and factors, and with coefficients $\mathbf{a}$ and $\mathbf{b}$, we then derive the distribution of the recovered randomness $\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}$ and of the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$

$$(f_{\mathbf{X},\mathbf{Z}}, \mathbf{a}, \mathbf{b}) \Rightarrow (f_{\mathbf{Y}}, f_{\mathbf{U}}). \tag{4}$$

Second, with the joint distribution $f_{\mathbf{Z},\mathbf{U}}$ of factors and residuals, and with coefficients $\mathbf{a}$ and $\mathbf{b}$, we then derive the distribution of the recovered randomness $\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}$ and of the market $\mathbf{X} \equiv \mathbf{Y} + \mathbf{U}$

$$(f_{\mathbf{Z},\mathbf{U}}, \mathbf{a}, \mathbf{b}) \Rightarrow (f_{\mathbf{Y}}, f_{\mathbf{X}}). \tag{5}$$

Notice that the market dimension N is not necessarily larger than the number of factors K. However, when $N \gg K$, LFM's represent a dimension reduction technique useful for instance to estimate the projected P&L of all the stocks in the world, see Section 3.1. In the opposite extreme case where $N = 1 < K$, LFM's represent a risk attribution tool useful for instance for optimal hedging and parsimonious interpretation of the P&L, see Section 3.5.

We emphasize that in our framework the market $\mathbf{X}$, the factors $\mathbf{Z}$, the recovered randomness $\mathbf{Y}$ and the residuals $\mathbf{U}$ in a LFM are simultaneous random variables with fully general distributions, that are by no means restricted to normal. The random variables $(\mathbf{X}, \mathbf{Z}, \mathbf{Y}, \mathbf{U})$, which we denote by capital letters, are not past observations of financial data

$\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t, \mathbf{u}_t\}_{t=1,\ldots,T}$, which we denote by lower-case letters. We expand on this confusion between forward-looking modeling versus backward-looking data manipulations in Section 4.1. Needless to say, the connection between random variables and data is important, and we will delve into it in the applications to estimation in Section 3.1.

2.1 Dominant-residual LFM's

A dominant-residual LFM for a market $\mathbf{X}$ is a decomposition $\mathbf{X} \equiv \mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}$ as in (2), where the recovered market randomness $\mathbf{Y} \equiv \mathbf{a}+\mathbf{b}\mathbf{Z}$ is optimized according to a fitness target or objective $\mathcal{T}\{\mathbf{X}, \mathbf{Y}\}$ function to explain the largest portion of the original randomness in the market $\mathbf{X}$ under a potential set of constraints $\mathcal{C}$.

Key definition. A *dominant-residual* LFM for a market $\mathbf{X}$ is a decomposition

$$\mathbf{X} \equiv \mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}, \tag{6}$$

where

$$(\mathbf{a}, \mathbf{b}, \mathbf{Z}) \equiv \underset{(\alpha,\beta,\bar{\mathbf{Z}})\in\mathcal{C}}{\operatorname{argmax}} \, \mathcal{T}\left\{\mathbf{X}, \alpha+\beta\bar{\mathbf{Z}}\right\}. \tag{7}$$

A dominant-residual LFM (6) is always specified as in (4), by first determining the joint distribution $f_{\mathbf{X},\mathbf{Z}}$ of the market and the factors and the coefficients $\mathbf{a}$ and $\mathbf{b}$ using (7), and then deriving the distribution of the recovered randomness $\mathbf{Y} \equiv \mathbf{a}+\mathbf{b}\mathbf{Z}$ and of the residuals $\mathbf{U} \equiv \mathbf{X}-\mathbf{Y}$

$$(f_{\mathbf{X},\mathbf{Z}}, \mathbf{a}, \mathbf{b}) \Rightarrow (f_{\mathbf{Y}}, f_{\mathbf{U}}). \tag{8}$$

In the dominant-residual formulation (7) the fitness target $\mathcal{T}$ is fully general. A useful target is the conditional-value-at-risk (CVaR), which has applications to hedging, see Section 3.5.

A second, fundamental, target is the multivariate r-square. To introduce it, we recall the standard definition of the univariate r-square provided by a "model" set of data $\{y\} \equiv \{y_j\}_{j=1,\ldots,J}$ to an "original" set of data $\{x\} \equiv \{x_j\}_{j=1,\ldots,J}$, namely $r^2(\{x\},\{y\}) \equiv 1-\widehat{\mathrm{V}}(\{x-y\})/\widehat{\mathrm{V}}(\{x\})$, where $\widehat{\mathrm{V}}$ denotes the sample variance of the data.

In our approach, the original market and the model market are not data, but rather random variables with a fully general distribution. Accordingly, we replace the sample variance $\widehat{\mathrm{V}}$ with the distributional variance V, thereby obtaining the distributional r-square $\mathrm{R}^2\{X,Y\} \equiv 1-\mathrm{V}\{X-Y\}/\mathrm{V}\{X\}$.

Then, we generalize the distributional r-square to a multivariate environment. Denoting by $\mathrm{Cv}\{\mathbf{X}\}$ the covariance matrix of the entries of $\mathbf{X}$ and by tr the trace of a matrix, i.e. the sum of the diagonal elements, we obtain the following definition, see also [Meucci, 2005]: the *multivariate distributional r-square* is a measure of fitness defined as

$$\mathrm{R}^2\{\mathbf{X},\mathbf{Y}\} \equiv 1-\frac{\mathrm{tr}\,(\mathrm{Cv}\{\mathbf{X}-\mathbf{Y}\})}{\mathrm{tr}\,(\mathrm{Cv}\{\mathbf{X}\})}. \tag{9}$$

Since $\mathrm{tr}\,(\mathrm{Cv}\{\mathbf{X}\})$ is proportional to the average variance among the N entries in the market $\frac{1}{N}\sum_{n=1}^{N}\mathrm{V}\{X_n\}$, the numerator in (9) is always positive, and only zero if the vector $\mathbf{Y}$

replicates $\mathbf{X}$ exactly. As in the univariate case, the denominator is a normalization term that makes the distributional r-square scale-independent.

In the dominant-residual formulation (7) not only the fitness target $\mathcal{T}$, but also the constraints $\mathcal{C}$ are fully general. For instance, the constraints can include the least absolute shrinkage and selection operator (LASSO), see e.g. [Hastie et al., 2009] for more on the subject.

Another useful constraint is the "few-out-of many", or cardinality constraints, used to select the most effective factors or hedging products, see the applications in Section 3.5.

Here we highlight a constraint that becomes useful when the fitness target in the dominant-residual LFM definition (6)-(7) is the r-square (9). In that case the constant vector $\mathbf{a}$ does not affect the optimization (7), and thus it cannot be determined by the optimization. To determine $\mathbf{a}$ we impose the constraint that the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{b}\mathbf{Z}$ have zero-expectation

$$\mathcal{C} \supseteq \{\mathbf{a} \equiv \mathrm{E}\{\mathbf{X}\} - \mathbf{b}\,\mathrm{E}\{\mathbf{Z}\}\} \,. \tag{10}$$

There are three different classes of dominant-residual LFM's widely used in the industry: time-series, which we discuss in Section 2.1.1, cross-sectional, which we cover in Section 2.1.2, and statistical, which we discuss in Section 2.1.3.

As we shall see, the differences between these three classes of LFM's are solely determined by the specification of the constraints $\mathcal{C}$ in the dominant-residual definition (7), refer to Figure 2. Researchers at times perceive these three classes of LFM's as recipes. By seeing them embedded in the general, flexible, modular dominant-residual framework (6)-(7) with general targets and constraints we can extend each class to a wealth of theoretically sound LFM's better suited for each situation.

2.1.1 Time-series LFM's

The first of the three sub-classes of dominant-residual LFM's are time-series LFM's.

Key definition. A standard *time-series* LFM for a market $\mathbf{X}$ is a dominant-residual decomposition $\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ with exogenously specified factors $\mathbf{Z}$, where the fitness target is the multivariate distributional r-square. Thus, a time-series model is defined by the general framework (7) as

$$(\mathbf{a}, \mathbf{b}, \mathbf{Z}) \equiv \underset{(\alpha,\beta,\bar{\mathbf{Z}})\in\mathcal{C}}{\mathrm{argmax}}\ \mathrm{R}^2\big\{\mathbf{X}, \alpha + \beta\bar{\mathbf{Z}}\big\} \,, \tag{11}$$

under the no-expectation constraint (10) on the residual and the additional constraint that the factors $\mathbf{Z}$ be fully specified exogenously

$$\mathcal{C} : \begin{cases} \alpha \equiv \mathrm{E}\{\mathbf{X}\} - \beta\,\mathrm{E}\big\{\bar{\mathbf{Z}}\big\} \\ \bar{\mathbf{Z}} \equiv \mathbf{Z} \end{cases} \tag{12}$$

As we show in Appendix A.1, the standard time-series optimization (11)-(12) can be solved explicitly for $\mathbf{b}$, yielding

$$\mathbf{b} \equiv \mathrm{Cv}\{\mathbf{X}, \mathbf{Z}\}\,\mathrm{Cv}\{\mathbf{Z}\}^{-1} \,. \tag{13}$$

Notice how to compute the optimal loadings $\mathbf{b}$ in (13) and thus specify the time-series LFM we need the joint distribution $f_{\mathbf{X},\mathbf{Z}}$ of the market $\mathbf{X}$ and the factors $\mathbf{Z}$.

Then, using the expression for $\mathbf{b}$ in (13), we obtain $\mathbf{a}$ from (12). Thus, the factor-recovered market $\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}$ becomes

$$\mathbf{Y} \equiv \mathrm{E}\{\mathbf{X}\} + \mathrm{Cv}\{\mathbf{X},\mathbf{Z}\}\,\mathrm{Cv}\{\mathbf{Z}\}^{-1}(\mathbf{Z} - \mathrm{E}\{\mathbf{Z}\})\,. \tag{14}$$

This equation shows that the recovered randomness lives in a K-dimensional plane, embedded in the N-dimensional space of the original market $\mathbf{X}$, see Figure 3.

As we prove in Appendix A.1, the distributional r-square provided by the recovered market (14) reads

$$\mathrm{R}^2\{\mathbf{X},\mathbf{Y}\} = \tfrac{1}{N}\,\mathrm{tr}\left(\mathrm{Cr}\{\mathbf{X},\mathbf{Z}\}\,\mathrm{Cr}\{\mathbf{Z}\}^{-1}\,\mathrm{Cr}\{\mathbf{Z},\mathbf{X}\}\right), \tag{15}$$

where $\mathrm{Cr}\{\mathbf{Z},\mathbf{X}\}$ are the$K \times N$ correlations between factors and market and $\mathrm{Cr}\{\mathbf{Z}\}$ are the $N \times N$ correlations among factors. This expression supports the intuition that a good model must display high overall correlations between the factors $\mathbf{Z}$ and the market $\mathbf{X}$. Furthermore, the factors should be as uncorrelated among each other as possible. If there are high correlations among the factors, or if there is high collinearity, the matrix $\mathrm{Cr}\{\mathbf{Z}\}^{-1}$ would be ill-defined. Geometrically, this means that the plane of the recovered market (14) is not properly defined, see Figure 3.

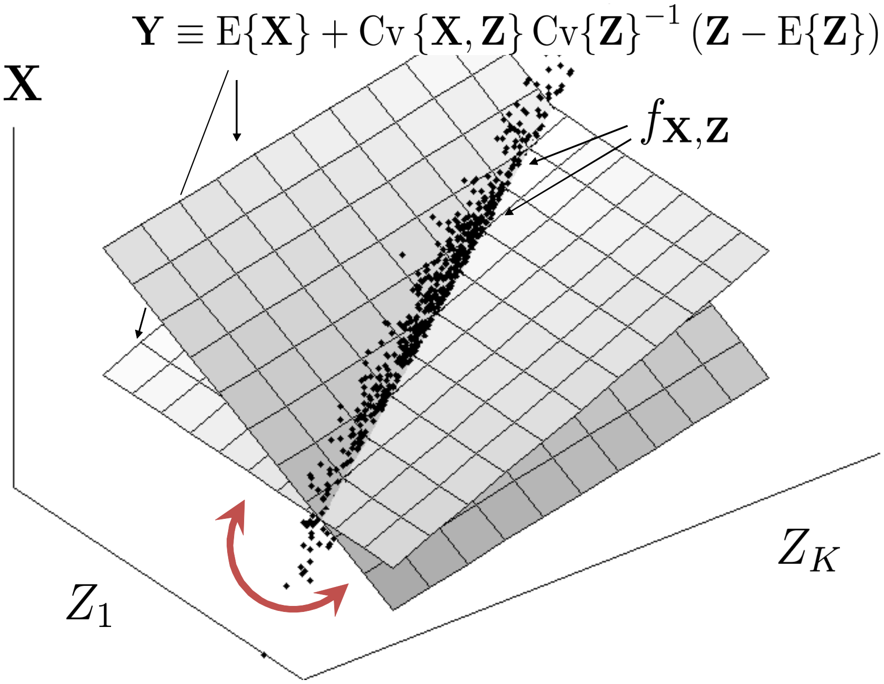

$\mathbf{Y} \equiv \mathrm{E}\{\mathbf{X}\} + \mathrm{Cv}\{\mathbf{X},\mathbf{Z}\}\,\mathrm{Cv}\{\mathbf{Z}\}^{-1}(\mathbf{Z} - \mathrm{E}\{\mathbf{Z}\})$
$\mathbf{X}$
$f_{\mathbf{X},\mathbf{Z}}$
Z_K
Z_1

Figure 3: Factors in time-series factor model should not be collinear.

With the expression for the randomness $\mathbf{Y}$ recovered by the LFM (14) we can compute the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$. As we show in Appendix A.1, the residuals are uncorrelated with the factors, i.e. $\mathrm{Cr}\{U_n, Z_k\} = 0$. However, the residuals are not uncorrelated among each other, i.e. $\mathrm{Cr}\{U_n, U_m\} \neq 0$ and thus the residuals are not idiosyncratic. We discuss further this pitfall, which can give rise to incorrect risk estimates, in Section 4.5.

To illustrate the concept of standard time-series models, we assume a two-dimensional market with one external factor

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0.7 \\ 0.6 & 0.7 & 1 \end{pmatrix} \right). \tag{16}$$

We maximize the distributional r-square as in (11)-(12). Then we obtain

$$\mathbf{a} = (0,0)', \quad \mathbf{b} = (0.6, 0.7)' \tag{17}$$

Then the joint distribution of the residuals and the LFM-recovered randomness reads

$$\begin{pmatrix} U_1 \\ U_2 \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.64 & 0.08 & 0 \\ 0.08 & 0.51 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right). \tag{18}$$

Notice that $\mathrm{Cr}\{U_n, Z\} = \mathrm{Cr}\{U_n, Y\} = 0$, but $\mathrm{Cr}\{U_1, U_2\} = 0.08 \neq 0$. We refer to the code at symmys.com/node/336 for more details.

Using the general dominant residual framework (6)-(7) we can construct *generalized time-series* LFM's by replacing in (11)-(12) the r-square with arbitrary fitness targets $\mathcal{T}$ and by adding arbitrary constraints $\mathcal{C}$. These generalizations make the correlations of the factors with the residuals non-zero, $\mathrm{Cr}\{U_n, Z_k\} \neq 0$, in addition to the correlations among residuals being non-zero, $\mathrm{Cr}\{U_n, U_m\} \neq 0$. We discuss this pitfall further in Section 4.5.

To illustrate a generalized time-series model, we continue with the above case study, but in addition to (12) we include the constraint that the loadings be bound from below by $\underline{b} \equiv 0.8$ and from above by $\overline{b} \equiv 1.2$

$$\underline{b} \leq b_1, b_2 \leq \overline{b}. \tag{19}$$

Now

$$\mathbf{b} = (0.8, 0.8)' \tag{20}$$

and the joint distribution of residuals and recovered market reads

$$\begin{pmatrix} U_1 \\ U_2 \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.68 & 0.1 & -0.2 \\ 0.1 & 0.52 & -0.1 \\ -0.2 & -0.1 & 1 \end{pmatrix} \right). \tag{21}$$

Notice that now $\mathrm{Cr}\{U_n, Z\} = \mathrm{Cr}\{U_n, Y\} \neq 0$. We refer to the code at symmys.com/node/336 for more details.

Time-series LFM's bear this name because in some applications the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is estimated as the empirical distribution stemming from a time-series of data $f_{\mathbf{X},\mathbf{Z}} \Longleftrightarrow \{\mathbf{x}_t, \mathbf{z}_t\}_{t=1,\ldots,T}$. In this case the loadings (13) become the coefficients of an ordinary least square regression, as in the application (61) below.

However, the term "time-series" is a bit of a misnomer, because all the other approaches, including the cross-sectional approach of Section 2.1.2, ultimately rely on time-series analysis for estimation purposes.

Time-series LFM's are also known as macroeconomic LFM's, because in some applications the factors are macroeconomic variables, such as interest rates, stock market returns, etc.

2.1.2 Cross-sectional LFM's

The second of the three sub-classes of dominant-residual LFM's are cross-sectional LFM.

Key definition. A standard *cross-sectional* LFM for a market $\mathbf{X}$ is a decomposition $\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ where the loadings $\mathbf{b}$ are specified exogenously. Then, the factors $\mathbf{Z}$ are extracted from the market by means of a linear transformation defined by a $K \times N$ matrix $\mathbf{c}$ as $\mathbf{Z} \equiv \mathbf{c}\mathbf{X}$, in such a way to maximize the multivariate distributional r-square of the LFM.

Accordingly, a cross-sectional LFM is defined by the general framework (7) as

$$(\mathbf{a}, \mathbf{b}, \mathbf{Z}) \equiv \underset{(\alpha, \beta, \bar{\mathbf{Z}}) \in \mathcal{C}}{\operatorname{argmax}} \mathrm{R}^2 \left\{ \mathbf{X}, \alpha + \beta \bar{\mathbf{Z}} \right\}, \tag{22}$$

under the usual no-expectation constraint on the residual (10), the exogenous specification of the loadings $\mathbf{b}$, and the requirement that the factors be extracted by a linear transformation

$$\mathcal{C} : \begin{cases} \alpha \equiv \mathrm{E}\{\mathbf{X}\} - \beta\, \mathrm{E}\left\{\bar{\mathbf{Z}}\right\} \\ \beta \equiv \mathbf{b} \\ \bar{\mathbf{Z}} \equiv \mathbf{c}\mathbf{X} \end{cases} \tag{23}$$

where $\mathbf{c}$ must be optimized.

As we show in Appendix A.2, the optimization (22) under the constraints (23) can be solved explicitly and reads[2]

$$\mathbf{c} \equiv \left(\mathbf{b}'\mathbf{b}\right)^{-1} \mathbf{b}'. \tag{24}$$

Using the factor-extraction matrix (24) and the constraint for $\mathbf{a}$ in (23) we can write the factor-recovered market $\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}$ as

$$\mathbf{Y} = \mathrm{E}\{\mathbf{X}\} + \mathbf{b} \left(\mathbf{b}'\mathbf{b}\right)^{-1} \mathbf{b}' \left(\mathbf{X} - \mathrm{E}\{\mathbf{X}\}\right). \tag{25}$$

As we show in Appendix A.2, the distributional r-square provided by the recovered market reads

$$\mathrm{R}^2\{\mathbf{X}, \mathbf{Y}\} = \frac{\mathrm{tr}\left(\mathrm{Cv}\{\mathbf{Y}\}\right)}{\mathrm{tr}\left(\mathrm{Cv}\{\mathbf{X}\}\right)}. \tag{26}$$

This result is intuitive: the r-square is the ratio of the average variance of the factor-recovered randomness $\mathbf{Y}$ over the average variance of the original market $\mathbf{X}$.

After obtaining $\mathbf{Y}$ in (25) we can compute the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$. Contrary to common belief, the residuals are not uncorrelated with the factors, i.e. $\mathrm{Cv}\{U_n, Z_k\} \neq 0$, and thus the

[2] Typically, in the cross-sectional approach the market $\mathbf{X}$ is normalized beforehand to display similar volatility $\mathbf{X} \to \quad \sigma^{2^{-\frac{1}{2}}} \mathbf{X}$, where $\sigma^{2^{-\frac{1}{2}}}$ is either a diagonal matrix of inverse standard deviations, or it is a root of the inverse covariance $(\sigma^{2^{-\frac{1}{2}}})^2 \equiv \quad \sigma^{2^{-1}}$. This yields the extraction matrix $\mathbf{c} \equiv (\mathbf{b}'\sigma^{2^{-1}}\mathbf{b})^{-1}\mathbf{b}'\sigma^{2^{-1}}$

factors $\mathbf{Z}$ are not systematic. Furthermore, the residuals are not uncorrelated among each other, i.e. $\mathrm{Cv}\{U_n, U_m\} \neq 0$, and thus the residuals $\mathbf{U}$ are not idiosyncratic. We discuss this issue further in Section 4.5.

To illustrate the concept of cross-sectional models, we consider a two-dimensional market $\mathbf{X}$.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} \right). \tag{27}$$

Then, instead of imposing factors exogenously, we specify the loadings $\mathbf{b} \equiv (1,1)'$ and we extract the one factor Z from the market, in such a way to maximize the distributional r-square as in (22)-(23).

Then the extraction coefficients in (24) read $\mathbf{c} = (0.5, 0.5)$ and the factor distribution follows as $Z \sim \mathrm{N}(0, 0.35)$. Then $\mathbf{a} \equiv (0,0)'$ follows from the constraint in (23). Finally the joint distribution of the recovered randomness $\mathbf{Y} \equiv \mathbf{a} + \mathbf{bcX}$ and of the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$ reads

$$\begin{pmatrix} U_1 \\ U_2 \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & -0.25 & 0.2 \\ -0.25 & 0.25 & -0.2 \\ 0.2 & -0.2 & 0.35 \end{pmatrix} \right). \tag{28}$$

Notice that $\mathrm{Cr}\{U_n, Z\} = \mathrm{Cr}\{U_n, Y\} \neq 0$ and $\mathrm{Cr}\{U_1, U_2\} \neq 0$. We refer to the code at symmys.com/node/336 for more details.

As we did for time-series LFM's, using the general dominant-residual framework (6)-(7) we can construct *generalized cross-sectional* LFM's by replacing in (22)-(23) the r-square with arbitrary fitness targets $\mathcal{T}$ and by adding arbitrary constraints $\mathcal{C}$.

To illustrate a generalized cross-sectional model, we continue with the above case study. For illustration, we add the arbitrary constraint that the factor Z be uncorrelated with the market spread, i.e. $\mathrm{Cv}\{X_2 - X_1, Z\} \equiv 0$. Since $X_2 - X_1 = (-1, 1)\,\mathbf{X}$ and since from the last constraint in (23) we have $Z = \mathbf{cX}$ this new constraint reads

$$(-1, 1) \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} \mathbf{c} \equiv 0, \tag{29}$$

The constraint (29) is linear in $\mathbf{c}$ and therefore the computation of the optimal cross-sectional factor becomes an instance of quadratic programming, which can be easily solved numerically. The result reads $\mathbf{c} = (0, 0.75)'$ and

$$\begin{pmatrix} U_1 \\ U_2 \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .9625 & -.0125 & -.0375 \\ -.0125 & .0125 & .0375 \\ -.0375 & .0375 & .1125 \end{pmatrix} \right). \tag{30}$$

We refer to the code at symmys.com/node/336 for more details.

The LFM (22)-(23) is known as "cross-sectional" because the factor-extracting coefficients (24) are the same as those of a cross-sectional regression. However, we emphasize that at no point did we run a regression on data. Instead, we maximized the r-square of the *distribution* of the factor model.

Cross-sectional LFM's are also known as "fundamental" LFM's, because, when applied to the equity market, the loadings are often defined in terms of fundamental book variables, see the applications in Section 3.1.2.

2.1.3 Statistical LFM's

The third of the three sub-classes of dominant-residual LFM's are statistical LFM's.

Key definition. A *principal component analysis* (PCA) LFM for a market $\mathbf{X}$ is a decomposition $\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ that maximizes the multivariate distributional r-square without constraints on the loadings $\mathbf{b}$, and the only constraint on the factors $\mathbf{Z}$ that they be extracted from the market by a linear transformation, defined by a $K \times N$ matrix $\mathbf{c}$ as $\mathbf{Z} \equiv \mathbf{c}\mathbf{X}$.

Thus the parameters $(\mathbf{a}, \mathbf{b}, \mathbf{Z})$ of a PCA LFM are defined according to the general framework (7) as

$$(\mathbf{a}, \mathbf{b}, \mathbf{Z}) \equiv \underset{(\alpha,\beta,\bar{\mathbf{Z}})\in\mathcal{C}}{\operatorname{argmax}} \mathrm{R}^2\left\{\mathbf{X}, \alpha + \beta\bar{\mathbf{Z}}\right\}, \tag{31}$$

under the usual no-expectation constraint on the residual (10) and the factor-extraction constraint

$$\mathcal{C} : \begin{cases} \alpha \equiv \mathrm{E}\{\mathbf{X}\} - \beta\,\mathrm{E}\left\{\bar{\mathbf{Z}}\right\} \\ \bar{\mathbf{Z}} \equiv \mathbf{c}\mathbf{X} \end{cases} \tag{32}$$

where $\mathbf{c}$ must be optimized.

As we show in Appendix A.3, the optimization (31) under the constraints (32) can be solved explicitly. To write the solution, first, we compute the covariance matrix of the market and we perform its spectral decomposition

$$\mathrm{Cv}\{\mathbf{X}\} \equiv \mathbf{e}\,\mathrm{diag}(\boldsymbol{\lambda}^2)\,\mathbf{e}'. \tag{33}$$

In this expression $\boldsymbol{\lambda}^2$ are the decreasing, positive eigenvalues

$$\boldsymbol{\lambda}^2 \equiv \left(\lambda_1^2, \ldots, \lambda_N^2\right)'; \tag{34}$$

and $\mathbf{e}$ is the juxtaposition of the respective eigenvectors

$$\mathbf{e} \equiv (\mathbf{e}_1| \ldots |\mathbf{e}_N). \tag{35}$$

The eigenvectors are orthogonal, i.e. $\mathbf{e}'_m\mathbf{e}_n = 0$, and they are normalized to have length 1, i.e. $\mathbf{e}'_n\mathbf{e}_n = 1$. Therefore $\mathbf{e}\mathbf{e}' = \mathbf{e}'\mathbf{e} = \mathbf{i}_N$, the identity matrix. Next, we define a $N \times K$ matrix as the juxtaposition of the first K eigenvectors

$$\bar{\mathbf{e}}_K \equiv (\mathbf{e}_1| \ldots |\mathbf{e}_K). \tag{36}$$

Then the loadings $\mathbf{b}$ and the statistical factors $\mathbf{Z}$ that solve the dominant-residual optimization (31)-(32) read

$$\mathbf{b} = \bar{\mathbf{e}}_K \tag{37}$$

$$\mathbf{Z} = \bar{\mathbf{e}}'_K\mathbf{X} \tag{38}$$

Using the expression for $\mathbf{b}$ in (37), for $\mathbf{Z}$ in (38), and for $\mathbf{a}$ in (32) we can write the factor-recovered market $\mathbf{Y} \equiv \mathbf{a} + \mathbf{bZ}$ as

$$\mathbf{Y} = \mathrm{E}\{\mathbf{X}\} + \bar{\mathbf{e}}_K \bar{\mathbf{e}}'_K (\mathbf{X} - \mathrm{E}\{\mathbf{X}\}) . \tag{39}$$

As we show in [Meucci, 2005] the recovered randomness $\mathbf{Y}$ has an intuitive geometrical interpretation: it is the orthogonal projection of the market $\mathbf{X}$ onto the hyperplane spanned by the first K eigenvectors, see Figure 4.

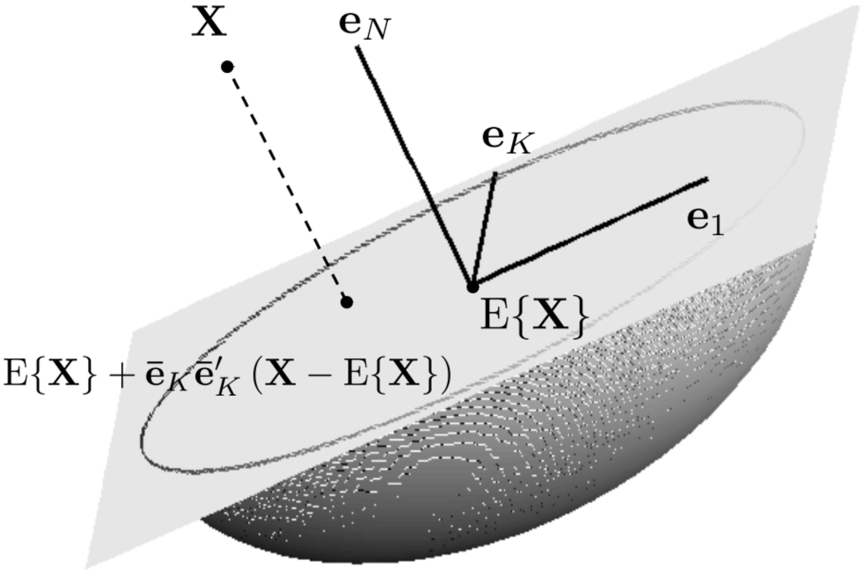

$\mathbf{X}$
$\mathbf{e}_N$
$\mathbf{e}_K$
$\mathbf{e}_1$
$\mathrm{E}\{\mathbf{X}\}$
$\mathrm{E}\{\mathbf{X}\} + \bar{\mathbf{e}}_K \bar{\mathbf{e}}'_K (\mathbf{X} - \mathrm{E}\{\mathbf{X}\})$

Figure 4: PCA as projection on location-dispersion ellipsoid.

We recall that the eigenvectors can be interpreted as the directions of the principal axes of the location-dispersion ellipsoid defined by the market expectation vector $\mathrm{E}\{\mathbf{X}\}$ and the market covariance matrix $\mathrm{Cv}\{\mathbf{X}\}$, and that the square root of the eigenvalues is the length of the principal axes, see Figure 4 and refer to [Meucci, 2010d] for more details. Therefore, the factor recovered market $\mathbf{Y}$ is the orthogonal projection of the market $\mathbf{X}$ onto the hyperplane spanned by the first K principal axes of the location-dispersion ellipsoid.

As we show in Appendix A.3, the distributional r-square provided by the recovered market (39) reads

$$\mathrm{R}^2\{\mathbf{X}, \mathbf{Y}\} = \frac{\sum_{k=1}^{K} \lambda_k^2}{\sum_{n=1}^{N} \lambda_n^2} . \tag{40}$$

As a consequence, the steeper the spectrum profile $K \to \sum_{k=1}^{K} \lambda_k^2$, i.e. the larger the discrepancy among the first eigenvalues, the lower the number K of factors required to obtain a large r-square.

The eigenvalues display high discrepancy in highly correlated markets, such as the changes of swap rates of different maturities. Therefore, PCA LFM's are more effective in achieving a large r-square with few factors in highly correlated markets.

With the expression for the randomness $\mathbf{Y}$ recovered by the LFM (39) we can compute the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$. The residuals are uncorrelated with the factors (38), i.e. $\mathrm{Cr}\{U_n, Z_k\} = 0$. However, contrary to common belief, the residuals are not uncorrelated among each other, i.e. $\mathrm{Cr}\{U_n, U_m\} \neq 0$ and thus they are not idiosyncratic. We discuss this issue further in Section 4.5.

> To illustrate the PCA factor models, we consider the two-dimensional market
>
> $$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4.1 & 1.2 \\ 1.2 & 3.4 \end{pmatrix}\right). \tag{41}$$
>
> We extract $K = 1$ factor Z. The extraction coefficients in (37) read $\mathbf{e}_1' = (0.8, 0.6)$ and the factor distribution follows from (38) as $Z \sim \mathrm{N}(0, 5)$. Then $\mathbf{a}$ follows from the constraint in (32) and reads $\mathbf{a} \equiv (0,0)'$. Finally the joint distribution of residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{e}_1\mathbf{e}_1'\mathbf{X}$ and recovered market $\mathbf{Y} \equiv \mathbf{a} + \mathbf{e}_1\mathbf{e}_1'\mathbf{X}$ reads
>
> $$\begin{pmatrix} U_1 \\ U_2 \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.9 & -1.2 & 0 \\ -1.2 & 1.6 & 0 \\ 0 & 0 & 5 \end{pmatrix}\right). \tag{42}$$
>
> Notice that $\mathrm{Cr}\{U_n, Z\} = \mathrm{Cr}\{U_n, Y\} = 0$ but $\mathrm{Cr}\{U_1, U_2\} \neq 0$.

As we did for time-series and cross-sectional LFM's, using the general dominant-residual framework (6)-(7) we can construct *generalized statistical* LFM's by replacing in (31)-(32) the r-square with arbitrary fitness targets $\mathcal{T}$ and by adding arbitrary constraints $\mathcal{C}$. In generalized statistical models the factors and the residuals will not be uncorrelated, i.e. $\mathrm{Cr}\{U_n, Z_k\} \neq 0$, in a way similar to the time-series example (21).

Before concluding the theory of statistical LFM's, we emphasize that factor analysis, which we discuss in Section 2.4, is not a statistical LFM's, although at times it is considered similar to PCA statistical LFM's. We expand on this pitfalls in Section 4.9.

2.2 Systematic-idiosyncratic LFM's

In Section 2.1 we introduced dominant-residual LFM's, which include time-series, cross-sectional, and statistical LFM's as special cases. Here we discuss a different class of models, systematic-idiosyncratic LFM's. These models, which play a fundamental role in the financial theory of asset pricing, are often incorrectly assumed to be the same as dominant-residual LFM's.

Key definition. A *systematic-idiosyncratic* LFM for a market $\mathbf{X}$ is a decomposition as in the general LFM (2)

$$\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}, \tag{43}$$

where the factors $\mathbf{Z}$ and the residuals $\mathbf{U}$ satisfy two types of constraints. First, the residuals are idiosyncratic, in that they are uncorrelated with each other

$$\mathrm{Cr}\{U_n, U_m\} = 0, \quad n \neq m = 1, \ldots, N. \tag{44}$$

Second, the factors $\mathbf{Z}$ are systematic, in that they are uncorrelated with the residuals

$$\mathrm{Cr}\{U_n, Z_k\} = 0, \quad n = 1, \ldots, N, \quad k = 1, \ldots, K. \tag{45}$$

Systematic-idiosyncratic LFM's are useful because they impose structure, especially in markets $\mathbf{X}$ of large dimension. This is particularly apparent in the covariance matrix, which, as a consequence of the conditions (44)-(45) is low-rank-diagonal

$$\underbrace{\mathrm{Cv}\{\mathbf{X}\}}_{N\times N} = \mathbf{b}\underbrace{\mathrm{Cv}\{\mathbf{Z}\}}_{K\times K}\mathbf{b}' + \mathrm{diag}(\underbrace{\mathrm{V}\{\mathbf{U}\}}_{N\times 1}). \tag{46}$$

Therefore the market covariance contains only $K\,(K+1)\,/2+N$ elements of stochastic nature, as opposed to general markets, which contain $N\,(N+1)\,/2$ such elements.

To illustrate the concept of systematic-idiosyncratic LFM, assume we estimated a bivariate market with one factor

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 3 & 3 \\ 3 & 5 & 3 \\ 3 & 3 & 3 \end{pmatrix}\right), \tag{47}$$

and with coefficients

$$\mathbf{a} = (0,0)', \quad \mathbf{b} = (1,1). \tag{48}$$

Then the joint distribution of the the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{b}Z$ and the factor Z reads

$$\begin{pmatrix} U_1 \\ U_2 \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}\right). \tag{49}$$

This shows that both the idiosyncratic condition (44) and the systematic condition (45) are satisfied. However, even the smallest deviation from the specification (47) will prevent this LFM from being systematic-idiosyncratic.

Systematic-idiosyncratic models do not exist in reality, because the requirements (44)-(45) are too strict. Therefore, systematic-idiosyncratic LFM's are only used as approximations of the true market distribution.

In typical approaches, researchers identify a parsimonious set of factors $\mathbf{Z}$ and coefficients $\mathbf{a}$ and $\mathbf{b}$, often by means of a dominant-residual optimization (7). Then, they model the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{bZ}$ as idiosyncratic, even though they are not, by truncating the correlations to zero. More formally, truncation turns any LFM, whether dominant-residual or not, into a systematic-idiosyncratic LFM by exogenously setting to zero the co-dependence among the residuals and between factors and residuals, in five steps.

First, a LFM $\mathbf{X} \equiv \mathbf{a} + \mathbf{bZ} + \mathbf{U}$ is specified as in (4) by an estimated joint distribution of the market and the factors $\overline{f}_{\mathbf{X},\mathbf{Z}}$ and by coefficients $\mathbf{a}$ and $\mathbf{b}$. Second, the distribution of the factors $\overline{f}_{\mathbf{Z}}$ alone is computed. Third, the distribution of the residuals $\overline{f}_{\mathbf{U}} \equiv \overline{f}_{\mathbf{X}-\mathbf{a}-\mathbf{bZ}}$ is computed. Fourth, the marginal distributions of the residuals $\{\overline{f}_{U_n}\}_{n=1,\ldots,N}$ are computed. Fifth, the joint distribution of factors and residuals $f_{\mathbf{Z},\mathbf{U}}$ is specified by assuming independence among the above marginal distributions

$$f_{\mathbf{Z},\mathbf{U}} \equiv \overline{f}_{\mathbf{Z}} \overline{f}_{U_1} \cdots \overline{f}_{U_N}. \tag{50}$$

Sixth, the final market distribution is determined by the LFM $\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ as in (5).

The truncation (50) ensures that the LFM $\mathbf{X} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ satisfies the systematic-idiosyncratic conditions (44)-(45) and thus the market covariance is low-rank-diagonal as in (46).

Truncations are performed routinely. However, truncating a LFM leads to incorrect estimates of risk, see the applications in Section 3.1.

To illustrate the truncated model, suppose we have estimated the joint distribution $f_{\mathbf{X},Z}$ of a bivariate market $\mathbf{X} \equiv (X_1, X_2)'$ (say, two bonds returns) and one single factor Z (say, change in 5-yr rate) by means of a joint normal model

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 9.5 & -1.8 \\ 9.5 & 25 & -4.5 \\ -1.8 & -4.5 & 1 \end{pmatrix} \right); \tag{51}$$

and we set the coefficients $\mathbf{a}$ (in this case the "carry") and $\mathbf{b}$ (in this case the durations) as

$$\mathbf{a} = (0,0)', \quad \mathbf{b} = (-2,-5)'. \tag{52}$$

Then the truncated residuals and factors are normally distributed

$$\begin{pmatrix} U_1 \\ U_2 \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right). \tag{53}$$

and the market distribution becomes

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 8.2 & 8 \\ 8 & 21 \end{pmatrix} \right). \tag{54}$$

A comparison of (54) with (51) shows that the truncation distorts the original market distribution. For more details, we refer to the code at symmys.com/node/336.

2.3 Pure exogenous LFM's

After the dominant-residual LFM's discussed in Section 2.1 and the systematic-idiosyncratic LFM's discussed in Section 3, the third and final broad category of LFM's used in practice are pure exogenous models.

Key definition. A *pure exogenous* LFM for a market $\mathbf{X}$ is decomposition $\mathbf{X} \equiv \mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}$ as in (2), where all elements of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{Z}$ are imposed exogenously. Therefore the randomness recovered by the LFM $\mathbf{Y} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z}$ is fully specified exogenously.

Note that, since $\mathbf{X}$ and $\mathbf{Y} \equiv \mathbf{a} + \mathbf{bZ}$ are fully specified, so is the residual, $\mathbf{U} \equiv \mathbf{X} - \mathbf{Y}$. A pure exogenous LFM imposes no structure on $\mathbf{X}$, because all the features of the generic distribution of $\mathbf{X}$ are absorbed into the distribution of the residual. In some sense a pure exogenous LFM is a "no model".

Pure exogenous LFM's are used in practice, often as the first of a two-step approach to market analysis. First, a pure exogenous LFM is used to extract the residuals. Then a different LFM is used to analyze and impose structure on the distribution of the residuals.

For instance, Pure exogenous LFM's are used in models for the bond market. For a set of bond returns $\mathbf{X}$, the carry $\mathbf{a}$, the key-rate durations $\mathbf{b}$, and the changes in key rates of the yield curve $\mathbf{Z}$, are subtracted from the returns to create the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{bZ}$. The residuals are then modeled by a second LFM.

Another use of pure exogenous LFM's is the traditional bottom-up approach to risk attribution, which we discuss in Section 3.5.

2.4 Factor analysis

Factor analysis is an approximation technique for covariances.

Key definition. *Factor analysis* aims at approximating the market covariance by a low-rank-diagonal covariance

$$\mathrm{Cv}\{\mathbf{X}\} \approx \mathbf{bb}' + \mathrm{diag}\left(\boldsymbol{\delta}^2\right), \tag{55}$$

where $\mathbf{b}$ is a full-rank $N \times K$ matrix and $\boldsymbol{\delta}^2 \equiv \left(\delta_1^2, \ldots, \delta_N^2\right)'$ is a vector with positive entries.

Several methods are available to perform factor analysis, see e.g. [Rencher, 2002]. These methods are routinely offered by software packages.

To illustrate factor analysis, consider the market

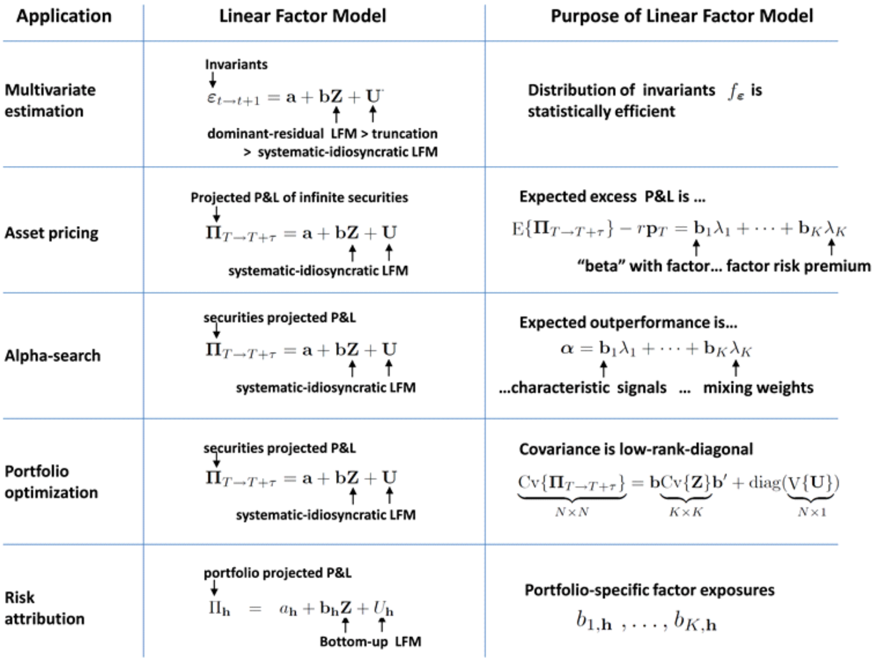
$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.06 \\ 0.3 & 0.06 & 1 \end{pmatrix}\right). \tag{56}$$

Factor analysis replaces the covariance matrix as follows

$$\begin{aligned} \mathrm{Cv}\{\mathbf{X}\} &\approx \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.0603 \\ 0.0603 & 0.3 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \end{pmatrix}' + \begin{pmatrix} 0.005 & 0 & 0 \\ 0 & 0.96 & 0 \\ 0 & 0 & 0.9 \end{pmatrix}. \end{aligned} \tag{57}$$

For more details, we refer to the code at symmys.com/node/336.

We emphasize that factor analysis is not a LFM, because it does not decompose the market as $\mathbf{X} = \mathbf{a} + \mathbf{bZ} + \mathbf{U}$. As a matter of fact, it is impossible to extract the hidden factors $\mathbf{Z}$ from the market.



Application	Linear Factor Model	Purpose of Linear Factor Model
Multivariate estimation	Invaliants ↓ $\varepsilon_{t\to t+1} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ ↑ ↑ dominant-residual LFM > truncation > systematic-idiosyncratic LFM	Distribution of invariants $f_{\boldsymbol{\varepsilon}}$ is statistically efficient
Asset pricing	Projected P&L of infinite securities ↓ $\mathbf{\Pi}_{T\to T+\tau} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ ↑ ↑ systematic-idiosyncratic LFM	Expected excess P&L is ... $\mathrm{E}\{\mathbf{\Pi}_{T\to T+\tau}\} - r\mathbf{p}_T = \mathbf{b}_1\lambda_1 + \cdots + \mathbf{b}_K\lambda_K$ ↑ ↑ "beta" with factor... factor risk premium
Alpha-search	securities projected P&L ↓ $\mathbf{\Pi}_{T\to T+\tau} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ ↑ ↑ systematic-idiosyncratic LFM	Expected outperformance is... $\boldsymbol{\alpha} = \mathbf{b}_1\lambda_1 + \cdots + \mathbf{b}_K\lambda_K$ ↑ ↑ ...characteristic signals ... mixing weights
Portfolio optimization	securities projected P&L ↓ $\mathbf{\Pi}_{T\to T+\tau} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ ↑ ↑ systematic-idiosyncratic LFM	Covariance is low-rank-diagonal $\underbrace{\mathrm{Cv}\{\mathbf{\Pi}_{T\to T+\tau}\}}_{N\times N} = \mathbf{b}\underbrace{\mathrm{Cv}\{\mathbf{Z}\}}_{K\times K}\mathbf{b}' + \mathrm{diag}(\underbrace{\mathrm{V}\{\mathbf{U}\}}_{N\times 1})$
Risk attribution	portfolio projected P&L ↓ $\Pi_{\mathbf{h}} = a_{\mathbf{h}} + \mathbf{b}_{\mathbf{h}}\mathbf{Z} + U_{\mathbf{h}}$ ↑ ↑ Bottom-up LFM	Portfolio-specific factor exposures $b_{1,\mathbf{h}}, \ldots, b_{K,\mathbf{h}}$

Figure 5: Applications of Linear Factor Models.

However, we introduced factor analysis here because at times it is confused with a statistical LFM, similar in nature to principal component analysis, and at times it is confused with a systematic-idiosyncratic LFM, because for such LFM's the covariance matrix is low-rank-diagonal as in (55). We expand on the confusion between factor analysis and LFM's in Section 4.9.

3 Applications of LFM's

The theoretical LFM's introduced in Section 2 play a prominent role in five different areas of risk and portfolio management: multivariate estimation, asset pricing theory, search for alpha, risk minimization, and risk attribution, refer to Figure 5. In Sections 3.1-3.5 we discuss these five applications respectively.

In reality, we argue that LFM's are not necessary, and actually they are better avoided, in most of the practical applications. The details of this unified no-LFM approach to quantitative finance are discussed in [Meucci, 2011a].

3.1 Multivariate estimation

Estimation is the process of fitting a distribution f_{ε} to the past observations of the invariants ε, i.e. the variables that drive the market prices, and that display an almost identical, independent distribution through time (for a stock, the invariant is its log-return; for an option the invariants include simultaneous changes in all the entries of the log-implied volatility surface, but never the return of the option price; for more on invariants, refer to [Meucci, 2005]).

Key point. Standard multivariate estimation techniques are statistically inefficient when the number N of joint simultaneous invariants $\varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_N)'$ is large. In this context, LFM's are used to impose structure on the distribution of the invariants, thereby achieving dimension reduction

$$\varepsilon \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}. \tag{58}$$

The most standard types of estimation LFM's (58) applied in the industry are truncated systematic-idiosyncratic that stem from time-series, cross-sectional, or principal component LFM's. Other types of estimation LFM's (58) include statistical-inspired models based on random-matrix-theory.

In the remainder of this section we discuss all the LFM's for multivariate estimation mentioned above. However, in the end, the following caveat will become apparent.

Warning. LFM's fail to provide accurate estimates of portfolio risk numbers. Fortunately, we can avoid using LFM's for estimation, and yet achieve efficient and very accurate risk estimates. For more details on no-LFM multivariate estimation refer to [Meucci, 2011a].

3.1.1 Truncated time-series LFM's

Time-series truncated LFM's are widely used in the industry, see e.g. [Straumann and Garidi, 2007]. In such models, the distribution f_ε of the invariants ε is estimated by means of the LFM (58), where the K factors $\mathbf{Z}$ are observable financial variables which also behave as invariants, such as industry sector returns, and where residuals and factors are made artificially independent by means of a truncation. More precisely, time-series truncated LFM's proceed as follows.

First, we collect the joint time-series $\{\epsilon_t, \mathbf{z}_t\}_{t=1,\ldots,T}$ of the invariants and the factors, which yields the joint empirical distribution of invariants and factors

$$\overline{f}_{\varepsilon,\mathbf{Z}} \iff \{\epsilon_t, \mathbf{z}_t\}_{t=1,\ldots,T}\,. \tag{59}$$

Now that we have the joint distribution of ε and $\mathbf{Z}$, we compute $\mathbf{b}$ in (58) as the coefficients of the dominant-residual model (11) relative to the joint distribution $\overline{f}_{\varepsilon,\mathbf{Z}}$

$$(\mathbf{a}, \mathbf{b}) \equiv \underset{(\alpha,\beta)\in\mathcal{C}}{\operatorname{argmax}}\, \mathrm{R}^2\{\varepsilon, \alpha + \beta\mathbf{Z}\}\,. \tag{60}$$

Without imposing constraints on β, (60) and using the empirical distribution (59), the time-series loadings (13) coincide with the standard ordinary least square (OLS) regression coefficients

$$\mathbf{b} = \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T} (\epsilon_t - \overline{\epsilon})\,(\mathbf{z}_t - \overline{\mathbf{z}})'\right)\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T} (\mathbf{z}_t - \overline{\mathbf{z}})\,(\mathbf{z}_t - \overline{\mathbf{z}})'\right)^{-1}, \tag{61}$$

where $\overline{\epsilon} \equiv \frac{1}{T}\sum_{t=1}^{T}\epsilon_t$ and $\overline{\mathbf{z}} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t$, refer to [Meucci, 2005] for the proof. Then we set $\mathbf{a} \equiv \overline{\epsilon} - \mathbf{b}\overline{\mathbf{z}}$ as in (12) to ensure that the residuals have zero expectation. The fact that we obtain the OLS coefficients (61) contributes to the confusion that LFM's are regression models, see the pitfalls in Section 4.

Next, we compute the past realizations of the residuals that follow from the coefficients $\mathbf{a}$ and $\mathbf{b}$ computed above, i.e. $\mathbf{u}_t \equiv \epsilon_t - \mathbf{a} - \mathbf{b}\mathbf{z}_t$. This yields the marginal empirical distribution of each residual

$$\overline{f}_{U_n} \iff \{u_{n,t}\}_{t=1,\ldots,T}. \tag{62}$$

Notice that so far all calculations, and in particular the loadings $\mathbf{b}$ in (60), have been performed according to the dominant-residual paradigm.

Finally, we impose by truncation as in (50) that $\mathbf{Z}$ and $\mathbf{U}$ be independent and that the entries of $\mathbf{U}$ be independent of each other, i.e.

$$f_{\mathbf{Z},\mathbf{U}} \equiv \overline{f}_{\mathbf{Z}}\overline{f}_{U_1} \cdots \overline{f}_{U_N}, \tag{63}$$

This concludes the full specification of the invariants distribution $f_\varepsilon \equiv f_{\mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}}$, which follows from the LFM (58) as in (5).

The truncation (63) guarantees that the systematic-idiosyncratic conditions (44)-(45) are satisfied by the invariant distribution f_ε. In particular, the global covariance becomes of low-rank-diagonal type as in (46), namely

$$\underbrace{\mathrm{Cv}\{\epsilon\}}_{N\times N} = \mathbf{b}\underbrace{\mathrm{Cv}\{\mathbf{Z}\}}_{K\times K}\mathbf{b}' + \mathrm{diag}\underbrace{(\mathrm{V}\{\mathbf{U}\})}_{N\times 1}. \tag{64}$$

Notice that, due to the truncation, the invariants distributions f_ε is *not* the empirical distribution $\overline{f}_\varepsilon$ that follows from the original time-series (59). As a result, the risk, as measured by the standard deviation, of a single invariant can be significantly different from the risk of the historical realizations of the invariant

$$\mathrm{Sd}\{\epsilon_n\} \neq [(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon_{t,n}^2 - \left(\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{t,n}\right)^2]^{\frac{1}{2}}. \tag{65}$$

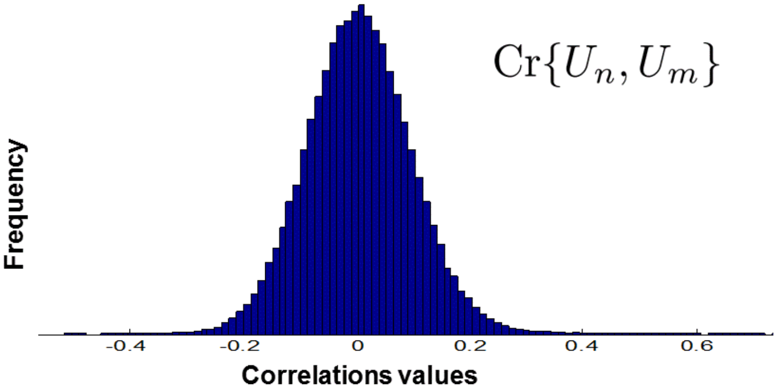


Figure 6: Cross-correlations among residuals in S&P 500 time-series LFM.

To illustrate the time-series approach, we consider as invariants ϵ_t the $N = 500$ weekly stock returns in the S&P 500. As factors $\mathbf{z}_t$ we set the weekly returns on the $K = 10$ MSCI US sector indices "Energy", "Materials", "Industrials", "Consumer Discretionary", "Consumer Staple", "Healthcare", "Financials", "IT", "Telecommunications", "Utilities".

We fit the OLS loadings $\mathbf{b}$ and the constant $\mathbf{a}$ as in (61) and thereafter and we compute the residuals $\mathbf{u}_t \equiv \epsilon_t - \mathbf{a} - \mathbf{b}\mathbf{z}_t$. The $N(N-1)/2 \approx 120{,}000$ cross-correlations of these residuals are not null, see Figure 6.

Then we truncate the model and compute the covariance (64). Next, we compute the risk of two portfolios, equal-weight and long-short pair, as it follows from the truncated covariance and the non-truncated historical covariance

	equal-weight $\mathbf{w}$	long/short $\mathbf{w}$
truncated $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0331	.1328
historical $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0333	.0796

(66)

As we see, the truncated LFM does a good job at modeling the risk of the equal-weight portfolio, whereas it does not estimate properly the long-short risk. For more details, we refer to the code at symmys.com/node/336.

We suggest here a simple enhancement of the time-series truncated approach. Using the flexible dominant-residual specification (60) we can apply useful constraints to the loadings $\mathbf{b}$. Of particular relevance in this context is the LASSO constraint $\mathcal{C} \supseteq \{\sum |b_{n,k}| \leq \xi\}$, that shrinks statistically non-significant loadings to zero.

3.1.2 Truncated cross-sectional LFM's

Truncated cross-sectional LFM's are especially popular in the industry, see e.g. [Menchero et al., 2008]. In these models, the distribution f_ε of the invariants ε is modeled by a LFM $\varepsilon \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$, where the K factors $\mathbf{Z}$ are not observable financial variables, but rather they are extracted from the invariants by a dominant-residual optimization. Then the residuals and the factors are made artificially independent by means of a truncation. More, precisely, cross-sectional truncated LFM's proceed as follows.

We start, as in the time-series approach, by collecting the past observations of the invariants ϵ_t to compute the empirical distribution, where the invariants are typically normalized to display similar volatility

$$\overline{f}_\varepsilon \iff \{\epsilon_t\}_{t=1,\ldots,T}\,. \tag{67}$$

We recall that the empirical distribution $\overline{f}_\varepsilon$ is not our desired final estimate f_ε, because it lacks structure and thus it is statistical inefficient.

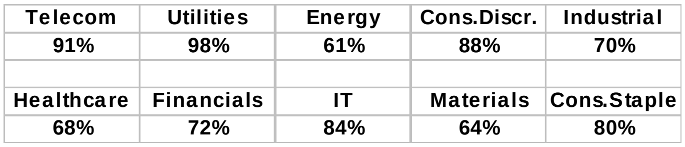
Then we specify exogenously the loadings $\mathbf{b}$ in (58) in terms of observable characteristics of the invariants. For instance, if the invariants are stock returns, the loadings are often defined in terms of accounting variables such as price-earnings rations, see e.g. [Fama and French, 1993].

Next, we extract the factors $\mathbf{Z} \equiv \mathbf{c}\varepsilon$ from the invariants ε as in the cross-sectional dominant-residual model (22)-(23)

$$(\mathbf{a}, \mathbf{c}) \equiv \underset{(\alpha,\gamma)\in\mathcal{C}}{\operatorname{argmax}}\, \mathrm{R}^2\{\varepsilon, \alpha + \mathbf{b}\gamma\varepsilon\} \tag{68}$$

Without imposing constraints on γ, (68) yields as in (24) the typical cross-sectional factor realizations[3]

$$\mathbf{z}_t = \mathbf{c}\epsilon_t = \left(\mathbf{b}'\mathbf{b}\right)^{-1}\mathbf{b}'\epsilon_t. \tag{69}$$



Telecom	Utilities	Energy	Cons.Discr.	Industrial
91%	98%	61%	88%	70%
Healthcare	Financials	IT	Materials	Cons.Staple
68%	72%	84%	64%	80%

Figure 7: Correlations of time-series and cross-section industry factors.

To illustrate, we consider again as invariants ϵ_t the returns of the $N = 500$ stocks in the S&P 500 and we extract sector-specific factors by means of binary exposures

$$b_{n,k} \equiv \begin{cases} 1 & \text{if stock } n \text{ is in sector } k \\ 0 & \text{if stock } n \text{ is not in sector } k, \end{cases} \tag{70}$$

where we partition the stocks into the same $K = 10$ sectors as in the time-series case study in Section 3.1.1. Then we extract the cross-sectional factor realizations 69. Given the structure of the loadings (70), the factors $\mathbf{z}_t$ must returns must resemble the returns of the industry indices that appear in the time-series truncated case study. Indeed, the correlations among the factors from the two LFM's is large, see Figure 7 and refer to the code at symmys.com/node/336 for more details.

Then we follow the theoretical cross-sectional specification (23) to set the coefficient $\mathbf{a} \equiv \bar{\epsilon} - \mathbf{b}\bar{\mathbf{z}}$, where $\bar{\epsilon} \equiv \frac{1}{T}\sum_{t=1}^{T}\epsilon_t$ and $\bar{\mathbf{z}} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t$. Next we compute the past realizations of the residuals $\mathbf{u}_t \equiv \epsilon_t - \mathbf{a} - \mathbf{b}\mathbf{z}_t$. From these, we compute the marginal distribution of each residual

$$\overline{f}_{U_n} \equiv \{u_{n,t}\}_{t=1,\ldots,T}\,. \tag{71}$$

Finally, we perform the truncation as in (50), setting the residuals to be independent of each other and of the factors

$$f_{\mathbf{Z},\mathbf{U}} \equiv \overline{f}_{\mathbf{Z}}\overline{f}_{U_1}\cdots\overline{f}_{U_N}. \tag{72}$$

This concludes the full specification of the invariants distribution $f_{\varepsilon} \equiv f_{\mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}}$, which follows from the LFM (58) as in (5).

Again, due to the truncation (72), the invariants distributions f_{ε} is not the empirical distribution (67). As in the time-series truncated model in Section 3.1.1, the truncation guarantees that the systematic-idiosyncratic conditions (44)-(45) are satisfied. As a result, the global covariance of the invariants achieves the desired low-rank-diagonal structure as in (64). However, the truncation creates problems in estimating the risk of non-diversified portfolios.

[3] Since the invariants ϵ_t are typically normalized before extracting the factor realizations by a suitable matrix σ^2, then $\mathbf{z}_t \equiv (\mathbf{b}'{\sigma^2}^{-1}\mathbf{b})^{-1}\mathbf{b}'{\sigma^2}^{-1}\epsilon_t$

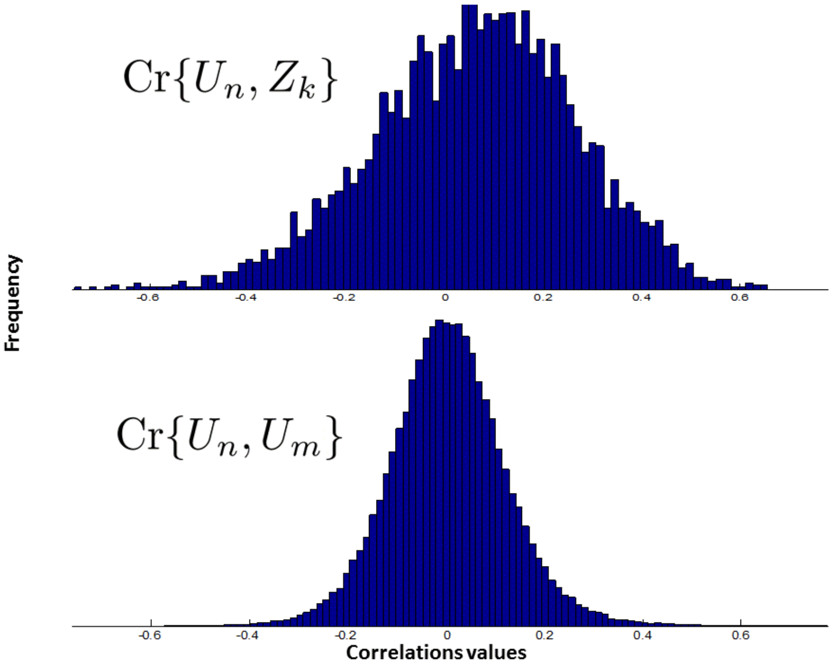


Figure 8: Correlations among residuals and between residuals and factors in cross-sectional LFM.

Continuing with our cross-sectional example in the S&P 500, we compute the constant $\mathbf{a}$ and the residuals $\mathbf{u}_t \equiv \epsilon_t - \mathbf{a} - \mathbf{b}\mathbf{z}_t$. As in the time-series case study, the $N(N-1)/2 \approx 120,000$ cross-correlations of these residuals are not null. Furthermore, the NK correlations of residuals and factors are not null, either, see Figure 6.

Then we truncate the model and compute the risk of two portfolios, equal-weight and long-short pair, as it follows from the truncated covariance and the non-truncated historical covariance

	equal-weight $\mathbf{w}$	long/short $\mathbf{w}$
truncated $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0334	.1430
historical $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0333	.0796

(73)

As it was the case for the truncated time-series LFM (66), the truncated cross-sectional LFM does a good job at modeling the portfolio risk, whereas it does not estimate correctly the risk of a long-short position. For more details, we refer to the code at symmys.com/node/336.

As for time-series LFM's, also for cross-sectional LFM's we can generalize the above approach. Using the flexible generalized dominant-residual specification (68) we can apply constraints to the extraction coefficients as in (29) or replace the r-square target with alternative optimality criteria, as in the "Factors on Demand" approach in [Meucci, 2010b].

3.1.3 Statistical LFM's: truncated-PCA and RMT-PCA

Statistical truncated LFM's are most commonly used in the fixed income market, where the joint invariants ε are the changes in yield to maturity for select tenors on a given yield or swap curve. In this context, the first $K \equiv 3$ principal components are responsible for a large percentage of total risk, as measured by the multivariate r-square (40). The residual is then completely truncated, as it accounts for only a few percent of the total randomness in the market.

For a detailed account of the PCA truncated model in fixed income, its geometrical interpretation in terms of the location-dispersion ellipsoid, and the parallel between the principal movements shift-slope-butterfly and oscillations with different frequencies over the curve, we refer the reader to the discussion in [Meucci, 2005].

A different breed of statistical LFM's that are not truncated is based on random matrix theory (RMT), which applies when the dimension of the invariants, typically stock returns, is large.

The idea behind RMT is to use the PCA approach with two enhancements: first, an advanced statistical test determines the number of significant principal factors K; second, the residuals are not truncated.

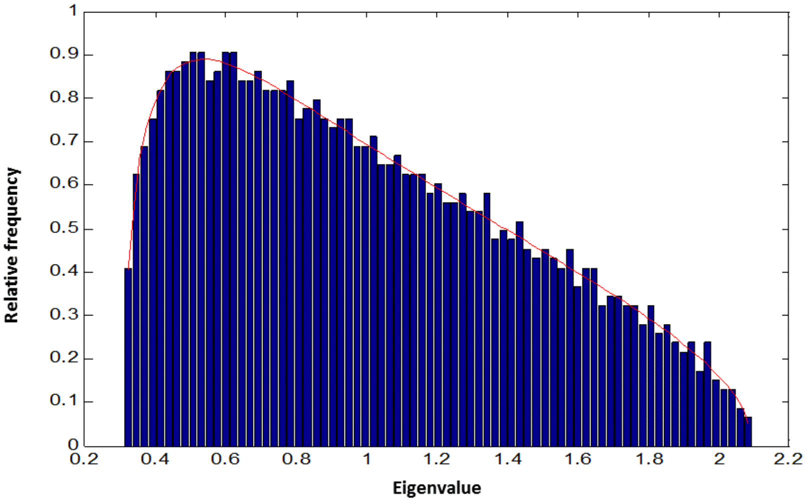


Figure 9: Marchenko-Pastur profile of pure-noise covariance eigenvalues.

To implement RMT, we start as usual from the realizations of the invariants, which we normalize to display unit variance, and we compute the empirical distribution

$$\overline{f}_{\varepsilon} \iff \{\epsilon_t\}_{t=1,\ldots,T} \, . \tag{74}$$

Then we compute the eigenvalues $\boldsymbol{\lambda}^2 \equiv \left(\lambda_1^2, \ldots, \lambda_N^2\right)'$ of the sample covariance matrix, which is actually a correlation matrix because the standard deviations are 1, and we compare them with the eigenvalues of a market of equal size with uncorrelated entries, i.e. purely noisy. The profile of the noisy eigenvalues can be computed due to the Marchenko-Pastur theorem, see Figure 9 and refer to the code at symmys.com/node/336. The deviations of the empirical eigenvalues λ_k^2 from the Marchenko-Pastur eigenvalues determines the number K of significant non-noisy factors, see [Potters et al., 2005] and [Gatheral, 2008] for more details. The benefits of this approach include the use of a lower number K of factors than in time-series or cross-sectional models.

Then we replace the smallest $N-K$ eigenvalues with their average $\tilde{\lambda}^2$

$$(\lambda_1^2,\ldots,\lambda_N^2) \mapsto \left(\lambda_1^2,\ldots,\lambda_K^2,\tilde{\lambda}^2,\ldots,\tilde{\lambda}^2\right). \tag{75}$$

The rationale of (75) is that the principal directions identified by the smallest eigenvalues are spurious, and thus it makes more sense to assume isotropy, refer to Figure 4.

Then we proceed as in (31) to extract the realizations of the first K principal factors by means of the PCA dominant-residual model $\mathbf{z}_t \equiv \mathbf{d} + \bar{\mathbf{e}}_K' \epsilon_t$, where $\bar{\mathbf{e}}_K \equiv (\mathbf{e}_1|\ldots|\mathbf{e}_K)$ is the matrix of the first K eigenvectors and $\mathbf{d}$ guarantees that the factors have zero mean. The scenarios $\mathbf{z}_t$ yield the empirical distribution of the principal factors

$$\overline{f}_{\mathbf{Z}} \iff \{\mathbf{z}_t\}_{t=1,\ldots,T}\,. \tag{76}$$

For the residuals, we can assume a normal distribution, due to the law of large numbers. To summarize

$$\mathbf{U} \sim \mathrm{N}\left(\mathbf{0},\tilde{\lambda}^2\tilde{\mathbf{e}}_K\tilde{\mathbf{e}}_K'\right), \tag{77}$$

where $\tilde{\mathbf{e}}_K \equiv (\mathbf{e}_{K+1}|\ldots|\mathbf{e}_N)$ is the matrix of the last $N-K$ eigenvectors.

Finally we specify the joint distribution of factors and residuals by imposing that they are independent

$$f_{\mathbf{Z},\mathbf{U}} \equiv \overline{f}_{\mathbf{Z}}\overline{f}_{\mathbf{U}}. \tag{78}$$

This concludes the full specification of the invariants distribution $f_\varepsilon \equiv f_{\mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}}$, which follows from the LFM (58) as in (5).

Notice that (78) does not constitute a truncation: the residuals preserve the correlations they inherited from the PCA decomposition (77). Since this statistical LFM is not truncated, i.e. it is not forced into a systematic-idiosyncratic LFM, it provides better estimates than time-series or cross-sectional truncated LFM's.

To illustrate the RMT-inspired statistical LFM, we consider again as invariants ϵ_t the returns of the $N \equiv 500$ stocks in the S&P 500. This market size is large enough to be treated by RMT.

First, we estimate with the Marchenko-Pastur filter that $K = 10$ significant factors should be used. Then we estimate the invariants distribution f_ε according to the above process.

Then we compute the risk of two portfolios, equal-weight and long-short-pair, as it follows from the RMT model and the historical covariance

	equal-weight $\mathbf{w}$	long/short $\mathbf{w}$
RMT $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0333	.0641
historical $\mathrm{Sd}\{\mathbf{w}'\varepsilon\}$	.0333	.0796

(79)

The RMT model does a good job at modeling the portfolio risk, whereas it does not estimate correctly the long-short risk. However, this effect is less pronounced than with truncated

LFM's both time-series, see (66), and cross-sectional , see (73). For more details, we refer to the code at symmys.com/node/336.

As we did for time-series and cross-sectional LFM's, we can generalize also statistical LFM's. We can apply constraints in the generalized dominant-residual specification to the extraction coefficients in (31), or replace the r-square target with alternative optimality criteria, see the "Factors on Demand" approach in [Meucci, 2010b].

3.2 Asset pricing theory

Asset pricing theory aims at establishing relationships among the returns, or, better, the P&L's generated by a set of securities. In this section we review the most popular results in asset pricing theory, namely the Capital Asset Pricing Model (CAPM) by [Sharpe, 1964] and [Lintner, 1965], and the Arbitrage Pricing Theorem (APT) by [Ross, 1976], and we discuss their connections with LFM's. For more details on CAPM and APT we refer the reader to [Ingersoll, 1987].

The following summarizes our discussion, which we state loosely here, but which we motivate precisely further below.

Key point. The CAPM is a result on the P&L's generated by N securities from the current time T to an investment horizon $T+\tau$ in the future, which we denote by $\boldsymbol{\Pi} \equiv (\Pi_{1,T\rightarrow T+\tau}, \ldots, \Pi_{N,T\rightarrow T+\tau})'$.

The derivation of CAPM-like results does not rely on LFM's at any point. Instead, such results always hold, regardless the structure or distributions of $\boldsymbol{\Pi}$. However, one can interpret CAPM-like results as a constraint on the coefficients $\mathbf{a}$ in a one-factor dominant-residual LFM $\boldsymbol{\Pi} \equiv \mathbf{a} + \mathbf{b}Z + \mathbf{U}$, which under no circumstance is systematic-idiosyncratic.

On the other hand, the derivation of the APT assumes a multi-factor systematic-idiosyncratic LFM $\boldsymbol{\Pi} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ to start with. The APT result can be interpreted as a constraint on the first moments of the variables in this LFM.

Note that the theory of asset pricing is typically formulated in terms of returns, rather than P&L's. However, the formulation in terms of P&L is more accurate and more general. Indeed, the P&L is an undisputable quantity that is defined for all securities. On the other hand, returns are not unequivocally defined, especially for leveraged assets such as swaps and futures. Furthermore, returns are often incorrectly mistaken for invariants. We expand on this pitfall in greater detail in Section 4.2.

3.2.1 The Capital Asset Pricing Model (CAPM)

Let us consider a market with N securities. We denote the projected P&L's of the securities from the current time T to the future investment horizon $T+\tau$ by $\boldsymbol{\Pi} \equiv (\Pi_{1,T\rightarrow T+\tau}, \ldots, \Pi_{N,T\rightarrow T+\tau})'$, and their current value by $\mathbf{p} \equiv (p_{1,T}, \ldots, p_{N,T})'$.

We denote an arbitrary portfolio by the holdings in each security $\mathbf{h} \equiv (h_1, \ldots, h_N)'$. Notice that holdings are number of shares or contracts, not portfolio weights. Also, we denote by d the current budget. Then the projected portfolio P&L is $\Pi_{\mathbf{h}} \equiv \mathbf{h}'\boldsymbol{\Pi}$. Similarly, the budget constraint reads $\mathbf{h}'\mathbf{p} \equiv d$.

With these inputs we can always compute the portfolio $\mathbf{h}_{SR}$ that maximizes the Sharpe ratio, i.e. the ratio of the portfolio expected P&L over the portfolio standard deviation

$$\mathbf{h}_{SR} \equiv \underset{\mathbf{h}'\mathbf{p}\equiv d}{\operatorname{argmax}} \left\{ \frac{\mathrm{E}\{\Pi_{\mathbf{h}}\}}{\mathrm{Sd}\{\Pi_{\mathbf{h}}\}} \right\}. \tag{80}$$

Let us denote by r the risk-free return over the investment horizon. Then the following identity, which links the expected outperformance of the securities to the expected outperformance of the maximum-Sharpe-ratio portfolio, is always true

$$\mathrm{E}\{\boldsymbol{\Pi}\} - r\mathbf{p} = \frac{\mathrm{Cv}\{\boldsymbol{\Pi}, \Pi_{\mathbf{h}_{SR}}\}}{\mathrm{V}\{\Pi_{\mathbf{h}_{SR}}\}} \left(\mathrm{E}\{\Pi_{\mathbf{h}_{SR}}\} - rd\right). \tag{81}$$

Some practitioners consider (81) to be the CAPM. Instead, (81) is a constraint on the distribution of the P&L's, which always holds true without any assumption on the size of the market N, the distribution of P&L's $\boldsymbol{\Pi}$, or the investors's preferences. This is a generalization of the critique in [Roll, 1977], on which we expand in Section 4.8.

On the other hand, the CAPM is an equilibrium result that requires assumptions on the investors's preferences. More precisely, the CAPM states that if all the investors maximize a subjective trade-off between a portfolio P&L expectation $\mathrm{E}\{\Pi_{\mathbf{h}}\}$ and its standard deviation $\mathrm{Sd}\{\Pi_{\mathbf{h}}\}$, then the equilibrium portfolio for all the investors $\mathbf{h}_e$ is the maximum-Sharpe-ratio portfolio $\mathbf{h}_{SR}$ defined in (80). Thus, if the assumptions of the CAPM are satisfied, then from the general result (81), the following equilibrium relationship must hold, i.e. the CAPM

$$\mathrm{E}\{\boldsymbol{\Pi}\} - r\mathbf{p} = \frac{\mathrm{Cv}\{\boldsymbol{\Pi}, \Pi_{\mathbf{h}_e}\}}{\mathrm{V}\{\Pi_{\mathbf{h}_e}\}} \left(\mathrm{E}\{\Pi_{\mathbf{h}_e}\} - rd\right). \tag{82}$$

Now, let us discuss the relationship between LFM's and CAPM. The reader will have noticed that no LFM is required to derive, or follows from, the general CAPM-like identity (81) or the specific CAPM result (82). However, we can interpret (81)-(82) as a constraint on the constant coefficient of a dominant-residual LFM.

To see this, first let us define as risk factor the P&L of the maximum-Sharpe-ratio portfolio, i.e. $Z \equiv \Pi_{\mathbf{h}_{SR}} = \mathbf{h}'_{SR}\boldsymbol{\Pi}$. Then, let us define a standard dominant-residual time-series LFM (11) for the projected P&L's $\boldsymbol{\Pi} = \mathbf{a} + \mathbf{b}\Pi_{\mathbf{h}_{SR}} + \mathbf{U}$, where $\mathbf{b}$ maximizes the r-square as in (13) and thus reads in this context $\mathbf{b} \equiv \mathrm{Cv}\{\boldsymbol{\Pi}, \Pi_{\mathbf{h}_{SR}}\} \,/\, \mathrm{V}\{\Pi_{\mathbf{h}_{SR}}\}$; and $\mathbf{a}$ makes the expectation of the residuals $\mathbf{U}$ null as in (12), i.e. $\mathbf{a} \equiv \mathrm{E}\{\boldsymbol{\Pi}\} - \mathbf{b}\,\mathrm{E}\{\Pi_{\mathbf{h}_{SR}}\}$. Using the general CAPM-like identity (81) as well as the maximum r-square $\mathbf{b}$ in the expression for $\mathbf{a}$ we obtain the following dominant-residual time series LFM

$$\boldsymbol{\Pi} \;=\; \underbrace{r(\mathbf{p} - d\tfrac{\mathrm{Cv}\{\boldsymbol{\Pi}, \Pi_{\mathbf{h}_{SR}}\}}{\mathrm{V}\{\Pi_{\mathbf{h}_{SR}}\}})}_{\mathbf{a}} \;+\; \underbrace{\tfrac{\mathrm{Cv}\{\boldsymbol{\Pi}, \Pi_{\mathbf{h}_{SR}}\}}{\mathrm{V}\{\Pi_{\mathbf{h}_{SR}}\}}}_{\mathbf{b}} \;\underbrace{\Pi_{\mathbf{h}_{SR}}}_{Z} \;+\; \mathbf{U}. \tag{83}$$

We emphasize that under no circumstance are the residuals in the time-series LFM (83) idiosyncratic, see also the pitfalls Section 4.5.

3.2.2 The Arbitrage Pricing Theory (APT)

Unlike in the CAPM, one of the premises of the APT is that the market at a specific horizon is distributed as systematic-idiosyncratic LFM

$$\boldsymbol{\Pi} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}, \tag{84}$$

where $\mathbf{\Pi} \equiv (\Pi_{1,T+\tau}, \ldots, \Pi_{N,T+\tau})'$ denotes the projected P&L's of the securities, the loadings $\mathbf{b}$ and the factors $\mathbf{Z} \equiv (Z_1, \ldots, Z_K)'$ can be hidden or observable, and the residuals $\mathbf{U}$ satisfy (44)-(45), namely

$$\mathrm{Cv}\{U_n, Z_k\} = 0, \quad \mathrm{Cv}\{U_n, U_m\} = 0. \tag{85}$$

Furthermore, without loss of generality, we can assume that the residuals have zero expectation

$$\mathrm{E}\{Z_k\} = 0, \quad \mathrm{E}\{U_n\} = 0. \tag{86}$$

Let us denote by $\mathbf{p} \equiv (p_{1,T}, \ldots, p_{N,T})'$ the current value of the securities. The APT states if the market is systematic-idiosyncratic as in (84)-(85) and large, i.e. $N \to \infty$, the following constraint holds on the expected outperformance of the securities in excess of a risk-free investment at the rate r

$$\mathrm{E}\{\mathbf{\Pi}\} - r\mathbf{p} = \mathbf{b}\lambda, \tag{87}$$

where the coefficients $\lambda \equiv (\lambda_1, \ldots, \lambda_K)'$ are known as risk premia.

To identify the risk premia in terms of the market (84), we construct a pseudo-inverse of $\mathbf{b}$, i.e. a $K \times N$ matrix $\mathbf{b}^+$ such that $\mathbf{b}^+\mathbf{b} = \mathbf{i}_K$ as follows

$$\mathbf{b}^+ \equiv \left(\mathbf{b}'\sigma^2\mathbf{b}\right)^{-1} \mathbf{b}'\sigma^2, \tag{88}$$

where σ^2 is an arbitrary symmetric and positive definite $K \times K$ matrix[4]. Then pre-multiplying the APT result (87) by $\mathbf{b}^+$ we readily obtain

$$\lambda = \mathrm{E}\{\mathbf{b}^+\mathbf{\Pi}\} - r\mathbf{b}^+\mathbf{p}. \tag{89}$$

From (89) we see that λ are called risk premia because they are the expected excess performance of K portfolios with holdings the rows of $\mathbf{b}^+$, which generate the K P&L's $\widetilde{\mathbf{Z}} \equiv \mathbf{b}^+\mathbf{\Pi}$. Such portfolios are called factor-replicating portfolios because in a large market the P&L's $\widetilde{\mathbf{Z}}$ replicate exactly the factors $\mathbf{Z}$, except for a constant term

$$\widetilde{\mathbf{Z}} \overset{N\to\infty}{\approx} \mathbf{Z} + \mathbf{b}^+\mathbf{a}. \tag{90}$$

In

Now, let us turn to the relationship between LFM's and APT. In this respect, the APT result (87) can be interpreted as the constraint that any systematic-idiosyncratic LFM (84)-(85) imposes, in the limit of large markets $N \to \infty$, on the constant coefficient $\mathbf{a}$ of the LFM, namely

$$\left(\mathbf{i}_N - \mathbf{b}\mathbf{b}^+\right)\mathbf{a} = \left(\mathbf{i}_N - \mathbf{b}\mathbf{b}^+\right) r\mathbf{p}. \tag{91}$$

3.3 Search for "alpha"

The search for alpha aims at identifying predictive signals in the market, in order to construct risky portfolios that outperform a risk-free investment.

[4] It is sufficient that σ^2 be invertible, but symmetric and postive definite matrices perform better in practice

Key point. Let us denote by $\mathbf{\Pi} \equiv (\Pi_{1,T+\tau}, \ldots, \Pi_{N,T+\tau})'$ the projected P&L's of the securities that a manager considers for his portfolio; by $\mathbf{p}$ their current price, and by $\alpha \equiv \mathrm{E}\{\mathbf{\Pi}\} - r\mathbf{p}$, the "alphas", i.e. the expected payoff of the securities in excess of the risk-free investment.

The mainstream approach to alpha search, popularized by [Rosenberg and Lanstein, 1985], [Fama and French, 1993], [Grinold and Kahn, 1999], postulates that P&L's follows a systematic-idiosyncratic LFM as in the APT (84)

$$\mathbf{\Pi} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}. \tag{92}$$

Then, as in (87) the alphas are a combination of signals $\mathbf{b}_k \equiv (b_{1,k}, \ldots, b_{N,k})'$ mixed by the risk premia λ_k

$$\alpha = \mathbf{b}_1\lambda_1 + \cdots + \mathbf{b}_K\lambda_K. \tag{93}$$

Detecting, constructing and testing the alpha-signals $\mathbf{b}_k$ and the mixing weights λ_k are the main tasks of quantitative portfolio managers.

The signals $\mathbf{b}_k$ are typically defined in the form of security-specific characteristics. To illustrate, consider the stock market, where for each stock $n = 1, \ldots, N$ we can observe accounting characteristics such as price-earnings ratios. Alternatively, consider the fixed-income market, where securities are swap contracts of different maturity and where the characteristic of each contract $n = 1, \ldots, N$ can be the historical z-score of the slope of the curve for the contract's maturity.

Let us denote by $c_{n,k}$ the value of a generic such characteristic for stock n. Then, the signal is set proportional to the characteristic, or, better, to a non-decreasing function s of the characteristic

$$b_{n,k} = \sigma_n s\left(c_{n,k}\right). \tag{94}$$

The security-specific proportionality constant σ_n is typically an estimate of the security's volatility. Two standard choices for the non-decreasing function s are: the sorting, which associates with the characteristic $c_{n,k}$ its cross-sectional ranking among $\{c_{k,1}, \ldots, c_{k,N}\}$; the cross-sectional z-score of $c_{n,k}$ within $\{c_{k,1}, \ldots, c_{k,N}\}$.

Once the signals $\mathbf{b}_k$ in (93) have been determined, the respective mixing weights λ_k, called information content in [Grinold and Kahn, 1999], can be set heuristically. Alternatively, the mixing weights are interpreted as in the APT (89) as risk-premia associated with the factor-replicating portfolios $\mathbf{b}^{+}$, defined as in (88), typically by means of an estimate of the covariance matrix $\sigma^2 \equiv \widehat{\mathrm{Cv}}\{\mathbf{\Pi}\}$. Then the risk premia are set by direct estimation from time-series analysis

$$\lambda \equiv \widehat{\mathrm{E}}\{\mathbf{b}^{+}\mathbf{\Pi}\} - r\mathbf{b}^{+}\mathbf{p}. \tag{95}$$

One further enhancement to define the mixing weights λ_k is to blend the views on the factor-replicating portfolios in (95) with a neutral prior $\lambda \equiv \mathbf{0}$ using the approach by [Black and Litterman, 1992]

Warning. In practice, it is impossible to detect all the systematic signals $\mathbf{b}$ in the LFM (92) in such a way that the residual $\mathbf{U}$ is truly idiosyncratic.

Furthermore, in typical applications the dimension of the market is too small for the asymptotic results of the APT to hold even in approximation, see also Section 4.10.

Fortunately, we do not need LFM's to search for alpha. As a matter of fact, we can perform alpha-searches by avoiding LFM's completely: the information contained in the predictive signals $\mathbf{b}$ and in the weights λ are weak relative ranking relationships, rather than direct statements on the alphas. This kind of information is more effectively embedded in the allocation process by the non-invasive Entropy Pooling approach in [Meucci, 2008]. For more details on no-LFM alpha search and how it fits with the other no-LFM applications we refer to [Meucci, 2011a].

3.4 Portfolio optimization

In its simplest form, portfolio optimization is the process of reducing the variance of a portfolio P&L, while satisfying constraints on expected performance and other investment constraints.

Key point. In the mainstream approach to portfolio optimization, the projected P&L's of the securities are modeled as a systematic-idiosyncratic LFM as in (43)

$$\mathbf{\Pi} \equiv \mathbf{a} + \mathbf{bZ} + \mathbf{U}. \tag{96}$$

Then as in (46) the covariance matrix of the P&L's is low-rank-diagonal

$$\underbrace{\mathrm{Cv}\{\mathbf{\Pi}\}}_{N\times N} = \mathbf{b}\underbrace{\mathrm{Cv}\{\mathbf{Z}\}}_{K\times K}\mathbf{b}' + \mathrm{diag}(\underbrace{\mathrm{V}\{\mathbf{U}\}}_{N\times 1}). \tag{97}$$

The low-rank-diagonal structure (97) allows for fast algorithms to minimize risk.

To illustrate the importance of the low-rank-diagonal structure, let us consider a simple case where risk is minimized with only one constraint on the portfolio performance in excess of the risk-free rate. Let us denote the holdings of each security in an arbitrary portfolio by $\mathbf{h} \equiv (h_1, \ldots, h_N)'$; let us denote the securities excess performance by $\alpha \equiv \mathrm{E}\{\mathbf{\Pi}\} - r\mathbf{p}$; and let us require an excess performance α_p from the portfolio. Then the optimization problem reads

$$\mathbf{h}^* \equiv \underset{\mathbf{h}'\alpha=\alpha_p}{\mathrm{argmin}} \left\{\mathbf{h}' \,\mathrm{Cv}\{\mathbf{\Pi}\}\, \mathbf{h}\right\}, \tag{98}$$

where α is typically specified as in (93). By applying the first order conditions to the Lagrangian of (98) we obtain

$$\mathbf{h}^* = \alpha_p \tfrac{\mathrm{Cv}\{\mathbf{\Pi}\}^{-1}\alpha}{\alpha'\,\mathrm{Cv}\{\mathbf{\Pi}\}^{-1}\alpha}. \tag{99}$$

Inverting the covariance matrix in (99) is challenging when the market is large. However, if the covariance is low-rank-diagonal as in (97), using the matrix inversion lemma we only need to invert two small $K \times K$ matrices

$$\mathrm{Cv}\{\mathbf{\Pi}\}^{-1} = \tfrac{1}{\mathrm{V}\{\mathbf{U}\}} - \tfrac{1}{\mathrm{V}\{\mathbf{U}\}}\mathbf{b}\left[\mathbf{b}'\tfrac{1}{\mathrm{V}\{\mathbf{U}\}}\mathbf{b} + \mathrm{Cv}\{\mathbf{Z}\}^{-1}\right]^{-1}\mathbf{b}'\tfrac{1}{\mathrm{V}\{\mathbf{U}\}}, \tag{100}$$

where $\frac{1}{\mathbb{V}\{\mathbf{U}\}}$ denotes a diagonal matrix of inverse variances. In real applications, the variance minimization (98) is more heavily constrained. However, the low-rank-diagonal structure (97) still allows for the implementation of fast algorithms, see e.g. [Boyd and Vandenberghe, 2004].

The standard approach to obtain the systematic-idiosyncratic LFM (96) that induces the low-rank-diagonal covariance (97) proceeds as follows. First, a global LFM is estimated for all the invariants, which include returns for stock, changes in yield for bonds, log-changes in implied volatilities surfaces for options, etc.

$$\varepsilon \equiv \overline{\mathbf{a}} + \overline{\mathbf{b}}\mathbf{Z} + \overline{\mathbf{U}}. \tag{101}$$

The LFM (101) is typically estimated by truncating a dominant-residual cross-sectional or time-series LFM into a systematic-idiosyncratic LFM, as discussed in Section 3.1. Then the invariants ε are mapped into the projected P&L's of the securities $\boldsymbol{\Pi}$ by means of linear pricing approximations

$$\Pi \approx \theta + \mathrm{diag}(\delta)\,\varepsilon, \tag{102}$$

where θ and δ are $N \times 1$ vectors of suitable coefficients. For instance, for stock-like securities, where the invariants are the compounded returns, (102) follows from a Taylor expansion of the exponential, see (110)-(112) below; for government bonds, where the invariants are changes in yields for relevant points of the curve, (102) is the "carry-duration" approximation; for vanilla options on stocks, where the invariants are the compounded returns of the underlying and the log-changes in implied volatility, (102) is the "theta-delta-vega" approximation; etc.

The coefficients of the linear pricing map the systematic-idiosyncratic LFM (101) into a LFM for the projected P&L's

$$\boldsymbol{\Pi} \equiv \underbrace{\theta + \mathrm{diag}(\delta)\,\overline{\mathbf{a}}}_{\mathbf{a}} + \underbrace{\mathrm{diag}(\delta)\,\overline{\mathbf{b}}}_{\mathbf{b}}\mathbf{Z} + \underbrace{\mathrm{diag}(\delta)\,\overline{\mathbf{U}}}_{\mathbf{U}}. \tag{103}$$

The LFM (103) has the systematic-idiosyncratic structure required by (96).

Warning. The two-step LFM (101)-(103) presents shortcomings. First, the systematic-idiosyncratic truncation distorts the risk numbers, see (54) for the theory and (66)-(73) for applications. Second, the linear pricing approximation (102) does not correctly model the randomness of non-linear securities, such as options. Furthermore, if the expected payoff of the securities is determined by a LFM as in (92), using a LFM for the covariance can create suboptimal portfolios, see [Lee and Stefek, 2008]

Fortunately, it is not necessary to rely on the two-step LFM for portfolio optimization. As a matter of fact, we can perform efficient risk minimization without resorting to truncations or linear approximations, by avoiding LFM's completely. We refer to [Meucci, 2011a] for more details on no-LFM portfolio optimization and how it fits with the other no-LFM applications.

3.5 Risk attribution

Risk attribution is the process of expressing the projected P&L of a portfolio into a few key drivers.

Key point. Let us denote by $\Pi_{\mathbf{h}}$ the projected P&L of a portfolio with holdings $\mathbf{h}$. Risk attribution is a LFM that expresses $\Pi_{\mathbf{h}}$ as a linear combination of risk factors $\mathbf{Z} \equiv (Z_1, \ldots, Z_K)'$ plus a possible residual

$$\Pi_{\mathbf{h}} = a_{\mathbf{h}} + \mathbf{b}_{\mathbf{h}}\mathbf{Z} + U_{\mathbf{h}}. \tag{104}$$

Risk factors can be industry sector returns or more general style factors.

In mainstream applications, the risk attribution (104) is achieved by means of a "bottom up" approach. First, a global truncated LFM $\varepsilon \equiv \bar{\mathbf{a}} + \bar{\mathbf{b}}\mathbf{Z} + \overline{\mathbf{U}}$ is estimated as in (101) for the invariants ε, such as returns for stock, log-changes in implied volatilities for options, etc., with suitable coefficients $(\bar{\mathbf{a}}, \bar{\mathbf{b}})$, factors $\mathbf{Z}$ and residuals $\overline{\mathbf{U}}$.

Next, the invariants LFM $\varepsilon \equiv \bar{\mathbf{a}} + \bar{\mathbf{b}}\mathbf{Z} + \overline{\mathbf{U}}$ is mapped into a second, separate LFM $\mathbf{\Pi} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ for the projected P&L by means of linear pricing approximations as in (103). Notice that in this second LFM for the P&L's the factors $\mathbf{Z}$ are the same as in the estimation LFM $\varepsilon \equiv \bar{\mathbf{a}} + \bar{\mathbf{b}}\mathbf{Z} + \overline{\mathbf{U}}$.

Finally, the portfolio P&L is computed as $\Pi_{\mathbf{h}} \equiv \mathbf{h}'\mathbf{\Pi}$, which yields bottom-up a third LFM

$$\mathbf{\Pi} \equiv \underbrace{\mathbf{h}'\mathbf{a}}_{a_{\mathbf{h}}} + \underbrace{\mathbf{h}'\mathbf{b}}_{\mathbf{b}_{\mathbf{h}}}\mathbf{Z} + \underbrace{\mathbf{h}'\mathbf{U}}_{U_{\mathbf{h}}}. \tag{105}$$

This way we obtain the desired attribution LFM (104).

Warning. The mainstream three-step bottom-up attribution LFM (101)-(103)-(105) presents shortcomings. First, it is suboptimal to use for attribution the same factors $\mathbf{Z}$ used for for estimation. Second, the attribution should be tailored top-down to the portfolio, instead of being inherited bottom-up from the securities. Third, the shortcomings of the systematic-idiosyncratic truncation for estimation discussed in Section 3.1 apply. Finally, the shortcomings of the linear approximation for pricing (102) apply.

Fortunately, we do not need to rely on an estimation LFM to perform attribution. Furthermore, we can use Factors on Demand (FoD) in [Meucci, 2010b] to widely broaden the spectrum of applications of attribution. For instance, we can set $\mathbf{Z}$ as the returns of potential hedging instruments, and compute the portfolio-specific hedges $\mathbf{b}_{\mathbf{h}}$ that minimize the downside CVaR of the hedged portfolio $\Pi_{\mathbf{h}} - \mathbf{b}_{\mathbf{h}}\mathbf{Z}$. Also, we can create an optimal, parsimonious LFM for portfolios of arbitrary non-linear instruments in terms of arbitrary on-the-fly risk factors $\mathbf{Z}$. We refer to [Meucci, 2011a] for more details on this no-LFM approach to risk attribution and how it ties to the other no-LFM applications.

4 Pitfalls of LFM's

As we saw in the discussion so far, LFM's are ubiquitous in finance. LFM's appear deceptively simple to grasp and to implement. As a result, numerous pitfalls lurk behind LFM's. This is evident in the content of the following box, which at first impression might appear correct, but which in reality is fraught with inaccuracies.

Pitfalls. A LFM is a *regression of the past*$^{(1)}$ *returns*$^{(2)}$ $\mathbf{r}_{t\rightarrow t+1}\equiv\ln(\mathbf{p}_{t+1}/\mathbf{p}_{t})$, or *equivalently*$^{(3)}$ $\mathbf{r}_{t\rightarrow t+1}\equiv\mathbf{p}_{t+1}/\mathbf{p}_{t}-1$, of a set of *stocks*$^{(4)}$ on a set of *factor returns*$^{(5)}$ $\mathbf{z}_{t\rightarrow t+1}$:

$$\mathbf{r}_{t\rightarrow t+1}=\mathbf{a}+\mathbf{b}\mathbf{z}_{t\rightarrow t+1}+\mathbf{u}_{t\rightarrow t+1}. \tag{106}$$

The factors $\mathbf{z}_{t\rightarrow t+1}$ are *systematic*$^{(6)}$ and the residuals $\mathbf{u}_{t\rightarrow t+1}$ are *idiosyncratic*$^{(7)}$. The coefficients $\mathbf{a}$ and $\mathbf{b}$ do not depend on the *time-step*$^{(8)}$.

LFM's are *dimension reduction techniques*$^{(9)}$, *validated*$^{(10)}$ by asset pricing theory. If $\mathbf{z}_{t\rightarrow t+1}$ is a broad stock index, (106) *is the CAPM*$^{(11)}$. If $\mathbf{z}_{t\rightarrow t+1}$ is a vector with multiple entries, (106) *is the APT*$^{(12)}$.

LFM's are *necessary*$^{(13)}$ for multivariate estimation, in the form of cross-sectional, time-series or statistical LFM's, such as principal components and *factor analysis*$^{(14)}$.

LFM's are *necessary*$^{(15)}$ for alpha search, to extract the alpha-generating factors $\mathbf{z}_{t\rightarrow t+1}$ from the securities returns $\mathbf{r}_{t\rightarrow t+1}$ by means of *factor-mimicking portfolios*$^{(16)}$.

LFM's are *necessary*$^{(17)}$ for portfolio optimization, to invert the covariance matrix of returns and thus implement optimization algorithms.

LFM's are *necessary*$^{(18)}$ for risk attribution, to see how the estimation factors affect the portfolio return.

Below we discuss the above pitfalls.

In order to properly analyze the above issues, we reproduce here the actual definition of LFM (2). A LFM for a variable $\mathbf{X}\equiv(X_1,\ldots,X_N)'$ is a decomposition of $\mathbf{X}$ as a linear combination of factors $\mathbf{Z}\equiv(Z_1,\ldots,Z_K)'$ plus residuals $\mathbf{U}\equiv(U_1,\ldots,U_N)'$, as follows

$$\mathbf{X}\equiv\mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}. \tag{107}$$

The LFM is fully specified by the vector $\mathbf{a}\equiv(a_1,\ldots,a_N)'$, the matrix $\mathbf{b}\equiv\{b_{n,k}\}$ and the structure in the joint distribution $f_{\mathbf{X},\mathbf{Z},\mathbf{U}}$ of the random variables $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{U}$.

To help the reader put the different pitfalls in perspective, we provide a subjective ranking of the relative importance of the pitfalls, which we denote as follows:

$$\begin{aligned} &[\star] && = \text{academic remark}\\ &[\star\star] && = \text{caveat}\\ &[\star\star\star] && = \text{crucial issue.}\end{aligned} \tag{108}$$

4.1 LFM's are not a regression on past data$^{(1)}$ $[\star\star]$

A LFM (107) is a predictive model for future yet-to-be realized simultaneous random variables $\mathbf{X}$, rather than models for a panel of past realized data $\{\mathbf{x}_t\}_{t=1,\ldots,T}$.

In particular, for estimation, LFM's model the next-step invariants $\mathbf{X}\equiv\epsilon_{T\rightarrow T+1}$, such as returns or log-yield changes. For asset pricing, alpha search, optimization, and attribution/hedging, LFM's models the projected P&L's $\mathbf{X}\equiv\mathbf{\Pi}_{T\rightarrow T+\tau}$ between the current time T and a future investment horizon $T+\tau$.

When LFM's are used for estimation, only one of the many possible applications of LFM's, LFM's do operate on past data. In particular, when the future distribution of the random invariants is modeled as the empirical distribution stemming from past data, one obtains the regressions (61) and (69). However, it is important to keep in mind that the empirical distribution is but one of the many possible ways to model the future distribution of the invariants: different distributional assumptions would give rise to non-regressive formulas for time-series and cross-sectional LFM's.

4.2 LFM's are not about securities returns$^{(2\text{-}3)}$ [$\star\star$]

Often LFM's are thought of as models for the securities returns. Instead, the variables $\mathbf{X}$ for the LFM's (107) are either invariants $\mathbf{X} \equiv \epsilon_{T\to T+1}$ or P&L's $\mathbf{X} \equiv \mathbf{\Pi}_{T\to T+\tau}$. Hence, $\mathbf{X}$ are not returns. Before we discuss further this pitfall, we need to define precisely the above variables, refer to [Meucci, 2011b] for more details.

Definition. The *profit and loss* (P&L) $\Pi_{t\to t+\tau}$ generated by a security or a portfolio over an arbitrary period is the difference between the value of the portfolio at the end of the period, $P_{t+\tau}$, and the value of the portfolio at the beginning of the period, P_t, plus the sum of any intermediate endogenous cashflows, $CF_{t\to t+\tau}$,

$$\Pi_{t\to t+\tau} \equiv P_{t+\tau} - P_t + CF_{t\to t+\tau}. \tag{109}$$

Definition. The *linear return* $R_{t\to t+\tau}$ of a security or, more in general, a portfolio over an arbitrary period is the P&L generated, $\Pi_{t\to t+\tau}$, divided by a normalizing quantity which is known at the beginning of the period, the basis denominator D_t

$$R_{t\to t+\tau} \equiv \frac{\Pi_{t\to t+\tau}}{D_t}. \tag{110}$$

For stocks, the basis denominator is typically the market value, i.e. $D_t \equiv P_t$. More in general, the basis allows us to define returns for leveraged products. For instance, for swaps a standard basis is the duration.

Definition. The *compounded return* $C_{t\to t+\tau}$ of a security or, more in general, a portfolio over an arbitrary period is the ratio of the portfolio value at the end of the period over the value at the beginning of the period, when defined

$$C_{t\to t+\tau} \equiv \ln \frac{P_{t+\tau}}{P_t}. \tag{111}$$

Notice a first pitfall: when the time to the horizon τ is short, the securities are not volatile, and the basis denominator is the market value, linear and compounded returns are similar

$$R_{t\to t+\tau} = \tfrac{P_{t+\tau}}{P_t} - 1 = e^{\ln \frac{P_{t+\tau}}{P_t}} - 1 = e^{C_{t\to t+\tau}} - 1 \approx C_{t\to t+\tau} \tag{112}$$

Definition. The *invariants* $\varepsilon_{t\to t+1} \equiv (\varepsilon_{1,t\to t+1}, \ldots, \varepsilon_{N,t\to t+1})'$ are a possibly large set of N simultaneous shocks that occur over a given unit time step $t \to t+1$ and that fully determine the joint dynamics of the prices over the given time step. The invariants are identically and independently distributed (i.i.d.) variables: the simultaneous invariants $\varepsilon_{t\to t+1} \equiv (\varepsilon_{1,t\to t+1}, \ldots, \varepsilon_{N,t\to t+1})'$ have the same joint N-variate distribution f_ε at all times t, and invariants $\varepsilon_{t\to t+1}$ and $\varepsilon_{s\to s+1}$ across different time steps are independent.

Notice a second pitfall: for a stock with price P_t the natural invariant is its compounded return $\varepsilon_{t\to t+1} \equiv C_{t\to t+1}$. For an option with price P_t the natural invariants include the log-changes of the whole implied volatility surface, but the compounded return of the option is not an invariant $\varepsilon_{t\to t+1} \neq C_{t\to t+1}$.

Now, let us turn to more pitfalls, related to LFM's.

For the application "multivariate estimation", discussed in Section 3.1, we must formulate the LFM's on the invariants $\mathbf{X} \equiv \varepsilon_{T\to T+1}$. For stocks, compounded returns are invariants and thus linear returns are invariants

$$C_{t\to t+1} \text{ i.i.d.} \Rightarrow R_{t\to t+1} = e^{C_{t\to t+1}} - 1 \text{ i.i.d.} \tag{113}$$

In general markets, such as options, returns are not invariants, whether compounded or linear. Hence, for most securities, no estimation can be performed based on past returns. Thus, it is incorrect to set $\mathbf{X} \equiv \mathbf{R}_{T\to T+1}$ in a LFM (107) for estimation, see also [Meucci, 2010c].

For the applications "asset pricing", "search for alpha", "portfolio optimization", and "risk attribution", discussed respectively in Sections 3.2-3.5, we advocated the formulation of LFM's in terms of future P&L's $\mathbf{X} \equiv \mathbf{\Pi}_{T\to T+\tau}$.

By choosing a suitable basis in the definition of returns (110), in principle any result on the future P&L's $\mathbf{\Pi}_{T\to T+\tau}$ can be transformed into a result for future linear returns $\mathbf{R}_{T\to T+\tau}$. However, it is better to avoid using linear returns for these applications, for four reasons.

First, returns are often confused with invariants, which is only true for stocks, see (113). Hence, one is tempted to link results on future P&L's $\mathbf{\Pi}_{T\to T+\tau}$ with past returns, which is incorrect. For instance, past returns on an option bear no connection with the future P&L of the option.

Second, linear returns are often confused with compounded returns, due to the approximation (112). However using the compounded returns $\mathbf{X} \equiv \mathbf{C}_{T\to T+\tau}$ in a LFM (107) for asset pricing, alpha search, optimization, or attribution must be avoided because, unlike linear returns, compounded returns do not aggregate across securities and they are not investable.

Third, commonly accepted results in terms of returns are not properly specified in general markets. For instance, the CAPM-like expression (81), which we report here in terms of returns

$$\mathrm{E}\{\mathbf{R}\} - r = \frac{\mathrm{Cv}\{\mathbf{R}, R_{\mathbf{h}_{SR}}\}}{\mathrm{V}\{R_{\mathbf{h}_{SR}}\}} \left(\mathrm{E}\{R_{\mathbf{h}_{SR}}\} - r\right), \tag{114}$$

is incorrect for, say, swaps, and one must instead use the P&L formulation (81).

Fourth, the formulation of a LFM in terms of P&L's $\mathbf{X} \equiv \mathbf{\Pi}_{T\to T+\tau}$ emphasizes the importance of the investment horizon in the definition of the model. This feature is often neglected when stating the model in terms of returns, mainly because the distribution of the returns is incorrectly assumed to be independent of the horizon, except for a square-root-of-horizon scale factor which does not affect the LFM, see also Section 4.6 below.

4.3 LFM's are not about stocks$^{(4)}$ [$\star\star$]

The different applications of LFM's (107) in finance, namely estimation, asset pricing, alpha search, optimization, and attribution, were originally developed for the most liquid market, namely the stock market, where $\mathbf{X} \equiv \mathbf{R}$, the return of the stocks, see also the pitfalls in Section 4.2.

The formulation $\mathbf{X} \equiv \varepsilon$ for estimation in terms of fully general invariants and the formulation $\mathbf{X} \equiv \mathbf{\Pi}$ for asset pricing, alpha search, optimization, and attribution allows us to extend the applications of LFM's to all tradable asset classes.

For instance, we can use a LFM to impose structure on the joint distribution of the log-changes of all the implied volatility surfaces $\varepsilon \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$, or we can use a LFM to model the joint distribution of the P&L of swap and futures contracts $\mathbf{\Pi} \equiv \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$.

4.4 LFM's "factors" are not "factors returns"$^{(5)}$ [$\star$]

The term $\mathbf{Z} \equiv (Z_1, \ldots, Z_K)'$ on the right hand side of a LFM decomposition (107) in general are "factors", i.e. fully general random variables, not "factor returns", i.e. the normalized P&L's of tradable instruments. To understand why, let us separate the case when $\mathbf{Z}$ is set exogenously, as in time-series LFM's, from the case when $\mathbf{Z}$ is implied endogenously from the market $\mathbf{X}$, as in cross-sectional LFM's.

When set exogenously, as in time series LFM, the factors $\mathbf{Z}$ can be in principle any random variables, including returns, linear or compounded, but also prices, square prices, etc. For estimation purposes, $\mathbf{Z}$ have to be invariants. Thus, for estimation, $\mathbf{Z}$ can be returns only if the underlying are stock-like securities.

When extracted endogenously from the market $\mathbf{X}$, as in cross-sectional LFM's or statistical LFM's, the factors $\mathbf{Z}$ can be interpreted as returns only when the market in the LFM (107) are linear returns of a set of securities, i.e. $\mathbf{X} \equiv \mathbf{R}$ Then the factors $\mathbf{Z}$ are extracted from the market $\mathbf{X}$ as in (23) and (32) by means of a $K \times N$ matrix $\mathbf{c}$

$$\mathbf{Z} \equiv \mathbf{cR}. \tag{115}$$

Then the K rows of $\mathbf{c}$ can be interpreted as the weights of K portfolios, and thus the factors $\mathbf{Z}$ can be interpreted as the returns of K factor portfolios.

Notice that we do not obtain factor returns if in the LFM (107) we set $\mathbf{X} \equiv \mathbf{C}$, the compounded returns, because linear combinations of compounded returns are not returns, see [Meucci, 2010c]. Also, we do not obtain factor returns if in the LFM (107) we set $\mathbf{X} \equiv \epsilon$, i.e. generic invariants, such as changes in log-implied-volatility.

4.5 LFM's are not systematic-idiosyncratic$^{(6\text{-}7)}$ [$\star\star\star$]

We recall that the generic LFM $\mathbf{X} = \mathbf{a} + \mathbf{bZ} + \mathbf{U}$ is systematic-idiosyncratic if, as in (44), the factors $\mathbf{Z}$ are uncorrelated with the residuals $\mathbf{U}$ and, as in (45), the residuals $\mathbf{U}$ are uncorrelated with each other

$$\mathrm{Cv}\{U_n, Z_k\} = 0, \quad \mathrm{Cv}\{U_n, U_m\} = 0. \tag{116}$$

Here we show that systematic-idiosyncratic LFM's cannot and should not be implemented in practice for two reasons: empirically, the assumptions underlying systematic-idiosyncratic LFM's deviate significantly from real markets; operationally, systematic-idiosyncratic LFM's call for inaccurate P&L computations.

From an empirical point of view, systematic-idiosyncratic LFM's do not accurately describe the markets. As we proceed to show, for a generic market $\mathbf{X}$, regardless what factors $\mathbf{Z}$ are isolated and what loadings $\mathbf{b}$ are used, the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{bZ}$ always appear correlated, in violation of (116). As a result, most commercially available LFM's, which are estimated with the dominant-residual criterion (time-series, cross-sectional, or statistical), and then treated as systematic-idiosyncratic LFM's after truncating the correlations of the residuals as discussed in Section 3.1, yields incorrect risk computations, see (66), (73) and (79).

First, let us start with a market $\mathbf{X}$ where the factors $\mathbf{Z}$ are specified exogenously, as in time-series LFM's. Is it possible to determine factor loadings $\mathbf{b}$ such that the residuals

$\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{bZ}$ satisfy the systematic-idiosyncratic requirements (116)? The answer is no. To do so, we would have to satisfy

$$\mathrm{Cv}\{X_n - \textstyle\sum_k b_{n,k} Z_k, X_m - \sum_j b_{m,j} Z_j\} \equiv 0 \text{ for all } n > m \tag{117}$$

$$\mathrm{Cv}\{X_n - \textstyle\sum_j b_{n,j} Z_j, Z_k\} \equiv 0 \text{ for all } n, k. \tag{118}$$

The constraints (117) are $N(N+1)/2$ equations and the constraints (118) are NK equations, whereas $\mathbf{b}$ only contains NK entries: the system is overdetermined and the solution does not exist. Of course, it is possible to satisfy a subset of all the requirements. For instance, remarkably the maximum r-square solution $\mathbf{b} \equiv \mathrm{Cv}\{\mathbf{X}, \mathbf{Z}\} \mathrm{Cv}\{\mathbf{Z}\}^{-1}$ obtained in (13) satisfies (118), as we prove in Appendix A.1. However, (117) cannot be satisfied and thus the residual is not idiosyncratic, as we saw in the time series simplified example (18), as well as in the stock market case study in Figure 6.

Second, let us now consider a market $\mathbf{X}$ where the factors $\mathbf{Z}$ are extracted from the market as in cross-sectional or statistical LFM's, i.e. $\mathbf{Z} \equiv \mathbf{cX}$, where $\mathbf{c}$ is a $K \times N$ matrix. Is it possible to determine extraction coefficients $\mathbf{c}$ and possibly factor loadings $\mathbf{b}$ such that the residuals $\mathbf{U} \equiv \mathbf{X} - \mathbf{a} - \mathbf{bcX}$ satisfy the systematic-idiosyncratic requirements (116)? The answer is, again, no. Indeed, to do so, we would have to satisfy

$$\mathrm{Cv}\{X_n - \textstyle\sum_{k,p} b_{n,k} c_{k,p} X_p, X_m - \sum_{j,q} b_{n,j} c_{j,q} X_q\} \equiv 0 \text{ for all } n > m \tag{119}$$

$$\mathrm{Cv}\{X_n - \textstyle\sum_{j,q} b_{n,j} c_{j,q} X_q, \sum_p c_{k,p} X_p\} \equiv 0 \text{ for all } n, k. \tag{120}$$

These $N(N+1)/2 + NK$ equations overdetermine the $2NK$ entries of $\mathbf{b}$ and $\mathbf{c}$. As a result, cross-sectional LFM's are not systematic-idiosyncratic, as we saw in the simplified example (28), as well as in the stock market case study in Figure 8. Similarly, PCA residuals are correlated, see (42)

From an operational point of view, systematic-idiosyncratic LFM's are problematic, because they are very sensitive to pricing transformations and thus force modelers to resort to inaccurate P&L proxies.

To illustrate this problem, let us consider the simple case of stock-like securities, where the compounded returns are invariants, i.e. $\mathbf{C} = \varepsilon$. Even if the compounded returns were to follow a systematic-idiosyncratic LFM, then the linear returns, which as in (112) satisfy $\mathbf{R} = e^{\mathbf{C}} - \mathbf{1}$, or the P&L's, which in terms of the current prices $\mathbf{p}$ read $\mathbf{\Pi} = \mathbf{pR}$, do not follow a systematic-idiosyncratic LFM, see [Meucci, 2009].

More in general, let us assume that the invariants ε of a given market are estimated using a systematic-idiosyncratic LFM, although we know from the introduction to this section that such LFM is inaccurate. Only the inaccurate linear pricing approximation (102) yields a systematic-idiosyncratic LFM for the projected P&L's of the securities.

Fortunately, systematic-idiosyncratic LFM's are not necessary, see Section 4.11 below and references therein. Therefore, we can safely avoid the truncations of Section 3.1 and of the linear approximations (102).

4.6 LFM's are not horizon-independent$^{(8)}$ [$\star\star$]

In standard applications, when specifying a LFM on the P&L's $\mathbf{\Pi}_{T \to T+\tau} \equiv \mathbf{a} + \mathbf{bZ} + \mathbf{U}$ typically little attention is paid to the fact that coefficients, factors and residuals all depend on the horizon.

In particular, the factor loadings $\mathbf{b}_{T\to T+\tau}$ depend on the whole span between the current time and the investment horizon. This dependence becomes important when evaluating a fund manager's performance based as the risk-adjusted excess P&L $\Pi_{T\to T+\tau} - b_{T\to T+\tau} Z_{T\to T+\tau}$, where the benchmark is, say, the return of a global equity index, see Kepos (2011).

In order to account for the horizon, one must embed a suitable projection step for all the random variables involved. For more details and code, refer to [Meucci, 2010a].

4.7 LFM's are not a dimension reduction technique$^{(9)}$ [$\star$]

LFM's (107) are dimension reduction techniques when used to estimate the distribution of the invariants $\mathbf{X} \equiv \epsilon$ in large dimensional markets.

LFM's are not a dimension reduction technique when used as in (104) to compute the optimal hedge or to attribute the risk in the portfolio P&L to a few key drivers $\Pi_h = a_{\mathbf{h}} + \mathbf{b}_{\mathbf{h}}\mathbf{Z} + U_{\mathbf{h}}$. Indeed, in this latter case, the LFM actually increases the dimension of the problem, from the univariate P&L $\Pi_{\mathbf{h}}$ of the portfolio to the $K > 1$ hedges/risk drivers $\mathbf{Z}$.

4.8 LFM's are not APT and CAPM$^{(10\text{-}11\text{-}12)}$ [$\star\star\star$]

The APT and CAPM are often loosely described as two conceptually similar LFM's, whereby the CAPM has one factor and the APT has several factors. In reality, APT and CAPM are profoundly different and their connections to LFM's present many pitfalls.

In the APT, one assumes that the securities projected P&L's follow a multi-factor systematic-idiosyncratic LFM $\mathbf{\Pi} = \mathbf{a}+\mathbf{b}\mathbf{Z}+\mathbf{U}$, and that there are infinitely many securities in the market. Then the APT result (87) follows, which can be interpreted as a constraint on the coefficients $\mathbf{a}$ of the LFM. As discussed in Sections 4.5 and 4.10 respectively the two assumptions of the APT are unrealistic.

In the CAPM, no assumptions are made on the distribution of the securities P&L's $\mathbf{\Pi}$. In particular, no assumption is made that $\mathbf{\Pi}$ follow a LFM. The general CAPM-like result (81) is thus always true, regardless the distribution of $\mathbf{\Pi}$ or the size of the market. Although no LFM is needed to derive CAPM-like results, such results can be interpreted as a constraint on the coefficients $\mathbf{a}$ of a suitable dominant-residual LFM $\mathbf{\Pi} = \mathbf{a} + \mathbf{b}Z + \mathbf{U}$, which is never systematic-idiosyncratic, see (83).

We emphasize that the financial theory of asset pricing does not endorse the use of LFM's, and in particular systematic-idiosyncratic LFM's: in the APT, systematic-idiosyncratic LFM's are an unrealistic assumption; in the CAPM, no systematic-idiosyncratic LFM's ever appear.

Aside from the above caveats on the relationship between asset pricing and LFM's, it is worth repeating how some general pitfalls highlighted for LFM's, also apply to CAPM and APT.

First, both CAPM and APT apply to the joint P&L's of a set of securities, which are unequivocally defined, rather than returns, whose definition for leveraged non-equity securities is subjective. It is possible to express CAPM and APT equivalently in terms of linear returns using suitable bases as in (110), but never in terms of compounded returns. Refer also to Section 4.2 for a similar pitfall in LFM's.

Second, despite similarities in appearance, neither CAPM nor APT bear any connection with estimation theory or with regression analysis. The variables involved are yet to be realized P&L's, or equivalently linear returns. The past, dynamics of such variables is absolutely inessential to the formulation of CAPM and APT. Thus, both CAPM and APT hold exactly regardless of any autocorrelation or any volatility clustering in the time series of the P&L's or the returns of the different securities. See also Section 4.1 for a similar pitfall in LFM's.

Third, both CAPM and APT are often mistakenly assumed to only hold for the stock market, mainly due to the wealth of empirical studies performed in this area and to the fact that for stocks returns are defined in a straightforward manner. However, both APT and CAPM hold for any tradable security, which include leveraged products such as swaps and futures, as well as nonstandard products such as exotic options. See also Section 4.3 for the same pitfall in LFM's.

Fourth, both CAPM and APT are one-period models between the current time T and a given investment horizon in the future $T+\tau$. Therefore, there exist an infinity of simultaneous, potentially completely different, versions of CAPM and APT, one for each horizon τ. For instance, the risk-free rates $r_{T\rightarrow T+\tau}$ in CAPM and APT are the deterministic, linear returns of different zero-coupon bonds that expire at the different horizons $T+\tau$. Similarly, the "betas" that appear in (83) are functions of the P&L's $\Pi_{T\rightarrow T+\tau}$, or the linear returns $R_{T\rightarrow T+\tau}$, between the current time T and the future horizon $T+\tau$. Refer to Section 4.6 for a similar pitfall in LFM's and for connections with performance attribution.

4.9 LFM's do not include factor analysis$^{(14)}$ [$\star\star$]

We recall from (55) that factor analysis (FA) aims at approximating a covariance as low-rank-diagonal

$$\mathrm{Cv}\{\mathbf{X}\} \approx \mathbf{b}\mathbf{b}' + \mathrm{diag}\big(\delta^2\big)\,. \tag{121}$$

FA is at times confused with a systematic-idiosyncratic statistical LFM for three reasons.

The first reason is that the FA covariance approximation (121) follows if one assumes a systematic-idiosyncratic LFM $\mathbf{X} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$, where the factors $\mathbf{Z}$ are uncorrelated and have unit variance, i.e. $\mathrm{Cv}\{\mathbf{Z}\} = \mathbf{i}_K$; the residuals $\mathbf{U}$ are uncorrelated with each other, i.e. $\mathrm{Cv}\{\mathbf{U}\} = \mathrm{diag}\big(\delta^2\big)$, and thus they are idiosyncratic, as in (44); and the factors are uncorrelated with the residuals, i.e. $\mathrm{Cv}\{\mathbf{Z}, \mathbf{U}\} = \mathbf{0}_{K\times N}$, and thus they are systematic as in (45). However, the existence of a systematic-idiosyncratic LFM is a sufficient, but by no means necessary, condition for the FA covariance approximation (121) to hold.

The second reason why FA is confused with a LFM, and in particular a statistical LFM, is that the FA covariance approximation (121) is purely statistical, i.e. no exogenous information enters the specification of the low-rank coefficients $\mathbf{b}$ or the diagonal coefficients δ^2. Furthermore, numerical techniques to obtain the FA covariance approximation often rely on principal components analysis, which determines statistical LFM's, see Section 2.1.3.

The third reason why FA is confused with a LFM is the very terminology "factor analysis", which misleads one to think that FA defines a LFM $\mathbf{X} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$.

Despite the three above reasons to associate FA with LFM's, no such thing as a FA-extracted systematic-idiosyncratic statistical LFM $\mathbf{X} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}$ can exist, because FA cannot extract the hidden factors $\mathbf{Z}$, and therefore it cannot identify the LFM, see the proof in Appendix A.4.

4.10 LFM's do not extract alpha-generating factors$^{(16)}$ [$\star\star$]

The APT (84) assumes that the projected P&L's of the securities are described by a systematic-idiosyncratic LFM

$$\mathbf{\Pi}_{T\rightarrow T+\tau} = \mathbf{a} + \mathbf{b}\mathbf{Z} + \mathbf{U}. \tag{122}$$

Let us make a second unrealistic hypothesis, namely that we observe the loadings $\mathbf{b}$ in the systematic-idiosyncratic LFM (122). Finally, let us make a third unrealistic hypothesis of a very large market, namely that the numbers of securities in the market.

Then, it is possible to recover the K systematic factors $\mathbf{Z}$ from the N securities P&L's $\mathbf{\Pi}_{T\rightarrow T+\tau}$ by means of the K factor-replicating portfolios, defined by the rows of the $K \times N$ holdings matrix $\mathbf{b}^{+} \equiv (\mathbf{b}'\sigma^{2}\mathbf{b})^{-1}\mathbf{b}'$, see (90)

$$\mathbf{b}^{+}\mathbf{\Pi}_{T\rightarrow T+\tau} \overset{N\rightarrow\infty}{=} \mathbf{Z} + \mathbf{b}^{+}\mathbf{a}. \tag{123}$$

In general, the above three assumptions do not hold: the systematic-idiosyncratic hypothesis is empirically unrealistic, see Section 4.5; we do not observe all the common signals; and the market considered for alpha generation have finite, and often limited, size.

4.11 LFM's are not necessary$^{(13\text{-}15\text{-}17\text{-}18)}$ [$\star\star\star$]

This is our epilogue. After much effort defining, studying and applying LFM's, it turns out that LFM's can, and should, be avoided in practice. As it turns out, without LFM's it is possible to achieve significantly better risk estimates, non-spurious alpha signals, efficient portfolio optimizations, and top-down, parsimonious, on-the-fly risk attribution. We refer the curious reader to [Meucci, 2011a] for the no-LFM approach to quantitative finance.

References

[Black and Litterman, 1992] Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analyst Journal*.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

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

[Gatheral, 2008] Gatheral, J. (2008). Random matrix theory and covariance estimation. *Working Paper*.

[Grinold and Kahn, 1999] Grinold, R. C. and Kahn, R. (1999). *Active Portfolio Management. A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, 2nd edition.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

[Ingersoll, 1987] Ingersoll, E. J. (1987). *Theory of Financial Decision Making*. Rowman and Littlefield.

[Lee and Stefek, 2008] Lee, J. and Stefek, D. (2008). Do risk factors eat alphas? *The Journal of Portfolio Management*, 34:12–25.

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

[Menchero et al., 2008] Menchero, J., Morozov, A., and Shepard, P. (2008). The Barra global equity model (GEM2). *Barra Research Notes*, pages 1–79.

[Meucci, 2005] Meucci, A. (2005). Risk and asset allocation. http://symmys.com. Springer Finance.

[Meucci, 2008] Meucci, A. (2008). Fully Flexible Views: Theory and practice. http://symmys.com/node/158. Risk, 21(10), 97-102.

[Meucci, 2009] Meucci, A. (2009). Exercises in advanced risk and portfolio management - with step-by-step solutions and fully documented code. http://symmys.com/node/170. Working Paper.

[Meucci, 2010a] Meucci, A. (2010a). Common misconceptions about 'beta' - hedging, estimation and horizon effects. http://symmys.com/node/165. GARP's Risk Professional Magazine.

[Meucci, 2010b] Meucci, A. (2010b). Factors on Demand - building a platform for portfolio managers risk managers and traders. http://symmys.com/node/164. Risk, 23(7), 84-89.

[Meucci, 2010c] Meucci, A. (2010c). Linear vs. compounded returns - common pitfalls in portfolio management. http://symmys.com/node/141. Garp Risk Professional "The Quant Classroom" series 2, 49-51.

[Meucci, 2010d] Meucci, A. (2010d). Visualizing Principal Component Analysis. http://ssrn.com/abstract=1650603. Working Paper.

[Meucci, 2011a] Meucci, A. (2011a). No-linear-factor-models riks and portfolio management. *Working Paper*.

[Meucci, 2011b] Meucci, A. (2011b). The Prayer: Ten-step checklist for advanced risk and portfolio management. http://symmys.com/node/63. Garp Risk Professional.

[Minka, 2003] Minka, T. P. (2003). Old and new matrix algebra useful for statistics. *Working Paper*.

[Potters et al., 2005] Potters, M., Bouchaud, J. P., and Laloux, L. (2005). Financial applications of random matrix theory: Old laces and new pieces. *arXiv:physics*, 0507111v1.

[Rencher, 2002] Rencher, A. C. (2002). *Methods of Multivariate Analysis*. Wiley, 2nd edition.

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

[Rosenberg and Lanstein, 1985] Rosenberg, B.and Reid, K. and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *Journal of Portfolio Management*, 11:9–17.

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

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

[Straumann and Garidi, 2007] Straumann, D. and Garidi, T. (2007). Developing an equity factor model for risk. *Riskmetrics Journal*, 7(1):89–128.

A Appendix

In this appendix we discuss technical results that can be skipped at first reading.

A.1 Results for the unconstrained "time-series" approach

Here we drop the boldface notation for vectors, we define

$$\Sigma_{X,Y} \equiv \mathrm{Cov}\left\{X,Y\right\} \equiv \mathrm{E}\left\{\left(X-\mathrm{E}\left\{X\right\}\right)\left(Y-\mathrm{E}\left\{Y\right\}\right)'\right\}, \tag{124}$$

and we use the convention $\Sigma_X \equiv \Sigma_{X,X}$.

A.1.1 The loadings

This proof follows [Meucci, 2005]. From the definition of the unconstrained time-series loadings (11) and of the r-square (9) we must minimize the target

$$\begin{aligned} m\left(b\right) &\equiv \mathrm{tr}\left(\Sigma_{\phi(X-bZ)}\right) = \mathrm{tr}\left(\phi\Sigma_{(X-bZ)}\phi'\right) \\ &= \mathrm{tr}\left(\Sigma_{(X-bZ)}\Phi\right) \\ &= \mathrm{tr}\left(\left(\Sigma_X + b\Sigma_Z b' - b\Sigma_{Z,X} - \Sigma_{X,Z}b'\right)\Phi\right), \end{aligned} \tag{125}$$

where $\Phi \equiv \phi'\phi$.

We will now use some matrix manipulations, see also [Minka, 2003]. The differential reads

$$dm\left(b\right) = \mathrm{tr}\left(d\left[b\Sigma_Z b'\right]\Phi - \left[db\right]\Sigma_{Z,X}\Phi - \Sigma_{X,Z}\left[db'\right]\Phi\right) \tag{126}$$

We use the identity

$$\begin{aligned} \mathrm{tr}\left(d\left[b\Sigma_Z b'\right]\Phi\right) &= \mathrm{tr}\left(\left[db\right]\Sigma_Z b'\Phi + b\Sigma_Z\left[db\right]'\Phi\right) \\ &= \mathrm{tr}\left(\left[db\right]\Sigma_Z b'\Phi + \left[db\right]\Sigma_Z b'\Phi\right), \end{aligned} \tag{127}$$

which follows from the equality $\mathrm{tr}\left(A'b\right) = \mathrm{tr}\left(Ab'\right)$, which holds for any $N \times K$ matrices A and b. Then the differential reads

$$\begin{aligned} dm\left(b\right) &= \mathrm{tr}\left(\left[db\right]\Sigma_Z b'\Phi\right) + \mathrm{tr}\left(\left[db\right]\Sigma_Z b'\Phi\right) \\ &\quad - \mathrm{tr}\left(\left[db\right]\Sigma_{Z,X}\Phi\right) - \mathrm{tr}\left(\left[db\right]\Sigma_{Z,X}\Phi\right) \\ &= 2\,\mathrm{tr}\left(\left[db\right]\Sigma_Z b'\Phi\right) - 2\,\mathrm{tr}\left(\left[db\right]\Sigma_{Z,X}\Phi\right). \end{aligned} \tag{128}$$

To find the maximum, we set the differential to zero

$$0 \equiv \mathrm{tr}\left(db\left(\Sigma_Z b' - \Sigma_{Z,X}\right)\Phi\right), \tag{129}$$

The solution reads

$$b = \Sigma_{X,Z}\Sigma_Z^{-1}, \tag{130}$$

which is the desired result (13).

A.1.2 The r-square

In general, for any two random vectors Z and X and any conformable matrix a we obtain

$$\begin{aligned} R^2_\phi \{aZ, X\} &\equiv 1 - \frac{\operatorname{tr}\left(\Sigma_{\phi(aZ-X)}\right)}{\operatorname{tr}\left(\Sigma_{\phi X}\right)} \\ &= 1 - \frac{\operatorname{tr}\left(\phi\Sigma_{aZ-X}\phi'\right)}{\operatorname{tr}\left(\phi\Sigma_X\phi'\right)} = 1 - \frac{\operatorname{tr}\left(\Sigma_{aZ-X}\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\Sigma_X\Phi\right) - \operatorname{tr}\left(\left(\Sigma_{aZ} - \Sigma_{aZ,X} - \Sigma_{X,aZ} + \Sigma_X\right)\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\left(\Sigma_{aZ,X} + \Sigma_{X,aZ} - \Sigma_{aZ}\right)\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\left(a\Sigma_{Z,X} + \Sigma_{X,Z}a' - a\Sigma_Z a'\right)\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \end{aligned} \tag{131}$$

For the unconstrained "time-series" case from (13) we have to set $Z \equiv Z$ and $a \equiv \Sigma_{X,Z}\Sigma_Z^{-1}$. Therefore (131) becomes

$$\begin{aligned} R^2_\phi &= R^2_\phi \left\{\Sigma_{X,Z}\Sigma_Z^{-1}Z, X\right\} \\ &= \frac{\operatorname{tr}\left(\left(\Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X} + \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X} - \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_Z\Sigma_Z^{-1}\Sigma_{Z,X}\right)\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X}\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\Sigma_{\phi X,Z}\Sigma_Z^{-1}\Sigma_{Z,\phi X}\right)}{\operatorname{tr}\left(\Sigma_{\phi X}\right)}, \end{aligned} \tag{132}$$

where $\Phi \equiv \phi'\phi$.

A.1.3 The residuals

Here we prove that factors and residuals are uncorrelated. The residuals read

$$U \equiv X - a - bZ = X - a - \Sigma_{X,Z}\Sigma_Z^{-1}Z \tag{133}$$

Therefore the covariance of factors and residuals reads

$$\begin{aligned} \Sigma_{U,Z} &= \Sigma_{X-\Sigma_{X,Z}\Sigma_Z^{-1}Z,Z} = \\ &= \Sigma_{X,Z} + \Sigma_{-\Sigma_{X,Z}\Sigma_Z^{-1}Z,Z} \\ &= \Sigma_{X,Z} - \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_Z \\ &= 0 \end{aligned} \tag{134}$$

Now we show that the residuals are not idiosyncratic. Indeed, the covariance of the residuals reads

$$\begin{aligned} \Sigma_U &= \Sigma_{X-\Sigma_{X,Z}\Sigma_Z^{-1}Z, X-\Sigma_{X,Z}\Sigma_Z^{-1}Z} \\ &= \Sigma_X + \Sigma_{X,-\Sigma_{X,Z}\Sigma_Z^{-1}Z} + \Sigma_{-\Sigma_{X,Z}\Sigma_Z^{-1}Z,X} + \Sigma_{-\Sigma_{X,Z}\Sigma_Z^{-1}Z,-\Sigma_{X,Z}\Sigma_Z^{-1}Z} \\ &= \Sigma_X - \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X} - \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X} + \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_Z\Sigma_Z^{-1}\Sigma_{Z,X} \\ &= \Sigma_X - \Sigma_{X,Z}\Sigma_Z^{-1}\Sigma_{Z,X}, \end{aligned} \tag{135}$$

which is not diagonal.

A.2 Results for the unconstrained "cross-sectional" approach

Here we drop the boldface notation for vectors, we define $\Sigma_{X,Y} \equiv \mathrm{Cov}\{X,Y\} \equiv \mathrm{E}\left\{(X - \mathrm{E}\{X\})(Y - \mathrm{E}\{Y\})'\right\}$, and we use the convention $\Sigma_X \equiv \Sigma_{X,X}$.

A.2.1 The factors

From the definition of the unconstrained cross-section factors (22) and of the r-square (9) we must minimize the target

$$\begin{aligned} m(c) &\equiv \mathrm{tr}\left(\Sigma_{\phi(X-bcX)}\right) = \mathrm{tr}\left(\phi\Sigma_{(X-bcX)}\phi'\right) \\ &= \mathrm{tr}\left(\Sigma_{(X-bcX)}\Phi\right) \\ &= \mathrm{tr}\left(\left(\Sigma_X + bc\Sigma_X c'b' - bc\Sigma_X - \Sigma_X c'b'\right)\Phi\right), \end{aligned} \tag{136}$$

where $\Phi \equiv \phi'\phi$.

We will now use some matrix manipulations, see also [Minka, 2003]. The differential reads

$$\begin{aligned} dm(c) = {} & \mathrm{tr}\left(d\left[bc\Sigma_X c'b'\right]\Phi\right) \\ & - \mathrm{tr}\left(b\left[dc\right]\Sigma_X\Phi\right) - \mathrm{tr}\left(\Sigma_X\left[dc'\right]b'\Phi\right) \end{aligned} \tag{137}$$

We use the identity

$$\begin{aligned} \mathrm{tr}\left(d\left[bc\Sigma_X c'b'\right]\Phi\right) &= \mathrm{tr}\left(bd\left[c\right]\Sigma_X c'b'\Phi\right) \\ &\quad + \mathrm{tr}\left(bc\Sigma_X\left[dc'\right]b'\Phi\right) \\ &= \mathrm{tr}\left(d\left[c\right]\Sigma_X c'b'\Phi b\right) + \mathrm{tr}\left(\left[dc'\right]b'\Phi bc\Sigma_X\right) \\ &= \mathrm{tr}\left(d\left[c\right]\Sigma_X c'b'\Phi b\right) + \mathrm{tr}\left(\left[dc\right]\Sigma_X c'b'\Phi b\right), \end{aligned} \tag{138}$$

which follows from the equality $\mathrm{tr}(A'b) = \mathrm{tr}(Ab')$, which holds for any $N \times K$ matrices A and b. Then the differential reads

$$\begin{aligned} dm(c) &= \mathrm{tr}\left(d\left[c\right]\Sigma_X c'b'\Phi b\right) + \mathrm{tr}\left(\left[dc\right]\Sigma_X c'b'\Phi b\right) \\ &\quad - \mathrm{tr}\left(\left[dc\right]\Sigma_X\Phi b\right) - \mathrm{tr}\left(\left[dc\right]\Sigma_X\Phi b\right) \\ &= 2\,\mathrm{tr}\left(d\left[c\right]\Sigma_X c'b'\Phi b\right) - 2\,\mathrm{tr}\left(\left[dc\right]\Sigma_X\Phi b\right) \end{aligned} \tag{139}$$

To find the maximum, we set the differential to zero

$$0 \equiv \mathrm{tr}\left(d\left[c\right]\Sigma_X\left(c'b'\Phi b - \Phi b\right)\right) \tag{140}$$

The solution reads

$$c = \left(b'\Phi b\right)^{-1}b'\Phi, \tag{141}$$

which is the desired result (24).

A.2.2 The r-square

To compute the r-square of the unconstrained "cross-section" factors (24) we must set in the general expression for the r-square (131) $Z \equiv X$ and

$$a \equiv b\left(b'\Phi b\right)^{-1} b'\Phi. \tag{142}$$

Then the r-square (131) becomes

$$R^2_\phi = \frac{\operatorname{tr}\left(a\Sigma_X\Phi\right) + \operatorname{tr}\left(\Sigma_X a'\Phi\right) - \operatorname{tr}\left(a\Sigma_X a'\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)}. \tag{143}$$

Using

$$\operatorname{tr}\left(a\Sigma_X\Phi\right) = \operatorname{tr}\left(\Sigma_X\Phi a\right) = \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right) \tag{144}$$

and

$$\operatorname{tr}\left(\Sigma_X a'\Phi\right) = \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right) \tag{145}$$

and

$$\begin{aligned}\operatorname{tr}\left(a\Sigma_X a'\Phi\right) &= \operatorname{tr}\left(\Sigma_X a'\Phi a\right) \\ &= \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right) \\ &= \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right),\end{aligned} \tag{146}$$

then (143) simplifies to

$$\begin{aligned}R^2_\phi &= \frac{\operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} \\ &= \frac{\operatorname{tr}\left(\Sigma_{\phi bZ}\right)}{\operatorname{tr}\left(\Sigma_X\Phi\right)} = \frac{\operatorname{tr}\left(\Sigma_{\phi bZ}\right)}{\operatorname{tr}\left(\Sigma_{\phi X}\right)},\end{aligned} \tag{147}$$

where we used where $\Phi \equiv \phi'\phi$ and

$$\begin{aligned}\operatorname{tr}\left(\Sigma_{\phi bZ}\right) &= \operatorname{tr}\left(\Sigma_{\phi b\left(b'\Phi b\right)^{-1} b'\Phi X}\right) \\ &= \operatorname{tr}\left(\phi b\left(b'\Phi b\right)^{-1} b'\Phi\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\phi'\right) \\ &= \operatorname{tr}\left(b\left(b'\Phi b\right)^{-1} b'\Phi\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right) \\ &= \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right) \\ &= \operatorname{tr}\left(\Sigma_X\Phi b\left(b'\Phi b\right)^{-1} b'\Phi\right).\end{aligned} \tag{148}$$

A.3 Results for the generalized PCA approach

Here we drop the boldface notation for vectors, we define $\Sigma_{X,Y} \equiv \operatorname{Cov}\left\{X,Y\right\} \equiv \operatorname{E}\left\{\left(X - \operatorname{E}\left\{X\right\}\right)\left(Y - \operatorname{E}\left\{Y\right\}\right)'\right\}$, and we use the convention $\Sigma_X \equiv \Sigma_{X,X}$.

A.3.1 The loadings and the factors

From the definition of the unconstrained statistical factors (31) and of the r-square (9) we must minimize the target

$$\begin{aligned} m(b,c) &\equiv \operatorname{tr}\left(\Sigma_{\phi(X-bcX)}\right) = \operatorname{tr}\left(\phi\Sigma_{(X-bcX)}\phi'\right) \\ &= \operatorname{tr}\left(\Sigma_{(X-bcX)}\Phi\right) \\ &= \operatorname{tr}\left(\left(\Sigma_X + bc\Sigma_X c'b' - bc\Sigma_X - \Sigma_X c'b'\right)\Phi\right), \end{aligned} \tag{149}$$

where $\Phi \equiv \phi'\phi$.

We will now use some matrix manipulations, see also [Minka, 2003]. Using the equality $\operatorname{tr}(A'b) = \operatorname{tr}(Ab')$, which holds for any $N \times K$ matrices A and b, the differential reads

$$\begin{aligned} dm(b,c) = {} & d\left[\operatorname{tr}\left(bc\Sigma_X c'b'\Phi\right)\right] - \operatorname{tr}\left(d[b]\,c\Sigma_X\Phi\right) \\ & - \operatorname{tr}\left(bd[c]\,\Sigma_X\Phi\right) - \operatorname{tr}\left(\Sigma_X d\left[c'\right]b'\Phi\right) - \operatorname{tr}\left(\Sigma_X c'd\left[b'\right]\Phi\right) \\ = {} & d\left[\operatorname{tr}\left(bc\Sigma_X c'b'\Phi\right)\right] - \operatorname{tr}\left(d[b]\,c\Sigma_X\Phi\right) \\ & - \operatorname{tr}\left(d[c]\,\Sigma_X\Phi b\right) - \operatorname{tr}\left(d[c]\,\Sigma_X\Phi b\right) - \operatorname{tr}\left(d[b]\,c\Sigma_X\Phi\right). \end{aligned} \tag{150}$$

Using

$$\begin{aligned} d\left[\operatorname{tr}\left(bc\Sigma_X c'b'\Phi\right)\right] &= \operatorname{tr}\left(d[bc]\,\Sigma_X c'b'\Phi\right) + \operatorname{tr}\left(bc\Sigma_X d\left[c'b'\right]\Phi\right) \\ &= \operatorname{tr}\left(d[bc]\,\Sigma_X c'b'\Phi\right) + \operatorname{tr}\left(d\left[c'b'\right]\Phi bc\Sigma_X\right) \\ &= \operatorname{tr}\left(d[bc]\,\Sigma_X c'b'\Phi\right) + \operatorname{tr}\left(d[bc]\,\Sigma_X c'b'\Phi\right) \\ &= 2\operatorname{tr}\left(d[bc]\,\Sigma_X c'b'\Phi\right) \\ &= 2\operatorname{tr}\left(d[b]\,c\Sigma_X c'b'\Phi\right) + 2\operatorname{tr}\left(bd[c]\,\Sigma_X c'b'\Phi\right) \\ &= 2\operatorname{tr}\left(d[b]\,c\Sigma_X c'b'\Phi\right) + 2\operatorname{tr}\left(d[c]\,\Sigma_X c'b'\Phi b\right) \end{aligned} \tag{151}$$

The differential reads

$$\begin{aligned} dm(b,c) \propto {} & \operatorname{tr}\left(d[b]\left(c\Sigma_X c'b'\Phi - c\Sigma_X\Phi\right)\right) \\ & + \operatorname{tr}\left(d[c]\left(\Sigma_X c'b'\Phi b - \Sigma_X\Phi b\right)\right). \end{aligned} \tag{152}$$

To find the maximum, we set the differential to zero, obtaining the two sets of equations

$$c'b'\Phi b = \Phi b, \quad c\Sigma_X c'b' = c\Sigma_X. \tag{153}$$

Defining

$$\widetilde{b} \equiv \phi b, \quad \widetilde{c} \equiv c\phi^{-1}, \quad \widetilde{\Sigma} \equiv \phi\Sigma_X\phi' \tag{154}$$

we can write (153) as

$$\widetilde{c}\widetilde{b}'\widetilde{b} = \widetilde{b}, \quad \widetilde{c}\widetilde{\Sigma}\widetilde{c}'\widetilde{b}' = \widetilde{c}\widetilde{\Sigma}. \tag{155}$$

Suppose that

$$c'b'b = b, \quad c\Sigma c'b' = c\Sigma. \tag{156}$$

Define

$$\widehat{b} \equiv bA, \widehat{c} \equiv A^{-1}c$$

then $\widehat{b}\widehat{c} = bc$. Then

$$\begin{aligned}
\widehat{c'}\widehat{b'}\widehat{b} &= c'A'^{-1}A'b'bA && (157)\\
&= c'b'bA = bA && (158)\\
&= \widehat{b} && (159)
\end{aligned}$$

and

$$\begin{aligned}
\widehat{c}\Sigma\widehat{c'}\widehat{b'} &= A^{-1}c\Sigma c'A'^{-1}A'b'\\
&= A^{-1}c\Sigma c'b'\\
&= A^{-1}c\Sigma\\
&= \widehat{c}\Sigma
\end{aligned}$$

The first equation implies

$$\widetilde{c} = \left(\widetilde{b'}\widetilde{b}\right)^{-1}\widetilde{b'}. \tag{160}$$

Substituting this result in the second one we obtain

$$\left(\widetilde{b'}\widetilde{b}\right)^{-1}\widetilde{b'}\widetilde{\Sigma}\widetilde{b}\left(\widetilde{b'}\widetilde{b}\right)^{-1}\widetilde{b'} = \left(\widetilde{b'}\widetilde{b}\right)^{-1}\widetilde{b'}\widetilde{\Sigma}. \tag{161}$$

Equation (161) is solved by

$$\widetilde{b} \equiv \widetilde{e}_K \tag{162}$$

where $\widetilde{e}_K$ is the $N \times K$ matrix obtained by juxtaposing *any* K eigenvectors, normalized to have length one (in particular, but not necessarily, the eigenvectors relative to the K largest eigenvalues). Indeed, in this case $\widetilde{b'}\widetilde{b} = I$ and (161) becomes

$$\widetilde{e'}_K\widetilde{\widetilde{\Sigma}}\widetilde{e}_K\widetilde{e'}_K = \widetilde{e'}_K\widetilde{\widetilde{\Sigma}}, \tag{163}$$

which is true, as can be verified by direct computation. Then (160) yields

$$\widetilde{c} = \widetilde{e'}_K. \tag{164}$$

The desired result (37) follows by substituting (162) and (164) back into (154).

A.3.2 The r-square

To compute the r-square of the PCA factors (38) we must set in the general expression for the r-square (131) $Z \equiv X$ and using the notation in Section A.3.1

$$A \equiv \phi^{-1}\widetilde{e}_K\widetilde{e'}_K\phi. \tag{165}$$

Then the r-square (131) becomes

$$R_{\phi}^{2}=\frac{\operatorname{tr}\left(A\Sigma_{X}\Phi\right)+\operatorname{tr}\left(\Sigma_{X}A'\Phi\right)-\operatorname{tr}\left(A\Sigma_{X}A'\Phi\right)}{\operatorname{tr}\left(\Sigma_{X}\Phi\right)}. \tag{166}$$

Using

$$\begin{aligned}\operatorname{tr}\left(A\Sigma_{X}\Phi\right)&=\operatorname{tr}\left(\Sigma_{X}\Phi A\right)=\operatorname{tr}\left(\Sigma_{X}\Phi\phi^{-1}\widetilde{e}_{K}\widetilde{e}'_{K}\phi\right)\\&=\operatorname{tr}\left(\Sigma_{X}\phi'\widetilde{e}_{K}\widetilde{e}'_{K}\phi\right)\\&=\operatorname{tr}\left(\widetilde{\Sigma}\widetilde{e}_{K}\widetilde{e}'_{K}\right)\end{aligned} \tag{167}$$

and

$$\begin{aligned}\operatorname{tr}\left(\Sigma_{X}A'\Phi\right)&=\operatorname{tr}\left(\Sigma_{X}\phi'\widetilde{e}_{K}\widetilde{e}'_{K}\phi'^{-1}\Phi\right)\\&=\operatorname{tr}\left(\Sigma_{X}\phi'\widetilde{e}_{K}\widetilde{e}'_{K}\phi\right)\\&=\operatorname{tr}\left(\widetilde{\Sigma}\widetilde{e}_{K}\widetilde{e}'_{K}\right)\end{aligned} \tag{168}$$

and

$$\begin{aligned}\operatorname{tr}\left(A\Sigma_{X}A'\Phi\right)&=\operatorname{tr}\left(\Sigma_{X}A'\Phi A\right)\\&=\operatorname{tr}\left(\Sigma_{X}\phi'\widetilde{e}_{K}\widetilde{e}'_{K}\phi'^{-1}\Phi\phi^{-1}\widetilde{e}_{K}\widetilde{e}'_{K}\phi\right)\\&=\operatorname{tr}\left(\Sigma_{X}\phi'\widetilde{e}_{K}\widetilde{e}'_{K}\phi\right)\\&=\operatorname{tr}\left(\widetilde{\Sigma}\widetilde{e}_{K}\widetilde{e}'_{K}\right)\end{aligned} \tag{169}$$

then (166) simplifies to

$$R_{\phi}^{2}=\frac{\operatorname{tr}\left(\widetilde{e}'_{K}\widetilde{\Sigma}\widetilde{e}_{K}\right)}{\operatorname{tr}\left(\widetilde{\Sigma}\right)}. \tag{170}$$

The numerator in (170) is the sum of the K eigenvalues relative to the K eigenvectors in $\widetilde{e}_{K}$ and the denominator is the sum of all the eigenvalues. Since in the original problem (31) we are maximizing the r-square, we must select as K eigenvectors those relative to the K largest eigenvalues. This is the desired result (40).

A.4 Results for factor analysis

The approximation (55) is consistent with a joint covariance of factors and residual that reads

$$\operatorname{Cv}\left\{\left(\begin{array}{c}Z\\U\end{array}\right)\right\}=\left(\begin{array}{cc}I_{K\times K}&0_{K\times N}\\\mathbf{0}_{N\times K}&\Delta^{2}\end{array}\right), \tag{171}$$

which has rank K (the number of factors Z) $+N$ (the dimension of the market X). However, assume we could find the extraction matrix $\mathbf{G}$ such that

$$Z\equiv gX. \tag{172}$$

Then from (107) we obtain

$$\begin{pmatrix} Z \\ U \end{pmatrix} = qX, \tag{173}$$

where

$$q \equiv \begin{pmatrix} g \\ I - bg \end{pmatrix}. \tag{174}$$

This implies that

$$\mathrm{Cv}\left\{ \begin{pmatrix} Z \\ U \end{pmatrix} \right\} = q \, \mathrm{Cv}\left\{ X \right\} q', \tag{175}$$

which has rank N and thus contradicts (171).