# Correlation and dependence measures:

# Academic literature and selected financial applications.

## Contents

# Introduction

The purpose of this document is to present and discuss the mathematical and statistical tools available to understand and quantify the comovement between financial variables. Understanding the dependence structure of assets or risk factors is of utmost important in many financial applications. Indeed, risk measurements constitute the first and most important inputs in any risk management process and these risk measurements rely on a good quantitative appreciation of the dependence structure of financial variables. Modern portfolio theory also requires accurate dependence measures in order to achieve efficient diversification.

Dependence measures have been developed by statisticians throughout the 20th century, independently of any financial application, and they have proved their universality. Linear correlation is certainly the most popular dependence measure. Other indicators, like rank correlation, belonging to the large family of concordance measures, have also been introduced in the first part of the 20th century in order to better apprehend non-linear dependence structures.

The description of multivariate probability distribution through the copula functions introduced in the fifties by Sklar, also have an important role in Finance, first in risk management and then in the rise and fall of multi-name credit derivatives.

Correlations and copula functions are now ubiquitous in Finance when describing dependence. Therefore, we had to make choices before writing this document. Our choice is to present the universal statistical tools used for risk management and then to focus on specific applications requiring additional and specific methods. These applications are in two fields: portfolio management and high-frequency trading. Portfolio management requires good covariance estimates, especially since the use of the most common estimators is known to provide very bad results. High-frequency trading is a more recent field and high-frequency correlation measures constitute a new and very active strand of research.

Multi-asset option pricing could have also been discussed (see for instance [14]), along with multi-name credit derivatives, but pricing constitutes a very different topic that often has more to do with calibration to a given model than with the choice of the correlation model itself.

The text is divided in four parts.

The first part is dedicated to an academic presentation of the commonly used tools to describe comovement between assets: copula functions, linear correlation, Kendall's $\tau$, Spearman's $\rho$, tail dependence, ... We particularly highlight the drawbacks of linear correlation whose use is practical but unsuited for financial variables that cannot be described by gaussian distributions. Statistical estimators are also provided along with their properties.

Focused on a financial application, the second part deals with the estimation of the large covariance/correlation matrices used in Markowitz model. We explain why sample covariance matrix leads to very bad outcomes in terms of portfolio choices. Methods to clean large covariance matrices from the noise they contain (curse of dimensionality) are then presented. Statistical methods are first reviewed before the eigenvalue cleaning methods proposed by econophysicists are introduced. The latter methods are based on random matrix theory which also has applications to the choice of the time horizon to compute correlation matrices.

Part 3 is dedicated to the graphical representations of dependence structure and to the associated clustering algorithms rooted to an old statistical literature on multivariate data analysis.

The last part of this document focuses on high-frequency data statistics, which is a recent strand of academic research. The usual estimators of covariance and correlation cannot be used on high-frequency data because the data is unevenly and asynchronously sampled, and because microstructure noise contaminates the data. Very recent statistical tools are presented to handle asynchronous data and filter out microstructure noise.

Along the text, recommendations based on both academic literature and interviews of practitioners are provided.

# Part I

# Dependence measures

## 1 Correlation vs. Dependence

Correlation is an ambiguous word. On one hand, it refers to the general issue of dependence and we often speak about the correlation structure of a finite number of random variables. On the other hand, (linear[1]) correlation refers to a precise dependence measure, defined for two random variables $X$ and $Y$ as:

$$r(X,Y) = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \frac{\mathbb{E}[(X - \mathbb{E}[X])\,(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}}$$

In the first case, the word correlation is an abuse of language for the general dependence structure and this abuse of language is rooted to the gaussian world in which any dependence structure boils down to a set of linear correlation coefficients as defined above in the second acceptation of the word correlation. If the predominance of gaussian variables in models, both in Finance and outside of Finance, usually generates potential misunderstandings, we present here measures of dependence in a general framework and the word correlation will always refer to a specific dependence measure, most often the linear correlation measure defined above.

In general, if we consider $n$ real-valued random variables $X_1, \ldots, X_n$, modeling for instance asset returns, the distribution (or law) of $X = (X_1, \ldots, X_n)$ is characterized by the joint (cumulative) distribution function:

$$F_X(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

This joint distribution function $F_X$ contains information about the law of each of the $n$ variables $X_1, \ldots, X_n$ and about the dependence structure of the variables. To disentangle the two, a probabilistic tool has been introduced by Sklar [113]: copula functions.

[1] Non-linear notions of correlation also exists and we will define rank correlation later in this text.

**Definition 1.** *A copula function is a function $C : [0,1]^n \to [0,1]$ that verifies the following properties:*

- *Marginals of $C$ are uniform:*

$$\forall 1 \leq i \leq N, \forall x_i \in [0,1], \quad C(1,\dots,1,\underbrace{x_i}_{i},1,\dots,1) = x_i$$

- *$C$ is $N$-increasing:*

$$\forall 1 \leq i \leq N, \forall 0 \leq x_i \leq y_i \leq 1, \sum_{\forall i, w_i \in \{x_i, y_i\}} \varepsilon(w) C(w_1,\dots,w_N) \geq 0$$

  *where*

$$\varepsilon(w) = \begin{cases} 1, & \text{if the number of indices } i \text{ such that } w_i = y_i \text{ is even} \\ 0, & \text{if } \exists i, x_i = y_i \\ -1, & \text{if the number of indices } i \text{ such that } w_i = y_i \text{ is odd.} \end{cases}$$

- *$\forall 1 \leq i \leq N, \forall (x_1,\dots,x_{i-1},x_{i+1},\dots,x_N) \in [0,1]^{N-1}$:*

$$C(x_1,\dots,x_{i-1},0,x_{i+1},\dots,x_N) = 0$$

Copula functions allow to write a joint distribution function $F_X$ using the cumulative distribution functions of the random variables $X_1,\dots,X_{n-1}$ and $X_n$, and a copula function which describes the dependence structure of $X = (X_1,\dots,X_n)$. More exactly we have Sklar's theorem:

**Theorem 1** (Sklar's Theorem). *Let us consider $X = (X_1,\dots,X_n)$ a tuple of real random variables. Then there exists a copula function $C$ such that:*

$$\forall (x_1,\dots,x_n) \in \mathbb{R}^n, \quad F_X(x_1,\dots,x_n) = C(F_1(x_1),\dots,F_n(x_n))$$

*where the $F_i$s are the marginal cumulative distribution functions.*

*Moreover, if we assume that the joint distribution function $F_X$ and the marginal cumulative distribution functions $F_i$s are continuous, then the copula function is unique and given by the formula:*

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)), \quad \forall (u_1, \ldots, u_n) \in [0,1]^n$$

*Conversely, let $F_1, \ldots, F_n$ be $n$ cumulative distribution functions.*
*If $C$ is a copula function then:*
*$(x_1, \ldots, x_n) \in \mathbb{R}^n \mapsto C(F_1(x_1), \ldots, F_n(x_n))$ defines a joint distribution with marginal distributions corresponding to the $F_i$s.*

The interested reader may refer to [30] for a formal introduction to copula functions. The point we want to focus on is that the dependence structure of the variables $X = (X_1, \ldots, X_n)$, assumed to be continuous for the copula function to be well-defined, is characterized by a copula function $C$. In particular, it must be noticed that, for any increasing functions $g_1, \ldots, g_n$, the copula functions associated to $(X_1, \ldots, X_n)$ and $(g_1(X_1), \ldots, g_n(X_n))$ are the same.

The information about the dependence structure contained in $C$ being exhaustive, it is often too complex to be used in practice, even in the simple case $n = 2$ we shall now consider. Therefore, in order to better understand the dependence structure, we shall present the classical measures of dependence that are used in practice as figures summing up the information contained in the copula function $C$. We shall also see that, outside of the gaussian world, the usual (linear) correlation coefficient cannot be understood as it usually is.

# 2 Usual dependence measures

## 2.1 Measures of concordance

If we consider two (continuous) random variables $X_1$ and $X_2$, the information about the two variables, considered independently of one another, is contained in their respective cumulative distribution functions $F_1$ and $F_2$ and the dependence between them is described by the copula function $C$ associated to $X = (X_1, X_2)$. The copula function describes the exact dependence between the two random variables but it is, in general, hard to estimate from a statistical point of view because a good (and almost always non-parametric) estimation requires many observations. Somehow, it contains too much information for practical use and simple indicators or dependence measures have been developed to describe dependence structure. The goal of a dependence measure is to describe through a single figure the nature of the dependence between two random variables and we shall present hereafter the most common ones.

To introduce what is considered a good dependence measure, namely a measure of concordance, we need to introduce first the notion of comonotonic and countermonotonic variables.

**Definition 2.** *A set $A \subset \mathbb{R}^2$ is comonotonic if:*

$$\forall (x_1, x_2), (\tilde{x}_1, \tilde{x}_2) \in A, (x_2 - x_1)(\tilde{x}_2 - \tilde{x}_1) \geq 0$$

**Definition 3.** *A couple of real-valued random variables $(X_1, X_2)$ is comonotonic (or equivalently $X_1$ and $X_2$ are comonotonic) if it has a comonotonic support, i.e. if there exists $A \subset \mathbb{R}^2$, a comonotonic set, such that $\mathbb{P}((X_1, X_2) \in A) = 1$.*

Another denomination for comonotonic is “perfectly positively dependent”. For two given cumulative distribution functions $F_1$ and $F_2$, a comonotonic couple $(X_1, X_2)$ with marginal distributions $F_1$ and $F_2$ is representative of the strongest possible positive dependence structure between a random variable with cumulative distribution function $F_1$ and a random variable with cumulative distribution function $F_2$. This is stated more formally in the following proposition:

**Proposition 1.** *Let us consider a couple of real-valued random variables* $X = (X_1, X_2)$*, assumed to be continuous. Then the following assertions are equivalent:*

- $(X_1, X_2)$ *is comonotonic*
- $F_X(x_1, x_2) = \min(F_1(x_1), F_2(x_2))$
- $\forall (u_1, u_2) \in [0,1]^2, C(u_1, u_2) = C^+(u_1, u_2) = \min(u_1, u_2)$
- $(X_1, X_2) =_{law} (F_1(F_2^{-1}(X_2)), F_2(F_1^{-1}(X_1)))$

In the copula language, this maximum positive dependence is associated to the Fréchet copula $C^+$ which is an upper bound for any copula function.

Similarly, we shall define the maximal negative dependence between two random variables with given marginal distributions (or countermonotonicity) in the following way:

**Definition 4.** *A set* $A \subset \mathbb{R}^2$ *is countermonotonic if:*

$$\forall (x_1, x_2), (\tilde{x}_1, \tilde{x}_2) \in A, (x_2 - x_1)(\tilde{x}_2 - \tilde{x}_1) \leq 0$$

**Definition 5.** *A couple of real-valued random variables* $(X_1, X_2)$ *is countermonotonic (or equivalently* $X_1$ *and* $X_2$ *are countermonotonic) if it has a countermonotonic support, i.e. if there exists* $A \subset \mathbb{R}^2$*, a countermonotonic set, such that* $\mathbb{P}((X_1, X_2) \in A) = 1$*.*

Another denomination for countermonotonic is “perfectly negatively dependent”. For two given cumulative distribution functions $F_1$ and $F_2$, a countermonotonic couple $(X_1, X_2)$ with marginal distributions $F_1$ and $F_2$ is representative of the strongest possible negative dependence structure between a random variable with cumulative distribution function $F_1$ and a random variable with cumulative distribution function $F_2$. This is stated more formally in the following proposition:

**Proposition 2.** *Let us consider a couple of real-valued random variables $X = (X_1, X_2)$, assumed to be continuous. Then the following assertions are equivalent:*

- $(X_1, X_2)$ *is countermonotonic*
- $F_X(x_1, x_2) = \max(F_1(x_1) + F_2(x_2) - 1, 0)$
- $\forall (u_1, u_2) \in [0,1]^2, C(u_1, u_2) = C^-(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$
- $(X_1, X_2) =_{law} (F_1(1 - F_2^{-1}(X_2)), F_2(1 - F_1^{-1}(X_1)))$

The Fréchet copula $C^-$ is a lower bound for any copula function and corresponds to the maximum negative dependence between two random variable.

We are now ready to present the properties that an ideal measure of dependence should have. Measures of dependence satisfying these properties are called concordance measures:

**Definition 6.** *We call a concordance measure a function $\delta(\cdot, \cdot)$ that verifies the following 5 axioms:*

1. $\delta(X_1, X_2) = \delta(X_2, X_1)$ *(symmetry)*
2. $-1 \leq \delta(X_1, X_2) \leq 1$
3. $\delta(X_1, X_2) = 1$ *if and only if* $X = (X_1, X_2)$ *is comonotonic*
4. $\delta(X_1, X_2) = -1$ *if and only if* $X = (X_1, X_2)$ *is countermonotonic*
5. $\forall g$ *strictly monotone,* $\delta(g(X_1), X_2) = \delta(X_1, X_2)$ *if $g$ is increasing and* $\delta(g(X_1), X_2) = -\delta(X_1, X_2)$ *if $g$ is decreasing*

## 2.2 Linear correlation and its drawbacks

The most famous and widely used dependence measure is the Pearson's product-moment coefficient $r$, also called linear correlation coefficient. It measures the linear relationship between variables.

**Definition 7.** *Pearson's coefficient $r$ of a couple of $L^2$ random variables $(X_1, X_2)$ is defined as the ratio of the covariance of $X_1$ and $X_2$ divided by the product of the standard deviations:*

$$r(X_1, X_2) = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathbb{V}[X_1]}\sqrt{\mathbb{V}[X_2]}} = \frac{\mathbb{E}[(X_1 - \mathbb{E}[X_1])\,(X_2 - \mathbb{E}[X_2])]}{\sqrt{\mathbb{V}[X_1]}\sqrt{\mathbb{V}[X_2]}}$$

By definition, Pearson's coefficient is symmetric and Cauchy-Schwarz inequality ensures that it takes values in $[-1, 1]$. However it is not invariant by increasing transformations but only invariant by increasing affine transformations of the random variables, hence the denomination of linear correlation. Similarly, linear correlation attains the bounds ±1 only if $X_1$ and $X_2$ are non-trivial affine transformations of one another, *i.e.*:

$$r(X_1, X_2) = 1 \iff X_2 = a + bX_1, \quad b > 0$$

$$r(X_1, X_2) = -1 \iff X_2 = a + bX_1, \quad b < 0$$

Hence Pearson's coefficient does not constitute a concordance measure. In particular, $r(X_1, X_2)$ does not only depend on the copula function $C$ that characterizes the dependence between $X_1$ and $X_2$ but depends also on the marginal distributions $F_1$ and $F_2$:

$$r(X_1, X_2) = \frac{1}{\sqrt{\mathbb{V}[X_1]\mathbb{V}[X_2]}} \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2)\, dF_1^{-1}(u_1) dF_2^{-1}(u_2)$$

A natural, but important, consequence of this formula, is that for any couple $(X_1, X_2)$ with given marginal distributions $F_1$ and $F_2$, the linear correlation coefficient $r(X_1, X_2)$ has bounds that may be dramatically smaller than 1 (in absolute value):

$$r^-_{F_1,F_2} \leq r(X_1, X_2) \leq r^+_{F_1,F_2}$$

where:

$$r^-_{F_1,F_2} = \frac{1}{\sqrt{\mathbb{V}[X_1]\mathbb{V}[X_2]}} \int_0^1 \int_0^1 \left(C^-(u_1, u_2) - u_1 u_2\right) dF_1^{-1}(u_1) dF_2^{-1}(u_2)$$

$$r^+_{F_1,F_2} = \frac{1}{\sqrt{\mathbb{V}[X_1]\mathbb{V}[X_2]}} \int_0^1 \int_0^1 \left(C^+(u_1, u_2) - u_1 u_2\right) dF_1^{-1}(u_1) dF_2^{-1}(u_2)$$

To exemplify these bounds and illustrate their importance in practice, one can consider the case of two random variables having log-normal distributions with parameters $(0, \sigma^2)$ and $(0, 2\sigma^2)$ respectively. Then, if $\sigma \to +\infty$, the two bounds $r^- \leq 0$ and $r^+ \geq 0$ tend towards 0. Hence, a linear correlation coefficient should never be analyzed without information about the marginals. In practice however, correlation coefficient is often compared to $-1$, 0 or 1. This widespread practice is not relevant in general but rooted to the gaussian case in which it makes sense. Pearson's coefficient indeed contains the whole information about the dependence between $X_1$ and $X_2$ when $(X_1, X_2)$ is a gaussian vector $\mathcal{N}\left(\mu, \sigma^2 \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right) = \mu + \sigma\mathcal{N}\left(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right)$. In that gaussian case, the linear correlation coefficient is equal to $r$ and it characterizes the dependence between $X_1$ and $X_2$, $r(X_1, X_2) = 1$ meaning a comonotone couple $(X_1, X_2)$, $r(X_1, X_2) = -1$ meaning a countermonotone couple $(X_1, X_2)$ and $r(X_1, X_2) = 0$ meaning independence between $X_1$ and $X_2$. Outside of the gaussian world this is not true in general and we have seen above that comonotone log-normal variables can have a Pearson's coefficient $r$ close to 0.

Although it is not a concordance measure, Pearson's coefficient is widely used in Finance and outside of Finance because:

- $r(X_1, X_2)$ describes the dependence structure between $X_1$ and $X_2$ if $(X_1, X_2)$ is a gaussian vector (as stated above).

- $r(X_1, X_2)$ has a natural interpretation in the $L^2$ space as the scalar product between centered and reduced versions of $X_1$ and $X_2$.

- $r(X_1, X_2)^2$ gives the part of the variance of $X_2$ that can be explained by a linear regression model of the form $X_2 = a + bX_1 + \epsilon$:

$$r(X_1, X_2)^2 = \frac{\mathbb{V}[X_2] - \min_{a,b} \mathbb{E}[(X_2 - a - bX_1)^2]}{\mathbb{V}[X_2]}$$

- for $X_1$ and $X_2$, two centered and reduced random variables, $\frac{1}{2}\mathbb{E}[|X_1 - X_2|^2] = 1 - r(X_1, X_2)$. Hence correlation is linked to the $L^2$ distance between two random variables. This will have applications in the penultimate part of this document about graph representations of dependence structure.

- it appears naturally in financial problems (portfolio choice, risk management, ...) when risk is measured through variance[2].
- it is central in the Capital Asset Pricing Model (CAPM).

In spite of the drawbacks of Pearson's coefficient, most of this document will be dedicated to the use of linear correlation coefficients in Finance. However, it is noteworthy that better measures exist[3] and we present them in the following paragraphs. These measures are concordance measures and therefore only depend on the copula function in accordance with the 5 points introduced earlier. In finance, these non-linear dependence measures are mainly used in risk analysis, for instance to calibrate a parametric copula function before a Monte-Carlo simulation is carried out (for Value at Risk computation for example[4]). When it comes to other areas of finance such as pricing or portfolio choice, nearly all models rely on (linear) correlations when a dependence measure is needed. Outside of finance, actuaries (see [36]) are the main practitioners using dependence measures such as Spearman's $\rho$ or Kendall's $\tau$ which are presented now.

## 2.3 Concordance measures: Spearman's $\rho$ and Kendall's $\tau$

Pearson's coefficient $r(X_1, X_2)$ is a measure of the linear dependence between $X_1$ and $X_2$. We have seen that it does not verify the axioms of a measure of concordance and that its value should not be interpreted relatively to the range $[-1, 1]$ because the real bounds of linear correlations strongly depend on the marginal distributions. In addition to these drawbacks, Pearson's coefficient $r(X_1, X_2)$ is *a priori* only defined for $X_1, X_2 \in L^2$, another restriction

[2] Variance is the most widely used risk measure although we know that it only describes completely the risk associated to a gaussian factor, ignoring for instance fat tails.

[3] For some applications, elliptical distributions are used and generalize the use of gaussian distributions (Student distributions for instance). In that case, the dependence structure is entirely described by the "covariance" matrix of the distribution and one may speak about non-linear correlation (except in the gaussian case), although it is an abuse of language.

[4] Copulas are also used in finance for credit derivatives pricing but copulas are often chosen to fit prices with no reference to any descriptive statistics. Option pricing in the case of multiple underlyings may also rely on copulas – see [29].

that may be important. Spearman's $\rho$ (see [114]) replaces the couple $(X_1, X_2)$ by the couple $(F_1(X_1), F_2(X_2))$ before considering linear correlation[5]. Hence, in line with Sklar's theorem, Spearman's $\rho$ only depends on the dependence structure between $X_1$ and $X_2$, *i.e.* on the copula function and not on the marginal distributions $F_1$ and $F_2$.

**Definition 8.** *The Spearman's $\rho$ of a couple of random variables $(X_1, X_2)$ is defined as:*

$$\rho(X_1, X_2) = r(F_1(X_1), F_2(X_2))$$

*where $F_1$ and $F_2$ are respectively the cumulative distribution functions of $X_1$ and $X_2$.*

This coefficient is sometimes called rank correlation.

It is straightforward to write the Spearman's $\rho$ of a couple $X = (X_1, X_2)$ with the copula function $C$ and we get:

$$\begin{aligned}\rho(X_1, X_2) &= 12 \int_{\mathbb{R}} \int_{\mathbb{R}} (F_X(x_1, x_2) - F_1(x_1)F_2(x_2)) dx_1 dx_2 \\ &= 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3\end{aligned}$$

Another representation of Spearman's $\rho$ compares the probabilities of concordance and discordance of $X = (X_1, X_2)$ and $X^{\perp} = (X_1^{\perp}, X_2^{\perp})$, where (i) $X_1^{\perp}$ and $X_2^{\perp}$ are two independent variables having respectively cumulative distribution functions $F_1$ and $F_2$ and (ii) $X^{\perp}$ is independent of $X$. Mathematically, this writes:

$$\rho(X_1, X_2) = 3 \left[\mathbb{P}((X_1 - X_1^{\perp})(X_2 - X_2^{\perp}) > 0) - \mathbb{P}((X_1 - X_1^{\perp})(X_2 - X_2^{\perp}) < 0)\right]$$

This representation of Spearman's $\rho$ helps in understanding the nature of this dependence measure. We can indeed easily verify that Spearman's $\rho$ is a concordance measure and we will see later in this text that Spearman's $\rho$ is part of a large class of concordance measure. In particular, Spearman's $\rho$ is equal to 1 if and only if $X_1$ and $X_2$ are comonotonic and

[5] Since $(F_1(X_1), F_2(X_2)) \in [0, 1]^2$, this linear correlation is always well defined.

Spearman's $\rho$ is equal to $-1$ if and only if $X_1$ and $X_2$ are countermonotonic. Also, if $X_1$ and $X_2$ are independent we have $\rho(X_1, X_2) = 0$.

Another widely cited dependence measure is Kendall's $\tau$ (see [65]). This coefficient can be defined in a similar way as Spearman's $\rho$, comparing the probabilities of concordance and discordance of $X = (X_1, X_2)$ and $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)$ where $\tilde{X}$ is a random variable distributed as $X$ and independent from $X$. More exactly, we give the following definition:

**Definition 9.** *The Kendall's $\tau$ of a couple of random variables $(X_1, X_2)$ is defined as:*

$$\tau(X_1, X_2) = \mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - \mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0)$$

*where $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)$ is a copy of $X$, independent of $X$.*

Intuitively, it measures the propensity of $X$ to be comonotone on a scale that goes from $-1$ for countermonotone $X$, to 1 for comonotone $X$ (in the case of continuous variables). Other measures related to Kendall's $\tau$ are also available such as Kendall's $\tau_b$ and Kendall's $\tau_c$ coefficients or Goodman-Kruskal $\gamma$ that are better suited to deal with potential ties, especially in the case of discrete variables.

Other expressions are available for Kendall's $\tau$ which prove that this measure only relies on the dependence structure between the two variables and not on their respective distributions:

$$\begin{aligned}\tau(X_1, X_2) &= 4\mathbb{E}[F_X(X_1, X_2)] - 1 \\ &= 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2)\end{aligned}$$

As for Spearman's $\rho$, Kendall's $\tau$ verifies the 5 axioms of a concordance measure. In addition, the Kendall's $\tau$ of two independent variables $X_1$ and $X_2$ is equal to 0.

Kendall's $\tau$ and Spearman's $\rho$ are both non-parametric dependence measures and are related to one another in the sense that the admissible couples $(\rho, \tau)$ must satisfy the following inequalities:

$$\begin{array}{rcl} \frac{3}{2}\tau - \frac{1}{2} \leq & \rho & \leq \frac{1}{2} + \tau - \frac{1}{2}\tau^2, \qquad \tau \geq 0 \\ -\frac{1}{2} + \tau + \frac{1}{2}\tau^2 \leq & \rho & \leq \frac{3}{2}\tau + \frac{1}{2}, \qquad \tau < 0. \end{array}$$

Although they are related, Kendall's $\tau$ possesses an important advantage over Spearman's $\rho$ because it can be calculated in closed-form for many commonly used copulas. For instance, the expression of Kendall's $\tau$ is simple for archimedean copulas (for instance Gumbel, Clayton and Frank copulas) whereas the expression of Spearman's $\rho$ is not available in closed-form or is at most complicated.

In the case of a gaussian vector $(X_1, X_2) \sim \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right)$ – or equivalently in the case of two random variables with a gaussian copula structure –, both Kendall's $\tau$ and Spearman's $\rho$ are known in closed-form:

$$\rho = \frac{2}{\pi} \arcsin\left(\frac{r}{2}\right)$$

$$\tau = \frac{2}{\pi} \arcsin(r)$$

This formula for Spearman's $\rho$ is specific to gaussian copulas. However, as far as Kendall's $\tau$ is concerned, the above formula is valid for all elliptical distributions if $r$ is the correlation[6] parameter of the bivariate elliptical distribution.

Spearman's $\rho$ and Kendall's $\tau$ can be generalized and they are in fact two special cases of the same more general notion. If we consider indeed two couples $X = (X_1, X_2)$ and $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)$ with:

- $X$ and $\tilde{X}$ independent
- $X_1$ and $\tilde{X}_1$ have cumulative distribution function $F_1$
- $X_2$ and $\tilde{X}_2$ have cumulative distribution function $F_2$

[6] It is called correlation by analogy with the gaussian case but we recall that it does not correspond to the linear correlation coefficient.

- the dependence structure between $X_1$ and $X_2$ is given by the copula function $C$
- the dependence structure between $\tilde{X}_1$ and $\tilde{X}_2$ is given by the copula function $\tilde{C}$

then, we can define the $Q$-concordance measure between $C$ and $\tilde{C}$:

$$Q(C, \tilde{C}) = \mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - \mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0)$$

$$= 4 \int_0^1 \int_0^1 C(u_1, u_2) d\tilde{C}(u_1, u_2) - 1$$

We have $\rho(X_1, X_2) = 3Q(C, C_I)$ where $C$ is the copula function associated to $X = (X_1, X_2)$ and where $C_I(u_1, u_2) = u_1 u_2$ is the copula associated to independence. Similarly, Kendall's $\tau$ is given by $\tau(X_1, X_2) = Q(C, C)$.

Other dependence measures based on this function $Q$ exist such as Gini coefficient but this coefficient is rarely used in Finance.

## 2.4 Tail dependence

The last dependence measure we present concerns tail dependence. For a couple $X = (X_1, X_2)$, it measures the concordance between the extreme events of $X_1$ and $X_2$. More precisely, we define the following indicators:

**Definition 10.** *The lower tail dependence of $X = (X_1, X_2)$ is defined, when the limit exists, as:*

$$\lambda_L(X_1, X_2) = \lim_{u \to 0^+} \mathbb{P}\left(F_1(X_1) \le u | F_2(X_2) \le u\right)$$

**Definition 11.** *The upper tail dependence of $X = (X_1, X_2)$ is defined, when the limit exists, as:*

$$\lambda_U(X_1, X_2) = \lim_{u \to 1-} \mathbb{P}\left(F_1(X_1) \ge u | F_2(X_2) \ge u\right)$$

These tail dependence measures only depend on the dependence structure of the couple $X = (X_1, X_2)$ and do not depend on $F_1$ nor $F_2$. We can indeed easily write the indicators as functions of the copula function $C$ describing the dependence between $X_1$ and $X_2$:

$$\lambda_L(X_1, X_2) = \lim_{u \to 0^+} \frac{C(u,u)}{u}$$

$$\lambda_U(X_1, X_2) = \lim_{u \to 1^-} \frac{1 - 2u + C(u,u)}{1 - u}$$

These indicators help to describe the concordance of tail events and we say that $X = (X_1, X_2)$ has lower (resp. upper) tail dependence when $\lambda_L > 0$ (resp. $\lambda_U > 0$).

In practice these indicators are mainly used in risk management in order to choose among a wide variety of parametric copula functions the most appropriate one to describe the dependence between risk factors. Gaussian copulas exhibit no tail dependence (except in the degenerate case of a correlation parameter equal to 1), Student copulas exhibit dependence tails (except in the degenerate case of a correlation parameter[7] equal to $-1$), both on the lower side and on the upper side. Asymmetric copulas are also available such as Gumbel copula which has upper tail dependence but no lower tail dependence and Clayton copula which has lower tail dependence but no upper tail dependence.

# 3 Statistical estimations and tests

We defined above several notions to measure the dependence between two random variables. We now present how to estimate them and specifically how to estimate linear correlation, Spearman's $\rho$ and Kendall's $\tau$. We focus, in particular, on the estimation of the linear correlation coefficient because most practitioners ignore the bias of the sample correlation coefficient whereas techniques to reduce this bias are available.

Let us consider two sequences $x_1 = (x_1^1, \ldots, x_1^T)$ and $x_2 = (x_2^1, \ldots, x_2^T)$ of i.i.d. realizations of the respective random variables $X_1$ and $X_2$. The sample correlation coefficient of $x_1$ and $x_2$ is defined by:

[7] We recall that the correlation coefficient for Student distributions does not correspond to linear correlation.

**Definition 12.**

$$\hat{r} = \frac{1}{T}\frac{\sum_{i=1}^{T}(x_1^i - \overline{x_1})(x_2^i - \overline{x_2})}{s_1 s_2}$$

*where the sample mean and standard deviation are given by:*

$$\overline{x_1} = \frac{1}{T}\sum_{i=1}^{T} x_1^i \qquad \overline{x_2} = \frac{1}{T}\sum_{i=1}^{T} x_2^i$$

$$s_1 = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_1^i - \overline{x_1})^2} \qquad s_2 = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_2^i - \overline{x_2})^2}$$

This definition is not the only one in the literature and the sample standard deviations are sometimes computed with $\frac{1}{T-1}$ instead of $\frac{1}{T}$. However, for both definitions the sample correlation coefficient is not an unbiased estimator of the true Pearson's coefficient $r = r(X_1, X_2)$ (see for instance [51]) and suffer from various defaults that are almost always ignored by practitioners, at least in Finance.

Regarding the bias of the estimator $\hat{r}$, in the case of a bivariate gaussian sample, a closed-form approximation of $\mathbb{E}[\hat{r}]$ is

$$\mathbb{E}[\hat{r}] \simeq r - \frac{r(1-r^2)}{2T}$$

Although it is asymptotically unbiased, the sample correlation coefficient is biased toward 0. Fisher thus proposed to replace the estimator $\hat{r}$ by $\hat{r}\left(1 + \frac{(1-\hat{r}^2)}{2T}\right)$ which has less bias – see also [93] for slightly different estimators of $r$.

Another important drawback of the sample correlation coefficient is that its distribution is skewed as soon as the true value $r$ is different from 0 and its variance depends strongly on $r$. To stabilize the variance, Fisher proposed to consider $\hat{z} = \frac{1}{2}\log\left(\frac{1+\hat{r}}{1-\hat{r}}\right)$. The main advantage of this non-linear transformation (still in the case of a bivariate gaussian sample) is that the distribution of $\hat{z}$ is approximately normal with mean $z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ and variance $\frac{1}{T-3}$ (that does not depend on $r$).

This transformation is particularly relevant when it comes to provide an estimation of a linear correlation coefficient along with a confidence interval. Similarly when one needs to test statistically the equality of a linear correlation coefficient to a specified value or the equality of the respective linear correlation coefficients of two random variables using two samples, a test based on $z$ rather than $r$ must almost always be preferred. The only exception is the test of the Null hypothesis $r = 0$. In that case, the test is usually made with $\hat{t} = \frac{\hat{r}\sqrt{T-2}}{\sqrt{1-\hat{r}^2}}$ and the values are compared to the quantiles of a Student distribution with $T - 2$ degrees of freedom.

Coming to Spearman's $\rho$, this concordance measure is nothing but the linear correlation coefficients of the ranks and the above analysis applies once the values $x_1 = (x_1^1, \dots, x_1^T)$ and $x_2 = (x_2^1, \dots, x_2^T)$ are replaced by the ranks $r_1 = (r_1^1, \dots, r_1^T)$ and $r_2 = (r_2^1, \dots, r_2^T)$. Interestingly, in the absence of ties, we get the following estimator for $\rho(X_1, X_2)$:

$$\hat{\rho} = \frac{\sum_{i=1}^T (r_1^i - \overline{r}_1)(r_2^i - \overline{r}_2)}{\sqrt{(\sum_{i=1}^T (r_1^i - \overline{r}_1)^2)(\sum_{i=1}^T (r_2^i - \overline{r}_2)^2)}} = 1 - 6\frac{\sum_{i=1}^T (r_1^i - r_2^i)^2}{T(T^2-1)}$$

Now, as far as Kendall's $\tau$ is concerned, the most natural estimator is:

$$\hat{\tau} = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j>i} L_{ij}$$

where $L_{ij} = \text{sign}(x_1^i - x_1^j)(x_2^i - x_2^j)$.

This estimator is unbiased. As we said above, the only possible issue comes from the cases $L_{ij} = 0$ corresponding to ties. However, in most financial cases, the variables at stake are continuous and ties are not an issue – the interested reader may read [28] about other indicators such as Kendall's $\tau_b$ / $\tau_c$ and Goodman-Kruskal $\gamma$ in the case of discrete variables subject to ties.

## Part II

# Correlation matrix cleaning for portfolio management

## 4 Introduction

### 4.1 Covariance and correlation matrices

In the first part of this document we described the different tools used in Finance and in other areas to measure dependence between random variables. In spite of its drawbacks Pearson's $r$, hereafter correlation, is in practice the most popular indicator for several reasons. Firstly, when asset returns are supposed to be gaussian – an assumption at odds with empirical data but often made in portfolio management –, correlation perfectly characterizes the dependence structure as explained above. Secondly, since risk is often measured through variance, it is natural to consider covariance and hence correlation. Also, when it comes to problems involving more than two assets, covariance matrices and correlation matrices are ubiquitous in Finance (risk management, portfolio choice, ...) because the simplest description of a dependence structure describes the dependence structure of every pair of variables. This part of the document is dedicated to the estimation of covariance matrices and correlation matrices with a special focus on portfolio management.

For the sake of completeness, we recall the definition of a covariance matrix and the definition of a correlation matrix:

**Definition 13.** *Let us consider $(X_1, \ldots, X_n)$ a tuple of $n$ random variables in $L^2$. The covariance matrix $\Sigma$ associated to $(X_1, \ldots, X_n)$ is defined as:*

$$\Sigma = (\mathrm{Cov}(X_i, X_j))_{1 \leq i,j \leq n}$$

*where* $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

**Definition 14.** *Let us consider* $(X_1, \ldots, X_n)$ *a tuple of* $n$ *random variables in* $L^2$*. The correlation matrix* $C$ *associated to* $(X_1, \ldots, X_n)$ *is defined as:*

$$C = (r(X_i, X_j))_{1 \leq i,j \leq n}$$

*where* $r(\cdot, \cdot)$ *is the linear correlation coefficient (Pearson's* $r$*).*

Now, if we are given $T$ i.i.d. realizations $(x_1^t, \ldots, x_n^t)_{1 \leq t \leq T}$ of the random variable $(X_1, \ldots, X_n)$, a natural estimator for the covariance matrix $\Sigma$ is the sample covariance matrix (or empirical covariance matrix) $\Sigma_{samp}$ defined by:

$$\Sigma_{samp} = \left( \frac{1}{T} \sum_{t=1}^{T} (x_i^t - \overline{x_i})(x_j^t - \overline{x_j}) \right)_{1 \leq i,j \leq n}$$

where $\forall i, \overline{x_i} = \frac{1}{T} \sum_{t=1}^{T} x_i^t$.

This estimator $\Sigma_{samp}$ inherits its properties from the properties of the empirical covariance coefficients and we have that $\frac{T}{T-1}\Sigma_{samp}$ is an unbiased estimator of $\Sigma$. Also, as $T$ tends to $+\infty$, the number of observations tends to $+\infty$ and the law of large numbers applies:

$$\Sigma_{samp} \rightarrow_{T \rightarrow +\infty} \Sigma$$

Turning to the correlation matrix $C$, a natural estimator is the sample correlation matrix (or empirical correlation matrix) $C_{samp}$ defined by:

$$C_{samp} = \left( \frac{1}{T} \frac{\sum_{t=1}^{T} (x_i^t - \overline{x_i})(x_j^t - \overline{x_j})}{s_i s_j} \right)_{1 \leq i,j \leq n}$$

where $\forall i, \overline{x_i} = \frac{1}{T} \sum_{t=1}^{T} x_i^t$ and $s_i = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_i^t - \overline{x_i})^2}$.

As for Pearson's $r$, we know the bias of the estimator and we can modify it to make it is approximately unbiased. Also, as above, as $T$ tends to $+\infty$, the law of large numbers applies and we get:

$$C_{samp} \rightarrow_{T \rightarrow +\infty} C$$

In spite of these nice properties in terms of bias and convergence for both the sample covariance matrix and the sample correlation matrix, the number of parameters to estimate is of order $\frac{n^2}{2}$ and we only have $nT$ observations. Therefore, if we are not in a situation where $T \gg n$, then there will be an important noise component in the empirical estimations of both $\Sigma$ and $C$.

In Finance, $n$ often represents the number of stocks or assets in a portfolio and $T$ represents the number of observations. Hence, if one considers 2 years of daily observations, *i.e.* around $T = 500$ observations for each stock, it is often the case that $n$ is of the same order as $T$. As a result, classical estimators are ill-conditioned for practical use and we are going to study cleaning methods for the covariance matrices and the correlation matrices. Cleaning methods identify the noise in sample covariance/correlation matrices and remove it. Several methods exist and their relevancy depends on the context. Methods have been developed whose use is adapted to genomics (see [15] or [78]) but not suited to financial problems. Here we present methods suited to portfolio optimization and we recall the Markowitz framework in the next paragraphs. Another field requiring good covariance matrix estimates is risk management, for VaR measurement for instance. It is, however, noteworthy that cleaning methods are of utmost importance for portfolio optimization but rather useless for VaR estimations (in a gaussian context) since the sample covariance matrix is the best estimator in the latter case.

## 4.2 The Markowitz framework and the importance of reliable covariance matrices estimates

The Markowitz mean-variance model [87] is a standard framework in modern finance for portfolio optimization. Theoretically appealing, it relies mainly (i) on the forecast of future returns of individual assets and (ii) on the estimation/forecast of the covariance matrix of assets' returns. Hence, it requires good estimates of the covariance matrix all the more since applying mean-variance optimization with the sample covariance matrix induces "estimation-error maximization" as [91] phrased it (see below).

Let us consider a risk-free asset and a set of $n$ risky assets in which an agent can invest. Let us write $w_1, \ldots, w_n$ the weights of the respective risky assets in the portfolio. Then, if $\mu = (\mu_1, \ldots, \mu_n)'$ is the vector of the expected excess returns of the risky assets and $\Sigma$ the covariance matrix of the risky assets' returns (supposed to be invertible), we recall that the mean-variance Markowitz model consists of choosing $w$ according to the following minimization principle:

$$\inf_{w'\mu \geq \mu_{min}} \quad w'\Sigma w$$

where $\mu_{min}$ is a return target.

In other words, we want to minimize the risk of the portfolio, measured as its variance, under the constraint of a minimum portfolio return.

The optimal weights $w^*$ are given by classical optimization techniques and we obtain:

$$w^* = \mu_{min} \frac{\Sigma^{-1}\mu}{\mu'\Sigma^{-1}\mu}$$

Using the spectral decomposition of $\Sigma$, we have:

$$w^* \propto \sum_{k=1}^{n} \frac{1}{\lambda_k} \langle V_k, \mu \rangle V_k$$

where $\lambda_1 \geq \ldots \geq \lambda_n$ are the eigenvalues of $\Sigma$ and $V_1, \ldots, V_n$ the associated eigenvectors. We see that the composition of the optimal portfolio is a linear combination of the eigenvectors with, *a priori*, more weight put on the eigenvectors corresponding to the smallest eigenvalues. However, the smaller the eigenvalues we consider, the larger the estimation errors: this is the "estimation-error maximization" phenomenon.

Consequently, we need a reliable estimate for $\Sigma$. This is clear when one computes the minimal risk according to three different points of view:

- Firstly, the theoretical minimal risk:

$$R^2_{theo} = w^{*\prime}\Sigma w^* = \frac{\mu^2_{min}}{\mu'\Sigma^{-1}\mu}$$

- Secondly, considering an estimator $\widehat{\Sigma}$ of $\Sigma$ and using this estimator to compute an optimal portfolio, the resulting estimated minimal (in-sample) risk is:

$$R^2_{estim} = \frac{\mu^2_{min}}{\mu'\widehat{\Sigma}^{-1}\mu}$$

- Thirdly, considering an estimator $\widehat{\Sigma}$ of $\Sigma$ and using this estimator to compute an optimal portfolio, the resulting minimal (out-of-sample) risk is:

$$R^2_{real} = \mu^2_{min}\frac{\mu'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\mu}{(\mu'\widehat{\Sigma}^{-1}\mu)^2}$$

If $\widehat{\Sigma}$ is an unbiased estimator of the true covariance matrix $\Sigma$, then Jensen's inequality gives:

$$R^2_{estim} \leq R^2_{theo}$$

and clearly, by optimality:

$$R^2_{theo} \leq R^2_{real}$$

In other words, using $\widehat{\Sigma} = \frac{T}{T-1}\Sigma_{samp}$ to compute our portfolio, we underestimate the theoretical minimal risk $R^2_{theo}$ and the real risk $R^2_{real}$ is obviously higher than this minimal risk.

To exemplify the extent of the risk estimate bias, Pafka and Kondor [98] considered the case $\Sigma = I_n$ and the limit $n, T \to +\infty$, holding $Q = \frac{T}{n} > 1$ constant. In that case we have a simple relation between the three risks:

$$R^2_{estim} = R^2_{theo}\sqrt{1-\frac{1}{Q}} = R^2_{real}\left(1-\frac{1}{Q}\right)$$

We see that the errors made depend strongly on the ratio $\frac{T}{n}$ that is linked to the ratio between the number of coefficients to estimate and the number of data points.

When $\frac{T}{n} < 1$, the situation is even worse and this is a general remark. Since there are, in that case, more assets than observations, the sample covariance matrix cannot be invertible. In particular, there exists a non-trivial portfolio that appears to be risk-free and provides positive excess return... In reality, this portfolio is risky and using the sample matrix in such a case (with a pseudo-inverse of $\Sigma_{samp}$ for instance) is unwise.

It is noteworthy that the framework we presented here is the Markowitz framework with a risk-free asset. We illustrated the importance of covariance estimates in this context but the problem is similar in the absence of a risk-free asset, when the trader can only invest in $n$ risky assets. In that case, the optimal portfolio is indeed a linear combination of the minimum variance portfolio $\Sigma^{-1}1$ and of the portfolio $\Sigma^{-1}\mu$.

This introduction focused on the Markowitz framework in order to illustrate the curse of dimensionality in covariance matrix estimation. The goal of this part is to provide better estimates than the naive ones whilst cleaning the sample covariance/correlation matrix from the noise it contains. Certain methodologies are suited to correlation matrix cleaning and others are naturally adapted to covariance matrices. In practice, this is not an issue since individual variances estimates are usually good and one can build a covariance matrix estimate from variances estimates and a correlation matrix estimate.

In what follows, we present the different methods proposed in the financial literature: factor models, shrinkage and random matrix theory. Other methods have been proposed, some of which provide bad results (*e.g.* bootstrapping – see [124]) and others which are efficient for non-financial applications (mainly in biology).

# 5 Factor models

Noise in covariance matrices is a curse of dimensionality *i.e.* when the number of coefficients to estimate is of the same order as the number of observations, one cannot expect sound estimates. To reduce the number of coefficients to estimate, one of the first ideas introduced in the literature has been to impose a structure to the covariance matrix through a factor model. In this approach, we choose $K \ll n$ observable factors $f_1, \dots, f_K$ that are presumably explanatory factors of the returns and we consider for each asset $i \in \{1, \dots, n\}$, the linear regression of this asset's returns on the factors. More precisely, if we consider the price $S_i^t$ of the asset $i$ at time $t$, then we can build the returns[8] $r_i^t = \log\left(\frac{S_i^t}{S_i^{t-1}}\right)$ and consider, for each $i$, the following model:

$$r_i^t = \sum_{k=1}^{K} b_{ik} f_k^t + \epsilon_i^t, \qquad \mathbb{E}[\epsilon|f] = 0$$

the covariance matrix of the idiosyncratic noises $\epsilon_1, \dots, \epsilon_n$ being a diagonal matrix $D$.

Then, we obtain the following covariance matrix in the model:

$$\Sigma_{model} = B\mathbb{V}(f)B' + D$$

where $B = (b_{ik})_{1\leq i\leq n, 1\leq k\leq K}$ is the $n \times K$ matrix of the model coefficients and $\mathbb{V}(f)$ is the covariance matrix of the factors.

A natural estimator of $\Sigma_{model}$ is then obtained considering:

- for $B$, the regression coefficients obtained with an optimal least squares procedure
- for $\mathbb{V}(f)$, the empirical covariance matrix of the factors

[8] In some models, excess returns are considered instead of returns. Also, one may consider returns normalized by their standard deviation to focus on the correlation matrix instead of the covariance matrix.

- for $D$, the diagonal matrix made of the empirical variance of each residual.

Then, the number of parameters to estimate is $nK + n$ for the regression coefficients and the variances of the residuals, in addition to the $\frac{K(K+1)}{2}$ coefficients for the covariance matrix of the factors. Therefore, the curse of dimensionality disappears as soon as $K \ll T$.

However, this raises two questions for practical use: Firstly, how to choose the number $K$ of factors and the factors themselves? Secondly, what are the properties of the resulting covariance matrix estimator?

Regarding the number of factors, it is clear that the fewer the factors, the stronger the structure. One of the first proposal made in the literature is rooted to the Capital Asset Pricing Model (CAPM). In that case, also often referred to as Sharpe one-factor model, $K = 1$ and the unique factor corresponds to the (excess) return of the market portfolio or to the market mode if we refer to the first principal component of the covariance matrix. This is the simplest case of a factor model and the covariance matrix then reduces to:

$$\Sigma_{CAPM} = \left(\beta_i \beta_j \sigma^2_{market} + \delta_{ij} \sigma^2_{\epsilon_i}\right)_{1 \leq i,j \leq n}$$

where $\beta_i$ is the $\beta$, as in the CAPM, of asset $i$, $\sigma_{market}$ is the volatility of the market factor and $\sigma^2_{\epsilon_i}$ is the variance of the idiosyncratic component of the return.

In practice, one considers the regression of the returns on market return and obtains the estimator:

$$\widehat{\Sigma}_{CAPM} = \left(\widehat{\beta}_i \widehat{\beta}_j \widehat{\sigma}^2_{market} + \delta_{ij} \widehat{\sigma}^2_{\epsilon_i}\right)_{1 \leq i,j \leq n}$$

Since $K = 1$, this estimator is crude and it does not constitute a reliable estimator *per se*. However, it is often used in the shrinkage approach as a target covariance matrix (see below and [73]) because of its constrained structure.

A better choice for the factors is to use Fama-French three-factor model [48, 49] which is an extension of the CAPM. In that case, $K = 3$ and the three factors are (i) the excess return of the market, (ii) SMB, a factor that takes account of market capitalization to differentiate large caps from small caps and (iii) HML, a factor that takes account of the book-to-market ratio to distinguish value stocks from growth stocks. Other models that extend the CAPM are available and the fourth factor usually considered accounts for a momentum effect.

To choose the number of factors and their nature, two other routes exist. The first one considers statistical factors such as principal components and explains the return with a small number of these factors. The second one uses both industry/sector factors and risk indices in order to explain the returns. In that latter case, the number of factors must be large enough to obtain a good fit. These two approaches are well instanced by the risk models offered by companies such as APT and BARRA.

Fan, Fan and Lv [50] compared the convergence speed of estimators based on factor models and the convergence speed of the sample covariance matrix toward the true covariance matrix and obtained remarkable results. Both theoretically and in practice (using a Fama-French framework), they showed that factor models do not provide any significant improvement when it comes to measuring the risk of a given portfolio (chosen independently of the covariance estimates). More generally, when the hypotheses of the $K$-factor model are verified and under mild additional hypotheses, they proved that for $\widehat{\Sigma}$ an estimator of $\Sigma$ based on a $K$-factor model:

$$\|\widehat{\Sigma} - \Sigma\| = \mathcal{O}\left(\frac{nK}{T^{\frac{1}{2}}}\right)$$

and

$$\|\Sigma_{samp} - \Sigma\| = \mathcal{O}\left(\frac{nK}{T^{\frac{1}{2}}}\right)$$

where the norm we consider on $n \times n$ matrices is the Frobenius norm $\|A\| = \mathrm{tr}(A'A)^{\frac{1}{2}}$. However, and this is of great interest for portfolio optimization, they obtain theoretical results in favor of factor models for problems involving $\Sigma^{-1}$. If indeed we compare the speed of convergence of $\widehat{\Sigma}^{-1}$ and $\Sigma_{samp}^{-1}$ toward $\Sigma^{-1}$ we obtain, under the same set of hypotheses, that:

$$\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\| = o\left(\frac{nK^2 \log(T)^{\frac{1}{2}}}{T^{\frac{1}{2}}}\right)$$

and

$$\|\Sigma_{samp}^{-1} - \Sigma^{-1}\| = o\left(\frac{n^2K \log(T)^{\frac{1}{2}}}{T^{\frac{1}{2}}}\right)$$

Hence, since it is natural to have $K \ll n$, substantial gains can be made using $\widehat{\Sigma}^{-1}$ instead of $\Sigma_{samp}^{-1}$ to estimate $\Sigma^{-1}$.

These theoretical results and the gains achieved in using $\widehat{\Sigma}^{-1}$ have been confirmed empirically in [50] through simulations within the Fama-French framework.

As a conclusion on factor models, we must emphasize four points. Firstly, covariance matrix estimates based on the factor models are simple estimates with a specific structure. Secondly, they do not provide improvement over the sample covariance matrix in terms of convergence speed toward the true covariance matrix. Thirdly, they contain less noise than the sample covariance matrix and may be considered as target matrices in the shrinkage approach (see below). Fourthly, as far as the estimation of the inverse of the covariance matrix is concerned, better estimators are obtained with a good factor model than with the inversion of the sample covariance matrix.

Now, before we turn to the presentation of the shrinkage approach, we have to say that some practitioners use a methodology based on sectoral factors and/or geographical factors which is different from the methodology presented above, although it also provides correlation matrices with a highly constrained structure. The idea is to consider a partition of the assets in $K$ groups. Within each group, a correlation matrix is computed. Across groups, the correlation matrices are constant correlation matrices calculated using the average correlations between pairs whose components belong to the concerned groups. This methodology produces estimates for the correlation matrices which are block matrices. It is based on a clustering of the assets and the clusters are often defined *a priori*. Another approach to define clusters uses clustering algorithms which are themselves based on correlation matrices. The interested reader may refer to [4] for classical clustering procedures and we also review some of them in the next part of this document since they are linked to graphical representations of dependence structure.

# 6 Shrinkage

## 6.1 Introduction to shrinkage

We have seen so far two types of estimators for the covariance matrix of asset returns: sample covariance matrix and estimators based on structural hypotheses. The main advantage of the sample covariance matrix is its unbiasedness. However, being unconstrained, this unbiased estimator contains a lot of noise when $T$ and $n$ are of the same order. Turning to the estimators of the previous section, they are biased but they have a very specific structure and they are definitely less noisy. Consequently, we face a trade-off that is rooted to Stein's work in statistics: a trade-off between estimation error and bias. The solution proposed by the shrinkage approach considers a linear combination of a noisy unbiased estimator (which is always the sample covariance matrix or the sample correlation matrix[9]) and a potentially biased estimator with a lot of structure called the target matrix.
This approach is called shrinkage because extreme values of the sample matrix are somehow "shrunk" toward the values of the target matrix that are assumed to be more reasonable.

This raises several questions. How should we choose the target matrix? How should we decide on the intensity of the shrinkage? In other words, what is the optimal structure we should impose to the estimator?

As far as the choice of the target matrix $F$ is concerned, three cases have been studied in the literature: the case of a matrix $\sigma^2 I_n$ (constant variance and no correlation) [74], the case

[9] The sample correlation matrix is slightly biased but the bias is small, well understood, and can be easily corrected.

of a constant correlation target [72] and the case of a covariance matrix based on Sharpe one-factor model [73].

Whatever the choice of the target covariance matrix $F$, the shrinkage approach considers:

$$\Sigma_{shrinkage} = (1-\theta)\Sigma_{samp} + \theta F$$

where $\theta$ is the shrinkage intensity: the case $\theta = 0$ corresponds to the sample covariance matrix and the case $\theta = 1$ corresponds to the target matrix $F$. To choose $\theta$, the usual criterion is to minimize

$$\mathbb{E}\left[\|\Sigma - ((1-\theta)\Sigma_{samp} + \theta F)\|^2\right]$$

where $\|\cdot\|$ is the Frobenius norm.

## 6.2 Shrinkage toward a covariance matrix of the form $F = \sigma^2 I_n$

In the Markowitz framework, the optimal portfolio is a linear combination of the eigenvectors of the covariance matrix with weights depending on the inverse of the eigenvalues of the covariance matrix. The largest eigenvalues of the covariance matrix, along with the corresponding eigenspaces, are usually well estimated. However, this is not true anymore when it comes to the smallest eigenvalues (see the applications of random matrix theory in the next section for a justification of this point). To mitigate the influence of very low eigenvalues corresponding to noise, the first shrinkage approach we present considers the target matrix $F = \sigma^2 I_n$ where $\sigma^2$ is the arithmetic average of the variances of the different assets' returns. In other words, $F = \sigma^2 I_n$ and $\Sigma$ have the same trace.

In this context, the shrinkage intensity $\theta$ which minimizes our objective function is $\theta^*$ given by:

$$\theta^* = \frac{a}{b}$$

where $a = \mathbb{E}\left[\|\Sigma_{samp} - \Sigma\|^2\right]$ and $b = \mathbb{E}\left[\|\Sigma_{samp} - \sigma^2 I_n\|^2\right]$.
This gives an estimator that is theoretical rather than practical since one must estimate $\sigma^2$ and $\theta^*$ before any practical use.

A consistent estimator has then been proposed in [74]:

- $\sigma^2$ is estimated by $\widehat{\sigma^2} = \frac{1}{n}\mathrm{tr}(\Sigma_{samp})$
- $b = \mathbb{E}\left[\|\Sigma_{samp} - \sigma^2 I_n\|^2\right]$ is estimated by $\widehat{b} = \|\Sigma_{samp} - \widehat{\sigma}^2 I_n\|^2$
- $a = \mathbb{E}\left[\|\Sigma_{samp} - \Sigma\|^2\right]$ is estimated by $\widehat{a} = \min\left(\widehat{b}, \frac{1}{T}\sum_{t=1}^{T} \|(x^t - \overline{x})'(x^t - \overline{x}) - \Sigma_{samp}\|^2\right)$

The resulting estimator of the covariance matrix is:

$$\widehat{\Sigma}_{shrinkage} = \left(1 - \frac{\widehat{a}}{\widehat{b}}\right)\Sigma_{samp} + \frac{\widehat{a}}{\widehat{b}}\frac{1}{n} tr\left(\Sigma_{samp}\right) I_n$$

Another possibility – although it is not proposed in [74] – could be to work on the correlation matrix instead of the covariance matrix. In that case, there would be no need to average the variances and the target correlation matrix would be $I_n$.

## 6.3 Shrinkage toward a constant correlation matrix $F$

The first shrinkage approach we presented misses a stylized fact of stock markets: asset returns tend to be positively correlated. There is no reason to consider a target matrix with zero correlations and [72] proposed to consider a target matrix $F$ with constant correlations. More precisely, we are going to consider the following matrices:

- $\Sigma = \left(r_{ij}\sigma_i\sigma_j\right)_{1\leq i,j\leq n}$ the true covariance matrix
- $\Sigma_{samp} = \left(\tilde{r}_{ij}s_i s_j\right)_{1\leq i,j\leq n}$ the sample covariance matrix
- A theoretical target $\Phi = \left(\phi_{ij}\right)_{1\leq i,j\leq n}$ with

$$\phi_{ii} = \sigma_i^2, \qquad \forall i \neq j, \phi_{ij} = \overline{r}\sigma_i\sigma_j$$

where $\overline{r} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} r_{ij}$

- The target matrix $F = (f_{ij})_{1 \le i,j \le n}$ with

$$f_{ii} = s_i^2, \qquad \forall i \neq j, f_{ij} = \overline{\tilde{r}} s_i s_j$$

where $\overline{\tilde{r}} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \tilde{r}_{ij}$

In that case, the shrinkage intensity $\theta^*$ that minimizes $\mathbb{E}\left[\|\Sigma - ((1-\theta)\Sigma_{samp} + \theta F)\|^2\right]$ is of the form $\theta^* = \alpha \frac{1}{T} + \mathcal{O}\left(\frac{1}{T^2}\right)$ where $\alpha$ can be obtained in closed-form:

$$\alpha = \frac{\pi - \rho}{\gamma}$$

with

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{n} \lim_{T \to +\infty} \mathbb{V}(\sqrt{T} \tilde{r}_{ij} s_i s_j)$$

$$\rho = \sum_{i=1}^{n} \sum_{j=1}^{n} \lim_{T \to +\infty} \mathrm{Cov}\left(\sqrt{T} f_{ij}, \sqrt{T} \tilde{r}_{ij} s_i s_j\right)$$

$$\gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} (\phi_{ij} - r_{i,j} \sigma_i \sigma_j)^2$$

Hence, once we have consistent estimators $\hat{\pi}$, $\hat{\rho}$ and $\hat{\gamma}$ of the above quantities (see [72] for the expressions of the estimators), a natural shrinkage estimator of the covariance matrix $\Sigma$ is:

$$\widehat{\Sigma}_{shrinkage} = \left(1 - \widehat{\theta}^*\right) \Sigma_{samp} + \widehat{\theta}^* F$$

where $\widehat{\theta}^* = \max\left(0, \min\left(1, \frac{\hat{\pi} - \hat{\rho}}{\hat{\gamma}} \frac{1}{T}\right)\right)$.

## 6.4 Shrinkage toward a covariance matrix $F$ based on Sharpe one-factor model

The third choice we present for the target matrix is a covariance matrix based on the simplest factor model we presented in the previous section. We recall that covariance matrices based on factor models have a well-defined structure and are therefore good candidates to be target matrices in the shrinkage procedure. Ledoit and Wolf proposed in [73] to consider Sharpe one-factor model, the unique factor being the return of a market portfolio[10]. As in the previous case, we need to consider 4 matrices:

- $\Sigma = \left(r_{ij}\sigma_i\sigma_j\right)_{1\leq i,j\leq n}$ the true covariance matrix.
- $\Sigma_{samp} = \left(\tilde{r}_{ij}s_i s_j\right)_{1\leq i,j\leq n}$ the sample covariance matrix
- A theoretical target $\Phi = \left(\phi_{ij}\right)_{1\leq i,j\leq n}$ with

$$\phi_{ii} = \beta_i^2\sigma_{market}^2 + \sigma_{\epsilon_i}^2, \qquad \forall i \neq j, \phi_{ij} = \beta_i\beta_j\sigma_{market}^2$$

- The target matrix $F = \left(f_{ij}\right)_{1\leq i,j\leq n}$ with

$$f_{ii} = \widehat{\beta_i}^2\widehat{\sigma}_{market}^2 + \widehat{\sigma}_{\epsilon_i}^2, \qquad \forall i \neq j, f_{ij} = \widehat{\beta}_i\widehat{\beta}_j\widehat{\sigma}_{market}^2$$

Then, the shrinkage intensity $\theta^*$ which minimizes $\mathbb{E}\left[\|\Sigma - ((1-\theta)\Sigma_{samp} + \theta F)\|^2\right]$ is:

$$\theta^* = \frac{\sum_{1\leq i,j\leq n} \mathbb{V}(\tilde{r}_{ij}s_i s_j) - \mathrm{Cov}(f_{ij}, \tilde{r}_{ij}s_i s_j)}{\sum_{1\leq i,j\leq n} \mathbb{V}(f_{ij} - \tilde{r}_{ij}s_i s_j) + (\phi_{ij} - r_{ij}\sigma_i\sigma_j)^2}$$

As for the previous case, we use in practice the asymptotic expansion in $\frac{1}{T}$ of $\theta^*$ which is of the same form:
$\theta^* = \alpha\frac{1}{T} + \mathcal{O}\left(\frac{1}{T^2}\right)$ where $\alpha = \frac{\pi-\rho}{\gamma}$ with

$$\pi = \sum_{i=1}^{n}\sum_{j=1}^{n}\lim_{T\to+\infty}\mathbb{V}(\sqrt{T}\tilde{r}_{ij}s_i s_j)$$

$$\rho = \sum_{i=1}^{n}\sum_{j=1}^{n}\lim_{T\to+\infty}\mathrm{Cov}\left(\sqrt{T}f_{ij}, \sqrt{T}\tilde{r}_{ij}s_i s_j\right)$$

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n}(\phi_{ij} - r_{i,j}\sigma_i\sigma_j)^2$$

[10] See the section on factor models for details and notations

Consistent estimators $\hat{\pi}$, $\hat{\rho}$ and $\hat{\gamma}$ of the above quantities are available in [73] and the resulting estimator of the covariance matrix is:

$$\widehat{\Sigma}_{shrinkage} = \left(1 - \widehat{\theta}^*\right) \Sigma_{samp} + \widehat{\theta}^* F$$

where $\widehat{\theta}^* = \max\left(0, \min\left(1, \frac{\hat{\pi}-\hat{\rho}}{\hat{\gamma}} \frac{1}{T}\right)\right)$.

### 6.5 Concluding remarks on shrinkage

We presented above the three most common shrinkage estimators for covariance matrices. Variants of these approaches exist and another shrinkage procedure is for instance considered in [101] where the target matrix $F$ is a constant covariance matrix. One can also build new estimators based on the same ideas, considering other factor models to build a target matrix or even considering a “portfolio” of estimators and averaging them. In practice, what is important for the shrinkage approach to be successful is that extreme values of the sample matrix are indeed shrunk toward more reasonable values.

To conclude on shrinkage, we should insist on three points: the shrinkage procedure is simple, it can be applied both when $T < n$ and $T \geq n$ and it provides an invertible covariance matrix as soon as the target matrix is itself a symmetric positive definite matrix. The latter point is important in the case $T < n$, since the use of pseudo-inverse matrices usually gives poor results.

## 7 Eigenvalue filtering based on random matrix theory

### 7.1 Spectral analysis

When the number of observations $n \times T$ is of the same order as the number of parameters to evaluate, that is $\frac{1}{2}n(n+1)$ for a covariance matrix, we claimed that the sample covariance matrix or the sample correlation matrix contains a lot of noise. To remove the noise from the sample correlation matrix (the approach presented in this section is better suited to correlation matrices), the idea developed by Bouchaud, Potters and their coauthors on one side

[22, 70, 71, 105] and by physicists from the Boston University [102, 103] on the other side, is to compare the sample correlation matrix to a random correlation matrix. More precisely, they propose to compare the spectrum of the sample correlation matrix to the spectrum of a so-called Wishart matrix (a random covariance matrix of the form $\frac{1}{T}YY'$ where $Y$ is an $n \times T$ matrix with random coefficients having mean 0 and variance 1). The spectrum of a Wishart matrix has been widely studied as one of the important topics of random matrix theory (see for instance [43] or [86]) and we exhibit below central properties for applications to correlation matrix cleaning.

At the limit when $T \to +\infty$ and $n \to +\infty$, the ratio $Q = \frac{T}{n} \geq 1$ being kept constant, we have a closed-form expression for the distribution of the eigenvalues of a random matrix of the Wishart form. Indeed, we know that the limit spectrum is deterministic and described by a Marcenko-Pastur [86] distribution function[11]:

$$\rho(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

where $\lambda_{min} = 1 + \frac{1}{Q} - 2\sqrt{\frac{1}{Q}}$ and $\lambda_{max} = 1 + \frac{1}{Q} + 2\sqrt{\frac{1}{Q}}$.

One of the most important characteristics of this distribution is its compact support since eigenvalues above the threshold $\lambda_{max}$ cannot correspond to noise. In fact, the situation is more complicated than this since we do expect that the true correlation matrix has a structure corresponding to the assets under scrutiny. At least, we know that there is a market mode corresponding to the first principal component. The relevant random matrix whose spectrum should be compared to the spectrum of the sample correlation matrix must take account of this market mode. In that case, instead of comparing the spectrum to the above distribution, one compares it to a modified one, still denoted $\rho(\cdot)$:

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

[11] If $Q < 1$ then a proportion $1-Q$ of the eigenvalues are equal to 0 and the others are distributed according to the truncated Marcenko-Pastur distribution.

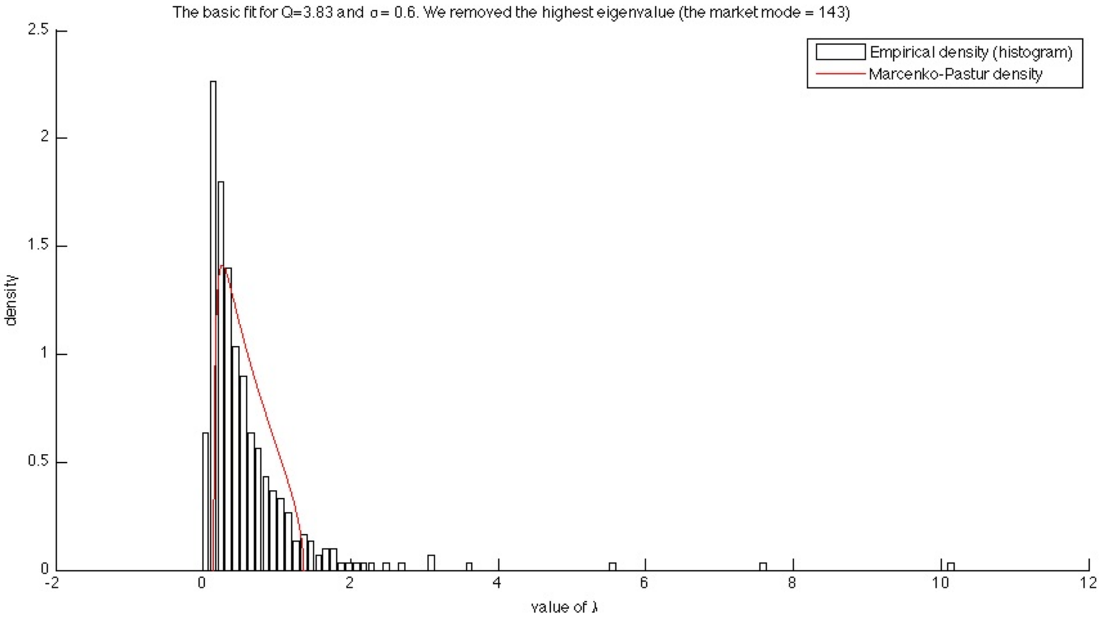

Figure 1: Spectrum of a sample correlation matrix vs. Marcenko-Pastur distribution

where $\lambda_{min} = \sigma^2 \left(1 + \frac{1}{Q} - 2\sqrt{\frac{1}{Q}}\right)$ and $\lambda_{max} = \sigma^2 \left(1 + \frac{1}{Q} + 2\sqrt{\frac{1}{Q}}\right)$, $\sigma^2$ being the part of the variance not explained by the market mode.

This is exemplified on Figure 1 for a few hundreds stocks belonging to the Eurostoxx 600 index – the market mode is not represented.
Figure 1 also shows that the lowest eigenvalues are the eigenvalues which most likely correspond to noise[12]. In the Markowitz framework, they also correspond to the most weighted investment directions. This is the main reason why using sample covariance matrices for portfolio optimization gives such bad results.
Also, a remarkable feature is that the proportions of eigenvalues lying in the interval $[\lambda_{min}, \lambda_{max}]$ is approximately 92%: less than 10% of the signal corresponds to trustworthy information! This has been also observed by Bouchaud and his coauthors who often obtained similar figures (*e.g.* 6% in [70]).

One may ask whether the approximation $n, T \to \infty$ is relevant when $n$ and $T$ are finite. This question has been tackled and we know the behavior of the largest eigenvalue of a Wishart

[12] The eigenvalues lying below the threshold $\lambda_{min}$ are often considered noisy in applications (see below).

matrix for finite values of $n$ and $T$. More precisely, when the variables are in $L^4$, we know, denoting $\lambda_1$ the largest eigenvalue of $C_{sample}$, that:

$$\mathbb{P}\left(\lambda_1 \leq \lambda_{max} + \sqrt{\frac{1}{Q}}\lambda_{max}^{\frac{2}{3}} T^{-\frac{2}{3}} u\right) \to_{T,n\to+\infty,T/n=Q} TW(u)$$

where $TW$ is the cumulative distribution function of the Tracy-Widom distribution (see for instance [42]).

In practice, the limit case is always considered, and, moreover, $\sigma^2$ is often calibrated ex-post (and sometimes also $Q$, though marginally) to better fit the so-called bulk of eigenvalues and thus identify which eigenvalues correspond to noise and which eigenvalues correspond to a real signal.
Hence, choosing the threshold $\lambda_{max}$ is on one hand a science, based on important results of random matrix theory, but it is also on the other hand an art[13]!

So far, we have only compared the eigenvalues of the sample correlation matrix to the eigenvalues of a (random) Wishart matrix and we explained why we can only be confident in the estimates of the eigenvalues which do not lie in the Marcenko-Pastur sea $[\lambda_{min}, \lambda_{max}]$. If we consider the Markowitz framework, eigenvalues are important because small noisy eigenvalues contaminate the estimation of the optimal portfolio. But eigenvectors are even more important because they determine the portfolios in which the agent should invest. We now turn to the comparison of the eigenvectors of the sample correlation matrix with the eigenvectors of a Wishart matrix (the $\|\cdot\|_2$-norm of the eigenvectors being normalized to $n$ in order to compare).

For a Wishart matrix, each component of any (appropriately normalized) eigenvector is distributed as a gaussian variable $\mathcal{N}(0,1)$. Hence, one can consider the eigenvectors of a sample matrix and compare the distribution of the coefficients with the distribution of a gaussian variable $\mathcal{N}(0,1)$ and/or use a normality test like Jarque-Bera. In practice (see for instance [102, 103]) eigenvectors corresponding to bulk eigenvalues (except sometimes near the thresh-

[13] In practice, the exact value of the threshold does not influence drastically the out-of-sample variance of a portfolio computed using the cleaning methods we present in this section.

olds) are typically in agreement with random matrix predictions and eigenvalues outside of the bulk usually exhibit a different pattern (the best example being the first principal component because all coefficients are usually of the same sign: this is the market mode). Therefore, the analysis of the eigenvectors allows one to refine the choice of the threshold. It also provides a sound interpretation of the threshold since the first principal components usually have economic meanings. In addition to this comparison to a vector whose components are normally distributed, authors also proposed to measure the degree of deviation from random matrix theory using the notion of inverse participation ratio (see [102][14]).

## 7.2 Cleaning methodologies

We described above the main approaches to separate the eigenvalues corresponding to noise from the eigenvalues carrying information. In practice, although the compact support of $\rho(\cdot)$ is bounded from below by a positive constant (except when $Q \leq 1$), all the approaches proposed in the literature consider the diagonalization $C_{samp} = \Omega D \Omega'$ of the sample correlation matrix where

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \lambda_n \end{pmatrix}, \quad \lambda_1 \geq \ldots \geq \lambda_n$$

and keep the eigenvalues above the upper threshold of the bulk $\lambda_{k^*} \simeq \lambda_{max}$ whilst replacing the eigenvalues below $\lambda_{k*}$ by new values. Three alternatives have been proposed for these new values and the final steps of the cleaning approach:

- Plerou et al. [102] proposed a brute force method and replaced the matrix $D$ by a new

[14] In particular the authors show that, for the smallest eigenvalues, below $\lambda_{min}$, the eigenvectors exhibit some non-random pattern. However, they are usually regarded as noisy vectors carrying no relevant information.

matrix $D_{new}$ given by:

$$D_{new} = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \lambda_{k^*} & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & 0 \end{pmatrix}$$

Then, the matrix $\Omega D_{new}\Omega'$ is computed and the cleaned matrix $C_{clean}$ is equal to $\Omega D_{new}\Omega'$ except that the diagonal terms are set to 1.

- Bouchaud, Potters and their coauthors [21, 70, 71] proposed a different method, often referred to as eigenvalue clipping. They replace the matrix $D$ by a new matrix $D_{new}$ given by:

$$D_{new} = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \lambda_{k^*} & \ddots & & \vdots \\ \vdots & & \ddots & \overline{\lambda} & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & \overline{\lambda} \end{pmatrix}$$

where $\overline{\lambda}$ is chosen so that $\text{tr}(D) = \text{tr}(D_{new})$ (trace preservation is important because it corresponds to the preservation of the total variance $n$ associated to a correlation matrix).

Then, the matrix $\Omega D_{new}\Omega' = (h_{ij})_{i,j}$ is computed and the cleaned matrix is $C_{clean} = \left(\frac{h_{ij}}{\sqrt{h_{ii}h_{jj}}}\right)_{i,j}$.

- A third methodology has been proposed in [111]. They replaced the matrix $D$ by a new matrix $D_{new}$ given by:

$$D_{new} = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \lambda_{k^*} & \ddots & & \vdots \\ \vdots & & \ddots & \epsilon + (n - k^* - 1)\Delta & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & \epsilon \end{pmatrix}$$

where $\epsilon > 0$ is a small positive number (the authors proposed $10^{-8}$ but several choices for $\epsilon$ have been tested – see [33] for instance) and where $\Delta$ is chosen so that $\mathrm{tr}(D) = \mathrm{tr}(D_{new})$. Then the authors considered $C_{clean} = \Omega D_{new} \Omega'$.

*Remark: Bouchaud and Potters also recently proposed in [22] a new method which replaces the eigenvalues below the threshold by eigenvalues whose distribution follow a power law, in order to better account for the underlying hierarchical structure. In our opinion, this method deserves additional and deeper study.*

To sum up, cleaning methods based on random matrix theory consists of three steps:

- Determine the eigenvalues and eigenvectors carrying information and those corresponding to noise
- Replacing the eigenvalues corresponding to noise by new values
- Building a new matrix to replace the sample correlation matrix $C_{samp}$

The third step deserves some comments. Whatever the approach retained, the matrix $\Omega D_{new} \Omega'$ usually does not satisfy the basic assumptions of a correlation matrix. For that reason, the first two approaches compute a new matrix so that the resulting matrix $C_{clean}$ looks like a correlation matrix. A better approach for the third step could be to consider the methodology developed by Rebonato and Jäckel [107] to create a valid correlation matrix from a given matrix or the new approach developed in [106] to find the nearest correlation matrix.

## 7.3 Complementary remarks and extensions

We compared the spectrum of a sample correlation matrix to the spectrum of a random covariance matrix and claimed that below a certain threshold $\lambda_{max}$ the eigenvalues may only correspond to noise. We discussed the implications of this for portfolio optimization and we presented cleaning methodologies for practical applications. The approaches can be extended and we now discuss several extensions: comparison to random matrices more general than the simple Wishart ones, covariance matrix cleaning instead of correlation matrix cleaning, the case of a sample matrix computed with an exponentially weighted moving average, the case of fat tails and the case of input/output matrices. Also, in addition to providing new cleaning methods, random matrix theory sheds light on the drawbacks of the other cleaning approaches presented above.

### Random matrix theory highlights a drawback of the shrinkage approach

We have seen that the eigenvalues belonging to the Marcenko-Pastur sea may correspond to noise. Hence the uncertainty on the eigenvalues belonging to the bulk is higher than the uncertainty on the few relevant eigenvalues larger then $\lambda_{max}$. Therefore, if we consider the shrinkage approach before mentioned, we see that the shrinkage intensity should not in fact be a constant: the bulk eigenvalues need more shrinkage than the largest ones. The approach developed by Bouchaud, Potters and their coauthors can be regarded as a non-uniform shrinkage approach: the eigenvalues above the threshold are not shrunk and the eigenvalues below the threshold are shrunk towards their mean.

### Comparison to a more general random matrix

In the above paragraphs, we considered a Wishart matrix as a prior in order to analyze the spectrum of the sample correlation matrix. We claimed that this choice was arguable and we considered a reduced variance in order to account for the variance explained by the market mode. In fact, one can consider a more general prior for the correlation matrix if there is a reason to do so and the interested reader may refer to the recent review article by Bouchaud and Potters [22] or to the papers by Burda et al. [23, 25] for a slightly different approach. In practice these generalizations are rarely used with non-trivial priors.

### Covariance matrix vs. Correlation matrix

All the methodologies we presented can be applied to both correlation and covariance matrices. However, when it comes to eigenvalue filtering, the choice of the random matrices whose spectrum is compared to the sample matrix is natural for correlation matrices but questionable as far as covariance matrices are concerned. For a sample covariance matrix, comparing the shape of the spectrum to a Marcenko-Pastur distribution $\rho(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max}-\lambda)(\lambda-\lambda_{min})}}{\lambda}$ is arguable as soon as the assets have different volatilities, since the theoretical results apply to variables with the same variance. Fitting $\sigma^2$ to the spectrum of the sample covariance matrix is always an option but it does not rely on a sound basis and it is not recommended (see below the empirical comparisons between the approaches).

### Exponentially-weighted moving average

The sample covariance matrix on a time window of width $T$ is defined by the formula $\Sigma_{samp} = \left(\frac{1}{T}\sum_{t=1}^{T}(x_i^t - \overline{x_i})(x_j^t - \overline{x_j})\right)_{1\leq i,j\leq n}$. The main assumption underlying this definition is that observations are i.i.d. realizations of a given random variable. In practice however, the underlying distribution of the returns evolves. Then, in order to mitigate the influence of past data, a common practice is to consider, instead of the above average, an exponential smoothing of the data using an exponentially weighted moving average.

Several definitions are possible and we propose here to consider the following estimate for the covariance matrix:

$$\Sigma_{ewma} = (h_{ij})_{1\leq i,j\leq n}\,, \qquad h_{ij} = \frac{1-\alpha}{1-\alpha^T}\sum_{t=1}^{T}\alpha^{T-t}(x_i^t - \overline{x_i})(x_j^t - \overline{x_j})$$

where $\overline{x_i}$ is an estimate of the mean of $X_i$, either the usual one or, to be coherent, an exponentially weighted moving average.

A similar formula is available to estimate the correlation matrix:

$$C_{ewma} = \left(\frac{h_{ij}}{\sqrt{h_{ii}h_{jj}}}\right)_{1\leq i,j\leq n}$$

Since these estimators are commonly used by practitioners (RiskMetrics volatility estimates have long time been of this form with $\alpha$ between 0.94 and 0.97), an eigenvalue filtering procedure for these matrices is necessary and can be found in [99]. Instead of comparing the spectrum of $C_{ewma}$ to the spectrum of a Wishart matrix, it is compared to the spectrum of a matrix of the form:

$$\left(\tilde{h}_{ij}\right)_{1\leq i,j\leq n}\,, \qquad \tilde{h}_{ij} = (1-\alpha)\sum_{t=-\infty}^{T}\alpha^{T-t}y_i^t y_j^t$$

where the variables $(y_i^t)_{1\leq i\leq n, t\leq T}$ are i.i.d. gaussian variables $\mathcal{N}(0,1)$.

As for Wishart matrices, the distribution of the spectrum of the above matrix is known when $n \to +\infty$ and $\alpha \to 1$ with $Q = \frac{1}{n(1-\alpha)}$ constant[15]. This distribution is $\rho(\lambda) = \frac{Q}{\pi}v$ where $v$ is the solution of:

$$\lambda - \frac{v\lambda}{\tan(v\lambda)} + \log(v) - \log(\sin(v\lambda)) = \frac{1}{Q}$$

or 0 if there is no solution. This distribution has a compact support and a shape very similar to Marcenko-Pastur distribution. Also, a threshold to separate the noise from the relevant part of the signal can be found as in the case of the sample correlation matrix[16].

[15] The limit $\alpha \to 1$ corresponds to an average with equal weights.

[16] If one wants to take account of the variance explained by the first principal components, the distribution is the same with $v$ solution of $\frac{\lambda}{\sigma^2} - \frac{v\lambda}{\tan(v\lambda)} + \log(v\sigma^2) - \log(\sin(v\lambda)) = \frac{1}{Q}$, where $\sigma^2$ is the part of the variance unexplained by the market mode, or as above, a parameter to calibrate.

**Other extensions**

The results we presented on the spectrum of random matrices are general. To insist on this universality, let us recall that they do not depend on gaussian assumptions but rather on the existence of a second moment as far as Marcenko-Pastur result is concerned, or on the existence of a fourth moment, for the results concerning the Tracy-Widom region. Although these results are fairly universal, extensions of Marcenko-Pastur result to Lévy matrices have been developed to deal with processes exhibiting very fat tails. We do not detail these extensions that are rarely used, we simply highlight that there is no upper bound for the distribution of the eigenvalues. The interested reader can refer to [13].

Another extension of the results concern input/output matrices which are rectangular matrices. The curse of dimensionality is similar in that case and a cleaning method based on random matrix theory has been developed in [20], with the reasoning on eigenvalues being replaced by a reasoning on the singular values of the matrix.

# 8 Comparison of the different approaches

Empirical comparisons have been carried out to compare the different approaches presented so far. Some articles compare shrinkage approaches between them (see for instance [40] and [73]), other papers mainly focus on eigenvalue filtering (see for instance [33]). The most complete study is perhaps [101] in which the authors considered the single-index estimator (Sharpe one-factor model), various shrinkage estimators, the eigenvalue filtering methods of both Plerou et al. and Bouchaud, Potters and their coauthors, and other cleaning methods based on hierarchical clustering (we present the hierarchical clustering approaches later in this text along with graphical representations of the covariance/correlation matrices).

Thinking of a best cleaning method depends entirely on the context. The relative performances of cleaning methods may be different depending on the considered market. Also, the choice of the criterion to compare cleaning methods is an issue: if considering out-of-sample risk is a natural option, evaluating the reliability of the risk estimate is also important. Instead of reviewing the large number of papers comparing the different approaches with different criterions, we do prefer to extract from these empirical studies a few important lessons.

Starting with shrinkage estimators, in nearly all cases, minimum variance portfolios[17] constructed using the sample covariance matrix exhibit a very large out-of-sample variance compared to minimum variance portfolios built using an estimator based on shrinkage, be it shrinkage toward the constant correlation matrix or shrinkage towards a one-factor model covariance matrix. Also, because the specification error of the one-factor model is important, shrinkage towards a one-factor model covariance matrix is often empirically preferred to the corresponding one-factor model covariance matrix. When comparing the targets in the shrinkage approach, the constant correlation target seems to provide results similar to those of the one-factor model covariance matrix. However, considering the optimal shrinkage intensity or a weight arbitrarily set at 50% seems to provide similar results in general.

Concerning the estimators based on eigenvalue filtering, they outperform the sample covariance matrix when the criterion is the out-of-sample variance of a portfolio built as to have minimum variance. As for the shrinkage approach, the performance does not depend strongly on the choice of the threshold. However, applying the approach to correlation matrices seems to provide better results than those obtained with covariance matrices. Finally, as far as the third filtering method presented above is concerned (see Daly et al. [33] and Sharifi et al. [111]), the performances seem to worsen when $\epsilon$ is chosen close to 0, certainly because the matrix is then not far from being singular. In practice, Bouchaud and Potter's approach is often preferred among all eigenvalue filtering approaches.

The comparison between the different methods appears to depend on the context and mainly on the ratio $Q = \frac{T}{n}$. When $Q$ is small, the improvement provided by the use of most cleaning methods is substantial. In that case, shrinkage toward either a constant correlation matrix

[17] In the absence of a risk-free asset.

or a single-index covariance matrix or the eigenvalue filtering approach of Bouchaud and Potter provide similar results. However, when $Q$ is large ($Q \geq 3$) the improvement provided by the cleaning approaches vanishes. The importance of cleaning is also less obvious when short-selling is forbidden. In that case, cleaning methods are only useful for $Q \leq 1$.

Overall, the different comparisons carried out between the cleaning approaches suggest the use of eigenvalue clipping. Discussions with practitioners reinforce this conclusion since this eigenvalue filtering approach appears to be only rarely beaten by other methods and performs similarly to the best approach in all cases.

# Appendix: The optimal choice of the time horizon

In this part, we presented a range of techniques that can be used by asset managers in order to better estimate correlations. These techniques usually start from an initial sample correlation matrix computed using a dataset on a time window of width $T$, but we also discussed above other initial estimates commonly used in finance and based on exponentially weighted moving averages. The main motivation for this exponential smoothing was that the data may not be stationary and recent observations should be given more weight than older ones. Whatever the initial estimate, one parameter must be chosen: the width $T$ of the time window for the sample correlation matrix, or the discount rate $\alpha \in (0, 1)$ in the case of an exponentially weighted moving average. If the time series are non-stationary, $T$ must be chosen small (equivalently, $\alpha$ must be chosen far from 1) so that the i.i.d. hypothesis is an acceptable approximation on the time window. However, when $T$ is small, the estimation is very noisy and there is a trade-off between the need to estimate the current value of correlations, a smaller $T$, and noise, a larger $T$.

Random matrix theory, whose applications in finance are reviewed in [22, 105], can help to choose the time horizon $T$. The idea, presented in [3], is to study the stability of the first eigenspaces of two sample correlation matrices calculated on two non-overlapping time windows, and to determine whether instability is due to noise or due to changes in the actual dependence structure.

Given two sets of orthonormal vectors in $\mathbb{R}^n$, $(v_1, \dots, v_p)$ and $(w_1, \dots, w_q)$ with $p \leq q$, Allez and Bouchaud [2, 3] proposed to consider the rectangular matrix of overlaps:

$$G = (\langle v_i, w_j \rangle)_{1 \leq i \leq p, 1 \leq j \leq q}$$

The singular values of this matrix are related to the overlap between the spaces respectively spanned by the vectors $(v_1, \dots, v_p)$ and by the vectors $(w_1, \dots, w_q)$. We define what the authors call a fidelity distance between the two spaces by:

$$D = -\frac{1}{2p} \log(\det(GG'))$$

To test stationarity, the idea is to consider two non-overlapping time windows of width $T$ and to consider the above fidelity distance for $(v_1, \dots, v_p)$ the first $p$ eigenvectors of the sample correlation matrix computed on the first time window and for $(w_1, \dots, w_q)$ the first $q$ eigenvectors of the sample correlation matrix computed on the second time window. The authors show that if the eigenvectors were stationary, $D$ would be on average given by (independently of the time interval between the two time windows):

$$D \simeq \frac{1}{2pT} \sum_{i=1}^{p} \sum_{j=q+1}^{n} \left( \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} + \frac{l_i l_j}{(l_i - l_j)^2} \right)$$

where $(\lambda_i)_i$ and $(l_i)_i$ stand for the eigenvalues of the two sample correlation matrices and where $T$ is supposed to be large.

In practice, this approximation does not hold for financial data and $D$ increases with the time interval separating the two time windows. Consequently, eigenvectors cannot be assumed to be stable with time.

To solve the trade-off and choose $T$, the authors proposed to compute the average $\overline{D}$ of $D$ for $p = 5$, $q = 10$, and consecutive time windows of size $T$. Then, the function $T \mapsto \overline{D}(T)$

is often U-shaped. It first decreases as $T$ increases because the noise decreases. Then, after a certain threshold $T^*$, the changes in the true eigenvectors are more important than the decrease of the noise and the function increases. This threshold $T^*$ is a natural candidate for an optimal $T$. Based on this approach, the authors recommend to consider $T^* = 600$ days to compute the correlation matrix of the Nikkei index components, $T^* = 400$ days for the CAC40, $T^* = 450$ days for the DAX and $T^* = 700$ days for the S&P.

# Part III

# Representations of correlation and clustering

## 9 Introduction

Correlation matrices provide figures describing the linear dependence between asset returns. However, for many applications, the information contained in correlation matrices is too complex to be easily interpreted and a simple representation of the linear dependence structure is needed.

The motivation is twofold. First, it is often assumed that understanding the correlation structure may help to understand the dynamics of shock. Consequently, academics and practitioners have developed numerical methods to represent the linear dependence structure of assets as a graph whose nodes are the assets under scrutiny. Another important advantage of representing the correlation structure in a simple way, be it graphical or not, is that we can expect the representation to be less noisy than the initial sample correlation matrix. From an academic point of view, classical techniques of graph theory have been used in Finance for the first time by Mantegna [85] who borrowed the concept of minimum spanning tree to describe the correlation structure of stock returns in the form of an asset tree. This concept has also been used to study a wide variety of assets and to understand systemic risk. In addition, other correlation-based graphs have been developed to complement the minimum spanning tree.

The second motivation is clustering. Clustering algorithms partition the assets in different classes, for instance to reduce the dimensionality of a problem. Contrary to *a priori* clustering, in which different classes are defined according to a geographical or industrial criterion, clustering algorithms use sample correlation matrices as inputs to determine endogenously the different clusters. It is remarkable that clustering procedures are useful in obtaining cleaned correlation matrices with a constrained structure, especially when the constrained structure is *a priori* unknown. Linked to clustering, the determination of a hierarchical structure between assets is another motivation. Hierarchical classification is a classical topic in multivariate data analysis and new (parsimonious) estimators for correlation matrices have been proposed using dendrograms resulting from classical hierarchical classification algorithms.

# 10 Correlation-based graphs

## 10.1 Minimum spanning tree and other graphs

In order to represent in the dependence structure between asset returns, the simplest idea is to consider a graph or a network representation of the correlation structure. A fully-connected graph whose nodes are the assets can be obtained from the correlation matrix: an edge is drawn between each couple of nodes and a weight is assigned to each edge according to the correlation coefficient associated to the two assets linked by the edge. This graph representation has the same complexity as the initial correlation matrix and correlation-based graphs are usually subgraphs of this fully-connected graph, providing simple but relevant and meaningful information. To extract a relevant subgraph, the first idea is to consider an asset graph (see [95]) in which one retains the $K$ edges corresponding to the $K$ highest correlation coefficients. The problem of this approach is that, for any reasonable choice of $K$ (that is $K$ and $n$ of the same order), the resulting graph may not be connected. A connected graph, however, is required to obtain a meaningful representation of the dependence structure, and the classical approach consists of building the so-called asset tree, which is the minimum spanning tree of the fully-connected graph. This asset tree is, by definition, a connected graph without loop built using the following procedure:

1. Start from a graph with no edge.

2. Order the empirical correlation coefficients $(\rho_{ij})_{1\leq i<j\leq n}$ in descending order to obtain $\rho_{i(1),j(1)} \geq \cdots \geq \rho_{i(k),j(k)} \geq \cdots$.

3. Set $k$ to 1.

4. Add an edge between the nodes $i(k)$ and $j(k)$ if it does not create a loop.

5. Increment $k$.

6. Go to step 4 if $n-1$ edges have not been inserted.

At the end of this procedure, a tree (with $n-1$ edges) is obtained and it provides an arrangement of the stocks only using the most relevant connections between them. We will see below that the structure of this tree is meaningful and that its shape is not the same during normal periods and a crash.

The reader may wonder why this tree is called a minimum spanning tree since we only consider the largest empirical correlation coefficients. In fact, different definitions have been proposed using different "distances" between assets. Mantegna, who was the first to consider the classical notion of minimum spanning tree for financial applications in [85], transformed the correlation matrix into a "distance" matrix in order to mimic the usual graph theory approach. The distance initially proposed by Mantegna was $d(i,j) = 1 - \rho_{ij}^2$. In that case, two stocks are close to one another if the absolute value of the associated correlation coefficient is close to 1. In fact, after Mantegna, many authors used a minimum spanning tree with a different definition for the distance, most often $d(i,j) = \sqrt{2(1-\rho_{ij})}$ (see for instance the papers by Bonanno and coauthors [16, 17, 18, 19] or the papers by Onnela and coauthors [96, 97]).

The minimum spanning tree, which may be different from the initial one, is then computed using the following procedure:

1. Start from a graph with no edge.

2. Order the distances $(d(i,j))_{1\leq i<j\leq n}$ in ascending order to obtain $d(i(1),j(1)) \geq \ldots \geq d(i(k),j(k)) \geq \ldots$.

3. Set $k$ to 1.

4. Add an edge between the nodes $i(k)$ and $j(k)$ if it does not create a loop.

5. Increment $k$.

6. Go to step 4 if $n-1$ edges have not been inserted.

With this definition, the tree is indeed the shortest tree connecting all the assets, hence the name minimum spanning tree.

Minimum spanning trees are the most popular graph representations of the linear correlation structure. Other relevant trees exist and the concept of average linkage minimum spanning tree will be presented in the next section as an output of a clustering algorithm – see also [120]. More complex subgraphs of the initial fully-connected graph have been proposed, for instance a Planar Maximally Filtered Graph (PMFG) presented in [119]. Such a graph is built using the same procedure as the minimum spanning tree. In the case of a minimum spanning tree, an edge was not inserted when it generated a loop. Now, the difference is that loops and small cliques (of 3 or 4 nodes) are authorized since the only requirement is that the graph must remain a planar graph when an edge is added. Any PMFG contains the minimum spanning tree along with additional information. However, in spite of being interesting as a generalization of the minimum spanning tree, it is rarely used in practice.

## 10.2 Usual indicators and analysis

The need for a simple graph representation of the correlation structure is motivated by the complexity of this correlation structure. In addition to providing a better understanding of the correlation structure, asset-based graphs or networks, and especially the minimum spanning tree, are often considered interesting tools to visualize the paths of shock propagation. Therefore, minimum spanning trees have been built for a wide variety of datasets. We shall not give a list of the studies but they usually concern stocks, foreign exchanges, interest rates in different countries or market indices.

Although we are skeptical in general about the interpretation of the minimum spanning tree as a description of the shortest paths for the propagation of shocks – because propagation has to do with causality whereas correlation only indicates simultaneous moves –, we believe that minimum spanning trees provide an important geometrical representation of the dependence between stocks. Certain stocks, like GE in the US, are always central in minimum spanning trees; other stocks are often on the leaves of the tree and can be considered as those providing maximal diversification. Also, the shape of the tree provides information about the market as a whole (number of clusters for example) and the evolution of the shape of the tree gives information about the evolution of the market state.

The shape of a minimum spanning tree is usually studied using two indicators. In the case of a distance-like matrix, the minimum spanning tree $\mathcal{T}$ is by definition the shortest connected subgraph which connects all the nodes. Hence, the tree length is an important indicator. If we normalize this length, dividing it by the number of edges, we obtain a first popular indicator: the normalized tree length defined as (see *e.g.* [96] or [97]):

$$L = \frac{1}{n-1} \sum_{(i,j)\in\mathcal{T}} d(i,j)$$

This length is a measure of the average correlation of the assets and it is a possible indicator to detect an overall increase in correlation levels, a phenomenon usually occurring during crashes. This indicator is strongly anticorrelated with the average of the empirical correlation coefficients. On one hand, it means that the minimum spanning tree is a good reduced representation of the correlation structure. On the other hand, why bother with this measure if it is only equivalent to the average correlation?

Obviously, the tree contains more information than its length. It can be a chain-like tree or a star-like tree and this dimension is often measured using the notion of mean occupation layer. This indicator is defined relatively to a node $c$ that is considered central[18] by:

[18] Several definitions of centrality have been proposed in graph theory. A central node may correspond to the node with the highest degree, it may also correspond to the center of mass (the node $c$ minimizing mean occupation layer). In practice, when it comes to stock returns, only a few central nodes are possible and

$$l = \frac{1}{n} \sum_{i=1}^{n} \mathrm{lev}(i, c)$$

where $\mathrm{lev}(i, c)$ is the level of $i$ with respect to $c$, that is the number of edges between $i$ and $c$. A low mean occupation layer indicates a star-like tree while a high mean occupation layer indicates a chain-like tree.

This indicator is important because the shape of an asset tree associated to stock returns is usually specific during a crash. This was observed for instance for Black Monday (see [94, 96]) since in that case the tree shrunk and the mean occupation layer dropped, the behaviour of the assets having been homogenous.
The same authors also proposed to compare the asset trees in periods of crashes and normal periods using the distribution of the vertex degrees.

In the definition of mean occupation layer, a central node must be chosen and the graph is usually drawn around this central node. This representation usually provides a natural clustering of the assets, especially when stocks are considered. Unsurprisingly, this clustering often coincides with the natural sectoral clustering of the assets (energy, utilities, technology, health care, finance, ...) or with a mix between geographical and sectoral clustering (see Figure 2). We shall see in the next section that this informal clustering can be made more rigorous using a parallel between the notion of minimum spanning tree which gives a geometrical/topological information and the single linkage cluster analysis which provides a taxonomy of the stocks.

To conclude on minimum spanning trees and other correlation-based graphs, it is important to notice that they provide a simple representation of the correlation structure. The geometry of the graph allows one to graphically define clusters (see next section for formal clustering algorithms). The evolution of this geometry[19] may help to anticipate and/or de-

they usually provide similar values for the mean occupation layer. Notions of centrality are also important in other applications of graph theory to Finance, in order to understand systemic risk and to identify systemic institutions for instance.

[19] The evolution of the edges may also be due to noise. Results on the stability of asset trees are usually

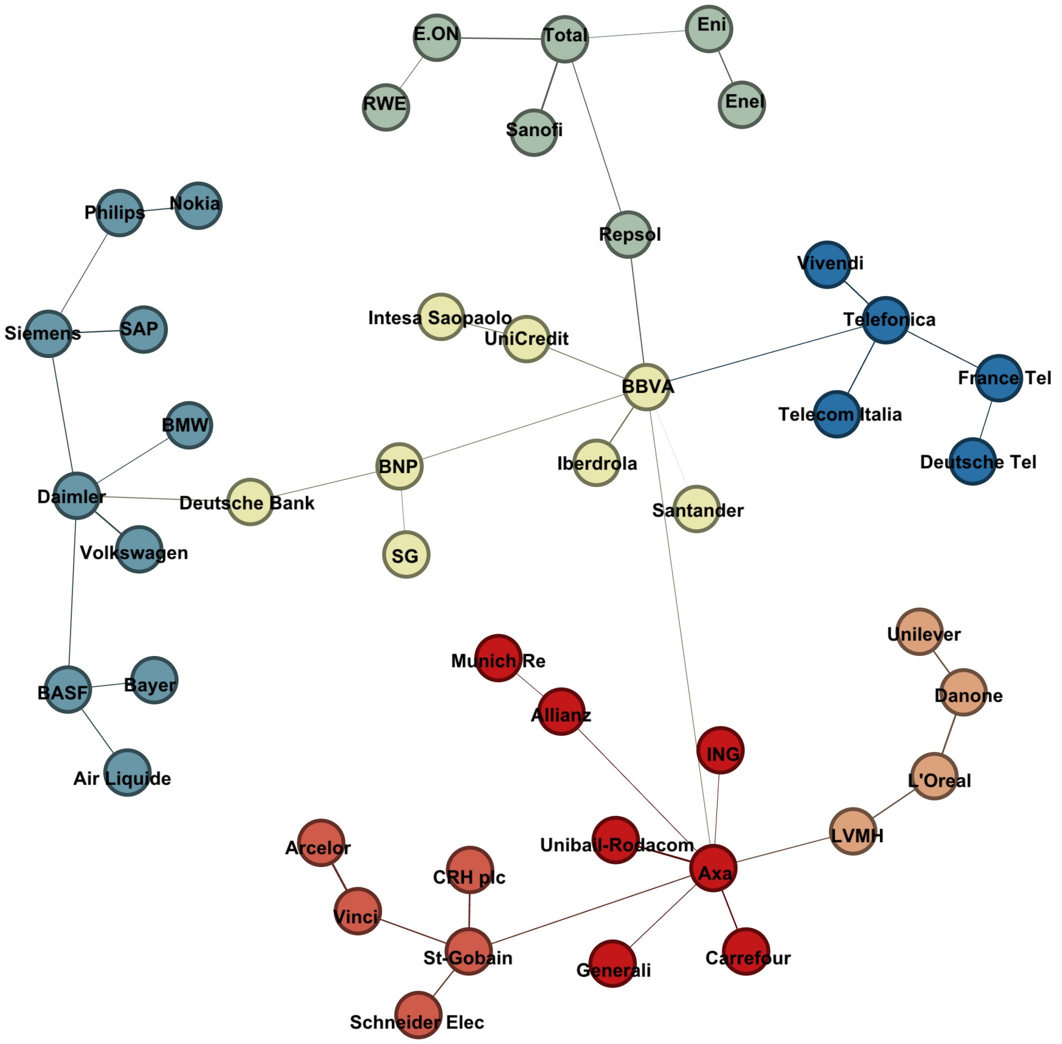

Figure 2: Example of a minimum spanning tree for 46 stocks of the EuroStoxx 50 index (2004-2011)

scribe crashes. However, we believe that the interpretation of correlation-based graphs is often wrong because the graph representation is misleading. Most empirical studies consider that the edges of the minimum spanning tree are transmission channels for shocks. However, this is questionable since minimum spanning trees are based on (static) correlation and not on (dynamic) causality between events.

# 11 Hierarchical structure and clustering

In many fields, complex systems are structured in a nested hierarchical fashion. The usual shape of minimum spanning trees suggests a hierarchical organization of the stocks and we present in this section some approaches to build a hierarchical tree (or dendrogram) from a sample correlation matrix. These approaches are part of the basic toolbox of multivariate data analysis. Their use in finance for clustering purposes is not new but the influence of minimum spanning trees and the link with the first clustering approach we present below have generated a new strand of academic literature. Interestingly, hierarchical clustering methods allow one to clean correlation matrices with a procedure that is different from the procedures described in the preceding part of this document.

## 11.1 Hierarchical clustering algorithms

Hierarchical clustering is a classical topic in multivariate data analysis with much literature associated with it (although it is not recent, a classical reference on clustering[20] is [4]). The approaches we present all share the same principle. They start with a cluster for each of the $n$ assets. Then, the procedure is based on pair grouping and the clusters are iteratively merged into new clusters according to a similarity measure calculated from the sample correlation matrix[21]. If the similarity measure between two assets is naturally the empirical correlation

computed using survival ratios as in [96]. Statistical results to disentangle the part of the evolution due to noise and the part due to actual evolution are not available...

[20] Econophysicists also proposed a clustering method based on Potts model [69].

[21] Interestingly, some authors compared the results obtained with a sample correlation matrix and the results obtained with an already cleaned correlation matrix. They concluded that eigenvalue cleaning removed part of the relevant hierarchical structure.

coefficient, the different methods differ in their definition of the similarity measure between an asset and a cluster of assets or more generally between two clusters. In the first method, called single linkage or single linkage cluster analysis (SLCA), the similarity measure between cluster $C_1$ and cluster $C_2$ is:

$$\rho_{C_1,C_2} = \max\{\rho_{ij} | i \in C_1, j \in C_2\}$$

The second method we present is either referred to as average linkage cluster analysis (ALCA) – see [116], [120] or [122] – or as the weighted pair group method with arithmetic mean (WPGMA) – see [101] –. In this second method, the similarity measure between cluster $C_1$ and cluster $C_2$ is the average of the empirical correlation coefficients:

$$\rho_{C_1,C_2} = \frac{1}{\text{card}(C_1)\text{card}(C_2)} \sum_{i \in C_1, j \in C_2} \rho_{ij}$$

Other possibilities exist and [101] proposed for instance an unweighted pair group method with arithmetic mean or the use of the Haussdorff distance.

A common skeleton for these clustering algorithms is:

1. Initialize a matrix $M = (m_{ij})_{1 \leq i,j \leq n}$ with the values of the sample correlation matrix.

2. Define $n$ singleton clusters.

3. Consider $m_{i^* j^*}$ the largest coefficient of the matrix $M$ for $i^*$ and $j^*$ belonging to different clusters. Denote $C(i^*)$ and $C(j^*)$ these clusters.

4. Link $C(i^*)$ and $C(j^*)$ in the dendrogram[22]

5. Redefine the matrix $M$ by:

$$m_{ij} = m_{i^* j^*} \quad \forall i \in C(i^*), j \in C(j^*) \qquad m_{ij} = m_{ij} \quad \forall i, j \in (C(i^*) \cup C(j^*))^c$$

[22] The dendrogram is usually drawn using heights linked to dissimilarity measures: the transformation of correlation into distances presented above is a good option.

and:

- In the case of single linkage:

$$m_{ij} = \max\{\rho_{kl}/k \in C(i), l \in C(j)\}, \forall i \in C(i^*) \cup C(j^*), \forall j \in (C(i^*) \cup C(j^*))^c$$

$$m_{ij} = \max\{\rho_{kl}/k \in C(i), l \in C(j)\}, \forall i \in (C(i^*) \cup C(j^*))^c, \forall j \in C(i^*) \cup C(j^*)$$

- In the case of average linkage:

$$m_{ij} = \frac{1}{\mathrm{card}(C(i))\mathrm{card}(C(j))} \sum_{k \in C(i), l \in C(j)} \rho_{kl}, \forall i \in C(i^*) \cup C(j^*), \forall j \in (C(i^*) \cup C(j^*))^c$$

$$m_{ij} = \frac{1}{\mathrm{card}(C(i))\mathrm{card}(C(j))} \sum_{k \in C(i), l \in C(j)} \rho_{kl}, \forall i \in (C(i^*) \cup C(j^*))^c, \forall j \in C(i^*) \cup C(j^*)$$

6. Merge the elements of $C(i^*)$ and $C(j^*)$ into a single cluster.

7. If there are still several clusters, go back to step 3.

This algorithm builds a hierarchical tree (see Figure 3). A partition of the assets can then be obtained by removing the branches of the tree when dissimilarity is above a certain threshold (or when the similarity measure is below a certain threshold). In addition to this clustering output, the algorithm can provide two outputs of different natures: a tree and a matrix. The tree may be the minimum spanning tree in the case of single linkage cluster analysis and it is in general a tree with properties similar to the minimum spanning tree. The matrix is the matrix $M$ at the end of the algorithm. With at most $n$ different values for the coefficient, the matrix $M$ is a good candidate to be a cleaned correlation matrix.

To obtain a correlation-based tree associated to the above clustering algorithms, one starts with a graph with no edge. Then, at step 6, in addition to merging the two clusters $C(i^*)$ and $C(j^*)$ into a single cluster, one chooses a node in the cluster $C(i^*)$ and a node in the cluster $C(j^*)$ and adds an edge between these two nodes. At the end of the algorithm, one obtains a tree which depends (i) on the clustering algorithm used and (ii) on the choice of the

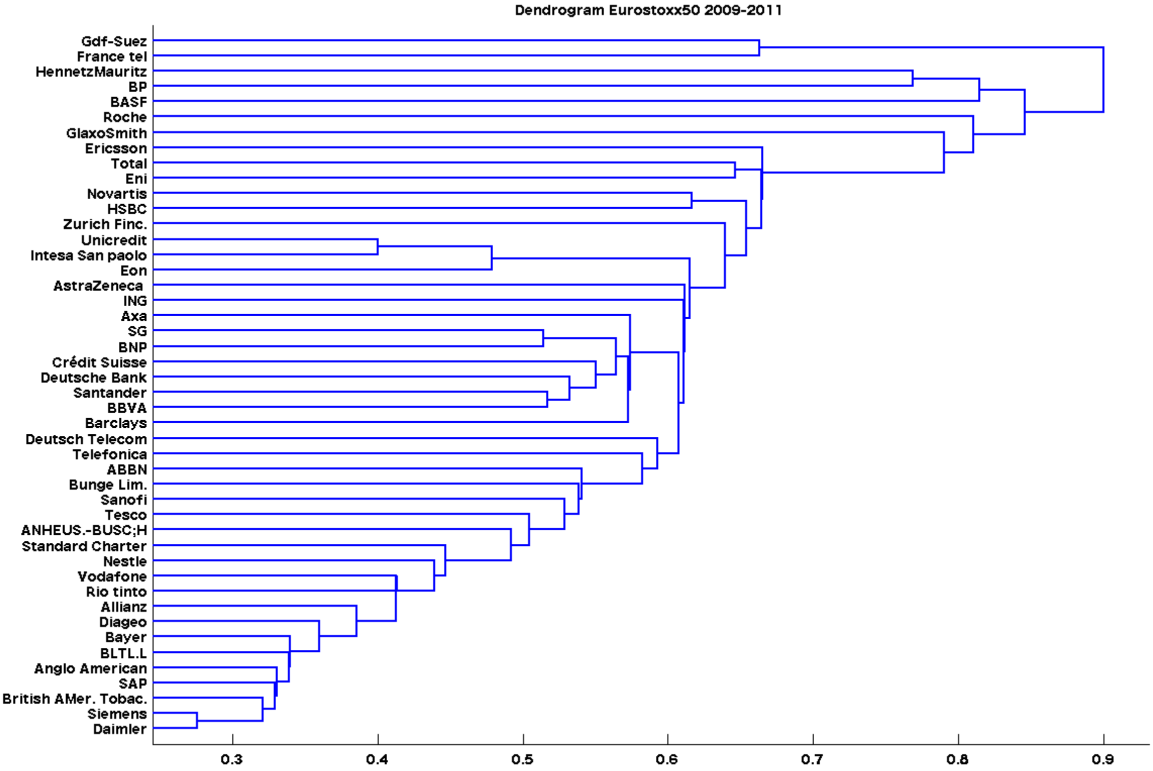

Figure 3: Example of a dendrogram for 46 stocks of the EuroStoxx 50 index (2009-2011). Average linkage with $d(i, j) = \sqrt{2(1 - \rho_{ij})}$

nodes when adding an edge. In the case of single linkage, if we choose to link $i \in C(i^*)$ and $j \in C(j^*)$ so that $\rho_{ij} = \max\{\rho_{kl} | k \in C(i^*), l \in C(j^*)\}$, the tree we obtain is the minimum spanning tree defined above. In turn, if we consider average linkage instead of single linkage, then the tree we obtain is different and called average linkage minimum spanning tree (see [120]).

The link between clustering algorithms and the construction of pruned correlation-based graphs is remarkable. However, hierarchical trees and correlation-based trees do not contain information of the same nature: hierarchical trees provide a taxonomy of the stocks whereas correlation-based trees provide a geometrical representation. In practice, the former defines clusters in a precise way and the latter may help distinguish central nodes from nodes located at the leaves and associated to higher diversification.

## 11.2 New cleaning methods

One of the outputs of the above clustering algorithms is the matrix $M$. The matrices produced by the algorithms contain at most $n$ different values based on the similarity measures between clusters. Specifically, the entry $m_{ij}$ is computed before $i$ and $j$ belong to the same cluster and corresponds to the similarity measure between the two clusters which merge at that step. Hence, the matrix $M$ contains relevant information about the dependence structure and it is regarded as an interesting constrained estimate of the correlation matrix. In particular, in the two cases[23] we presented, the matrix $M$ is a positive definite matrix as soon as the entries of the initial correlation coefficients are positive.

It is then interesting to notice that the matrix $M$ is related to a factor model and results in this direction are reviewed in [122]. However, what is important is to compare the matrix $M$ to other cleaned correlation matrices, especially when it comes to portfolio optimization. [101] and [116] compared the approaches and both articles conclude that the use of the ma-

[23] It is not true in general, see [101].

trices $M$ most often result in higher realized risk than the use of eigenvalue clipping, in the Markowitz optimization framework. However, and specifically in the case of average linkage, the estimation of realized risk is often accurate.

## 12 Concluding remarks

The methods we presented in this part are applications of either graph theory or classical multivariate data analysis tools to finance. The concept of minimum spanning tree is now widely used in finance to obtain a geometrical representation of the correlation structure. It gives meaningful information about the centrality of certain nodes and provides a first estimation of the relevant number of clusters. The graphical tools may complement other decision-making tools and they may also serve to graphically illustrate a point but the common interpretation of the edges as shock transmission channels should be avoided.

Regarding clustering methods, they are useful for many applications. Here, we add that the matrices produced by clustering algorithms should be tested as target matrices in the shrinkage approach. This seems not to have been done in the academic literature.

Finally, for correlation-based graphs and hierarchical trees presented above, the input is the correlation matrix of a set of $n$ assets. Linear correlation coefficients are used in the literature on graphical representation of dependence and clustering. However, the same reasoning can be applied to other dependence measures and it could be interesting to replace linear correlation by either Kendall's $\tau$ or Spearman's $\rho$. This is particularly relevant since the "distances" introduced above, for instance $d(i,j) = \sqrt{2(1-\rho_{ij})}$, are defined as if the upper bound of the correlation coefficient was always 1.

# Part IV

# High-frequency correlation

## 13 Introduction

The proportion of transactions resulting from high-frequency traders has skyrocketed over the last five years, both in Europe and in the US. Subsequently, practitioners express a real need of statistics to estimate high-frequency liquidity, high-frequency volatility and high-frequency dependence measures[24]. Unfortunately, the statistics presented in the first part of this document are not adapted to high-frequency data. Because of market rules and due to the resulting nature of high-frequency data (*e.g.* tick-by-tick data), increasing the sampling frequency also increases the so-called microstructure noise. Usual statistics turn out to be contaminated by the bid-ask bounce and by the impact of the discrete nature of prices (tick size). Moreover, the very definition of the statistics used so far in this document is an issue. We implicitly considered daily observations and as far as stocks were concerned, we considered daily returns, for instance calculated using the closing price of each trading day. Turning to high-frequency, since orders are sent to the market at random times (at least from a simple statistical point of view), statistics measuring dependence must then be suited to unevenly sampled and asynchronous data.

This part will be dedicated to the statistical methods developed to measure covariance or realized quadratic covariation using asynchronous observations potentially contaminated by microstructure noise. As opposed to the methods developed in the two preceding parts, which have been mainly developed by econophysicists, the methods we present now have been proposed by statisticians and econometricians. It is noteworthy that their first interest was in estimating volatility or realized quadratic variation (see for instance [5, 6] or [8, 126] and other

[24] The availability of high-frequency data may have seemed to reduce the curse of dimensionality. As we will see, new problems appear with high frequency data and, furthermore, high-frequency indicators do not account for what happens overnight. For that reason, we believe that high-frequency data should only be used to solve high-frequency issues and that the cleaning techniques presented above in this text remain the only way to deal with the curse of dimensionality.

references therein). It appears that, in that case, the presence of microstructure frictions in the data imposed limitations in the use of the popular realized quadratic variation estimators and that new theoretical apparatus was necessary in order to filter out microstructure noise or, at least, to limit its influence. The optimal sampling methodology or the two-scale estimator we discuss below for quadratic covariation are rooted to this early literature on high-frequency volatility.

If this issue concerning the upward bias of high-frequency volatility empirical estimator (associated to the famous volatility signature plot) has rapidly been understood and successfully tackled, the situation is different as far as dependence measures are concerned. The classical estimator of correlation is known to decrease as the sampling frequency increases. This stylized fact, referred to as Epps effect [47], has been studied from an economic viewpoint and is usually attributed to lead-lag relations between assets[25] and to asynchronicity of observations (see [117, 118]): correlation between stocks virtually disappears as the time interval to compute returns goes to 0 because there is an unavoidable lag between moves on different stocks and because transactions do not occur simultaneously. Interestingly, the correlation-based trees presented in the previous section also witness an Epps-like effect: when the sampling frequency increases, [17] showed that the minimum spanning tree of US stocks evolves from an highly-structured tree to a star-like one.

The simultaneous presence of microstructure noise and asynchronous data makes dependence measures more complicated to estimate than volatility[26]. We will first tackle the questions associated to asynchronicity: what is the bias of the classical quadratic covariation estimator in the case of asynchronous data? How should we sample the data? Can we design new estimators which are robust to asynchronous data?

Then, we shall discuss the issue of microstructure noise and present the estimators proposed to take simultaneously account of the two issues. It is noteworthy that this is still an active

[25] For lead-lag effects, see for instance [54] or [60].

[26] Microstructure would in fact deserve an entire review document and many issues common to all high-frequency estimators are not dealt with in this document. We indeed leave aside the question of the relevant "price" to consider (last price, mid-price, bid price, ask price) or the issue of the tick size. Market microstructure will in fact be modeled through a very simple form of microstructure noise.

strand of academic research and that no “best answer” is available.

# 14 Dealing with asynchronous data

## 14.1 Epps effect

In the above introduction, we discussed asynchronicity as one of the main source of the Epps effect. To understand theoretically the bias of correlation toward 0 due to asynchronicity, let us introduce the model with two stocks we shall use throughout this part. The log-prices of the two stocks are assumed to follow a bivariate brownian motion with variances $\sigma_1^2$ and $\sigma_2^2$ respectively and a correlation coefficient $\rho$. The prices $P_1$ and $P_2$ ($p_1$ and $p_2$ will stand for the log-prices) are recorded at observation times $t_0^1 = 0 \leq t_1^1 \leq \ldots \leq t_{M_1}^1$ for asset 1 and observations times $t_0^2 = 0 \leq t_1^2 \leq \ldots \leq t_{M_2}^2$ for asset 2. These observation times correspond most often to transaction times or best quotes updates.

To define prices at any time, one can use the “previous tick” method: the price at time $t$ is the last price recorded before time $t$. Under this natural and simple hypothesis, we compute the classical estimator of the realized quadratic covariation over the time interval $[0, 1]$, hereafter $RC$:

$$RC = \sum_{i=1}^{N} r_1\left(\frac{i}{N}\right) r_2\left(\frac{i}{N}\right)$$

where $r_1(t) = p_1(t) - p_1(t - \frac{1}{N})$ and $r_2(t) = p_2(t) - p_2(t - \frac{1}{N})$ are log-returns of assets 1 and 2 respectively, and where $N$ is the number of evenly distributed points considered to compute the above estimator.

To better understand the impact of asynchronous data, Griffin and Oomen [54] considered the case of two independent Poisson processes for the observation times and quantified the bias of the classical estimator:

$$\mathbb{E}\left[RC\right] = \rho\sigma_1\sigma_2\left(1 - \frac{N}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}(1-e^{-\frac{\lambda_2}{N}}) + \frac{\lambda_2}{\lambda_1}(1-e^{-\frac{\lambda_1}{N}})\right)\right)$$

where $\lambda_1$ and $\lambda_2$ are the intensities associated to the Poisson processes driving the observation times.

We see that the estimator is biased toward 0 and the Epps effect is well illustrated by the limit $\lim_{N\to\infty}\mathbb{E}\left[RC\right] = 0$ (for covariance or quadratic covariation instead of correlation). This confirms the intuition about the consequences of asynchronous data: because changes in price are not simultaneous, returns appear to be less correlated than they really are (this is especially true when one stock is more liquid than the other one, trading then at a faster pace).

This result raises the following questions:

- Is there a way to define the price process which provides better results than the "previous tick" approach?
- Is there a sampling methodology which reduces the above bias?
- Is there an estimator of quadratic covariation better suited to asynchronous data?

## 14.2 Previous tick and other hypotheses

In the above example, we considered a price process built using the "previous tick" hypothesis. This hypothesis reinforces the effect of asynchronicity since the resulting price processes are piecewise constant. Another hypothesis made in the literature (see [60]) is to consider a set of sampling times through the reunion of the two initial sets of observations times:

$$\{t_0,\dots,t_M\} = \{t_0^1,\dots,t_{M_1}^1\} \cup \{t_0^2,\dots,t_{M_2}^2\}$$

and to linearly interpolate log-prices $p_1$ (resp. $p_2$) at times $(t_k^2)_k$ (resp. $(t_j^1)_j$).
If we consider this method with $0 = t_0^1 \leq \dots \leq t_{M_1}^1 = 1$ and $t_0^2 = 0 \leq \dots \leq t_{M_2}^2 = 1$ two subdivisions of $[0,1]$, denoting $\tilde{p}_1$ and $\tilde{p}_2$ the resulting log-prices defined on the larger

subdivision $0 = t_0 \leq \ldots \leq t_M = 1$, a natural estimator or the quadratic covariation over $[0, 1]$ is:

$$RC' = \sum_{i=1}^{M} (\tilde{p}_1(t_i) - \tilde{p}_1(t_{i-1}))(\tilde{p}_2(t_i) - \tilde{p}_2(t_{i-1}))$$

Because we considered linear interpolation, this estimator can also be written with weights:

$$RC' = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \omega_{jk}(p_1(t_j^1) - p_1(t_{j-1}^1))(p_2(t_k^2) - p_2(t_{k-1}^2))$$

where:

$$w_{jk} = \begin{cases} \left(1 - \frac{t_{k-1}^2 - t_{j-1}^1}{t_j^1 - t_{j-1}^1}\right) \frac{t_j^1 - t_{j-1}^1}{t_k^2 - t_{k-1}^2}, & t_{j-1}^1 \leq t_{k-1}^2 \leq t_j^1 \leq t_k^2 \\ \left(1 - \frac{t_{j-1}^1 - t_{k-1}^2}{t_j^1 - t_{j-1}^1}\right) \frac{t_k^2 - t_{k-1}^2}{t_j^1 - t_{j-1}^1}, & t_{k-1}^2 \leq t_{j-1}^1 \leq t_k^2 \leq t_j^1 \\ \frac{t_k^2 - t_{k-1}^2}{t_j^1 - t_{j-1}^1}, & t_{j-1}^1 \leq t_{k-1}^2 \leq t_k^2 \leq t_j^1 \\ \frac{t_j^1 - t_{j-1}^1}{t_k^2 - t_{k-1}^2}, & t_{k-1}^2 \leq t_{j-1}^1 \leq t_j^1 \leq t_k^2 \\ 0, & [t_{j-1}^1, t_j^1] \cap [t_{k-1}^2, t_k^2] = \emptyset \end{cases}$$

This estimator artificially synchronizes asynchronous data. Unfortunately, it is a biased estimator.

Other approaches consist of withdrawing data points in order to avoid a spurious absence of correlation. Such an approach is referred to as the Refresh Time approach in [10] (see also [1] for a generalized framework) and consists of keeping prices artificially constant until at least one new piece of data for each stock has been obtained.

More precisely, the Refresh Time approach considers new observation times $\tau_0 \leq \ldots \leq \tau_N$ defined by $\tau_0 = \max(t_0^1, t_0^2)$ and:

$$\forall j \geq 0, \quad \tau_{j+1} = \max\left(t_{N_1(\tau_j)+1}^1, t_{N_2(\tau_j)+1}^2\right)$$

where $N_i(t) = \max\{l, t_l^i \leq t\}$

Then, at each newly defined observation time, we consider a price as in the “previous tick” approach.

This approach is interesting and it will be used in the next section to deal with microstructure noise. However, to deal with asynchronous data, a better approach has been introduced by Hayashi and Yoshida [57].

## 14.3 Hayashi-Yoshida estimator

We presented above several methods to artificially create synchronized data from asynchronous data. If these approaches seem natural, Hayashi and Yoshida defined an estimator of quadratic covariation which considers raw asynchronous data. This estimator is defined for $0 = t_0^1 \leq \ldots \leq t_{M_1}^1 = 1$ and $t_0^2 = 0 \leq \ldots \leq t_{M_2}^2 = 1$ two subdivisions of $[0,1]$ as:

$$HY = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} (p_1(t_j^1) - p_1(t_{j-1}^1))(p_2(t_k^2) - p_2(t_{k-1}^2)) 1_{(t_{j-1}^1, t_j^1] \cap (t_{k-1}^2, t_k^2] \neq \emptyset}$$

The most important property of this quadratic covariation estimator, introduced in [57] (see also [32]), is its unbiasedness. Asymptotic normality properties have been proved in [58] and an expression for the variance of the estimator is available in [54] in the case of observation times in $(0,1)$ driven by two independent Poisson processes.

This estimator can be rewritten in another way which aggregates returns as in the Refresh Time case, although differently (see [123]):

$$HY = \sum_{j=1}^{M_1} (p_1(t_j^1) - p_1(t_{j-1}^1)) \left( p_2(\min\{t_k^2, t_k^2 > t_j^1\}) - p_2(\max\{t_k^2, t_k^2 \leq t_{j-1}^1\}) \right)$$

Another aggregation method is presented in [100], which is to write the estimator in another way. Contrary to the above method, it does not always aggregate the returns corresponding to one specific asset (asset 2 in the above case). We shall not detail this aggregation algorithm here but the goal is to obtain a collection $(A_l^1)_{1 \leq l \leq L}$ of subsets of $\{t_1^1, \ldots, t_{M_1}^1\}$ and a collection $(A_l^2)_{1 \leq l \leq L}$ of subsets of $\{t_1^2, \ldots, t_{M_2}^2\}$ so that the Hayashi-Yoshida estimator can be written as:

$$HY = \sum_{l=1}^{L} \overline{r}_l^1 \overline{r}_l^2$$

where $\overline{r}_l^1 = \sum_{j/t_j^1 \in A_l^1} p_1(t_j^1) - p_1(t_{j-1}^1)$ and $\overline{r}_l^2 = \sum_{k/t_k^2 \in A_l^2} p_1(t_k^2) - p_1(t_{k-1}^2)$.

This unbiased estimator can be calculated easily using raw asynchronous data. However, we will show in the next section that it cannot be used in the presence of a microstructure noise. Before turning to this issue of microstructure noise which makes the above estimators inconsistent, we present a last approach to estimate quadratic covariation. This approach, introduced by Malliavin and Mancino in [83], relies on a Fourier analysis of the data and is therefore perfectly suited to asynchronous and unevenly sampled data.

## 14.4 The Fourier approach of Mancino-Malliavin

The above estimators of quadratic covariation are rooted to the classical definition of quadratic covariation. To estimate instantaneous volatility, instantaneous covariance, quadratic variation and quadratic covariation, Malliavin and Mancino considered a completely different route and proposed a Fourier approach. They indeed relate the Fourier coefficients of the log-price processes to the Fourier coefficients of the (now *a priori* non-constant) variance process. Therefore, quadratic covariation can be approximated using approximations of the Fourier coefficients of the log-prices.

We start with a log-price process $p = (p_1, p_2)$ following the dynamics $dp(t) = \sigma(t)dW(t)$ where $\sigma(t)$ is a $2 \times 2$ matrix and $W$ is a bidimensional brownian motion[27]. Then, the instantaneous covariance matrix is $\Sigma(t) = \sigma(t)\sigma'(t)$ and the quadratic covariation over the time interval under consideration, supposed to be $[0, 2\pi]$ to simplify, is:

[27] In the case of constant correlation and constant volatilities, the matrix $\sigma$ is simply

$$\begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{1-\rho^2}\sigma_2 \end{pmatrix}$$

.

$$\int_0^{2\pi} \Sigma_{12}(t)dt$$

Malliavin and Mancino [83] (see also [84]) expressed the Fourier coefficients of $\Sigma$ in terms of the Fourier coefficients of the log-price process $p$ (rescaled to be defined on $[0, 2\pi]$). Here, we shall only focus on quadratic covariation over the entire time window, *i.e.*:

$$\int_0^{2\pi} \Sigma_{12}(t)dt = 2\pi a_0(\Sigma_{12})$$

The central result is then:

$$a_0(\Sigma_{12}) = \lim_{N\to+\infty} \frac{\pi}{2N} \sum_{n=1}^{N} a_n(dp_1)a_n(dp_2) + b_n(dp_1)b_n(dp_2)$$

where for $i \in \{1, 2\}$:

$$a_n(dp_i) = \frac{1}{\pi} \int_0^{2\pi} cos(nt)dp_i(t) = \frac{p_i(2\pi) - p_i(0)}{\pi} + \frac{n}{\pi} \int_0^{2\pi} sin(nt)p_i(t)dt$$

$$b_n(dp_i) = \frac{1}{\pi} \int_0^{2\pi} sin(nt)dp_i(t) = -\frac{n}{\pi} \int_0^{2\pi} cos(nt)p_i(t)dt$$

These Fourier coefficients can easily be estimated using either the “previous tick” hypothesis or a linear interpolation as above. One then obtains numerical approximations $\widehat{a_n}(dp_i)$ and $\widehat{b_n}(dp_i)$ of the Fourier coefficients and the resulting estimator for the quadratic covariation over $[0, 2\pi]$ is:

$$\frac{\pi^2}{N} \sum_{n=1}^{N} \widehat{a_n}(dp_1)\widehat{a_n}(dp_2) + \widehat{b_n}(dp_1)\widehat{b_n}(dp_2)$$

where $N$ needs to be sufficiently large for the approximation to be accurate (the choice of $N$ is not discussed by the authors although it influences the accuracy of the approximation).

This Fourier approach is completely different from the above Hayashi-Yoshida estimator. In particular, the consequence of the integrations by parts carried out in the computation of the Fourier coefficients is that the above Malliavin-Mancino estimator does not rely on any return computation, whereas the Hayashi-Yoshida estimator is computed using differences of log-prices. Independent academic comparisons between the two estimators are still needed to clarify the superiority of one or the other although [60], using Monte-Carlo simulations, supports the use of the Hayashi-Yoshida estimator.

## 15 Microstructure noise and statistical tools

The above estimators have been developed to deal with asynchronous data. When using either Hayashi-Yoshida estimator or Malliavin-Mancino estimator, one does not indeed to artificially synchronize the data beforehand. However, one has to deal with a second issue concerning high-frequency data: the influence of microstructure. High-frequency data is subject to the influence of the tick size and to the subsequent discreteness of prices. The bid-ask bounce also adds an artificial noise to the data if we consider the last price as our reference price. These microstructure effects can be studied independently of one another but, for statistical purposes, the influence of microstructure is often modeled in a simple way, through a noise contaminating the price processes we observe.

To take account of microstructure, most models consider an observed log-price $p_i$ for asset $i$ which is the sum of the real underlying (unobserved) log-price $p_i^*$ and of an independent noise $\epsilon_i$ (the microstructure noises associated to the different assets being *a priori* independent). The influence of microstructure noise was first studied in the univariate context of volatility estimation because of the famous volatility signature plot (see [89] for a review on high-frequency volatility). For covariance or quadratic covariation it has been shown (see for instance [54] and [100]) that microstructure noise makes the classical RC estimator and the Hayashi-Yoshida estimator inconsistent, the variance of both estimators increases to infinity as the number of observations goes to infinity. However, contrary to what happens with volatility, no bias is added by an *i.i.d.* microstructure noise in the case of quadratic covariation.

## 15.1 Optimal sampling, subsampling and the two-scale method

Since sampling at ultra-high-frequency increases the variance of estimators, it may not be optimal to use all available data. There is indeed a trade-off between the use of additional data that reduces mean squared error in the absence of noise and the accumulation of noise resulting from their use. Bandi and Russell [7] provide an optimal sampling frequency for the RC estimator that minimizes mean squared error. However, we have seen that this estimator is biased and that a better estimator was introduced by Hayashi and Yoshida.

We know that Hayashi-Yoshida estimator of the quadratic covariation over $[0,1]$ is unbiased in the presence of independent microstructure noises and its variance, in the case of observation times following independent Poisson processes of intensity $\lambda_1$ and $\lambda_2$, is:

$$2\sigma_1^2\sigma_2^2\left(\frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2}+\frac{\rho^2}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}+\frac{\lambda_2}{\lambda_1}\right)\right)+2\sigma_1^2\mathbb{V}(\epsilon_1)+2\sigma_2^2\mathbb{V}(\epsilon_2)+4\mathbb{V}(\epsilon_1)\mathbb{V}(\epsilon_2)\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}$$

We can consider a Hayashi-Yoshida estimator based on a reduced set of observations. If we keep a proportion $p$ of the observations for each asset, the variance (or equivalently the mean squared error of the estimator) is, within the above Poissonian framework:

$$2\sigma_1^2\sigma_2^2\left(\frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2}+\frac{\rho^2}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}+\frac{\lambda_2}{\lambda_1}\right)\right)\frac{1}{p}+2\sigma_1^2\mathbb{V}(\epsilon_1)+2\sigma_2^2\mathbb{V}(\epsilon_2)+4\mathbb{V}(\epsilon_1)\mathbb{V}(\epsilon_2)\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}p$$

Therefore, the proportion $p^*$ which minimizes mean squared error of the estimator is given by:

$$p^*=\sigma_1\sigma_2\sqrt{\frac{(1+\rho^2)(\lambda_1^2+\lambda_2^2)+2\lambda_1\lambda_2}{2\lambda_1^2\lambda_2^2\mathbb{V}(\epsilon_1)\mathbb{V}(\epsilon_2)}}$$

This result is interesting but optimally sampling the data requires estimates about the variance of the noise.

Other methods have been developed which do not require any knowledge about the noise. The first one is referred to as subsampling. For the classical estimator that uses evenly sampled data indexed from 1 to $N$, the subsampling approach considers $K$ different estimators indexed by $i \in \{1, \dots, K\}$, the $i^{th}$ estimator being the classical RC estimator computed with sampling times $i, i+K, i+2K, \dots, i+(\lfloor N/K \rfloor - 1)\, K$. Then, these $K$ estimators are averaged to obtain a new estimator of quadratic covariation[28] with less influence of microstructure noise because of the averaging:

$$RC^K = \frac{1}{K} \sum_{i=1}^{K} \sum_{l=1}^{\lfloor N/K \rfloor - 1} (p_1(i+lK) - p_1(i+(l-1)K))(p_2(i+lK) - p_2(i+(l-1)K))$$

When it comes to the Hayashi-Yoshida estimator, aggregation techniques such as those presented above can be used to apply the same techniques since the estimator $HY$ can be written as $\sum_{l=1}^{L} \overline{r}_l^1 \overline{r}_l^2$ (see [100], [123] and subsection 14.3 above). However, the computation of the scaling factors to be applied to each estimator is not well defined and highly arguable.

In fact this subsampling approach is more adapted to the usual RC estimator and it serves as a basis to build a second estimator which is consistent in the presence of both microstructure noise and asynchronous data. This estimator is based on a two-scale approach. The idea, proposed by Zhang [125] and following a methodology developed for quadratic variation by Zhang and coauthors (see also [92] for a similar approach), is to consider two estimators $RC^J$ and $RC^K$ with different sampling frequencies and to eliminate part of the noise using a linear combination of these two estimators.

Zhang proposed to choose $N$ and $M_1 + M_2$ of the same order and to consider, for $1 \leq J \ll K = O\left(N^{\frac{2}{3}}\right)$ the following two-scale estimator:

$$RC^K - \frac{N-K+1}{N-J+1} \frac{J}{K} RC^J$$

[28] A multiplicative factor is necessary to scale the estimation to the inverval $[0, 1]$. This multiplicative factor, which is voluntarily omitted in the formula for $RC^K$, is $\frac{N}{(\lfloor N/K \rfloor - 1)K}$

Under mild hypotheses, this estimator is consistent, in the presence of both microstructure noise and asynchronous data. Moreover, Zhang proved that the rate of convergence is $\mathcal{O}(N^{-\frac{1}{6}})$.

In practice this estimator is often computed with $J = 1$ so that $RC^J$ reduces to $RC$. We will present below other consistent estimators with better rates of convergence. As above, the next one does not rely on the Hayashi-Yoshida estimator although it is robust to the presence of asynchronous data.

## 15.2 Kernel estimator

The above estimator was based on the "previous tick" hypothesis. We will now introduce an estimator which uses the Refresh Time methodology to synchronize the data. Let us recall that, in the Refresh Time approach, artificial observation times $0 \leq \tau_1 \leq \ldots \leq \tau_N \leq 1$ are considered so that at least one new piece of data is observed for each stock between two successive times.

Then, to eliminate problems near the boundary, [10] proposes to define new log-prices $\tilde{p_1^0}, \ldots, \tilde{p_1^n}$ and $\tilde{p_2^0}, \ldots, \tilde{p_2^n}$ by:

$$\tilde{p_1^i} = p_1(\tau_{i+m}) \qquad \tilde{p_2^i} = p_2(\tau_{i+m}), \qquad \forall i \in \{1, \ldots, n\}$$

and

$$\tilde{p_1^0} = \frac{1}{m}\sum_{i=1}^{m} p_1(\tau_i) \qquad \tilde{p_2^0} = \frac{1}{m}\sum_{i=1}^{m} p_2(\tau_i)$$

$$\tilde{p_1^n} = \frac{1}{m}\sum_{i=1}^{m} p_1(\tau_{n-m+i}) \qquad \tilde{p_2^n} = \frac{1}{m}\sum_{i=1}^{m} p_2(\tau_{n-m+i})$$

where $m$ and $n$ are two integers ($m$ being small) so that $n + 2m - 1 = N$.
From these prices, we build returns for asset 1 and asset 2:

$$\tilde{r_1^i} = \tilde{p_1^i} - \tilde{p_1^{i-1}} \qquad \tilde{r_2^i} = \tilde{p_2^i} - \tilde{p_2^{i-1}}$$

The kernel estimator of quadratic covariation proposed in [10] is then:

$$K = \sum_{l=0}^{n} g\left(\frac{l}{L+1}\right) \sum_{i=l+1}^{n} \tilde{r}_1^i \tilde{r}_2^{i-l} + \sum_{l=1}^{n} g\left(-\frac{l}{L+1}\right) \sum_{i=l+1}^{n} \tilde{r}_1^{i-l} \tilde{r}_2^i$$

where $g$ is a smooth kernel function with $g(0) = 1$ and $g'(0) = 0$, and where $L$ is the bandwidth of the kernel estimator.

Under mild hypotheses (which allow for a very general form of microstructure noise), one can prove that for a bandwidth parameter $L \propto n^\eta$ ($\eta \in (0.5, 1)$), the kernel estimator is consistent. Moreover, the optimal bandwidth is known to be $L \propto n^{\frac{3}{5}}$ and in that case the rate of convergence of the kernel estimator is[29] $\mathcal{O}(n^{-\frac{1}{5}})$.

This estimator has a better rate of convergence than the preceding one and it can be used in practice, as exemplified in [10]. It is noteworthy that the convergence result holds for a general form of microstructure noise and that the optimal bandwidth can easily be estimated (see [10]). However, the rate of convergence is not the best possible and we present in the next section a better estimator.

## 15.3 Pre-averaging

The last approach we present resembles the preceding one but a kernel is used to pre-average the data.

In the case of evenly sampled data at times $0, \frac{1}{N}, \ldots, 1$, this approach, proposed in [31], starts with a pre-averaging of the data and defines (using the "previous tick" hypothesis):

$$\overline{r_i^1} = \sum_{k=1}^{k_N - 1} g\left(\frac{k}{k_N}\right) \left(p_1\left(\frac{i+k}{N}\right) - p_1\left(\frac{i+k-1}{N}\right)\right)$$

$$\overline{r_i^2} = \sum_{k=1}^{k_N - 1} g\left(\frac{k}{k_N}\right) \left(p_2\left(\frac{i+k}{N}\right) - p_2\left(\frac{i+k-1}{N}\right)\right)$$

[29] We have to impose that $m \gg n^{\frac{1}{5}}$

where $k_N$ is a bandwidth parameter that may be chosen of the form[30] $k_N = \lfloor \theta\sqrt{N} \rfloor$, where $g$ is a positive function with $g(0) = g(1) = 0$ ($g(x) = \max(x, 1-x)$ is a good example). Then, a natural estimator is:

$$RC^{avg} = \frac{N}{N - k_N + 2} \frac{\sum_{k=0}^{N-k_N+1} \overline{r_k^1 r_k^2}}{\sum_{k=1}^{k_N-1} g\left(\frac{k}{k_N}\right)^2}$$

If, as above, the microstructure noises $\epsilon_1$ and $\epsilon_2$ are independent, then $RC^{avg}$ is a consistent estimator of quadratic covariation over $[0, 1]$ and the rate of convergence is $\mathcal{O}(N^{-\frac{1}{4}})$. It is also noteworthy that the formula can be generalized to the case of dependent microstructure noises: in that case, the above estimator is biased but the bias can be corrected (see [31]).

Coming now to the case of asynchronous data, the same approach can be used to build a pre-averaged Hayashi-Yoshida estimator. If we consider a set of observation times for each stock, namely $\{t_0^1 = 0, t_1^1, \ldots, t_{M_1}^1 = 1\}$ and $\{t_0^2 = 0, t_1^2, \ldots, t_{M_2}^2 = 1\}$, then one can define pre-averaged returns by:

$$\overline{r_{t_j^1}^1} = \sum_{l=1}^{l_M - 1} g\left(\frac{l}{l_M}\right) \left(p_1\left(t_{j+l}^1\right) - p_1\left(t_{j+l-1}^1\right)\right)$$

$$\overline{r_{t_k^2}^2} = \sum_{l=1}^{l_M - 1} g\left(\frac{l}{l_M}\right) \left(p_2\left(t_{k+l}^2\right) - p_2\left(t_{k+l-1}^2\right)\right)$$

where $M = M_1 + M_2$ and $l_M = \lfloor \theta\sqrt{M} \rfloor$.

Then, under technical hypotheses (implying in particular that $M$, $M_1$ and $M_2$ are of the same order), the pre-averaged Hayashi-Yoshida estimator defined by

$$HY^{avg} = \frac{1}{l_M^2 \left(\int_0^1 g(x)dx\right)^2} \sum_{j=0}^{M_1 - l_M + 1} \sum_{k=0}^{M_2 - l_M + 1} \overline{r_{t_j^1}^1 r_{t_k^2}^2} 1_{(t_j^1, t_{j+l_M}^1] \cap (t_k^2, t_{k+l_M}^2] \neq \emptyset}$$

is a consistent estimator of the quadratic covariation over $[0, 1]$ and the rate of convergence is $\mathcal{O}(N^{-\frac{1}{4}})$.

[30] See [31] for the choice of $\theta$.

To conclude, pre-averaging reduces the impact of microstructure noise and an augmented version of the Hayashi-Yoshida estimator is proposed using pre-averaging. This estimator can be directly implemented on asynchronous data contaminated by microstructure noise and it is consistent with a better rate of convergence than the above estimates. Nevertheless, empirical studies are still needed to determine the respective advantages and drawbacks of the different estimators. In view of the available results, we recommend the use of this pre-averaging Hayashi-Yoshida estimator.

# Conclusion

The goal of this document was to present the contributions of academic research to the central financial issue of correlation measurement. Several routes and viewpoints have been considered, starting from the definition of dependence measures and ending with the recent contributions of academics to the estimation of high-frequency correlation. Several scientific fields have contributed to this body of knowledge.

The definitions of copulas and concordance measures are rooted in the work of statisticians in the $20^{th}$ century. The notions they defined and the tools they crafted turned out to be of utmost importance half a century after their first introduction in a pre-computer world. Moreover, we believe that non-linear correlations like Spearman's $\rho$ or Kendall's $\tau$ have not yet had the deserved popularity amongst practitioners, mainly because the tools applied to fat-tailed financial time series are still those inherited from the gaussian world.

Following the statisticians, physicists and econophysicists solved one of the main issues when applying Markowitz portfolio theory in practice: denoising correlation matrices. Statisticians' contributions have been presented, but the significant breakthrough was made by Bouchaud and his physicists coauthors who introduced eigenvalue cleaning using mathematical results of random matrix theory.

Econophysicists also contributed to new representations of correlations in the form of graphs or trees. These new representations are useful to visualize a dependence structure. Also, graphical representations of the dependence between stocks put forward old tools from multivariate data analysis such as clustering techniques based on correlation matrices and we believe that these tools should be used instead of *a priori* clustering.

Finally, statisticians and econometricians recently contributed to the new strand of research surrounding the estimation of high-frequency correlation. We presented the quadratic covariation estimators they introduced and discussed the issues associated to microstructure and to the asynchronicity of data. The pre-averaged Hayashi-Yoshida estimator is, to our knowledge, the best estimator introduced so far. It is noteworthy that research is active in this direction because of the increasingly important role played by high-frequency traders.

In summary, we believe that:

- The drawbacks of linear correlation and especially the existence of bounds different from $-1$ and 1 should be common knowledge.
- Non-linear correlation measures should be used more often by practitioners, at least as descriptive statistics.
- Asset managers' awareness must increase regarding the curse of dimensionality surrounding the estimation of large correlation matrices. The use of commercial models for correlation is to some extent an issue since eigenvalue cleaning should be a more popular practice than (commercial) factor models.
- Techniques inherited from graph theory should be used moderately because the outcomes are often misinterpreted.
- Correlation-based clustering techniques should be used instead of *a priori* clustering.
- Classical covariance/correlation estimators should not be used on high-frequency data. Several alternative estimators have been proposed and the most recent ones, which can be used with raw asynchronous and unevenly sampled, data filter out the effects of microstructure.

# References

[1] Y. Aït-Sahalia, J. Fan, and D. Xiu. High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.

[2] R. Allez and J.P. Bouchaud. Eigenvector dynamics: theory and some applications. *Arxiv preprint arXiv:1108.4258*, 2011.

[3] R. Allez and J.P. Bouchaud. Eigenvector dynamics: general theory and some applications. *Arxiv preprint arXiv:1203.6228*, 2012.

[4] M.R. Anderberg. Cluster analysis for applications. Technical report, DTIC Document, 1973.

[5] T. Andersen, T. Bollerslev, F.X. Diebold, and P. Labys. The distribution of exchange rate volatility. Technical report, National Bureau of Economic Research, 1999.

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

[7] F.M. Bandi and J.R. Russell. Realized covariation, realized beta and microstructure noise. *Unpublished paper, Graduate School of Business, University of Chicago*, 2005.

[8] F.M. Bandi and J.R. Russell. Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3):655–692, 2006.

[9] F.M. Bandi, J.R. Russell, and Y. Zhu. Using high-frequency data in dynamic portfolio choice. *Econometric Reviews*, 27(1-3):163–198, 2008.

[10] O.E. Barndorff-Nielsen, P.R. Hansen, A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 2011.

[11] O.E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925, 2004.

[12] D. Bartz, K. Hatrick, C.W. Hesse, K.R. Müller, and S. Lemm. Directional variance adjustment: a novel covariance estimator for high dimensional portfolio optimization. *Arxiv preprint arXiv:1109.3069*, 2011.

[13] S. Belinschi, A. Dembo, and A. Guionnet. Spectral measure of heavy tailed band and covariance random matrices. *Communications in Mathematical Physics*, 289(3):1023–1055, 2009.

[14] L. Bergomi. Correlations in asynchronous markets. *Social Science Research Network*, 2010.

[15] P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[16] G. Bonanno, G. Caldarelli, F. Lillo, and R.N. Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68(4):046130, 2003.

[17] G. Bonanno, G. Caldarelli, F. Lillo, S. Miccichè, N. Vandewalle, and R.N. Mantegna. Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):363–371, 2004.

[18] G. Bonanno, F. Lillo, and R.N. Mantegna. High-frequency cross-correlation in a set of stocks. 2001.

[19] G. Bonanno, N. Vandewalle, and R.N. Mantegna. Taxonomy of stock market indices. *Physical Review E*, 62(6):7615–7618, 2000.

[20] J.P. Bouchaud, L. Laloux, M.A. Miceli, and M. Potters. Large dimension forecasting models and random singular value spectra. *The European Physical Journal B-Condensed Matter and Complex Systems*, 55(2):201–207, 2007.

[21] J.P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management.* Cambridge Univ Pr, 2003.

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

[23] Z. Burda, A. Görlich, A. Jarosz, and J. Jurkiewicz. Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications*, 343:295–310, 2004.

[24] Z. Burda, A.T. Görlich, and B. Wacław. Spectral properties of empirical covariance matrices for data with power-law tails. *Physical Review E*, 74(4):041129, 2006.

[25] Z. Burda and J. Jurkiewicz. Signal and noise in financial correlation matrices. *Physica A: Statistical Mechanics and its Applications*, 344(1):67–72, 2004.

[26] A. Chakraborti. An outlook on correlations in stock prices. *Econophysics of Stock and other Markets*, pages 13–23, 2006.

[27] L.K.C. Chan, J. Karceski, and J. Lakonishok. On portfolio optimization: Forecasting covariances and choosing the risk model. Technical report, National Bureau of Economic Research, 1999.

[28] P.Y. Chen and P.M. Popovich. *Correlation: Parametric and nonparametric measures*, volume 139. Sage Publications, 2002.

[29] U. Cherubini and E. Luciano. Bivariate option pricing with copulas. *Applied Mathematical Finance*, 9(2):69–85, 2002.

[30] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*, volume 42. John Wiley & Sons Chichester, 2004.

[31] K. Christensen, S. Kinnebrock, and M. Podolskij. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133, 2010.

[32] F. Corsi and F. Audrino. *Realized correlation tick-by-tick*. Departement of Economics, University of St. Gallen, 2007.

[33] J. Daly, M. Crane, and H.J. Ruskin. Random matrix theory filters in portfolio optimisation: a stability and risk assessment. *Physica A: Statistical Mechanics and its Applications*, 387(16):4248–4260, 2008.

[34] F. De Jong and T. Nijman. High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2):259–277, 1997.

[35] M. De Pooter, M. Martens, and D. Van Dijk. Predicting the daily covariance matrix for s&p 100 stocks using intraday data – but which frequency to use? *Econometric Reviews*, 27(1-3):199–229, 2008.

[36] M. Denuit and A. Charpentier. Mathématiques de l'assurance non-vie. tome i: Principes fondamentaux de théorie du risque. 2004.

[37] D.K. Dey and C. Srinivasan. Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, pages 1581–1591, 1985.

[38] J. Dhaene, M. Denuit, M.J. Goovaerts, R. Kaas, and D. Vyncke. The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33, 2002.

[39] D. Disatnik and S. Katz. Portfolio optimization using a block structure for the covariance matrix. 2011.

[40] D.J. Disatnik and S. Benninga. Shrinking the covariance matrix. *The Journal of Portfolio Management*, 33(4):55–63, 2007.

[41] B. Efron and C. Morris. Stein's paradox in statistics. 1977.

[42] N. El Karoui. Tracy–widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, 2007.

[43] N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

[44] N. El Karoui. High-dimensionality effects in the markowitz problem and other quadratic programs with linear equality constraints: risk underestimation. *Preprint*, 2009.

[45] P. Embrechts, A. McNeil, and D. Straumann. Correlation: pitfalls and alternatives. *Risk*, 12:69–71, 1999.

[46] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223, 2002.

[47] T.W. Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, pages 291–298, 1979.

[48] E.F. Fama and K.R. French. The cross-section of expected stock returns. *Journal of finance*, pages 427–465, 1992.

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

[50] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[51] R.A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[52] P.A. Frost and J.E. Savarino. An empirical bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis*, 21(3):293–305, 1986.

[53] L. Giada and M. Marsili. Data clustering and noise undressing of correlation matrices. *Physical Review E*, 63(6):061101, 2001.

[54] J. Griffin and R. Oomen. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. 2006.

[55] T. Guhr and B. Kälber. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General*, 36:3009, 2003.

[56] LR Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 8(3):586–597, 1980.

[57] T. Hayashi and N. Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379, 2005.

[58] T. Hayashi and N. Yoshida. Asymptotic normality of a covariance estimator for non-synchronously observed diffusion processes. *Annals of the Institute of Statistical Mathematics*, 60(2):367–406, 2008.

[59] T. Heimo, K. Kaski, and J. Saramäki. Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks. *Physica A: Statistical Mechanics and its Applications*, 388(2-3):145–156, 2009.

[60] T. Hoshikawa, K. Nagai, T. Kanatani, and Y. Nishiyama. Nonparametric estimation methods of integrated multivariate volatilities. *Econometric Reviews*, 27(1-3):112–138, 2008.

[61] J.Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[62] P. Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(03):279–292, 1986.

[63] N.E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.

[64] H.M. Kat. The dangers of using correlation to measure dependence. *The Journal of Alternative Investments*, 6(2):54–58, 2003.

[65] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[66] G. Kimeldorf and A.R. Sampson. Monotone dependence. *The Annals of Statistics*, pages 895–903, 1978.

[67] WJ Krzanowski. Sensitivity of principal components. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 558–563, 1984.

[68] L. Kullmann, J. Kertesz, and K. Kaski. Time-dependent cross-correlations between different stock returns: A directed network of influence. *Physical Review E*, 66(2):026125, 2002.

[69] L. Kullmann, J. Kertesz, and RN Mantegna. Identification of clusters of companies in stock indices via potts super-paramagnetic transitions. *Physica A: Statistical Mechanics and its Applications*, 287(3):412–419, 2000.

[70] L. Laloux, P. Cizeau, J.P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, 1999.

[71] L. Laloux, P. Cizeau, M. Potters, and J.P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–398, 2000.

[72] O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. 2003.

[73] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

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

[75] J.H. Lee, D. Stefek, R. Xu, and J. Yao. Mitigating risk forecast biases of optimized portfolios. 2011.

[76] Y. Lee. Noise detection from financial correlation matrices. *Academic Report, Massachusetts Institute of Technology, US*, 2001.

[77] E.L. Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, pages 1137–1153, 1966.

[78] E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

[79] F. Lindskog. Linear correlation estimation. *Preprint, ETH Zürich*, 2000.

[80] F. Lindskog, A. Mcneil, and U. Schmock. Kendall's tau for elliptical distributions. *Credit risk: Measurement, evaluation and management*, pages 149–156, 2003.

[81] R. Litterman and K. Winkelmann. Estimating covariance matrices. *Risk Management Series, Goldman Sachs*, 2, 1998.

[82] A.W. Lo and A. Craig MacKinlay. An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1-2):181–211, 1990.

[83] P. Malliavin and M.E. Mancino. Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics*, 6(1):49–61, 2002.

[84] P. Malliavin and M.E. Mancino. A fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics*, 37(4):1983–2010, 2009.

[85] R.N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[86] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967.

[87] H. Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[88] M. Martens. Estimating unbiased and precise realized covariances. In *EFA 2004 Maastricht Meetings Paper No. 4299*, 2004.

[89] M. McAleer and M.C. Medeiros. Realized volatility: A review. *Econometric Reviews*, 27(1-3):10–45, 2008.

[90] S. Miccichè, G. Bonanno, F. Lillo, and R. N Mantegna. Degree stability of a minimum spanning tree of price return and volatility. *Physica A: Statistical Mechanics and its Applications*, 324(1):66–73, 2003.

[91] R.O. Michaud. The markowitz optimization enigma: is "optimized" optimal? *Financial Analysts Journal*, pages 31–42, 1989.

[92] I. Nolte and V. Voev. *Estimating high-frequency based (co-)variances: A unified approach.* Bibliothek der Universität Konstanz, 2007.

[93] I. Olkin and J.W. Pratt. Unbiased estimation of certain correlation coefficients. *The annals of mathematical statistics*, pages 201–211, 1958.

[94] J.P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz. Dynamic asset trees and black monday. *Physica A: Statistical Mechanics and its Applications*, 324(1):247–252, 2003.

[95] J.P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto. Asset trees and asset graphs in financial markets. *Physica Scripta*, 2003:48, 2003.

[96] J.P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5):056110, 2003.

[97] J.P. Onnela, A. Chakraborti, K. Kaski, and J. Kertiész. Dynamic asset trees and portfolio analysis. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(3):285–288, 2002.

[98] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization ii. *Physica A: Statistical Mechanics and its Applications*, 319:487–494, 2003.

[99] S. Pafka, M. Potters, and I. Kondor. Exponential weighting and random-matrix-theory-based filtering of financial covariance matrices for portfolio optimization. *Arxiv preprint cond-mat/0402573*, 2004.

[100] A. Palandri. Consistent realized covariance for asynchronous observations contaminated by market microstructure noise. *Unpublished manuscript*, 2006.

[101] E. Pantaleo, M. Tumminello, F. Lillo, and R.N. Mantegna. When do improved covariance matrix estimators enhance portfolio optimization? an empirical comparative study of nine estimators. *Quantitative Finance*, 11(7):1067–1080, 2011.

[102] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, and H.E. Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[103] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and HE Stanley. A random matrix theory approach to financial cross-correlations. *Physica A: Statistical Mechanics and its Applications*, 287(3):374–382, 2000.

[104] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A. Nunes Amaral, and H.E. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471–1474, 1999.

[105] M. Potters, J.P. Bouchaud, and L. Laloux. Financial applications of random matrix theory: old laces and new pieces. *Arxiv preprint physics/0507111*, 2005.

[106] H. Qi and D. Sun. A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications*, 28(2):360, 2006.

[107] R. Rebonato and P. Jackel. The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. *Journal of Risk*, 2:17–28, 2000.

[108] R. Renò. A closer look at the epps effect. 2002.

[109] M. Scarsini. On measures of concordance. *Stochastica: revista de matemática pura y aplicada*, 8(3):201–218, 1984.

[110] J. Schäfer, K. Strimmer, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

[111] S. Sharifi, M. Crane, A. Shamaie, and H. Ruskin. Random matrix theory for portfolio optimization: a stability approach. *Physica A: Statistical Mechanics and its Applications*, 335(3):629–643, 2004.

[112] K. Sheppard. Realized covariance and scrambling. *Unpublished manuscript*, 2006.

[113] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):11, 1959.

[114] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[115] J. Svensson. The asymptotic spectrum of the ewma covariance estimator. *Physica A: Statistical Mechanics and its Applications*, 385(2):621–630, 2007.

[116] V. Tola, F. Lillo, M. Gallegati, and R.N. Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

[117] B. Toth and J. Kertesz. On the origin of the epps effect. *Physica A: Statistical Mechanics and its Applications*, 383(1):54–58, 2007.

[118] B. Toth and J. Kertesz. The epps effect revisited. *Quantitative Finance*, 9(7):793–802, 2009.

[119] M. Tumminello, T. Aste, T. Di Matteo, and RN Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421, 2005.

[120] M. Tumminello, C. Coronnello, F. Lillo, S. Miccichè, and R.N. Mantegna. Spanning trees and bootstrap reliability estimation in correlation based networks. *Arxiv preprint physics/0605116*, 2006.

[121] M. Tumminello, F. Lillo, and R.N. Mantegna. Shrinkage and spectral filtering of correlation matrices: a comparison via the kullback-leibler distance. *Arxiv preprint arXiv:0710.0576*, 2007.

[122] M. Tumminello, F. Lillo, and R.N. Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58, 2010.

[123] V. Voev and A. Lunde. Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics*, 5(1):68–104, 2007.

[124] M. Wolf. Resampling vs. shrinkage for benchmarked managers. 2004.

[125] L. Zhang. Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47, 2011.

[126] L. Zhang, P.A. Mykland, and Y. Ait-Sahalia. A tale of two time scales. *Journal of the American Statistical Association*, 100(472):1394–1411, 2005.

[127] G. Zumbach. The empirical properties of large covariance matrices. 2009.