Statistical inference of Ornstein-Uhlenbeck processes : generation of stochastic graphs, sparsity, applications in finance

Gustaw Matulewicz

NNT : 2017SACLX066

Thèse de doctorat de l'Université Paris-Saclay préparée à l'École Polytechnique

Ecole doctorale n°574
Ecole doctorale de mathématiques Hadamard
Spécialité de doctorat: Mathématiques
par

M. Gustaw Matulewicz

Inférence statistique de processus d'Ornstein-Uhlenbeck: génération de graphes stochastiques, sparsité, applications en finance

Thèse présentée et soutenue à Palaiseau, le 15 Décembre 2017.

Composition du Jury :

M.	Mathieu Rosenbaum	Professeur École Polytechnique	(Président du jury)
Mme.	Marina Kleptsyna	Professeur Université du Maine	(Rapporteur)
M.	Markus Reiß	Professeur Humboldt-Universität zu Berlin	(Rapporteur)
M.	Mohamed Ben Alaya	Professeur Université de Rouen	(Examinateur)
M.	Sylvain Delattre	Maître de Conférences Université Paris Diderot	(Examinateur)
M.	Stéphane Gaïffas	Professeur Chargé de Cours École Polytechnique	(Directeur de thèse)
M.	Emmanuel Gobet	Professeur École Polytechnique	(Directeur de thèse)

Acknowledgements

First and foremost, I would like to thank my thesis advisors, Stéphane Gaïffas and Emmanuel Gobet, for giving me the opportunity to do this thesis at CMAP. Thanks to their support and patience I have been able to complete successfully this difficult task. I am also grateful for all that they taught me and helped me discover.

My thanks go to Marina Kleptsyna and Markus Reiß who gracefully accepted to review my thesis. I would also like to thank Mohamed Ben Alaya, Sylvain Delattre and Mathieu Rosenbaum who accepted to be part of the thesis jury.

I wish to thank *Chaire Risques Financiers* and the Natixis Foundation for Quantitative Research which have financed the work in this thesis.

A thesis is also a human experience and I have met during the past three years many wonderful people. I wanted to take this opportunity to thank all the PhD students in applied mathematics at Polytechnique, all the staff of CMAP and all those who participated in the weekly football game.

Finally, I want to thank my parents and family, my friends and Marion, who have been there for me no matter what.

Contents

Introduction

The topic of this thesis is to study inference for multi-dimensional Ornstein-Uhlenbeck processes, in the context of their relation to graph structures. The resulting analysis of the generation of stochastic graphs and the estimation under a sparsity assumption is motivated by a problem from systemic risk analysis, which we present below.

Motivations

The recent financial crisis showcased the limits of the usual approach of risk management, which concentrated on the risk of a single institution facing the market. Indeed, reducing individual risks of failure can actually increase the probability of systemic failure [FL13]. This sparkled the accelerated development of the field of *systemic risk* analysis. For instance, the simultaneous default of several institutions is completely unlikely if they are considered as independent events, but phenomena such as default cascades create such dependencies [ACM12, SC10]. Among the vectors of propagation of systemic risk, we take the example of liquidity risk [FL13].

In the practice of fractional reserve banking, which applies in most countries in the world, banks are required to hold adequate amounts of liquid assets such as cash [SC10]. To fulfill the regulatory requirements, banks sometimes need relatively short-term lending on a short notice. The common solution is to borrow the money from other banks which have excess reserves, a process called interbank lending. The resulting interbank lending market is a decentralized over-the counter (OTC) market and is the main source of liquidity for banks [AKS11]. Although the configurations change between countries, their structure is similar and is well described in [CL93].

During the crisis, increased uncertainty about the health of major institutions as well as increased risk of being subject to a bank run was followed by many lenders leaving the interbank market [GG14, ANP09]. As a result, the interbank lending market dried-up and even borrowers with sound positions had trouble getting the liquidity they needed. This shows that the interbank lending market is a vector of systemic

risk [AKS11, GG14, IP17]. Therefore, we see the need to understand the structure of the interbank lending market, in order to be able to predict possible contagions.

According to recent research, see [AKS14], the interbank-lending market has typically a stable structure. It is a general belief that there are long-term relationships that are created between some institutions and that most transactions are made through these channels. These relationships are typically informal and can be known, at best, by bank insiders. Because of this, the structure is not of common knowledge and is hard to observe. Moreover, one cannot simply assume structure from cumulative transactions. For example, a strong relationship may be inactive for a longer period of time simply because both parties are not in simultaneous need for borrowing and lending. That is why we ask the following question:

Question 1. *How to infer the structure of the interbank lending system from the observation of lending activity?*

We model interbank lending activity through a multi-dimensional Ornstein-Uhlenbeck (OU) process. We use it for the following reasons. The Brownian Motion that drives the OU models the random variations of the reserves of a bank, while the deterministic drift models the flow of reserves. From its linear form follows that increased imbalances create increased lending activity. Similar models have already been used for that same application. For instance, [CFS15] uses a model that fits exactly an OU process except for an added contribution that models the activity of a central bank. Also, [FI13] considers a similar model, but with a drift matrix parameter that depends on the current state of the reserves and in the setting of a Feller diffusion.

Being an OTC market, the lending activity is not directly observable. This leads us to assume that we have only partial information concerning the lending activity. Partial observation can be modelled using a non-bijective function applied to the process - each observation corresponds to a set of possible values for the underlying process. In the most radical case, which will be the one we assume, the function takes only two values: 0 or 1. This brings us to the following question:

Question 2. *How to infer the parameters of a multi-dimensional Ornstein-Uhlenbeck process from binary observations?*

While for some actors the assumption of partial observation is justified, for others it is superfluous. Indeed, the European Central Bank (ECB) has at its disposal the TARGET2 system, using which they observe with good precision the interbank lending activity between any two banks of the eurozone [AHH$^+$16, GSV15]. However, such an actor has to cope with another issue which is the large number of institutions it supervises. The resulting problem of parameter inference for a high-dimensional

OU process is then plagued by the curse of dimensionality, see [BvdG11]. Fortunately, at the scale of the eurozone, the interbank lending structure is highly sparse: any bank is typically interacting only with a restricted number of other banks [SC10]. The estimation of sparse parameters has attracted a lot of attention in high-dimensional statistics, see [Gir14a]. For instance, it has been suggested to use penalization and in particular ℓ^1 penalization [Tib96].

Question 3. *Can we improve the inference of Ornstein-Uhlenbeck parameters under a sparsity assumption?*

Summary

Questions 2 and 3 above are covered in two separate parts of the thesis, while Question 1 provides the context that guides our answers. As our work is on the estimation of Ornstein-Uhlenbeck processes, we start by presenting the properties of the process and fundamental results on its estimation.

1 Estimation and properties of the Ornstein-Uhlenbeck process

Definition and properties of the Ornstein-Uhlenbeck process Throughout the thesis, let $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P}\right)$ be a filtered space and $(W_t)_{t\in\mathbb{R}^+}$ a q-dimensional Brownian motion with respect to $\mathcal{F}$. For an integer $d \geq 1$, we define the d-dimensional Ornstein-Uhlenbeck process as the solution of the Stochastic Differential Equation (SDE) below

$$\mathrm{d}X_t = -\boldsymbol{A}X_t\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t, \tag{1}$$

where the initial value X_0 is a random value independent of W.

The Ornstein-Uhlenbeck process has been historically introduced with $d = 1$ in order to model the motion of a particle, as considered in the Brownian Motion, subject to friction [UO30]. This creates an attraction point for the process, which is 0 in our setting. Since then, the model gathered a lot of attention and found many applications. For instance, in mathematical finance, the Ornstein-Uhlenbeck is used to model interest rates and is known as the Vasicek model [Vas77]. It expresses that interest rates may change in the short-term, but revert in a longer term to some typical value. The Ornstein-Uhlenbeck process has also been used for pairs trading [LL15]. Pairs trading consists in trading two financial assets that are expected to follow similar price trends in the long term, but may experience short term deviations. The difference is then mean-reverting and the Ornstein-Uhlenbeck process helps in quantifying the

volatility and the speed of return to the mean. Also in mathematical finance, the Cox-Ingersol-Ross model [CIR85] is a close evolution of the Ornstein-Uhlenbeck. In neuroscience, the OU has been used to model membrane voltages [LL09]. It has been also used for population models [Bla14].

In the applications above, the Ornstein-Uhlenbeck is considered as a relevant model because of its mean-reverting property. This is achieved in dimension $d = 1$ by choosing a positive real parameter in place of $\boldsymbol{A}$ in Equation (1.5.2). For larger dimensions, the equivalent assumption is that $\boldsymbol{A}$ has a spectrum with strictly positive real parts. We will work under this assumption all along the thesis:

(H1) The spectrum of $\boldsymbol{A}$ has strictly positive real parts:

$$a_0 := \min_{\lambda \in \mathrm{Sp}(\boldsymbol{A})} \Re(\lambda) > 0.$$

In that case, the process is ergodic and possesses a stationary distribution which is a centered Gaussian, see [KS91], and in particular Section 5.6 herein. Using Ito's lemma, it is easy to prove that

$$X_t = e^{-\boldsymbol{A}t} X_0 + \int_0^t e^{-\boldsymbol{A}(t-s)} \boldsymbol{\Sigma} \mathrm{d}W_s. \tag{2}$$

The process X_t is Gaussian and we compute its mean and variance at any t using Equation (2) above:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\boldsymbol{A}t} X_0, \boldsymbol{C}_t\right) \tag{3}$$

$$\boldsymbol{C}_t = \int_0^t e^{-\boldsymbol{A}(t-s)} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top (t-s)} \mathrm{d}s. \tag{4}$$

It is desirable to have a non-degenerate distribution for X_t, hence an invertible $\boldsymbol{C}_t$. For this we take a new assumption:

(H2) The matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is invertible.

We see in the Equations (3) and (4) above why we need $\boldsymbol{A}$ to have a spectrum with strictly positive real parts. Indeed, it is the condition under which $e^{-\boldsymbol{A}t}$ converges to 0, which from one side gives

$$\mathbb{E}\left[X_t \mid X_0\right] \xrightarrow{t \to +\infty} 0$$

and from the other provides convergence of $\boldsymbol{C}_t$. However, contrary to the one-dimensional setting, $\boldsymbol{A}$ doesn't need to have only positive entries. Let us write Equation (1) as it applies to a fixed coordinate i:

$$\mathrm{d}X_t^i = -\sum_{j=1}^{d} \boldsymbol{A}^{ij} X_t^j \mathrm{d}t + (\boldsymbol{\Sigma}\mathrm{d}W_t)^i.$$

Some of the $\boldsymbol{A}^{ij}$ may be equal to zero, which we interpret as the absence of influence of component j on component i. Of course, this interpretation has its limits. For instance, i and j can be linked through correlated noise terms resulting from a non-diagonal $\boldsymbol{\Sigma}$, or influence each other through a third component.

The non-zero entries of $\boldsymbol{A}$ are interpreted as influences which can be positive or negative. However, because of the minus sign in Equation (1), the negative entries of $\boldsymbol{A}$ correspond to the positive influences, and *vice versa.* By positive influence we understand that an above-zero value of component j favors an increase of process i and a decrease when j takes a negative value. The reverse is true for negative influence.

This participates to the richness of the multi-dimensional Ornstein-Uhlenbeck process. In the one-dimensional case, at any point in time the process is expected to be attracted by 0. In the multi-dimensional case, the expected variations depend on the other components which can take values that will push a component away from 0. However, as long as assumption (H1) holds, in the long run the whole process, and hence all its components, is attracted by 0 and will take values according to the stationary distribution mentioned above. Indeed, observe that $\boldsymbol{C}_t$ possesses a limit when $t \to +\infty$:

$$\boldsymbol{C}_t = \int_0^t e^{-\boldsymbol{A}s}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top s}\mathrm{d}s \xrightarrow{t\to+\infty} \int_0^{+\infty} e^{-\boldsymbol{A}s}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top s}\mathrm{d}s =: \boldsymbol{C}_\infty.$$

The limit $\boldsymbol{C}_\infty$ is the covariance matrix of the stationary distribution. We will take in this thesis the assumption that the process is initialized in the stationary distribution:

(H3) The initial value X_0 is a Gaussian centered random variable with covariance matrix $\boldsymbol{C}_\infty$:

$$X_0 \sim \mathcal{N}(0, \boldsymbol{C}_\infty).$$

The density of the stationary distribution $\mathcal{N}(0, \boldsymbol{C}_\infty)$ will be denoted μ_∞.

The reasons for this assumption are twofold. First, the applications we had in mind, such as interbank lending or stock returns modeling, are essentially homogeneous in time, in the sense that the starting point of the observation is not any different from any other point in time. Therefore, it is natural to assume that the process

has reached stationarity, which means X_0 should follow the stationary distribution. Second, the impact of the initial value is quite limited, because of the exponential falloff of the factor preceding X_0 in Equation (3). After a short period the process is virtually indistinguishable from a stationary process. Hence, we assumed stationarity which allowed us to get clean and concise results. For instance, we have for $t \geq s$:

$$\mathbb{C}\mathrm{ov}\left(X_t, X_s\right) = e^{-\boldsymbol{A}(t-s)}\mathbb{C}\mathrm{ov}\left(X_s\right) = e^{-\boldsymbol{A}(t-s)}\boldsymbol{C}_\infty.$$

Observe also that the set of Ornstein-Uhlenbeck processes is stable by linear transformation, in the sense that applying a linear transformation of $\mathbb{R}^d$ to a OU process X gives another OU process. Consider for this an invertible matrix $\boldsymbol{L}$:

$$\mathrm{d}(\boldsymbol{L}X_t) = -(\boldsymbol{L}\boldsymbol{A}\boldsymbol{L}^{-1})(\boldsymbol{L}X_t)\mathrm{d}t + (\boldsymbol{L}\boldsymbol{\Sigma})\mathrm{d}W_t.$$

Thanks to this observation, we can simply broaden the applicability of any result concerning X to any linear transformation of X.

Parameter inference. We call parameter inference the task of estimating the values of $\boldsymbol{A}, \boldsymbol{\Sigma}$ from the observation of a single trajectory of X. Let us start by the simplest setting, where we have complete observation of a trajectory on the interval $[0, T]$, without further assumptions as in [LB76]. In that case, the estimation of $\boldsymbol{\Sigma}$ has a simple solution. By computing the squared variations integral, we recover exactly the value of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$:

$$\frac{1}{T}\int_0^T \langle \mathrm{d}X_t, \mathrm{d}X_t \rangle = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top.$$

This computation can be done for any interval length $T > 0$ and has no error. In probabilistic terms, the probability distributions resulting from $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \neq \boldsymbol{\Sigma}'\boldsymbol{\Sigma}'^\top$ are mutually singular, see [Jac01, LB77]. However $\boldsymbol{\Sigma}$ is non-identifiable as the map $\boldsymbol{\Sigma} \to \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is not one-to-one. For this reason we will only look into the estimation of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$.

In order to estimate $\boldsymbol{A}$ the usual approach is to construct the likelihood ratio using Girsanov's theorem [LB76, KS91]:

$$\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}W} = \exp\left(-\int_0^T (\boldsymbol{A}X_t)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathrm{d}X_t - \frac{1}{2}\int_0^T (\boldsymbol{A}X_t)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{A}X_t\mathrm{d}t\right).$$

Maximizing it yields the Maximum Likelihood Estimator (MLE), which is equal to:

$$\widehat{\boldsymbol{A}}^{MLE} := \left(\int_0^T \mathrm{d}X_t X_t^\top\right)\left(\int_0^T X_t X_t^\top \mathrm{d}t\right)^{-1}.$$

Because of assumption (H3), the likelihood in our setting includes also a term due to the initial value being non-zero, but we can neglect it as T increases. xThe MLE converges to the true value of the parameter $\boldsymbol{A}$ and we have the following convergence in law, see [Jac01]:

$$\sqrt{T}(\text{vec } \widehat{\boldsymbol{A}}^{MLE} - \text{vec } \boldsymbol{A}) \xrightarrow{\text{d}} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{-1} \otimes \text{Id}\right).$$

The rate of convergence with respect to T is $1/\sqrt{T}$. It is optimal in the sense that we have the Local Asymptotic Normality property, with rate $1/\sqrt{T}$ and $\boldsymbol{C}_\infty \otimes \text{Id}$ as the information matrix. This optimality result and the popularity of this method makes it a natural benchmark for the methods that we present in this thesis.

2 Estimation from a stochastic graph

In this part, we build a stochastic graph from an Ornstein-Uhlenbeck process. Consider for instance $\boldsymbol{Y}$ defined for any t by

$$\boldsymbol{Y}_t^{ij} := \mathbb{1}_{X_t^i - X_t^j \geq 1} \tag{5}$$

and see $\boldsymbol{Y}_t^{ij}$ as the adjacency value between nodes i and j at time t. Then $\boldsymbol{Y}_t$ encodes the adjacency matrix of a directed graph with d nodes. We see therefore $\boldsymbol{Y}$ as a stochastic graph process. This definition finds particular interest in the modeling of interbank lending. Indeed, in the models from [CFS15, FI13] mentioned in the Motivations Section above, reserve flows are proportional to reserve imbalances, hence the condition $\boldsymbol{Y}_t^{ij} = 1$ is interpreted as a large flow of reserves from i to j. For other applications, the definition from Equation (5) above can be easily extended to a more general setting. For any pair of nodes (i, j), consider a set S^{ij} and define

$$\boldsymbol{Y}_t^{ij} := \mathbb{1}_{X_t \in S^{ij}}. \tag{6}$$

The notion of stochastic graph links therefore the topic of stochastic processes to graphs and in particular random graphs.

Random graphs The topic of random graphs dates back to [ER59]. The Erdös-Renyi model, commonly seen as the most basic model for random graphs, assumes that all adjacencies are independent Bernouilli variables with parameter p, where p is the same for all adjacencies. The topic has attracted a large attention, but it has become clear that the model can't reproduce the characteristics of real-world graphs.

Given a random graph, it is a difficult and somewhat subjective question whether the graph ressembles some real-world graph [CF06]. To compare graphs, the typical

strategy has become the comparison of *features* of the graph [LCK+10]. An example of such feature is the *density* of a graph, which is the ratio of the number of edges to the number of possible edges, which is $\binom{d}{2}$ for a non-directed graph without self-loops. This feature can be reproduced by the Erdös-Renyi model, in the following sense: given a density δ, by choosing $p = \delta$, the expected density of the resulting Erdös-Renyi graph will be equal to δ. A more complex feature is the *degree distribution*, which is the frequency of degrees among the nodes. The Erdös-Renyi graph can reproduce only binomial degree distributions, see [Bol01], but in order to get a given degree distribution, one can use for example the Chung-Lu model [CL02].

One can continue this reasoning. Given some selected compatible features, it is entirely possible to consider theoretically a random graph generator that verifies all features. For example, one can restrict the space of graphs to those that have the right features, then choose one at random with a uniform probability. A lot of effort has been put in finding the "relevant" features, but it remains unclear at which point adding new features becomes irrelevant, see the conclusion of [CF06]. Among the most popular features that are used we have for example:

- diameter, the maximum distance between two nodes, see [Har69];
- clustering coefficients, a measure of density of the graph around a node introduced in [WS98];
- the Laplacian eigenvalue distribution, related to the spectral clustering of graphs, see [vL07];
- hop-plots, which represent the proportion of nodes that can be reached after following k edges, see [CF06], and measures the *small-world* property [WS98].

Instead of fitting features, one may want to develop an intuitively interesting framework. For example, the Barabási-Albert model [Bar99], or preferential attachment model, formalises the *rich-get-richer* mechanic, which is observed for instance in the development of the world wide web. It is observed that new websites send links more frequently to websites that are more connected. In the model, starting from some initial graph we add nodes and link them randomly to a fixed number of nodes with probabilities that are proportional to the degrees of the existing nodes. This model has the additional interest of modelling the growth of a graph.

In the models above, the edges are static, in the sense that they are not created and destroyed independently of the nodes. In the preferential attachment model, the evolution of the graph comes from the addition of new nodes, but once we have two

nodes, their connection status is static. We are interested here in analysing the variations of a graph with respect to another variable which will be seen as time. The resulting objects are called *evolving graphs* [AŁL+16, CZ16], *evolving networks* [BFG+08] or *temporal networks* [HS12, HS13]. We will use the first name. The topic has received increased attention in the past decade. The usual approach to model a network using a graph is static, but this may be unrealistic in some cases. For instance, the observation of large internet networks, such as the World Wide Web and social networks, is typically done using web crawlers and APIs [AŁL+16] which can observe only a part of the network in a given time. Hence, by the time that one would observe the whole network, it would necessarily change. There is no single definition for these evolving graphs. Sometimes time is included in the graph itself [Kos09]. However, there are two main approaches to evolving graphs [HKW16]. In the first, the network changes randomly and we count the changes. In the second, we consider regular snapshots of the graph that evolves between them. Both approaches generally lead to time that is represented by integers. Some work has been also done on non-integral time, such as in [CZ16].

Existing work on evolving graphs shares many commonalities. First, the objectives are typically to construct efficient methods and algorithms that compute some graph features of an evolving graph [AŁL+16, BKMU12, CZS+07]. Second, the evolution of the graph is either not modeled [BFG+08, CZ16] or modeled by independent rearrangements of edges and nodes [BKMU12, AŁL+16].

In our model, the graph is a function of a stochastic process. Instead of computing graph features, we want to estimate the dynamics of the underlying process. The fact that the process has a non-zero auto-covariance leads to non-independent graph changes. Our contribution in the topic of evolving graphs are hence first, the creation of a new model with correlated changes at random times and second, the proposition of a new point of view on questions that can be asked about evolving graphs.

Estimation from partial observation We have presented above the problem of parameter inference in the setting of complete observation in continuous time. We have outlined the estimation using the Maximum Likelihood Estimator, but other methods are also considered, see [Kut04]. While this setting is interesting from a theoretical perspective, it is more realistic to consider discrete observations. In general, one fixes an integer $n > 0$ and non-negative real times $(t_1, \dots t_n)$, and assumes that the process is observed only at these n time points [Sor97], with $n \to +\infty$. One can in principle use any increasing sequence of times, such as in [GCJ93], or even consider them as random, as in [GCJ94] or [ASM04]. It is however also natural to consider equidistant timestamps. We take then $\Delta_n > 0$, and choose $t_i = i\Delta_n$. We can then

broadly find three classes of settings.

- **The high-frequency fixed-time setting (HF-FT)**. We fix an observation length $T > 0$ and take $\Delta_n = T/n$, such that the observation is done on a fixed-length interval, but the observations become denser. One hopes to recover the parameters using increased granularity. However, the information one can use in that setting is bounded by what is included in the continuous intrval, and we know from what we explained above that then $\boldsymbol{\Sigma}$ is known but the estimation of $\boldsymbol{A}$ relies on $T \to +\infty$. Therefore we know that in the HF-FT setting, we can asymptotically recover only $\boldsymbol{\Sigma}$. In the homogeneous one-dimensional case, the LAMN property is proved in [Doh87]. The inhomogeneous multi-dimensional case is dealt with in [GCJ93]. Additionally, the authors work without the assumption of equidistant timestamps. Instead, they use the assumption that the empirical distribution of observation times converges weakly to some measure. The average time-step is converging to 0, which is the reason we classify it as a HF-FT setting. The random sampling for observation times is considered in [GCJ94]. The minimax rates of convergence, for one-dimensional processes, are obtained in [Hof99].

- **The low-frequency long-time setting (LF-LT)**. We take a constant time-step $\Delta_n = \Delta$. The observation has now a fixed granularity, but the observation length is increasing to $+\infty$. Therefore, we hope to recover the parameters using the ergodic properties of the process. The observation is in this case akin to a discrete time Markov chain. A first idea is to approximate the calculations of the continuous case with discrete equivalents. However, as pointed out in [Ped95], this method leads to inconsistent estimators [Flo89]. Instead, other methods have been proposed [Jac01], for instance [Kes97] and [KS99] provide methods that are both consistent and asymptotically normal.

- **The high-frequency long-time setting (HF-LT)**. We take here $\Delta_n \to 0$ and assume $n\Delta_n \to +\infty$. Although the granularity of the observation increases, the window of observation expands to $+\infty$. Here we can recover information both from decreasing time-step - by keeping the first T/Δ_n observations we recover the HF-FT setting - and increasing observation length, with the associated ergodic properties. The non-parametric estimation of the diffusion coefficient for one-dimensional processes has been done in [Flo93]. Also for one-dimensional processes, [Kes97] does the simultaneous parametric estimation of the drift and diffusion coefficients. The Local Asymptotic Normality property is proven in [Gob02]. The case of random observation times is the subject of [ASM04]. The paper works with the limit $\mathbb{E}[\Delta_n] \to 0$, which is why we classify it as a HF limit.

It is worth reminding that in practice, a time-step will be fixed as well as the observation length. However, if one assumes the time-step to be *small*, then the HF-LT setting provides a way to quantify jointly the impact of the size of the time-step and of the observation length. Therefore, we place the study in Section 2.1 in that setting.

We have started this part by introducing stochastic graphs. However, it is easy to see in Equation (6) that the setting can be extended to many cases of partial observation. Indeed, observations can be limited by the precision of the output, like a price that can't change by less than a cent. We can think then of the observed price as a discrete function of an underlying real-valued process. In Equation (6), we take the heaviest assumption of a binary function. This isn't particularly restrictive, as one can combine several binary functions to get many discrete functions. For example, observing $\mathbb{1}_{x\geq 0}$ and $\mathbb{1}_{x\geq 1}$ gives the knowledge about 3 intervals $]-\infty, 0[$,$[0, 1[$ and $[1, +\infty[$.

Binary information is similar to the observation of the crossing of some threshold. The threshold is typically zero for diffusions in $\mathbb{R}$, without loss of generality. In [Flo86, Flo88], it is assumed that a process is observed in continuous time, but one records only the zero-crossings of more than ϵ, which is a positive constant. It is proven that under some assumptions on the drift parameter, this is enough information to estimate the drift parameter if the volatility parameter is known and constant. Furthermore, the performance is significant compared to the complete information MLE, in the sense that their Fischer information ratio is around 0.618 for Ornstein-Uhlenbeck estimation.

The problem in discrete time is considered in [Flo87, Flo91]. The diffusion that is considered has a known volatility and the drift is a function $b(X_t, \theta)$ of the process and the parameter that is inferred. Interestingly, this paper uses two time-discretization settings. First, an estimator of θ is constructed using quasi-likelihoods and it is shown that it is asymptotically efficient in the LF-LT setting. Second, it is proved that the number of sign changes $N_\Delta(T)$ verifies

$$\sqrt{\pi\Delta/2}N_\Delta(T) \xrightarrow{L^2} L_t(0)$$

where $L_t(0)$ is the local time in zero of X. The result is extended to smooth approximations of the process in [Aza89]. As the local time is proportional to the density of the stationary density in the ergodic setting [KS91], this teaches us that we can expect to learn the volatility parameter in the HF setting. However, the article does not tackle simultaneously the HF and LT limits, and is limited to dimension $d = 1$.

Our contributions in this topic are the following:

- we prove mixing properties for Ornstein-Uhlenbeck processes;
- we provide a simple method to prove the convergence in L^2 of sums of functionals of an Ornstein-Uhlenbeck process;
- we completely characterize the convergence of occupation time statistics, including a CLT property for $d = 1$;
- we prove the convergence in L^2 of a normalization of the crossing statistic, thus extending results on the convergence of crossing statistics to the local time in zero, see [Flo87, Flo91] above, to the HF-LT limit;
- we show how the results can be used for parameter inference from a stochastic graph generated from an Ornstein-Uhlenbeck process;
- we conjecture an optimal rate of convergence and the validity of a CLT for the crossing statistic.

2.1 Parameter estimation of Ornstein-Uhlenbeck process generating a stochastic graph

Our estimation method consists in observing two quantities: how often is $\boldsymbol{Y}_t$ equal to 1 and how often does it jump between 0 and 1. This approach is not unlike the one from [Flo91]. We show several convergence properties for both statistics and we conclude with the applications for parameter inference. As mentioned above, we assume the HF-LT setting. A typical issue with assuming that $\Delta_n \to 0$ is that the observations are taken at times that are inconsistent between two values of n. Indeed, the Markov chain $(X_{k\Delta_n}, k \geq 0)$ depends on n through Δ_n, which is why we can't apply general mixing properties for Markov chains. This pushed us to examine more closely the rate at which the autocorrelation decreases with increasing time distance. The following Theorem expresses that this decorrelation rate is exponential and is based on the Gebelein inequality [Jan97], also known as Lancaster inequality.

Theorem 1 (Theorem I.1). *Assume (H1) – (H3). There exists a finite constant $C_{(7)}$, depending only on the stationary distribution covariance matrix $\boldsymbol{C}_\infty$, such that for any $t \geq s \geq 0$ and for any functions $\varphi : \mathcal{C}^0([0,s], \mathbb{R}^d) \to \mathbb{R}, \phi : \mathcal{C}^0([t, +\infty), \mathbb{R}^d) \to \mathbb{R}$ such that φ, ϕ are square-integrable w.r.t. the distribution of X, we have*

$$\left|\mathbb{C}\mathrm{ov}\left(\varphi\left((X_u)_{u\leq s}\right), \phi\left((X_v)_{v\geq t}\right)\right)\right| \leq C_{(7)} e^{-a_0|t-s|} \sqrt{\mathbb{V}\mathrm{ar}\left(\varphi\left((X_u)_{u\leq s}\right)\right) \mathbb{V}\mathrm{ar}\left(\phi\left((X_v)_{v\geq t}\right)\right)} \tag{7}$$

where a_0 is defined in Equation 1 from assumption (H1).

Theorem 1 above enables us to prove the following corollary, which applies to quite generic functionals g.

Corollary 1 (Corollary I.1). *Consider a measurable function* $g : \mathbb{N}\times\mathbb{N}\times\mathcal{C}^0([0,1],\mathbb{R}^d) \to \mathbb{R}$ *such that* $\mathbb{E}\left[g(k,n,(X_s)_{k\Delta_n\le s\le(k+1)\Delta_n})^2\right] < +\infty$ *for any* $k,n\in\mathbb{N}$. *For* $n\in\mathbb{N}$ *define*

$$v_n^2 = \sup_{k<n} \mathbb{V}\mathrm{ar}\left(g(k,n,(X_s)_{k\Delta_n\le s\le(k+1)\Delta_n})\right),$$
$$\xi_k^{(n)} = \sqrt{\frac{\Delta_n}{n}}\, g(k,n,(X_s)_{k\Delta_n\le s\le(k+1)\Delta_n}).$$

Then, there is a finite constant $C_{(7)}$*, dependent only on the parameters* $\boldsymbol{A},\boldsymbol{\Sigma}$ *of the model, such that:*

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=0}^{n-1}\xi_k^{(n)}\right) \le C_{(7)} v_n^2. \tag{8}$$

2.1.1 Occupation time statistic

We fix here a subset $S\subset\mathbb{R}^d$ and define the occupation time statistic by

$$\mathrm{OT}_n^S = \frac{1}{n}\sum_{k=0}^{n-1} Y_{k\Delta_n}^S = \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{1}_{X_{k\Delta_n}\in S}.$$

Since we have assumed that the process is stationary, see assumption (H3), we easily compute $\mathbb{E}\left[\mathrm{OT}_n^S\right] = \mu_\infty(S) := \int_S \mu_\infty(\mathrm{d}x)$. Then, in order to get convergence in L^2, we take interest in the variance. Here, we apply Corollary 1 and get :

Theorem 2 (Theorem I.2). *Assume (H1) - (H3). For any measurable set* $S\subset\mathbb{R}^d$, OT_n^S *converges to* $\mu_\infty(S)$ *in* L^2 *and*

$$\mathbb{E}\left[(\mathrm{OT}_n^S - \mu_\infty(S))^2\right] = O\left(\frac{1}{n\Delta_n}\right).$$

Theorem 2 above gives an upper bound on the convergence speed in L^2. In order to ensure the optimality of the rate and to better characterize the estimation error, we wish to provide a Central Limit Theorem (CLT) type property. We state exactly that with dimension $d=1$, hence $(\boldsymbol{A},\boldsymbol{C}_\infty) = (a,v_\infty)\in\mathbb{R}^2$, and $S=[1,+\infty[$ in Theorem 3 below.

Theorem 3 (Theorem I.3). *Assume (H1) - (H3). As* $n\to+\infty$*, we have*

$$\sqrt{n\Delta_n}\left(\mathrm{OT}_n^{[1,+\infty[} - \mu_\infty([1,+\infty[)\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0,\mu_\infty(\sigma^2 F'^2)\right)$$

where F is such that

$$F'(x) = \frac{2}{\sigma^2} \frac{N\left(\frac{x\wedge 1}{\sqrt{v_\infty}}\right) - N\left(\frac{1}{\sqrt{v_\infty}}\right) N\left(\frac{x}{\sqrt{v_\infty}}\right)}{\mu_\infty(x)} \in \left[0, 2\sqrt{\frac{\pi}{a\sigma^2}}\right].$$

We restrict the dimension to $d = 1$ in order to be able to solve explicitly a Poisson equation. The choice of S is motivated by our application in interbank lending, see Equation (5). We also use it to get explicit and clean limit variance. A slight modification of the proof allows to obtain results for other sets S. From this result follows a CLT for the resulting estimator of a assuming knowledge of σ.

Finally, we confirm the rate of convergence in extensive numerical tests, which we present in Section 4 of Chapter I, in which we estimate the expected values and standard deviations of the statistic using a regular Monte Carlo technique.

2.1.2 Crossing statistic

We define the crossings statistic by:

$$\mathcal{C}_n = \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{Y^S_{k\Delta_n} \neq Y^S_{(k+1)\Delta_n}}$$

where S is a half-space $S = \{x : x^1 \geq 1\}$. As we mentioned before, using linear transformations we can replace S by any half-space that doesn't contain 0. Using Corollary 1, we prove convergence in L^2:

Theorem 4 (Theorem I.4). *Assume (H1) – (H3). Assume that $n\Delta_n^{3/2} \to +\infty$. We have the following convergence:*

$$\frac{\mathcal{C}_n}{\sqrt{\Delta_n}} \xrightarrow{L^2} 2\sqrt{\frac{\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{11}}{2\pi}} \mu_{\boldsymbol{C}^{11}_\infty}(1),$$
$$\mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right) = O\left(\frac{1}{n\Delta_n^{3/2}}\right),$$

where $\mu_{\boldsymbol{C}^{11}_\infty}(1)$ denotes the density of a centered Gaussian with variance $\boldsymbol{C}^{11}_\infty$.

The result is again backed by Monte Carlo methods. However, although the results are consistent with the limit of the expectation that we provide in Theorem 4 above, simulations suggest it is possible to get a better rate of convergence in L^2. We conjecture therefore that the correct rate is $n\Delta_n$ instead of $n\Delta_n^{3/2}$; it also seems that rescaling using this rate gives in the limit a normal distribution, hence we conjecture that a CLT is true for $\mathcal{C}_n$.

2.1.3 Application to parameter inference

Using the occupation time statistics, we can estimate the measures of sets S under the stationary distribution, which depends on $\boldsymbol{C}_\infty$. By choosing several sets S, for example all sets $S^{ij} = \{x : x^i \geq 1, x^j \geq 1\}$, we can get to a situation where the vector of expected measures $\mathbb{E}\left[(\mathrm{OT}_n^{S^{ij}})\right]_{ij}$ is one-to-one with the set of covariances $\boldsymbol{C}_\infty$. This gives a natural candidate for an estimator of $\boldsymbol{C}_\infty$. As we prove that the measure $\mu_\infty(S)$ is a continuous function of $\boldsymbol{C}_\infty$, we conclude that this estimator has the right limit.

The covariance $\boldsymbol{C}_\infty$ is however a non-bijective function of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. To tackle this, we use the fact that in Theorem I.4, the crossing statistic converges to a product of, on one side, a function of $\boldsymbol{C}_\infty$ and, on the other, of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. As $\boldsymbol{C}_\infty$ is already estimated thanks to the occupation statistic, this enables estimating $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ independently of $\boldsymbol{C}_\infty$, which is followed by the estimation of $\boldsymbol{A}$.

2.2 Properties of the stochastic graph

The multi-dimensional Ornstein-Uhlenbeck process can be used in order to model interbank lending, but to make intuitive sense the flows of reserves must conserve the sum of all reserves. This condition writes $\mathbb{1}^\top \boldsymbol{A} = 0$ and hence $\boldsymbol{A}$ is non-invertible. We deal with this degeneracy by describing the dynamic of the d components using $d-1$ deviations from the mean, which follow a slightly changed OU dynamic, and an independent mean process. For the Carmona-Fouque model, which we take from [CFS15, FI13], the OU dynamic for the deviations has a scalar drift parameter.

By simulating small interbank-lending networks, using both the Carmona-Fouque homogeneous model and a sparse model with a central institution, we show that the observed flows of money can be completely different from the actual structure of the model. Observing the stochastic graph over an interval helps in finding which nodes are connected, but it isn't straightforward to infer the lending structure from the stochastic graph.

Taking larger interbank lending networks, we analyze the stochastic graph from the viewpoint of evolving graphs, as presented in [BFG+08]. First, we look at how some basic measures of the graph evolve with time. The graph density is varying around a mean value, in a manner similar to a one-dimensional Ornstein-Uhlenbeck process. The degree distributions show a fast falloff for the tail distribution, but most graphs we simulated had some nodes that were connected to the majority of other nodes. Further, the graphs are very concentrated, with small diameters and average shortest path lengths, but have almost no clustering. Second, we look at properties of edges

across time. For instance, the edge durations have an exponential falloff of their tail distribution. Third, we look at the connectedness of the central node in the sparse model. Its degree is much lower than what is observed for other nodes.

3 Sparse estimation

In this section, we assume a complete observation on an interval $[0, T]$. Thanks to this assumption, as we have pointed out earlier, $\mathbf{\Sigma}$ can be considered as known, which is why we simplify the question by further assuming $\mathbf{\Sigma} = \mathrm{Id}$. From this follows from instance that Assumption (H2) is true. Finally, we add a row-sparsity assumption:

(H4) For a fixed $s \leq d$, the matrix $\boldsymbol{A}$ is row-s-sparse, which we define as

$$\forall i \leq d, \|\boldsymbol{A}^{i\bullet}\|_0 \leq s,$$

where $\boldsymbol{A}^{i\bullet}$ is the i-th line of $\boldsymbol{A}$. This assumption signifies that any coordinate of X is impacted only by at most s coordinates:

$$\mathrm{d}X_t^i = -\sum_{j:\boldsymbol{A}^{ij}\neq 0} \boldsymbol{A}^{ij} X_t^j \mathrm{d}t + \mathrm{d}W_t^i.$$

Sparse estimation is a problem that has gathered a lot of attention in statistics [BvdG11, Gir14a]. The idea is to find the best estimator which verifies a given sparsity condition, i.e. that has a limited number of non-zero values. Unfortunately, this task is computationally very expensive. The reason for this is that the set of parameters under a sparsity assumption is non-convex. Indeed, the set of matrices that verify Assumption (H4) is the union of all linear sub-spaces that are obtained by selecting s indices per line and restricting the non-zero values to these indices. In order to do the estimation, one would need to do the estimation on each of the subspaces and then choose the solution with minimal estimation error. However, there are $\binom{d}{s}^d$ such subspaces, hence the problem is non-tractable for large d and even moderate s.

The Lasso estimator. Another approach is to decrease the number of non-zero values by penalizing them. This would lead naturally to penalizing the objective function with $\|\boldsymbol{A}\|_0$, where $\|\boldsymbol{A}\|_0 = \#\{(i,j) : \boldsymbol{A}^{ij} \neq 0\}$. The pseudo-norm $\|\boldsymbol{A}\|_0$ is however non-convex which makes the problem intractable. The idea of convex relaxation suggests then replacing the 0-norm by $\|\boldsymbol{A}\|_1 = \sum_{ij} |\boldsymbol{A}^{ij}|$. The resulting estimator is the Lasso estimator [Tib96]:

$$\widehat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} -\frac{1}{T} \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}W} + \lambda \|\boldsymbol{A}\|_1. \tag{9}$$

In front of the ℓ^1 norm we add a parameter λ which can be tuned in order to balance between fit to data and sparsity.

Lasso estimation and more generally sparse estimation using convex penalization has been the subject of extensive work in the recent past [BvdG11, FBO12, Gir14a]. A popular area of application is the Gaussian regression. In this problem, one considers the problem of estimating a vector parameter a given a series of observations y and a series of vectorial variables that form a matrix $\boldsymbol{X}$, and assuming that

$$y = \boldsymbol{X}a + w$$

where w is a centered Gaussian noise term. The setting is somewhat analogous to the one we consider in Equation (1). In sparse estimation, one usually takes a sparsity assumption applied to the vector a, for instance that the number of non-zero values is bounded by a constant s. Sparse oracle properties consist in asserting how estimation errors react to this sparsity assumption. For instance, it can be shown that the norm of the estimation error is bounded by a function of s. Such oracle properties are shown for the Lasso in [BTW07] and [BRT09] in the linear regression model. An issue of the Lasso that has been raised is that the penalization applies equally to all terms. For instance, some wanted to induce structured sparsity, resulting in an estimation called the Group Lasso [YL06], which was also used for reduced rank estimation [BSW12].

Sparse estimation using the lasso has also been applied to time series. In a particularly relevant case for our work, it has been used for the estimation of the parameters of a Vector Auto-Regressive process. Indeed, a discretized Ornstein-Uhlenbeck process can be seen as an AR(1) time-series. For instance [BM15] gives error bounds for the Lasso estimator under a sparsity assumption, stating that the equivalent of the drift parameter has a limited number of non-zero entries. The method is using spectral densities and is relying heavily on the discrete scheme, as it lets us see a d-dimensional process observed at n times as a $n \times d$-dimensional Gaussian variable. In the context of Ornstein-Uhlenbeck estimation, the Lasso estimation has been considered in [Sok13], but without providing any theoretical results such as error bounds. Our approach provides such error bounds, under a sparsity assumption which states that each row of $\boldsymbol{A}$ has a limited number of non-zero entries.

The Adaptive Lasso estimator. Unfortunately, Lasso estimation has its downsides. Indeed, sparse estimation has two objectives: to estimate correctly the true parameter and to get the right level of sparsity. The first objective means we wish the estimator to be asymptotically unbiased and with optimal convergence rate, such that with in-

creasing observation time we can expect a decrease to zero of the estimation error. The second objective means we want the estimator to set to zero the entries that correspond to true zero values such that we recover asymptotically the support of the true matrix. In the case of Gaussian regression, [Zou06] shows that one can't get both properties simultaneously for the Lasso: the values of λ that lead to consistent estimation make also sure that there can't be asymptotically consistent variable selection. To solve this, the article proposes the Adaptive Lasso, which consists again in a ℓ^1 norm, but now each entry is weighted by the inverse of the corresponding entry in the MLE.

Our contributions in this part are the following:

- we develop sparse estimation using Lasso in the continuous time setting;
- we show results in the setting of the row-sparsity assumption;
- we provide a complete theoretical basis for sparse estimation of Ornstein-Uhlenbeck processes, including an upper bound for the estimation error;
- we give a minimax lower bound on the estimation error under a row-sparsity assumption;
- we prove a Restricted Eigenvalue property for the Ornstein-Uhlenbeck process;
- we present asymptotic optimality properties of the Adaptive Lasso;
- we show the advantages of the penalized methods through an application to real-world financial data.

3.1 Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process

In order to estimate the parameter $\boldsymbol{A}$, of which the true value will be denoted $\boldsymbol{A}_0$ in this Section, from a trajectory governed by Equation (1) with $\boldsymbol{\Sigma} = \mathrm{Id}$, we consider two methods based on the ℓ^1 penalization: the Lasso and the Adaptive Lasso. We consider first the Lasso in a non-asymptotic setting. Indeed, it is noteworthy that we can derive error bounds in the non-asymptotic setting, which apply with high probability. The non-asymptotic setting enables us to better quantify the effect of the sparsity s, dimension d and observation length T.

We start by rephrasing the problem in terms of matrices:

$$\mathcal{L}_T(\boldsymbol{A}) := \mathrm{tr}\left[\boldsymbol{A}^\top \boldsymbol{\varepsilon}_T + \frac{1}{2}(\boldsymbol{A} - \boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T(\boldsymbol{A} - \boldsymbol{A}_0)^\top\right]$$

$$\text{with} \quad \widehat{\boldsymbol{C}}_T = \frac{1}{T}\int_0^T X_T X_t^\top \mathrm{d}t \quad \text{and} \quad \boldsymbol{\varepsilon}_T = \frac{1}{T}\int_0^T \mathrm{d}W_T X_t^\top.$$

In the definition above, $\mathcal{L}_T$ is a normalized negative approximate log-likelihood, as we neglect terms of order $O\left(1/T\right)$ coming from the likelihood of the initial value X_0.

3.1.1 Preliminary results

The convergence of any estimator based on the negative log-likelihood $\mathcal{L}_T$ can be retraced to the behaviour of the two quantities $\widehat{\boldsymbol{C}}_T$ and $\boldsymbol{\varepsilon}_T$. Indeed $\boldsymbol{\varepsilon}_T$ can be seen as a noise term that needs to be sufficiently small, while $\widehat{\boldsymbol{C}}_T$ needs to be far enough from being degenerate.

In the setting of Gaussian regression, the equivalent of $\widehat{\boldsymbol{C}}_T$ is a Gram matrix, which is computed from a given design matrix. It becomes necessary to make the assumption that its eigenvalues are far enough from 0 in order to ensure fast rates of convergence. To the contrary, in the OU setting the matrix $\widehat{\boldsymbol{C}}_T$ is computed from the values of X, which are obtained from the model in Equation (1). Thanks to the ergodicity of X, the empirical covariance $\widehat{\boldsymbol{C}}_T$ converges in probability to $\boldsymbol{C}_\infty$. As $\boldsymbol{C}_\infty$ has strictly positive eigenvalues, which comes from assumption (H1), this shows that for large T, $\widehat{\boldsymbol{C}}_T$ will also have a strictly positive spectrum. In Theorem 5 below, we quantify that property, but before we need to introduce a fifth assumption. It relates to the speed of convergence of the empirical covariance $\widehat{\boldsymbol{C}}_T$ to the stationary covariance $\boldsymbol{C}_\infty$.

(H5) There exists a non-decreasing function H, positive on $\mathbb{R}^+$, such that for any vector u verifying $\|u\|_2 \leq 1$, we have:

$$\mathbb{P}\left[|u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u| \geq R\right] \leq 2\exp(-TH(R)).$$

We conjecture the validity of that assumption. Indeed, for reversible Ornstein-Uhlenbeck processes, which corresponds here to having a symmetric $\boldsymbol{A}_0$, it follows from the application of a Theorem from [CG07]. This Theorem states that the probability that, for a real-valued function V, the integral $T^{-1}\int_0^T V(X_t)\mathrm{d}t$ deviates from its expectation $\mu_\infty(V)$ exceeds any threshold decreases exponentially with T. Unfortunately, we haven't managed to apply the Theorem for non-symmetric $\boldsymbol{A}_0$. However, we believe that the property should apply to other settings, since [HHMS93] shows that the rate of convergence of $\widehat{\boldsymbol{C}}_T$ to $\boldsymbol{C}_\infty$ is faster for non-symmetric $\boldsymbol{A}_0$ than for symmetric ones.

Based on this assumption and on a union bound that enables going from Assumption (H5) to the convergence of eigenvalues, we prove Theorem 5 below. It can be seen

as the statement of the validity of a Restricted Eigenvalue property [BRT09], which is typically needed to be assumed in settings such as the Gaussian Regression [ZWJ14].

Theorem 5 (Theorem III.3). *Assume (H1), (H3) – (H5). Set* $s \leq d$ *and* $c_0 > 0$. *Define* $C(s, c_0) := \{u \in \mathbb{R}^d : \|u\|_1 \leq (1 + c_0)\|u_{|\mathcal{I}_s(u)}\|_1\}$ *where* $\mathcal{I}_s(u)$ *stands for the set of indices of the* s *largest entries of* $|u|$. *Consider* $\epsilon_0 \in (0, 1)$ *and assume that*

$$T \geq T_0 := H\left(\frac{\kappa^2}{9(c_0+2)^2}\right)^{-1}\left(s\log\left(21d \wedge \frac{21ed}{s}\right) + \log\left(\frac{4}{\epsilon_0}\right)\right).$$

Then,

$$\mathbb{P}\left[\inf_{u\in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa\right] \geq 1 - \frac{\epsilon_0}{2} \tag{10}$$

where $\kappa := \sqrt{\frac{\min \mathrm{Sp}(\boldsymbol{C}_\infty)}{2}}$.

We continue with a property which states in a practical way that $\boldsymbol{\varepsilon}_T$ is relatively small with high probability.

Theorem 6 (Theorem III.8). *Assume (H1) and (H3). Define for* $x > 0$*:*

$$\theta(x, (X_t)) := \sqrt{4eT^{-1}|\mathrm{diag}\,(\widehat{\boldsymbol{C}}_T)|_\infty\left(x + \log(2 + |\log T\,\mathrm{diag}\,(\widehat{\boldsymbol{C}}_T)|_\infty)\right)},$$

where we denote diag *the extraction of the diagonal of a matrix and* log *applies naturally to each term (which are all positive). For any matrix* $\boldsymbol{U}$*, we have*

$$\mathbb{P}\left[\langle \boldsymbol{U}, \boldsymbol{\varepsilon}_T\rangle_F \leq \theta(x, (X_t))\|\boldsymbol{U}\|_1\right] \geq 1 - \frac{\pi^2}{3}\exp(-2x + 2\log d).$$

3.1.2 Non-asymptotic properties of the Lasso

Building on the results of Section 3.1.1 above, we prove Theorem 7 below. It provides a bound on the estimation error, in terms of ℓ^1 and L^2 norms.

Theorem 7 (Theorem III.1). *Assume (H1), (H3) – (H5). Set* $\gamma > 1$, $0 \leq \tau < \gamma - 1$, $\epsilon_0 \in (0, 1)$ *and define*

$$\lambda_T := \gamma\sqrt{\frac{4e|\hat{\delta}_T|_\infty}{T}\left(\frac{1}{2}\log\frac{2\pi^2 d^2}{3\epsilon_0} + \log(2 + |\log(T\hat{\delta}_T)|_\infty)\right)}$$

where $\hat{\delta}_T := \operatorname{diag} \widehat{\boldsymbol{C}}_T$ *and* $\log$ *is applied entrywise on* $T\hat{\delta}_T$. *Set* $c_0 := \frac{\gamma+\tau+1}{\gamma-\tau-1}$, *recall the definitions of* κ *and* T_0 *from Theorem 5 and assume that* $T \geq T_0$. *Then, for any row-*s*-sparse matrix* $\boldsymbol{A}$, *the lasso estimator* $\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{A}}_{\lambda_T}$ *verifies*

$$2\tau\gamma^{-1}\lambda_T\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 \leq \|(\boldsymbol{A} - \boldsymbol{A}_0)X\|_{L^2}^2 + \left(\frac{1+\gamma+\tau}{\gamma\kappa}\right)^2 \lambda_T^2 ds$$

with probability at least $1 - \epsilon_0$.

From the Theorem follow naturally error bounds for the ℓ^1 and L^2 norm. Then, combined with the Restricted Eigenvalue property from Theorem 5, we get an error bound in Frobenius norm and finally, in q-norm.

The bound is, up to constant factors, of order $\lambda_T^2 ds$. After we prove that $\hat{\delta}_T$ can be in the asymptotic setting seen as a constant, we get that the asymptotic equivalent of the error bound is

$$\frac{ds(\log d + \log\log T)}{T}.$$

We compare this equivalent to the minimax lower bound that is given in Theorem 8 below. We acknowledge that the two are not equal, but the difference of $\log s$ and $\log\log T$ is rather small. We believe it would be possible to improve the estimation in terms of upper bound on the error by considering alternative estimators. For instance, SLOPE achieves the minimax rate [BLT16].

Theorem 8 (Theorem III.2). *For some constants* $c > 0$ *and* $c' > 0$, *we have:*

$$\inf_{\widehat{\boldsymbol{A}}} \sup_{\boldsymbol{A}\in\Gamma_s} \mathbb{E}_{\boldsymbol{A}}\left[\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\right] \geq \frac{c' ds\log(cd/s)}{T},$$

where Γ_s *is the set of row-*s*-sparse matrices and the infimum is taken over all possible estimators.*

The proof of the lower bound uses a Gilbert-Varshamov bound applied to the set of row-s-sparse matrices Γ_s. The volumes that are needed to compute the lower bound are computed using a result on the counting of regular graphs.

3.1.3 Asymptotic properties of the Adaptive Lasso

We know that in Gaussian regression the Lasso can't achieve asymptotically both the optimal rate of convergence to the true parameter $\boldsymbol{A}_0$ and the exact support recovery [Zou06]. We prove in the following that this applies also to parameter inference of the Ornstein-Uhlenbeck process. Indeed, λ is a parameter that essentially shrinks all entries proportionally to λ. Unfortunately, this means that support recovery necessitates λ to be of an order larger than $1/\sqrt{T}$, while optimal speed is achieved only

if λ is of an order smaller than $1/\sqrt{T}$.

The idea of the Adaptive Lasso is simple: it is to penalize more the entries which we believe that are probably zeros and less those that are probably non-zero. As a way to judge that, we use the MLE - if its value is low, then the entry probably should be a zero. This leads to a penalization given by

$$\sum_{i,j} \frac{|\boldsymbol{A}^{ij}|}{|(\widehat{\boldsymbol{A}}^{MLE})^{ij}|^{\gamma}}.$$

In this definition, γ can take any positive real value, and we use $\gamma = 1$ in our experiments. Theorem 9 below shows that this estimator achieves both the optimal rate of convergence and has the property of asymptotic support recovery. The dependence of λ and of the estimators with respect to T has been hidden to reduce clutter.

Theorem 9 (Theorem III.4). *Assume (H1) - (H3). Fix $\gamma > 0$ and assume that λ verifies $\lambda T^{1/2} \to 0$ and $\lambda T^{(\gamma+1)/2} \to +\infty$ when $T \to +\infty$. Then, we have the following properties.*

1. *Consistency of the variable selection: as $T \to +\infty$, we have*

$$\mathbb{P}\left[\mathrm{supp}(\widehat{\boldsymbol{A}}_{ad.}) = \mathrm{supp}(\boldsymbol{A}_0)\right] \to 1.$$

2. *Asymptotic normality: as $T \to +\infty$, we have*

$$\sqrt{T}\left((\widehat{\boldsymbol{A}}_{ad.})_{|\mathcal{A}_0} - (\boldsymbol{A}_0)_{|\mathcal{A}_0}\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0 \times \mathcal{A}_0}\right)^{-1}\right),$$

where $\mathcal{A}_0 = \mathrm{supp}(\boldsymbol{A}_0)$ is the support of the parameter $\boldsymbol{A}_0$.

3.1.4 Numerical comparison of MLE, Lasso and Adaptive Lasso

The theoretical results above show some of the strengths of the penalized methods, but it remains to be seen how they perform in numerical experiments compared to the MLE.

The first point we need to address is the choice of λ. Although Theorem 7 provides a theoretical value for λ, the theoretical penalization parameters tend to be pessimistic [BvdG11]. Instead, we use a cross-validation technique [AC10]: we use the first 80% of a trajectory to compute the log-likelihood in Equation (8) (the training set) and the last 20% of the trajectory for tuning of λ (the validation set). In this Section, we assume the Lasso and Adaptive Lasso estimators have been obtained through that

cross-validation method.

Our experiments on simulated data confirm the advantages of the penalized methods compared to the MLE in terms of estimation errors. They also confirm the advantages of the Adaptive Lasso in terms of support recovery accuracy. Finally, we apply the estimation methods to real world data, a 120-dimensional time-series of exponential moving averages of returns of SP500 components.

3.2 Sparse estimation applied to finance

We apply the methods presented above in two separate use cases. The first application is to US stock daily close prices, where we wish to test the hypothesis that recent over-performing stocks are more likely to under-perform. The second application is to dividend future prices, for which we wish to test if their variations are compatible with the link between stock price and expected dividends coming from a discounted cash-flow approach.

3.2.1 Mean-reversion in SP500 stocks

We take here a subset of the SP500 components for which we have a long enough price history. The subset we use has 466 stocks and 4305 daily close prices for each stock, spanning the period between January 3, 2000 and February 10, 2017. The stocks are attributed to groups with common GICS industrial sector. For each stock, we compute the exponential moving average (EMA) of its returns, with a parameter $\alpha = .1$ (the time is counted in days). If its value is above its average value, we will say the stock is a recent over-performer, while under-performers have an EMA below said average value. By construction, the EMA is mean-reverting:

$$E_t := \alpha \int_0^t e^{-\alpha(t-s)} \mathrm{d}M_s, \qquad \mathrm{d}E_t = -\alpha E_t \mathrm{d}t + \alpha \mathrm{d}M_t.$$

We will henceforth analyse the EMA instead of the returns data, as the EMA is continuous and any prediction on the EMA can be transformed into a prediction on future returns. We perform an analysis in several levels, both one-dimensional and multi-dimensional. In each case, we fit the data to an Ornstein-Uhlenbeck process with parameters $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ (a and σ for one-dimensional processes). Then, we compare the results to a benchmark we call the Brownian Motion Exponential Moving Average (BMEMA), the EMA of serially uncorrelated returns with covariance $\boldsymbol{\Sigma}$. Estimating for it the mean-reversion parameter $\boldsymbol{A}$, we get the mean-reversion strength expected for serially uncorrelated data.

As we believe that the strength of the mean-reversion can vary, we do the estimation on multiple sub-intervals of the total interval. We take in practice 30 regularly separated intervals of length 1000 (4 years of daily prices), which give us an idea of typical mean-reversion strengths. Then, we use the first 800 days for the estimation and use the last 200 days in order to verify the prediction power of our model. The verification is made using likelihood ratios.

We do an analysis on several levels. When analysing one-dimensional data, we only use the maximum likelihood estimator. For multi-dimensional data, we compare the Lasso, Adaptive Lasso, MLE and the Diagonal estimator. The Diagonal estimator is the estimator with only diagonal entries, all computed from independent one-dimensional analysis of the components.

- **Mean-reversion of sector averages**. We compute for each industrial sector the equal-weighted average of the EMAs. An analysis in one-dimension of the individual sector averages points to the existence of mean-reversion for most industrial sectors, with typical values of a between 0.11 and 0.14, compared to an expected value for serially uncorrelated data of 0.105 ± 0.015. The multi-dimensional approach shows possible cross-sectorial influences, as the non-diagonal Lasso estimator beats in average the Diagonal estimator. However, both for one-dimensional and multi-dimensional approaches, we can't beat consistently the benchmark in terms of likelihood ratios.

- **Mean-reversion of stocks to their sector average**. We subtract here the sector average from each individual stock data series. The resulting mean-reversion parameters are typically between 0.1 and 0.15 and indicate the mean-reversion strength typically exceeds what we expect from serially uncorrelated data. The multi-dimensional approach shows the added value of the penalized methods, which perform better than both the Diagonal and the non-penalized MLE estimators. However, both for one-dimensional and multi-dimensional approaches, we fail to beat the benchmark in terms of likelihood ratios.

The consistent observation of mean-reversion exceeding what is expected from uncorrelated data and the failure to verify it using independent likelihood ratio testing needs an explanation. For this, we apply our methods to data that we know is mean-reverting: a autoregressive process with negative parameters. We observe that our methods correctly see mean-reversion in the estimation of a, but the likelihood ratios fail to confirm the result. We conclude that the likelihood ratio test is not a good method to verify mean-reversion. Given what we observed, we assert the existence of mean-reverting properties, but can't statistically confirm or reject the significance of the observation.

3.2.2 Mean-reversion in EURO STOXX 50 Dividend Futures

Using a discounted cash flow approach, we can see a stock as equivalent to the cash flow of all future dividends. With some approximations, we get

$$\tilde{S}_t \simeq \sum_{n \geq 0} D(t,n) F_t^{(n)}$$

where $\tilde{S}$ is the stock price adjusted for dividends, D denotes discount factors and F dividend futures prices. Dividend futures are contracts with underlying an index counting the amounts of paid dividends during a calendar year. As a result, if we fix the stock price, the variations of the future prices are constrained by a linear relation, which could create mean-reversion properties.

We take here the EURO STOXX 50 Index, an Index which tracks the value of 50 European blue chip stocks. The Eurex exchange created the EURO STOXX 50 Dividend Points index, which tracks the total dividends paid by the EURO STOXX 50 companies, and uses it as the underlying for EURO STOXX 50 Dividend Futures. The index can then be seen as a regular stock. We do then the regression of the Dividend Future prices versus the Index. The slopes change from year to year, but conserve a common shape as functions of maturities. The remainder is the part of the variations of the futures prices that isn't explained by the price and is the subject of our further analysis.

We first try to fit the data to a multi-dimensional Ornstein-Uhlenbeck process. The result is mixed, but we observe that the prices tend to cluster around their average. We decide to use a two-step approach. In a first step, we analyze mean-reversion of the average. It is significant and enables prediction of future variations, but one has to be careful with the length of the time-spans used for estimation. In a second step, we analyze the movements of the individual prices around the average, with again mixed results. We conclude that dividend future prices, controlled by the index variations, are gathering around their average value which is relatively slowly mean-reverting.

Introduction

Le sujet de cette thèse est l'étude de la façon dont une structure de graphe se manifeste dans des systèmes qui évoluent de manière aléatoire avec le temps. Ceci mène à l'analyse des interactions entre processus stochastiques et graphes, que nous faisons dans le cadre des processus d'Ornstein-Uhlenbeck. Nous commencerons par préciser les motivations des questions que nous abordons dans cette thèse.

Motivations

La crise financière récente a montré les limites de l'approche habituelle de l'analyse de risque, qui se concentrait sur les risques d'une entité faisant face à un marché. En effet, il s'avéra que les décisions visant à réduire les risques de défaut d'entités isolées peuvent mener à une augmentation du risque d'effondrement du système financier entier [FL13]. Ceci a mené vers une accélération du développement du domaine de l'analyse de *risque systémique*. Par exemple, le défaut simultané de nombreuses institutions est complètement improbable si ces défauts sont considérés comme des événements indépendants. Cependant, des phénomènes comme les cascades de défauts créent ces dépendances [ACM12, SC10]. Parmi les vecteurs de propagation de risque systémique, nous nous intéresserons en particulier au risque de liquidité [FL13].

Dans un système bancaire à réserves fractionnaires, qui est le système partagé par la majorité des pays dans le monde, il est exigé des banques d'avoir en réserve des quantités adéquates d'actifs liquides tels que de l'argent liquide [SC10]. Afin de satisfaire cette réglementation, les banques se retrouvent parfois en besoin urgent d'emprunter à relativement court terme. Pour ceci, les banques s'adressent généralement à d'autres banques qui se retrouvent en situation d'excès de réserves. Ce mécanisme de prêt interbancaire crée un marché décentralisé dit OTC (*over-the counter* ou "de gré à gré"). Il forme la source primaire de liquidité pour les banques [AKS11]. Bien que les configurations et législations diffèrent entre les pays, la structure du marché reste similaire et est décrite par exemple dans [CL93].

Pendant la crise, l'incertitude concernant la santé financière des institutions ainsi que les peurs de panique bancaire ont poussé certains fournisseurs de liquidités à se rétracter du marché de prêt interbancaire [GG14, ANP09]. Par suite, le marché s'est retrouvé *asséché*, c'est à dire que même les institutions de bonne santé et avec des positions solides avaient du mal à emprunter sur le marché [AKS11, GG14, IP17]. Pour mieux comprendre les risques créés par ces situations, nous concluons de la nécessité de mieux comprendre la structure du marché de prêt interbancaire, pour pouvoir ainsi mieux prédire les contagions liées au risque de liquidité.

Selon des études récentes, comme [AKS14], le marché de prêt interbancaire possède généralement une structure stable dans le temps. Il est communément admis que le marché est structuré par des relations de longue durée qui se forment entre certaines institutions et que la majorité des transactions se font à travers ces relations. Cependant, ces relations sont généralement informelles et ne sont connues par personne d'autre que certains employés. On peut donc affirmer que la structure du marché n'est pas publiquement connue et n'est pas facilement observable. De plus, il n'est pas nécessairement pertinent d'inférer la structure des volumes échangés totaux. En effet, une relation forte que nous souhaitons découvrir peut être inactive pendant une durée prolongée simplement parce que les deux partenaires ne se retrouvent pas simultanément en besoin et excès de réserves. Pour cette raison, nous posons la question suivante :

Question 1. *Comment inférer la structure du marché de prêt interbancaire à partir de l'observation de l'activité de prêt ?*

Nous modélisons le prêt interbancaire à travers un processus d'Ornstein-Uhlenbeck (OU) multidimensionnel pour les raisons suivantes. Le Mouvement Brownien qui y figure permet de modéliser les variations aléatoires des réserves des banques. Le drift permet de modéliser le flux de capitaux. Sa forme linéaire en X implique que l'activité augmente lorsque les déséquilibres en termes de réserves augmentent. Des modèle similaires ont déjà été proposés pour modéliser le marché de prêt. En particulier [CFS15] utilise un modèle qui peut s'écrire comme un OU auquel est ajouté une contribution qui représente l'activité de la banque centrale. Similairement, le modèle de [FI13] prévoit des flux proportionnels aux déséquilibres de réserve, en ajoutant une dépendance de $\boldsymbol{A}$ en X et en se placant dans le cadre d'une diffusion de Feller.

En plus de ce que nous avons déjà mentionné sur le marché de prêt interbancaire, nous remarquons ensuite qu'étant donné qu'il s'agit d'un marché OTC il n'est pas directement observable. Nous choisissons de modéliser une observation partielle à travers une fonction non-bijective que nous appliquons au processus. Ainsi, chaque

observation correspond à un ensemble de valeurs possibles pour le processus. Nous supposons la version la plus extrême de ceci, où la fonction de prend que les valeurs de 0 ou 1. Ceci nous mène à la question suivante

Question 2. *Comment inférer les paramètre d'un processus d'Ornstein-Uhlenbeck multidimensionnel à parti d'observations binaires ?*

Bien que pour certaines institutions l'hypothèse d'observation partielle est justifiée, pour d'autres elle est superflue. En effet, la Banque Centrale Européenne (BCE) dispose des informations sur les transactions faites à travers le système TARGET2 à partir desquelles ils savent reconstruire avec grande précision l'activité de prêt dans la zone euro [AHH+16, GSV15]. Un acteur comme la BCE se retrouve face à un autre problème qu'est celui du grand nombre d'institutions supervisées. En effet, le problème d'inférence de paramètres d'OU multidimensionnel devient un problème à grande dimension, qui est sujet au fléau de la dimension [BvdG11]. Heureusement, nous pouvons faire une hypothèse de sparsité sur la structure du marché. En effet, toutes les banques sont généralement en relation qu'avec un nombre restreint d'autres banques [SC10]. L'estimation de paramètres sparses a attiré beaucoup d'attention dans le domaine de la statistique des hautes dimensions, voir [Gir14a]. Une des idées est celle d'utiliser la pénalisation et en particulier la pénalisation ℓ^1 [Tib96].

Question 3. *Peut-on améliorer l'inférence des paramètres de processus d'Ornstein-Uhlenbeck sous une hypothèse de sparsité ?*

Résumé

Les questions 2 et 3 seront les sujets de deux parties indépendantes de la thèse. La question 1 donne le contexte qui guide nos réponses. Notre travail étant sur l'estimation de paramètres de processus d'Ornstein-Uhlenbeck, nous commençons par présenter quelques propriétés de ces processus et rappeler des résultats fondamentaux sur l'estimation de leurs paramètres.

1 Estimation et propriétés des processus d'Ornstein-Uhlenbeck

Définition et propriétés du processus d'Ornstein-Uhlenbeck À travers la thèse, soit $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P}\right)$ un espace filtré et $(W_t)_{t\in\mathbb{R}^+}$ un mouvement Brownien q-dimensionnel adapté à la filtration $\mathcal{F}$. Pour un entier $d \geq 1$, nous définissons le

processus d'Ornstein-Uhlenbeck d-dimensionnel comme la solution de l'Équation Différentielle Stochastique (EDS) suivante

$$\mathrm{d}X_t = -\boldsymbol{A}X_t\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t, \tag{1}$$

où la valeur initiale X_0 est une variable aléatoire indépendante de W.

Le processus d'Ornstein-Uhlenbeck a été introduit en dimension $d = 1$ afin de modéliser le la vitesse de particules de pollen, comme dans le mouvement Brownien physique, soumises à des frictions [UO30]. Celles-ci créent un point d'attraction pour le processus, qui est 0 dans notre cadre. Depuis, le modèle a attiré une attention intense et a trouvé des applications dans de nombreux domaines. En particulier en finance, le modèle d'Ornstein-Uhlenbeck est utilisé pour modéliser les variations de taux d'intérêt et est connu sous le nom de modèle de Vasicek [Vas77]. Le modèle exprime que le taux d'intérêt peut varier à court terme, mais revient à long terme à une valeur habituelle. Le processus d'Ornstein-Uhlenbeck a aussi été utilisé pour le trading par paires [LL15]. Cette pratique consiste à exploiter des différences de prix entre actifs qui devraient évoluer ensemble d'après le trader. La différence de prix est alors modélisée par un OU, ce qui permet de quantifier la volatilité et la vitesse de retour à la moyenne. Aussi en mathématiques financières, le modèle de COX-Ingerol-Ross [CIR85] est une évolution proche du modèle d'Ornstein-Uhlenbeck. En neurosciences, l'OU est utilisé pour modéliser la tension de membranes [LL09]. Il a aussi été utilisé pour des modèles de dynamique populations [Bla14].

Dans les applications que nous venons de citer, l'Ornstein-Uhlenbeck tient son intérêt de sa propriété de retour à la moyenne. Ceci est obtenu en dimension $d = 1$ en choisissant un réel strictement positif à la place de $\boldsymbol{A}$ dans l'équation (1). Pour des dimensions pus grandes, l'hypothèse équivalent est celle d'avoir une matrice $\boldsymbol{A}$ dont le spectre est de parties réelles strictement positives. Nous faisons cette hypothèse dans l'ensemble de la these :

(H1) Le spectre de $\boldsymbol{A}$ a des parties réelles strictement positives :

$$a_0 := \min_{\lambda \in \mathrm{Sp}(\boldsymbol{A})} \Re(\lambda) > 0.$$

Dans ce cas, le processus est ergodique et possède une distribution stationnaire qui est une Gaussienne centrée, voir la Section 5.6 de [KS91]. En utilisant le lemme d'Ito, il est facile de montrer que

$$X_t = e^{-\boldsymbol{A}t}X_0 + \int_0^t e^{-\boldsymbol{A}(t-s)}\boldsymbol{\Sigma}\mathrm{d}W_s. \tag{2}$$

Le processus X_t est Gaussien est nous calculons sa moyenne et variance a tout instant t en utilisant l'Équation (2) au-dessus :

$$X_t \sim \mathcal{N}\left(e^{-\boldsymbol{A}t}X_0, \boldsymbol{C}_t\right) \tag{3}$$
$$\boldsymbol{C}_t = \int_0^t e^{-\boldsymbol{A}(t-s)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top(t-s)}\mathrm{d}s.$$

Il est souhaitable que X_t ait une distribution non-dégénérée, et donc que $\boldsymbol{C}_t$ soit inversible. Pour ceci, nous ajoutons une nouvelle hypothèse :

(H2) La matrice $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ est inversible..

À partir des Équations (3) et (4), nous voyons pourquoi est-ce qu'il est nécessaire que $\boldsymbol{A}$ ait un spectre aux parties réelles strictement positives. En effet, il s'agit de la condition nécessaire à ce que $e^{-\boldsymbol{A}t}$ converge vers 0, ce qui donne d'un côté

$$\mathbb{E}\left[X_t \mid X_0\right] \xrightarrow{t\to+\infty} 0$$

et de l'autre implique la convergence de $\boldsymbol{C}_t$. Cependant, contrairement au cas uni-dimensionnel, la matrice $\boldsymbol{A}$ peut avoir des entrées négatives. Écrivons donc l'Équation (1) comme elle s'applique à une coordonnée i de X :

$$\mathrm{d}X_t^i = -\sum_{j=1}^{d} \boldsymbol{A}^{ij}X_t^j\mathrm{d}t + (\boldsymbol{\Sigma}\mathrm{d}W_t)^i.$$

Certaines entrées parmi les $\boldsymbol{A}^{ij}$ seront égales à 0, ce que nous interprétons comme une absence d'influence de la composante j sur la composante i. Nien sûr, cette interprétations a ses limites. Par exemple i et j peuvent être liées à cause d'une matrice $\boldsymbol{\Sigma}$ non-diagonale ou encore à travers une troisième composante.

Les entrées non-nulles de $\boldsymbol{A}$ seront interprétées comme des influences, qui peuvent être positives ou négatives. À cause du signe moins dans l'Équation (1), les négatives entrées de $\boldsymbol{A}$ correspondent aux influences positives et *vice-versa*. Nous appelons une influence positive la situation où la valeur au-dessus de 0 de j pousse i à croître, et à décroître si i est au-dessous de 0. Une influence négative fonctionne de manière opposée.

Ceci participe à la richesse du cadre des processus d'Ornstein-Uhlenbeck multi-dimensionnels. Dans le cas uni-dimensionnel, à tout moment le processus va en espérance se rapprocher de 0. Dans le cas multi-dimensionnel, la variation espérée dépend de la valeur du vecteur entier. Celui-ci peut prendre des valeurs qui pousseront une composante à s'éloigner de 0. Par contre, sous l'hypothèse (H1), à

long terme le processus et donc toutes ses composantes sont attirés par 0 pour finir par être distribués selon la loi stationnaire que nous avons mentionné auparavant. En effet, observons que $\boldsymbol{C}_t$ converge vers une limite lorsque $t \to +\infty$:

$$\boldsymbol{C}_t = \int_0^t e^{-\boldsymbol{A}t}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top t}\mathrm{d}s \xrightarrow{t\to+\infty} \int_0^{+\infty} e^{-\boldsymbol{A}s}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top s}\mathrm{d}s =: \boldsymbol{C}_\infty.$$

La limite $\boldsymbol{C}_\infty$ est la matrice de covariance de la distribution stationnaire. Nous travailleront ans cette thèse sous l'hypothèse que l'initialisation du processus se fait selon cette distribution :

(H3) La valeur initiale X_0 est une variable Gaussienne centrée de matrice de covariance $\boldsymbol{C}_\infty$:

$$X_0 \sim \mathcal{N}\left(0, \boldsymbol{C}_\infty\right).$$

La densité de la distribution stationnaire $\mathcal{N}\left(0, \boldsymbol{C}_\infty\right)$ sera notée μ_∞.

Nous faisons cette hypothèse pour deux raisons. Premièrement, les applications que nous considérons, comme le prêt interbancaire ou la modélisation de prix d'actions, sont essentiellement homogènes en temps, au sens où le début de l'observation ne diffère pas des autres points de l'observation. Il est alors naturel de supposer que le système est arrivé à stationnarité, ce qui impose X_0 suivant la distribution stationnaire. Deuxièmement, l'impact de la valeur initial est limité à cause de la décroissance exponentielle du facteur précédant X_0 dans l'Équation (3). Après un court laps de temps, le processus est pratiquement impossible à distinguer d'un processus stationnaire. Ainsi, l'hypothèse de stationnarité n'a pas de coût important mais permet d'obtenir des résultats plus clairs et plus concis. Par exemple, nous avons pour tout $t \geq s$:

$$\mathbb{C}\mathrm{ov}\left(X_t, X_s\right) = e^{-\boldsymbol{A}(t-s)}\mathbb{C}\mathrm{ov}\left(X_s\right) = e^{-\boldsymbol{A}(t-s)}\boldsymbol{C}_\infty.$$

Observons enfin que l'ensemble de processus d'Ornstein-Uhlenbeck est stable par transformation linéaire, au sens où l'application d'une transformation linéaire de $\mathbb{R}^d$ au processus X donne un nouveau OU. Prenons par exemple une matrice inversible $\boldsymbol{L}$:

$$\mathrm{d}(\boldsymbol{L}X_t) = -(\boldsymbol{L}\boldsymbol{A}\boldsymbol{L}^{-1})(\boldsymbol{L}X_t)\mathrm{d}t + (\boldsymbol{L}\boldsymbol{\Sigma})\mathrm{d}W_t.$$

Grâce à cette observation, nous pouvons élargir le cadre d'application de certains résultats d'un processus X à des transformations linéaires de X.

Inférence de paramètres. Nous appelons inférence de paramètres le problème d'estimation des valeurs de $\boldsymbol{A}, \boldsymbol{\Sigma}$ à partir de l'observation d'une trajectoire unique de X. Nous illustrerons ceci par le cas fondamental de l'observation complète d'une

trajectoire sur l'intervalle $[0, T]$, présentée dans [LB76]. Danse ce cas, l'estimation de $\mathbf{\Sigma}$ est résolue rapidement. En calculant l'intégrale des variations quadratiques, nous retrouvons exactement la valeur de $\mathbf{\Sigma}\mathbf{\Sigma}^\top$:

$$\frac{1}{T}\int_0^T \langle \mathrm{d}X_t, \mathrm{d}X_t\rangle = \mathbf{\Sigma}\mathbf{\Sigma}^\top.$$

Ce calcul peut être fait pour tout intervalle de longueur $T > 0$ et n'a pas d'erreur. En termes probabilistes, les distributions de probabilités obtenues avec $\mathbf{\Sigma}\mathbf{\Sigma}^\top \neq \mathbf{\Sigma}'\mathbf{\Sigma}'^\top$ sont mutuellement singulières, voir [Jac01, LB77]. Par contre la matrice $\mathbf{\Sigma}$ n'est pas identifiable car la fonction $\mathbf{\Sigma} \to \mathbf{\Sigma}\mathbf{\Sigma}^\top$ n'est pas injective. Pour cette raison, nous ne chercherons désormais qu'à estimer la matrice $\mathbf{\Sigma}\mathbf{\Sigma}^\top$.

Afin d'estimer $\boldsymbol{A}$, l'approche habituelle consiste à construire le ration de vraisemblance en utilisant le théorème de Girsanov [LB76, KS91] :

$$\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}W} = \exp\left(-\int_0^T (\boldsymbol{A}X_t)^\top (\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}\mathrm{d}X_t - \frac{1}{2}\int_0^T (\boldsymbol{A}X_t)^\top (\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}\boldsymbol{A}X_t\mathrm{d}t\right).$$

Sa maximisation donne le Estimateur du Maximum de Vraisemblance (MLE de l'anglais ***Maximum Likelihood Estimator***) :

$$\widehat{\boldsymbol{A}}^{MLE} := \left(\int_0^T \mathrm{d}X_t X_t^\top\right)\left(\int_0^T X_t X_t^\top \mathrm{d}t\right)^{-1}.$$

Le MLE converge vers la vraie valeur du paramètre $\boldsymbol{A}_0$ et on a la convergence en loi suivante, voir [Jac01] :

$$\sqrt{T}(\mathrm{vec}\ \widehat{\boldsymbol{A}}^{MLE} - \mathrm{vec}\ \boldsymbol{A}_0) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{-1} \otimes \mathrm{Id}\right).$$

La vitesse de convergence en T est $1/\sqrt{T}$. Il s'agit d'une vitesse optimale étant donné que nous avons da propriété LAN (***Local Asymptotic Normality***) avec vitesse $1/\sqrt{T}$ et matrice d'information de Fischer $\boldsymbol{C}_\infty \otimes \mathrm{Id}$. Ce résultat d'optimalité et la popularité de cette méthode en font une référence à laquelle il est naturel de comparer des méthodes concurrentes.

2 Estimation à partir d'un graphe stochastique

Dans cette partie, nous générons un graphe stochastique à partir d'un processus d'Ornstein-Uhlenbeck. Définissons par exemple

$$Y_t^{ij} := \mathbb{1}_{X_t^i - X_t^j \geq 1}, \tag{4}$$

alors Y_t^{ij} peut être vu comme l'adjacence entre deux nœuds au temps t. Ainsi Y_t est la matrice d'adjacence d'un graphe dirigé avec d nœuds et nous interprétons Y comme un processus de graphes stochastiques. La définition de l'Équation (4) est particulièrement adaptée aux modèles de prêt interbancaire. En effet, dans les modèles de [CFS15, FI13], les flux de réserves sont proportionnels aux déséquilibres de réserves. La condition Y_t^{ij} sera alors vue comme un déséquilibre important de réserves au profit de i et sera accompagné dans le modèle par un grand flux de réserves de i vers j. Pour d'autres modèles et applications, la définition de l'Équation (4) peut être facilement élargie à un cadre plus général. Pour ceci, pour toute paire de nœuds (i, j), considérons S^{ij} un sous-ensemble de $\mathbb{R}^d$

$$Y_t^{ij} := \mathbb{1}_{X_t \in S^{ij}}. \tag{5}$$

La notion de graphe stochastique est donc sur le croisement entre les processus stochastiques et le graphes aléatoires.

Graphes aléatoires La notion de graphe aléatoire existe depuis [ER59]. Le modèle d'Erdös-Renyi, considéré généralement comme le modèle le plus basique de graphe aléatoire, suppose que toutes les adjacences sont des variables aléatoires de Bernouilli indépendantes de paramètre p, où p est fixé pour toutes les adjacences. Le modèle a attiré une attention importante qui a fait cependant apparaître certains défauts du modèle. En effet, le modèle ne reproduit pas les caractéristiques habituelles des graphes observés dans le monde réel.

Étant donné un graphe aléatoire, la question de savoir s'il ressemble à un graphe réel est difficile et toute réponse sera en partie subjective [CF06]. Afin de comparer deux graphes, la stratégie habituelle était de comparer des ***traits*** caractéristiques [LCK+10]. Un exemple de trait est la densité du graphe, c'est-à-dire le rapport du nombre de liens existants et du nombre de liens possibles, qui est $\binom{d}{2}$ dans le cas d'un graphe non-dirigé sans boucles. Ce trait peut être reproduit par le modèle d'Erdös-Renyi, au sens suivant : étant donné un objectif de densité δ, en choisissant $p = d$ la densité du graphe aléatoire sera d'espérance δ. Un exemple de trait plus complexe est la ***distribution de degrés***, c'est à dire la fréquence des degrés parmi les nœuds. Les graphes d'Erdös-Renyi ne peuvent que reproduire des distributions binomiales, voir [Bol01]. Pour obtenir une distribution de degrés donnée, il est possible d'utiliser par exemple le modèle de Chung-Lu [CL02].

Il est possible de continuer ce raisonnement et chercher les modèles qui possèdent des traits de plus en plus complexes. Étant donnés un ensemble de traits souhaités non-concurrents, il est théoriquement possible de considérer un générateur de graphe aléatoire qui vérifie l'ensemble des traits. Pour ceci, il suffit de considérer une distribution quelconque, par exemple uniforme, sur l'ensemble de graphes restreint à ceux

qui vérifient l'ensemble des traits. Ceci n'est généralement pas possible en pratique, mais la question des traits souhaitables reste intéressante. Un effort considérable a été porté à la sélection de de ces traits, mais il ne reste pas clair à quel point l'ajout de nouveaux traits devient superflu, voir pour ceci la conclusion de [CF06]. Parmi les traits populaires nous notons :

- le diamètre du graphe, c'est à dire la distance maximale entre deux nœuds [Har69] ;
- coefficients de *clustering*, une mesure de densité du graphe au voisinage d'un nœud introduite dans [WS98] ;
- la distribution de valeurs propres du Laplacien du graphe, trait intimement liée au *clustering* spectral de graphes [vL07] ;
- *hop-plots*, qui représentent la proportion de nœuds atteignables en k sauts selon des liens du graphe, voir [CF06], et mesurent la propriété de *petit monde* [WS98].

Au lieu de faire en sorte que le modèle vérifie des traits particuliers, il est parfois souhaitable de chercher plutôt à obtenir un cadre intuitif. Par exemple, le modèle de Barabási-Albert [Bar99], ou encore de rattachement préférentiel s'inspire du mécanisme où *les riches deviennent plus riches* (*rich-get-richer*). Celui-ci est observé par exemple dans la croissance d'internet. En effet, les nouveaux sites internet ont tendance à renvoyer vers des sites qui sont déjà populaires. Dans le modèle de rattachement préférentiel, on part d'un graphe initial auquel on ajoute successivement des nœuds rattachés chacun à un nombre fixé de nœuds, la probabilité de chaque lien étant proportionnelle au degré du nœud cible. Ceci permet de modéliser la croissance d'un graphe.
Dans tous les modèles évoqués ci-dessus nous remarquons cependant que les liens sont statiques, au sens où ils ne sont ni détruits ni crées indépendamment des nœuds. Par exemple, le modèle de rattachement préférentiel, l'évolution du graphe ne repose que sur l'addition de nouveaux nœuds, les liens existant initialement ne disparaissent jamais. Nous nous intéresserons au cas où les graphe évoluera avec une nouvelle variable réelle que nous interpréterons comme le temps. Les objets résultant de ceci sont appelés *graphes évolutifs* [AŁL$^+$16, CZ16], *réseaux évolutifs* [BFG$^+$08] ou encore *réseaux temporels* [HS12, HS13]. Nous utiliserons le premier nom. Le sujet des graphes évolutifs a attiré une attention accrue dans la décennie passée. L'approche habituelle des réseaux consiste à les modéliser comme un graphe statique, mais ceci n'est pas réaliste dans certains cas. Par exemple, l'observation de grands réseaux, tels que internet ou les réseaux sociaux, se fait à travers l'utilisation de *crawlers web* ou APIs [AŁL$^+$16] qui ne peuvent observer à un instant donné qu'une partie du graphe. Ainsi, avant d'avoir fini l'observation complète du graphe, nous pouvons être sûrs que le graphe aura changé, d'où l'intérêt de graphes évolutifs. Ceux-cis n'ont pas de définition arrêtée. La

variable temps peut même être inclue dans le graphe [Kos09]. Les approches typiques peuvent cependant être classées dans deux groupes [HKW16]. Dans le premier, nous supposons que le graphe subit des changements que nous comptons. Ce décompte fait alors office d'une variable temps. Dans le deuxième, on suppose des observations régulières du graphe entre lesquelles le graphe peut changer à de nombreux endroits. Les deux approches mènent cependant à une variable temps assimilable aux entiers. Les temps non-entiers sont utilisés par exemple dans [CZ16].

Les travaux existants sur le sujet des graphes évolutifs partagent quelques points récurrents. Premièrement, l'objectif est souvent d'obtenir des algorithmes et méthodes efficaces pour calculer les traits du graphe évolutif [AŁL+16, BKMU12, CZS+07]. Deuxièmement, l'évolution du graphe n'est soit pas modélisée [BFG+08, CZ16] ou modélisée de manière simple, en particulier avec réarrangements indépendants de nœuds et liens.

Dans notre modèle, le graphe évolutif est une fonction d'un processus stochastique. Au lieu de calculer les traits du graphe, nous nous intéressons particulièrement à la dynamique du processus sous-jacent. Le graphe peut changer à des temps réels quelconques, qui sont d'ailleurs aléatoires. L'auto-covariance non-nulle du processus a comme suite que les changements du graphe ne sont pas indépendants. Notre travail contribue alors au domaine des graphes évolutifs en introduisant un nouveau modèle de graphe, avec changements dépensants à des temps aléatoires et en proposant une nouvelle question à poser sur l'observation d'un graphe évolutif, celle de s'intéresser à des variables qui peuvent guider les variations du graphe.

Estimation from partial observation Nous avons présenté au-dessus le problème de l'inférence de paramètres dans le cadre d'observation complète en temps continu. Nous avons résumé la méthode d'estimation par Maximum de Vraisemblance, en notant que d'autres méthodes existent aussi, voir [Kut04]. Bien que ce cadre soit intéressant dans une perspective théorique, il nest plus réaliste de considérer des observations discrètes. En général, ceci consiste à fixer un entier $n > 0$ et une suite de temps $(t_1, \dots t_n)$ qui sont des réels positifs, et de supposer ensuite que le processus n'est observé qu'à ces n points dans le temps [Sor97] avec $n \to +\infty$. Il est possible d'utiliser une suite quelconque de temps, comme dans [GCJ93], ou encore une suite de temps aléatoires comme dans [GCJ94] ou [ASM04]. Il est cependant au moins aussi naturel de considérer des observations régulièrement séparées dans le temps. Dans ce cas, nous fixons une durée $\Delta_n > 0$ puis choisissons $t_i = i\Delta_n$. Nous pouvons distinguer alors distinguer globalement 3 régimes.

- **Le régime haute-fréquence temps-fixé (HF-FT).** Nous fixons ici une durée d'observation $T > 0$ puis choisissons $\Delta_n = T/n$, de manière à ce que l'obser-

vation devienne de plus en plus dense dans $[0, T]$ lorsque $n \to +\infty$. On espère alors d'estimer les paramètres grâce à la précision grandissante de l'observation de la trajectoire. La limite $n \to +\infty$ est l'observation complète en temps continu sur intervalle fixé, que nous avons étudié au-dessus. Rappelons qu'alors l'estimation de $\boldsymbol{\Sigma}$ est possible, mais celle de $\boldsymbol{A}$ nécessite l'hypothèse $T \to +\infty$. Nous en déduisons que dans le régime (HF-FT) il sera au plus possible d'estimer le paramètre lié à la volatilité. En effet, la propriété LAMN est prouvé dans le cas homogène uni-dimensionnel dans [Doh87]. Le cas non-homogène multidimensionnel est traité dans [GCJ93]. De plus, les auteurs n'utilisent pas l'hypothèse d'observations régulières. Il la remplacent par une hypothèse plus faible, celle que la distribution empirique des temps d'observation converge en loi vers une distribution sur l'intervalle d'observation. Nous classons ceci dans le régime HF-FT car l'espérance de l'espacement en temps converge vers 0. De manière similaire, nous classons dans ce régime les résultats de [GCJ94] qui considèrent des temps d'observation aléatoires. Enfin, les vitesses minimax de convergence, pour les processus uni-dimensionnel, sont obtenus dans ![Hof99].

- **Le régime basse-fréquence temps-long (LF-LT)**. Nous fixons une durée entre observations à une constante $\Delta_n = \Delta$. La granularité de l'observation reste constante, mais nous supposons le temps d'observation $T = n\Delta \to +\infty$. Ainsi, nous espérons retrouver les paramètres grâce aux propriétés d'ergodicité du processus. L'observation ici forme alors une chaîne de Markov. Une première approche consiste à approcher les formules du cas continu par des calculs discrets, mais cette méthode crée des estimaturs inconsistants [Flo89], comme préciser par exemple dans [Ped95]. Au lieu de ceci, d'autres méthodes ont été proposées [Jac01], par exemple [Kes97] et [KS99] donnent des estimateurs consistents et vérifiant une propriété de normalité asymptotique.

- **Le régime haute-fréquence temps-long (HF-LT)**. Nous supposons ici que $\Delta_n \to +\infty$ mais qu'en même temps la durée de l'observation diverge $T = n\Delta_n \to +\infty$. Dans ce régime, il est possible d'exploiter et la granularité croissante - en effet, il suffit d'utiliser les premiers T/Δ_n points pour retrouver le régime HF-FT - et les propriétés ergodiques venant de $T \to +\infty$. L'estimation non-paramétrique du coefficient de diffusion est présentée pour le cas uni-dimensionnel dans [Flo93]. Encore pour les processus uni-dimensionnels [Kes97] présnete l'estimation paramétrique simultanée des paramètres de drift et de la diffusion. La propriété LAN est prouvée dans [Gob02]. Le cas des temps d'observation aléatoires est le sujet de [ASM04]. Ces observations vérifient cependant $\mathbb{E}[\Delta_n] \to 0$, ce qui nous permet de classer ces travaux dans le cadre HF-LT.

Nous souhaitons rappeler qu'en pratique, un pas de temps devra être fixé de même

que lea durée d'observation. En supposant que le pas de temps est petit, le régime HF-LT permet de quantifier simultanément l'impact de sa longueur et de la durée d'observation. Pour cette raison, nous plaçons notre travail dans le cadre du régime HF-LT.

Nous avons commencé cette partie en introduisant les graphes stochastiques. Le cadre présenté dans l'Équation (5) permet cependant d'étendre notre étude à de nombreux cas d'observation partielle. En effet, prenons l'exemple des observations qui sont limitées par la précision de leur affichage, comme un prix qui ne peut varier que d'au moins un centime. Nous pouvons interpréter un prix observer comme une fonction discrète d'un processus latent qui prend des valeurs dans $\mathbb{R}$. Dans l'équation (5), nous prenons l'hypothèse la plus forte d'observation binaire, mais celle-ci peut être contournée. Il suffit pour ceci de décrire une fonction discrète à l'aide de fonctions binaires. Par exemple, l'observation de $\mathbb{1}_{x\geq 0}$ et $\mathbb{1}_{x\geq 1}$ nous donne l'information sur 3 intervalles $]-\infty, 0[$,$[0, 1[$ and $[1, +\infty[$.

Une information binaire peut facilement représenter le dépassement d'un seuil. Pour des diffusions dans $\mathbb{R}$, il est naturel de placer ce seuil à zéro. Dans [Flo86, Flo88] il est supposé que le processus est observé en temps continu, mais que l'estimation ne se fait qu'à partir de la donnée des dépassements de 0 de plus de ϵ, qui est une constante positive. Il est ensuite montré que sous certaines hypothèses sur le paramètre de drift et sous l'hypothèse de paramètre de drift constant et connu, cette information est suffisante pour estimer le paramètre de drift. De plus la performance de l'estimateur ainsi construit est significative en comparaison au MLE, au sens où leur ratio d'informations de Fischer est autour de 0.618 pour l'estimation d'un processus d'Ornstein-Uhlenbeck.

Le même problème en temps continu est le sujet de [Flo87, Flo91]. Il y est de nouveau admis que le paramètre de la diffusion est connu, puis que le drift est une fonction $b(X_t, \theta)$ qui dépend d'un paramètre θ à estimer. Le papier se place dans deux régimes de discrétisation. Premièrement, un estimateur de θ est construit à partir de quasi-vraisemblances et il est montré qu'il est asymptotiquement efficace dans le régime LF-LT. Deuxièmement, il est prouvé que dans le régime HF-FT le nombre de changements de signes $N_\Delta(T)$ de la discrétisation par pas Δ vérifie

$$\sqrt{\pi\Delta/2}N_\Delta(T) \xrightarrow{L^2} L_t(0)$$

où $L_t(0)$ est le temps local du processus en zéro de X. Ce résultat est ensuite étend aux approximations continues du processus dans [Aza89]. Étant donné que le temps local est en espérance proportionnel à la densité stationnaire dans le cas ergodique [KS91], nous apprenons de ceci que nous devrions pouvoir estimer le pa-

ramètre de volatilité.

Nos contributions dans cette partie sont :

- nous démontrons des propriétés de *mixing* pour les processus d'Ornstein-Uhlenbeck ;
- nous donnons une méthode pour prouver la convergence L^2 de sommes de fonctions d'un processus d'Ornstein-Uhlenbeck ;]item nous caractérisons la convergence de statistiques de temps d'occupation, y compris avec un théorème de limite centrale en dimension $d = 1$;
- nous montrons la convergence L^2 d'une normalisation de la statistique de croisement de seuil, effectivement étendant les résultats de convergence vers le temps local de [Flo87, Flo91] mentionné ci-dessus au régime HF-LT ;
- nous montrons comment utiliser ces propriétés dans un problème d'inférence de paramtètres à partir d'un graphe stochastique généré par un processus d'Ornstein-Uhlenbeck ;
- nous conjecturons une vitesse optimale de convergence L^2 et la validité d'un théorème de limite centrale pour la statistique de croisement de seuil.

2.1 Estimation de paramètres d'un processus d'Ornstein-Uhlenbeck générant un graphe stochastique

Notre méthode d'estimation consiste en l'utilisation de deux quantités : le temps passé par Y_t à être égal à 1 et la fréquence de saut entre 0 et 1. Cette approche est similaire à celle de [Flo91]. Nous montrons plusieurs propriétés de convergence pour les deux statistiques et nous concluons en les appliquant à l'inférence de paramètres. Comme dit auparavant, nous travaillons sous le régime HF-LT. Un problème typique de cette approche est que les temps d'observation ne sont pas partagés pour différentes valeurs de n à cause de la variation de Δ_n. En effet, la chaîne de Markov $(X_{k\Delta_n}, k \geq 0)$ dépend de n à travers Δ_n, ce qui nous empêche d'appliquer les propriétés de *mixing* valides pour les chaînes de Markov. Ceci nous a poussé à examiner la vitesse avec laquelle l'auto-corrélation du processus décroît avec la distance en temps. Le Théorème suivant exprime que cette vitesse de décorrélation est exponentielle et est fondé sur une inégalité dite de Gebelein [Jan97] ou encore de Lancaster.

Theorem 1 (Théorème I.1). *Supposons (H1) – (H3). Il existe une constante $C_{(6)}$, dépendant uniquement de la matrice de covariance $\boldsymbol{C}_\infty$ de la distribution stationnaire, telle que pour*

tout $t \geq s \geq 0$ *et pour toutes fonctions* $\varphi : \mathcal{C}^0([0,s],\mathbb{R}^d) \to \mathbb{R}, \phi : \mathcal{C}^0([t,+\infty),\mathbb{R}^d) \to \mathbb{R}$ *telles que* φ, ϕ *soient de carré intégrables pour la distribution de* X*, on ait*

$$|\mathbb{C}\mathrm{ov}\left(\varphi\left((X_u)_{u\leq s}\right), \phi\left((X_v)_{v\geq t}\right)\right)| \leq C_{(6)} e^{-a_0|t-s|} \sqrt{\mathbb{V}\mathrm{ar}\left(\varphi\left((X_u)_{u\leq s}\right)\right) \mathbb{V}\mathrm{ar}\left(\phi\left((X_v)_{v\geq t}\right)\right)} \tag{6}$$

où a_0 *est défini à l'Équation 1 de l'hypothèse (H1).*

Le Théorème 1 ci-dessus nous permet de prouver le corollaire suivant.

Corollary 1 (Corollaire I.1). *Soit* $g : \mathbb{N}\times\mathbb{N}\times\mathcal{C}^0([0,1],\mathbb{R}^d) \to \mathbb{R}$ *une fonction mesurable telle que* $\mathbb{E}\left[g(k,n,(X_s)_{k\Delta_n\leq s\leq (k+1)\Delta_n})^2\right] < +\infty$ *pour tout* $k,n \in \mathbb{N}$*. Pour* $n \in \mathbb{N}$ *notons*

$$v_n^2 = \sup_{k<n} \mathbb{V}\mathrm{ar}\left(g(k,n,(X_s)_{k\Delta_n\leq s\leq (k+1)\Delta_n})\right),$$
$$\xi_k^{(n)} = \sqrt{\frac{\Delta_n}{n}} g(k,n,(X_s)_{k\Delta_n\leq s\leq (k+1)\Delta_n}).$$

Il existe alors une constante $C_{(7)}$*, ne dépendant que des paramètres* $\boldsymbol{A}, \boldsymbol{\Sigma}$ *du modèle, telle que :*

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=0}^{n-1} \xi_k^{(n)}\right) \leq C_{(7)} v_n^2. \tag{7}$$

2.1.1 Statistique de temps d'occupation

Nous fixons ici un sous-ensemble $S \subset \mathbb{R}^d$ puis définissons la statistique de temps d'occupation

$$\mathrm{OT}_n^S = \frac{1}{n}\sum_{k=0}^{n-1} Y_{k\Delta_n}^S = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X_{k\Delta_n}\in S}.$$

Vu l'hypothèse (H3), qui implique la stationnarité, nous obtenons facilement $\mathbb{E}\left[\mathrm{OT}_n^S\right] = \mu_\infty(S) := \int_S \mu_\infty(\mathrm{d}x)$. Ensuite, afin de prouver une convergence L^2, nous nous intéressons à la variance. Au vu du Corollaire 1 on a :

Theorem 2 (Théorème I.2). *Supposons (H1) – (H3). Pour tout ensemble mesurable* $S \subset \mathbb{R}^d$*,* OT_n^S *converge vers* $\mu_\infty S$ *dans* L^2 *et*

$$\mathbb{E}\left[(\mathrm{OT}_n^S - \mu_\infty S)^2\right] = O\left(\frac{1}{n\Delta_n}\right).$$

Le Théorème 2 ci-dessus donne une borne supérieure pour la convergence L^2. Pour assurer l'optimalité du résultat et de mieux caractériser l'erreur d'estimation, nous souhaitons prouver un Théorème de Limite Centrale. C'est exactement ce qu'on fait dans le Théorème 3 qui suit, en dimension $d = 1$, d'où $(\boldsymbol{A}, \boldsymbol{C}_\infty) = (a, v_\infty) \in \mathbb{R}^2$, et avec $S = [1, +\infty[$.

Theorem 3 (Théorème I.3). *Supposons (H1) – (H3). Lorsque* $n \to +\infty$*, on a*

$$\sqrt{n\Delta_n}\left(\mathrm{OT}_n^{[1,+\infty[} - \mu_\infty([1,+\infty[)\right) \xrightarrow{d} \mathcal{N}\left(0, \mu_\infty(\sigma^2 F'^2)\right)$$

où F *vérifie*

$$F'(x) = \frac{2}{\sigma^2}\frac{N\left(\frac{x\wedge 1}{\sqrt{v_\infty}}\right) - N\left(\frac{1}{\sqrt{v_\infty}}\right)N\left(\frac{x}{\sqrt{v_\infty}}\right)}{\mu_\infty(x)} \in \left[0, 2\sqrt{\frac{\pi}{a\sigma^2}}\right].$$

Nous nous restreignons ici à $d = 1$ afin de pouvoir résoudre explicitement une équation de Poisson. Le choix de S est motivé par l'application au marché de prêt interbancaire exposée dans l'Équation (4). Ce choix nous permet aussi d'obtenir une écriture explicite pour la variance de la Gaussienne. Une légère modification de la preuve permet d'obtenir des résultats analogues pour des ensembles S différents.

Enfin, nous confirmons la vitesse de convergence dans des expériences numériques de grande échelle, que nous présentons dans la Section 4 du Chapitre I. Nous y estimons les espérances et écarts-types des deux statistiques en utilisant une méthode de Monte-Carlo habituelle.

2.1.2 Statistique de croisements

Nous définissons la statistique de croisements

$$\mathcal{C}_n = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{Y^S_{k\Delta_n} \neq Y^S_{(k+1)\Delta_n}}$$

où nous choisissons pour S le demi-espace $S = \{x : x^1 \geq 1\}$. Comme expliqué précédemment, nous pouvons obtenir des résultats similaires pour d'autres demi-espaces ne contenant pas 0 en utilisant des transformations linéaires de $\mathbb{R}^d$. En utilisant le Corollaire 1, nous démontrons la convergence L^2 suivante :

Theorem 4 (Théorème I.4). *Supposons (H1) – (H3). Supposons aussi que* $n\Delta_n^{3/2} \to +\infty$*. Alors on la convergence :*

$$\frac{\mathcal{C}_n}{\sqrt{\Delta_n}} \xrightarrow{L^2} 2\sqrt{\frac{\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{11}}{2\pi}}\mu_{\boldsymbol{C}^{11}_\infty}(1),$$
$$\mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right) = O\left(\frac{1}{n\Delta_n^{3/2}}\right),$$

où $\mu_{\boldsymbol{C}^{11}_\infty}(1)$ *est la densité d'une variable Gaussienne centrée de variance* $\boldsymbol{C}^{11}_\infty$*.*

Le résultat est confirmé dans des simulations Monte Carlo. Cependant, bien que les résultats sont cohérents avec notre prévision de la limite, les simulations suggèrent que la vitesse de convergence excède celle que nous donnons. En effet, nous faisons la conjecture que la vitesse de convergence L^2 est $1/n\Delta_n$ au lieu de $1/n\Delta_n^{3/2}$. Il semble aussi que la normalisation en utilisant cette vitesse fait apparaître une loi Gaussienne, nous conjecturons donc aussi la validité d'un Théorème de Limite Centrale pour $\mathcal{C}_n$.

2.1.3 Application à l'inférence de paramètres

En utilisant la statistique de temps d'occupation, nous pouvons estimer les mesures d'ensembles S par rapport à la distribution stationnaire, une fonction de $\boldsymbol{C}_\infty$. En choisissant plusieurs ensembles S, par exemple les ensembles $S^{ij} = \{x : x^i \geq 1, x^j \geq 1\}$, on peut obtenir une situation où le vecteur de mesures espérées $\mathbb{E}\left[(\mathrm{OT}_n^{S^{ij}})\right]_{ij}$ est en bijection avec l'ensemble de covariances $\boldsymbol{C}_\infty$. L'inversion de cette bijection appliquée au vecteur de mesures estimées est un candidat naturel pour estimer $\boldsymbol{C}_\infty$. Nous prouvons aussi que $\mu_\infty(S)$ est une fonction continue de $\boldsymbol{C}_\infty$, ce qui confirme que cet estimateur a la bonne limite.

La matrice de covariance $\boldsymbol{C}_\infty$ est cependant une fonction non-bijective de $\boldsymbol{A}$ et $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. Nous utilisons donc la statistique de croisement et en particulier le fait que le Théorème 4 converge vers un produit de deux termes, d'un côté une fonction de $\boldsymbol{C}_\infty$ et de l'autre une fonction de $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. En remplaçant $\boldsymbol{C}_\infty$ par son estimation obtenue grâce à la statistique de temps d'occupation, on obitient facilement une estimation de termes de $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ indépendante de $\boldsymbol{C}_\infty$, ce qui permet d'estimer ensuite $\boldsymbol{A}$.

2.2 Propriétés du graphe stochastique

Le processus d'Ornstein-Uhlenbeck multi-dimensionnel peut-être utilisé afin de modéliser le prêt interbancaire, mais il faut imposer alors que les flux de réserves résultants conservent la somme totale de réserves. Cette condition s'écrivant $\mathbb{1}^\top \boldsymbol{A}_0 = 0$, la matrice $\boldsymbol{A}_0$ n'est pas invertible. Pour éviter cette dégénérescence, nous décrivons la dynamique des d réserves avec $d-1$ déviations de la moyenne, qui elles suivent un OU modifié, et avec la moyenne décrite comme processus indépendant. Pour le modèle de Carmona-Fouque [CFS15, FI13], la dynamique des déviations utilise un paramètre de drift scalaire.

En simulant de petits réseaux de prêt interbancaire, suivants le modèle de Carmona-Fouque ou encore un modèle sparse avec institution centrale, nous observons que les flux de réserves observées dans Y sont souvent très différentes de la vraie structure du réseau. En moyennant les observations sur un intervalle, il est possible de savoir

quels sont les liens dans le réseau, mais il n'est pas simple d'en déduire la structure du système de prêt, telle que décrite par la matrice $\boldsymbol{A}_0$.

Pour des modèles de réseaux de prêt plus larges, nous analysons les propriétés des graphes stochastiques du point de vue des graphes évolutifs, en suivant l'approche de [BFG+08]. Premièrement, nous regardons l'évolution de mesures de graphe typiques dans le temps. En particulier, la densité du graphes évolue continument autour de sa moyenne, de façon similaire à un processus d'Ornstein-Uhlenbeck uni-dimensionnel. Les distributions de degré ont des queues à décroissance rapide, mais ceci n'empêche pas l'existence habituelle de nœuds connectés à la majorité du graphe. Le graphe est typiquement très concentré, avec de petits diamètres et distances moyennes, mais à faible *clustering*. Deuxièmement, nous regardons des propriétés de liens dans le temps. En particulier, la durée d'existence des liens est généralement petite, avec une distribution à queue à décroissance rapide. Troisièmement, nous regardons la connectivité du nœud central dans le modèle sprase. Son degré est largement inférieur à ceux observés pour les autres nœuds.

3 Estimation sparse

Dans cette partie, nous supposons que l'observation d'une trajectoire se fait de manière complète sur l'intervalle $[0, T]$. Dans ce cadre, comme nous l'avons précisé auparavant, la matrice $\boldsymbol{\Sigma}$ peut être considérée comme donnée. Pour cette raison nous simplifions davantage le problème en supposant $\boldsymbol{\Sigma} = \mathrm{Id}$, ce qui permet entre autres d'affirmer que l'hypothèse (H2) est satisfaite et donc superflue. Enfin, nous ajoutons une hypothèse de sparsité par lignes :

(H4) Pour un $s \leq d$ fixé, la matrice $\boldsymbol{A}$ s-sparse par lignes, c'està dire telle que

$$\forall i \leq d, \|\boldsymbol{A}^{i\bullet}\|_0 \leq s,$$

où $\boldsymbol{A}^{i\bullet}$ est la i-ème ligne de $\boldsymbol{A}$, vue comme un vecteur. L'hypothèse (H4) signifie que les coordonnées de X sont influencées dans la dynamique par au plus s autres coordonnées :

$$\mathrm{d}X_t^i = -\sum_{j:\boldsymbol{A}^{ij}\neq 0} \boldsymbol{A}^{ij} X_t^j \mathrm{d}t + \mathrm{d}W_t^i.$$

L'estimation sparse est un problème qui a attiré une attention importante en statistiques [BvdG11, Gir14a]. Le but est de trouver le meilleur estimateur qui vérifie une certaine condition de sparsité, c'est à dire une restriction sur le nombre de valeurs non-nulles de la matrice. Malheureusement, la résolution de ce problème a une complexité numérique extrêmement grande. La raison de ceci est la non-convexité de l'ensemble de paramètres à considérer. En effet, l'ensemble de matrices qui vérifient

l'hypothèse (H4) est l'union de tous les sous-espaces linéaires, définis par la sélection de s indices par ligne. Pour résoudre le problème, il faudrait effectuer l'estimation restreinte à chaque sous-espace, puis comparer les estimateurs. Le problème est que le nombre de ces sous-espaces est $\binom{d}{s}^d$, ce qui rend le problème intraitable pour d grand et même petit s.

L'estimateur Lasso. Une approche alternative est de réduire le nombre de valeurs non-nulles de $\boldsymbol{A}$ en les pénalisant, c'est à dire en ajoutant un terme supplémentaire dans la fonction dont la minimisation donne l'estimateur. Il serait naturel d'utiliser la pseudo-norme $\|\boldsymbol{A}\|_0 := \#\{(i,j) : \boldsymbol{A}^{ij} \neq 0\}$ comme pénalisation. Le problème es que cette pseudo-norme n'est pas convexe, ce qui rend le problème à nouveau intraitable. La stratégie de relaxation convexe mène alors à remplacer $\|\boldsymbol{A}\|_0$ par $\|\boldsymbol{A}\|_1 := \sum_{ij} |\boldsymbol{A}^{ij}|$. L'estimateur que nous obtenons est l'estimateur Lasso [Tib96] :

$$\widehat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} -\frac{1}{T} \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}W} + \lambda \|\boldsymbol{A}\|_1. \tag{8}$$

Nous plaçons un paramètre λ devant la norme ℓ^1 qui sert ensuite à régler l'intensité de la pénalisation, ce qui permet de trouver un compromis entre sparsité et accord avec les données.
L'estimation lasso et plus généralement l'estimation sparse à l'aide de pénalisation convexe a été le sujet de recherches approfondies dans le passé récent [BvdG11, FBO12, Gir14a]. Un sujet d'application typique est la regression Gaussienne. Dans ce cas, nous considérons le problème consistant à estimer un paramètre vectoriel a étant donnés une série d'observations y et une série de variables vectorielles qui forment une matrice $\boldsymbol{X}$, et en supposant que

$$y = \boldsymbol{X}a + w$$

où w est un vecteur Gaussien centré modélisant le bruit. Ce modèle est en quelque sorte analogue au modèle d'Ornstein-Uhlenbeck présenté dans l'équation (1). En estimation sparse, le modèle est généralement accompagné d'une hypothèse de sparsité qui s'applique au vecteur a. Ceci consiste par exemple à supposer que le nombre d'entrées non-nulles est majoré par un entier fixé s. Les ***propriétés d'oracle*** sont des propriétés des erreurs d'estimation qui dépendent de l'hypothèse de sparsité. Par exemple, il peut s'agir d'une majoration de la norme de l'erreur d'estimation par une fonction de s. Dans le cas de la régression linéaire, ces propriétés d'oracle sont montrées pour le Lasso dans [BTW07] et [BRT09]. Des développement proches du Lasso ont été proposés pour l'adapter à des cas nécessitants plus de flexibilité. Par exemple, le Lasso applique une pénalisation égale à tous les paramètres. Afin d'introduire de

la sparsité structurée, comme de la sparsité par blocs, un développement populaire est le Lasso par groupes (*Group Lasso*), qui a aussi e´té utilisé pour l'estimation de paramètres à rang réduit.

L'estimation sparse à l'aide du Lasso a aussi été appliquée à l'étude de série temporelles. En particulier, il a été utilisé pour l'estimation de paramètres d'un Vecteur Autoregressif (VAR). La discrétisation d'un processus d'Ornstein-Uhlenbeck peut être vu comme un série temporelle AR(1), ce qui permet de voir un processus VAR comme une version discrète d'un processus d'Ornstein-Uhlenbeck. Des majorations d'erreurs de l'estimation Lasso dans le cas des VAR sont donnés par [BM15]. La méthode de preuve utilise des densités spectrales et dépend fortement de l'hypothèse d'observation discrète. En effet, celle-ci permet de voir un processus d-dimensionnel observé n fois comme une variable Gaussienne de dimension $n \times d$. Ensuite, l'estimation Lasso dans le cadre des processus d'Ornstein-Uhlenbeck est considéré dans [Sok13], mais le travail ne donne pas de résultats théoriques comme des majorations d'erreurs. Notre travail donne ces majorations et utilise une hypothèse de sparsité différente qui s'applique indépendamment à toutes les lignes de $\boldsymbol{A}$.

L'estimateur Lasso Adaptif. Malheureusement, l'estimation Lasso possède ses défauts. En effet, l'estimation sparse cherche à atteindre deux objectifs : d'estimer correctement le vrai paramètre et d'obtenir le bon niveau de sparsité. Le premier objectif signifie que nous cherchons à avoir un estimateur non-biaisé qui converge avec une vitesse optimale. Le deuxième signifie que nous voulons que l'estimateur assigne une valeur nulle aux zéros de la vraie matrice paramètre, de sorte qu'on obtienne asymptotiquement une classification parfaite des valeurs nulles et non-nulles. Dans le cas de la régression Gaussienne, nous savons d'après [Zou06] qu'il n'est pas possible d'obtenir les deux propriétés pour le Lasso : les plages de valeurs de λ qui donnent une estimation consistante et celles qui assurent une sélection de variables consistante sont incompatibles. Pour ceci, l'article propose la méthode de Lasso Adaptatif, qui consiste à utiliser une norme ℓ^1 mais avec des poids inversement proportionnels au MLE.

Nos contributions dans cette partie sont les suivants :

- nous développons l'estimation sparse à l'aide du Lasso dans le cadre de temps continu ;
- nous produisons des résultats utilisants une hypothèse de sparsité par lignes ;
- nous créons une base théorique complète pour l'estimation sparse de processus d'Ornstein-Uhlenbeck, y compris des majorations de l'erreur d'estimation ;

- nous donnons une borne inférieure au sens minimax pour l'erreur d'estimation sous hypothèse de sparsité par lignes ;
- nous prouvons une propriété de valeurs propres restreintes (*Restricted Eigenvalue*) pour le processus d'Ornstein-Uhlenbeck ;
- nous présentons des propriétés d'optimalité asymptotique du Lasso Adaptatif.

3.1 Inférence sparse du drift d'un processus Ornstein-Uhlenbeck en grande dimension

Afin d'estimer le paramètre de drift $\boldsymbol{A}$ de l'équation (1) avec $\boldsymbol{\Sigma} = \mathrm{Id}$, nous considérons deux méthodes d'estimation reposant sur la pénalisation ℓ^1 : le Lasso et le Lasso Adaptatif. Nous considérons dans un premier temps le Lasso dans un cadre non-asymptotique. Nous prouvons en effet que dans ce cadre il est possible de donner des bornes sur les erreurs d'estimation qui s'appliquent avec grande probabilité. Le cadre non-asymptotique nous permet de mieux quantifier les effets de la sparsité s, de la dimension d et de la longueur d'observation T. L'avantage principal du Lasso Adaptatif est d'obtenir des résultats plus forts par rapport au Lasso dans le cadre asymptotique. C'est pour ceci que nous analysons placer l'analyse de cet estimateur dans le cadre asymptotique.
Commençons par écrire le problème à résoudre en termes de matrices :

$$\mathcal{L}_T(\boldsymbol{A}) := -\frac{1}{T}\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}W} = \mathrm{tr}\left[\boldsymbol{A}^\top \boldsymbol{\varepsilon}_T + \frac{1}{2}(\boldsymbol{A}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T(\boldsymbol{A}-\boldsymbol{A}_0)^\top\right] + \mathrm{Const}$$

$$\text{with} \quad \widehat{\boldsymbol{C}}_T = \frac{1}{T}\int_0^T X_T X_t^\top \mathrm{d}t \quad \text{and} \quad \boldsymbol{\varepsilon}_T = \frac{1}{T}\int_0^T \mathrm{d}W_T X_t^\top.$$

3.1.1 Résultats préliminaires

La convergence de tout estimateur construit à partir de la log-vraisemblance négative $\mathcal{L}_T$ repose sur le comportement des quantités matricielles $\widehat{\boldsymbol{C}}_T$ et $\boldsymbol{\varepsilon}_T$. En effet, $\boldsymbol{\varepsilon}_T$ peut être vu comme un terme de bruit qui doit être suffisamment petit, alors que $\widehat{\boldsymbol{C}}_T$ doit être suffisamment loin d'être dégénéré.

Dans le cadre de la regression linéaire Gaussienne, l'équivalent de $\widehat{\boldsymbol{C}}_T$ est la matrice de Gram, qui se calcule à partir de la matrice de conception (*design matrix*). Celle-ci étant donnée il devient nécessaire de prendre l'hypothèse que la matrice de Gram a ses valeurs propres, toutes positives, suffisamment loin de 0 afin d'assurrer des vitesses de convergence rapides. Au contraire, dans notre modèle, la matrice $\widehat{\boldsymbol{C}}_T$ est calculée à partir de la trajectoire X qui suit la dynamique du modèle de l'Équation (1).

Grâce à ceci, nous pouvons parler de propriétés aléatoires de $\widehat{\boldsymbol{C}}_T$. Grâce à l'ergodicité de X, la covariance empirique $\widehat{\boldsymbol{C}}_T$ convrge en probabilité vers $\boldsymbol{C}_\infty$. Comme la covariance $\boldsymbol{C}_\infty$ est définie positive, ce qui vient de l'hypothèse (H1), nous obtenons que pour de grandes valeurs de T, la matrice $\widehat{\boldsymbol{C}}_T$ aura aussi des valeurs strictement positives. Dans le théorème 5 ci-dessous, nous quantifions cette propriété, mais celle-ci nécessitera une hypothèse supplémentaire. Elle décrit la vitesse avec laquelle est-ce qu'on a convergence de $\widehat{\boldsymbol{C}}_T$ vers $\boldsymbol{C}_\infty$.

(H5) Il existe une fonction H, strictement positive sur $\mathbb{R}^+$, telle que pour tout vecteur u vérifiant $\|u\|_2 \leq 1$, on a :

$$\mathbb{P}\left[|u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u| \geq R\right] \leq 2\exp(-TH(R)).$$

Nous conjecturons la validité de cette hypothèse. En effet, pour des processus d'Ornstein-Uhlenbeck réversibles, qui correspondent ici au cas de matrice $\boldsymbol{A}_0$ symétrique, l'inégalité est une application d'un Théorème de [CG07]. Le théorème dit que, pour toute fonction réelle V, l'intégrale $T^{-1}\int_0^T V(X_t)\mathrm{d}t$ dévie de son espérance $\mu_\infty(V)$ de plus que tout seuil donné avec une probabilité qui décroît exponentiellement avec T. Malheureusement, nous n'avons pas réussi à étendre ce théorème au matrices $\boldsymbol{A}_0$ non-symétriques. Cependant, nous avons de bonnes raisons de croire que la propriété reste valide, vu que [HHMS93] montre que la vitesse de convergence de $\widehat{\boldsymbol{C}}_T$ vers $\boldsymbol{C}_\infty$ est asymptotiquement plus rapide pour $\boldsymbol{A}_0$ non-symétrique.

En utilisant cette hypothèse et une borne d'union qui permet de déduire une vitesse de convergence des valeurs propres à partir de l'hypothèse (H5), nous prouvons le théorème 5 ci-dessous. Celui-ci peut être vu comme une propriété de valeurs propres restreintes [BRT09], qui est habituellement une hypothèse dans des cadres tels que la régression Gaussienne [ZWJ14].

Theorem 5 (Théorème III.3). ***Supposons (H1), (H3) – (H5). Soit*** $s \leq d$ ***et*** $c_0 > 0$. ***Définissons*** $C(s, c_0) := \{u \in \mathbb{R}^d : \|u\|_1 \leq (1 + c_0)\|u_{|\mathcal{I}_s(u)}\|_1\}$ ***où*** $\mathcal{I}_s(u)$ ***est l'ensemble des indices des*** s ***plus grandes entrées de*** $|u|$. ***Soit*** $\epsilon_0 \in (0, 1)$ ***et supposons que***

$$T \geq T_0 := H\left(\frac{\kappa^2}{9(c_0+2)^2}\right)^{-1}\left(s\log\left(21d \wedge \frac{21ed}{s}\right) + \log\left(\frac{4}{\epsilon_0}\right)\right).$$

Alors,

$$\mathbb{P}\left[\inf_{u\in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa\right] \geq 1 - \frac{\epsilon_0}{2} \tag{9}$$

où $\kappa := \sqrt{\frac{\min \mathrm{Sp}(\boldsymbol{C}_\infty)}{2}}$.

Nous poursuivons avec une propriété qui décrit que $\boldsymbol{\varepsilon}_T$ est relativement petit avec grande probabilité.

Theorem 6 (Théorème III.8). *Supposons (H1) et (H3). Définissons pour $x > 0$:*

$$\theta(x,(X_t)) := \sqrt{4eT^{-1}|\mathrm{diag}\,(\widehat{\boldsymbol{C}}_T)|_\infty \Big(x + \log(2 + |\log T \mathrm{diag}\,(\widehat{\boldsymbol{C}}_T)|_\infty)\Big)},$$

où diag *est l'extraction de la diagonale d'une matrice et* log *s'applique terme par terme (qui sont tous positifs). Pour toute matrice $\boldsymbol{U}$, nous avons*

$$\mathbb{P}\left[\langle \boldsymbol{U}, \boldsymbol{\varepsilon}_T\rangle_F \le \theta(x,(X_t))\|\boldsymbol{U}\|_1\right] \ge 1 - \frac{\pi^2}{3}\exp(-2x + 2\log d).$$

3.1.2 Propriétés non-asymptotiques du Lasso

À partir des résultats de la Section 3.1.1, nous prouvons le Théorème 7 ci-dessous. Il donne une borne sur l'erreur d'estimation, exprimée avec des normes ℓ^1 et L^2.

Theorem 7 (Théorème III.1). *Supposons (H1), (H3) - (H5). Soit $\gamma > 1$, $0 \le \tau < \gamma - 1$, $\epsilon_0 \in (0,1)$ et définissons*

$$\lambda_T := \gamma\sqrt{\frac{4e|\hat{\delta}_T|_\infty}{T}\left(\frac{1}{2}\log\frac{2\pi^2 d^2}{3\epsilon_0} + \log(2 + |\log(T\hat{\delta}_T)|_\infty)\right)}$$

où $\hat{\delta}_T := \mathrm{diag}\,\widehat{\boldsymbol{C}}_T$ et log *s'applique terme par terme à $T\hat{\delta}_T$. Soit $c_0 := \frac{\gamma+\tau+1}{\gamma-\tau-1}$, et utilisant les définitions de κ et T_0 du Théorème 5 et supposons $T \ge T_0$. Alors, pour toute matrice s-sparse par ligne $\boldsymbol{A}$, l'estimateur Lasso $\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{A}}_{\lambda_T}$ vérifie*

$$2\tau\gamma^{-1}\lambda_T\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 \le \|(\boldsymbol{A} - \boldsymbol{A}_0)X\|_{L^2}^2 + \left(\frac{1+\gamma+\tau}{\gamma\kappa}\right)^2 \lambda_T^2 ds$$

with probability at least $1 - \epsilon_0$.

Le Théorème est alors suivi en corollaire par des bornes d'erreur en norme ℓ^1 et L^2. Ensuite, en utilisant la propriété de valeurs propres restreintes du Théorème 5, on obtient aussi un majoration en term de norme Frobenius puis en norme q, où $q \in [1,2]$.
La borne a un ordre de grandeur de $\lambda_T^2 ds$. En utilisant une technique déjà utilisée dans le Théorème 5, nous montrons que $\hat{\delta}_T$ peut être asymptotiquement considéré comme une constante. Ainsil'équivalent asymptotique de l'ordre de grandeur de l'erreur d'estimation est

$$\frac{ds(\log d + \log\log T)}{T}.$$

Nous comparons cet équivalent à la borne inférieure minimax donnée par le Théorème 8 ci-dessous. Sans que les deux soient égales, la différence d'ordre $\log s$ et $\log\log T$ est petite. Nous pensons qu'il est possible d'améliorer la borne supérieure en utilisant d'autres estimateurs. En particulier SLOPE satisfait une borne supérieure de l'ordre de la borne inférieure minimax dans le cas de la régression linéaire [BLT16].

Theorem 8 (Théorème III.2). *Pour des constantes $c > 0$ et $c' > 0$, on a :*

$$\inf_{\widehat{\boldsymbol{A}}} \sup_{\boldsymbol{A}\in\Gamma_s} \mathbb{E}_{\boldsymbol{A}} \left[\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2 \right] \geq \frac{c' d s \log(cd/s)}{T},$$

où Γ_s est l'ensemble des matrices s-sparses par ligne et la borne inférieure est caalculée parmi tous les estimateurs possibles.

La preuve de la borne inférieure utilise une borne de Gilbert-Varshamov appliqueé à l'ensemble des matrices s-sparses par ligne Γ_s. Les volumes nécessaires au calcul de la borne inférieure sont obtenus grâce à un résultat de comptage de graphes réguliers.

3.1.3 Propriétés asymptotiques du Adaptive Lasso

Nous savons qu'en régression Gaussienne, le Lasso ne permet pas d'avoir simultanément une vitesse de convergence optimale et une détection de support exacte [Zou06]. Ceci suggère que ce soit le cas aussi pour l'estimation de processus d'Ornstein-Uhlenbeck. En effet, la pénalisation du Lasso fonctionne de façon qu'elle réduit toutes les entrées de la matrices proportionnellement à λ. Malheureusement, les valeurs de λ qui permettent d'obtenir une détection de support asymptotique sont au minimum d'ordre $1/\sqrt{T}$, alors que celles qui permettent d'obtenir une vitesse de convergence maximale sont d'ordre de grandeur strictement inférieur à $1/\sqrt{T}$.

L'idée derrière le Lasso Adaptatif est simple : il s'agit de pénaliser plus les entrées que nous pensons être des zéros. Pour juger de ceci, nous utilisons les valeurs du MLE. Les valeurs du MLE proches de zéro seront des indicateurs d'un vrai paramètre égal à zéro. Ceci nous mène à utiliser une pénalisation égale à

$$\|\boldsymbol{A} \circ |\widehat{\boldsymbol{A}}^{MLE}|^{-\gamma}\|_1 = \sum_{i,j} \frac{|\boldsymbol{A}^{ij}|}{|(\widehat{\boldsymbol{A}}^{MLE})^{ij}|^\gamma}.$$

Dans cette définition, γ peut être choisi parmi tous les réels strictement positifs et nous choisissons $\gamma = 1$ dans no tests numériques. Le Théorème 9 ci-dessous montre que l'estimateur qui en résulte permet d'avoir simultanément une vitesse de convergence optimale et une détection de support exacte. Pour éviter une notation trop lourde, nous cachons la dépendance en T de λ et des estimateurs.

Theorem 9 (Théorème III.4). *Supposons (H1) – (H3). Fixons $\gamma > 0$ et supposons que λ vérifie $\lambda T^{1/2} \to 0$ et $\lambda T^{(\gamma+1)/2} \to +\infty$ lorsque $T \to +\infty$. Alors nous avons les propriétés suivantes.*

1. *Sélection de variable asymptotiquement exacte : lorsque $T \to +\infty$, on a*

$$\mathbb{P}\left[\mathrm{supp}(\widehat{\boldsymbol{A}}_{ad.}) = \mathrm{supp}(\mathcal{A}_0)\right] \to 1.$$

2. *Normalité asymptotique : lorsque $T \to +\infty$, on a*

$$\sqrt{T}\left((\widehat{\boldsymbol{A}}_{ad.})_{|\mathcal{A}_0} - (\boldsymbol{A}_0)_{|\mathcal{A}_0}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0 \times \mathcal{A}_0}\right)^{-1}\right),$$

où $\mathcal{A}_0 = \mathrm{supp}(\boldsymbol{A}_0)$ est le support (l'ensemble d'indices des entrées non-nulles) du paramètre $\boldsymbol{A}_0$.

3.1.4 Comparaison numérique du MLE, Lasso et Adaptive Lasso

Les résultats théoriques montrent certaines qualités des méthodes pénalisées. Nous concluons le chapitre avec une analyse des performances des estimateurs sur des expériences numériques, ou nous comparerons en particulier les méthodes pénalisées au MLE.
La mise en place pratique des estimateurs nécessite en particulier de décider de la méthode utilisée pour choisir la valeur de λ. Le Théorème 7 donne une valeur théorique pour λ, mais nous savons que les paramétres de pénalisation typiques sont souvent pessimistes [BvdG11]. Ainsi nous préférerons utiliser une méthode par cross-validation [AC10] : nous utiliserons les premiers 80% de la trajectoire pour calculer l'estimateur pour différentes valeurs de λ, puis nous choisissons la valeur de λ qui minimise la log-vraisemblance négative calculée sur les derniers 20% de la trajectoire. Nous utilisons cette méthode pour fixer les valeurs de λ et les estimateurs de Lasso et Lasso Adaptatif qui en résultent.
Nos expériences confirment les avantages des méthodes pénalisées comparées au MLE en termes d'erreurs d'estimation. Elles montrent aussi que le Lasso Adaptatif permet de mieux estimer le support du vrai paramètre. Enfin, nous appliquons nos méthodes d'estimation à des données boursières, une série temporelle de dimension 120 composée de moyennes glissantes exponentielles des retours journaliers.

3.2 Estimation sparse appliquée aux données financières

Nous appliquons les méthodes discutés ci-dessus dans deux cas séparés. Premièrement, nous analysons les retours journaliers d'actions américaines, ou nous cherchons à tester l'hypothèse selon laquelle les actions ayant eu une bonne performance récente

ont tendance à avoir de moindres performances dans le futur. Deuxièmement, nous analysons les prix de futures sur dividendes, pour lesquelles nous cherchons à voir si leurs variations sont compatibles avec le lien entre les prix d'actions et l'estimation de ses dividendes futures, que nous construisons avec une approche par flux de trésorerie escomptés.

3.2.1 Retours à la moyenne dans l'indice SP500

Nous prenons ici un sous-ensemble d'actions incluses dans l'indice SP500 avec un historique de données suffisamment long. Notre ensemble de données contient les prix de 466 actions, mesurés à la fermeture de 4305 jours, du 3 Janvier 2000 au 10 Février 2017, séparés par secteurs industriels GICS. Pour chaque action, nous calculons la moyenne roulante exponentielle (EMA pour *Exponential Moving Average*) de ses retours journaliers. Lorsque la moyenne roulante dépasse sa valeur habituelle, nous disons que l'action sur-performe, et sous-performe dans le cas contraire. La moyenne roulante retourne vers la moyenne par construction, au sens suivant :

$$E_t := \alpha \int_0^t e^{-\alpha(t-s)} \mathrm{d}M_s, \qquad \mathrm{d}E_t = -\alpha E_t \mathrm{d}t + \alpha \mathrm{d}M_t.$$

Nous utilisons dorénavant la moyenne roulante E au lieu des données de retours, la moyenne roulante étant continue et toute prévision des variations de celle-ci permet d'obtenir une prévision pour les retours journaliers. Nous analysons alors E de diverses manières, utilisant des modèles uni-dimensionnels et multi-dimensionnels. Dans tous les cas, nous estimons les paramètres $\boldsymbol{A}$ et $\boldsymbol{\Sigma}$ (a et σ en dimension 1) de processus d'Ornstein-Uhlenbeck correspondants aux données. Ensuite, nous comparons les résultats aux valeurs de référence que nous mesurons avec une moyenne roulante sur retours indépendants stationnaires.

Étant donné que l'intensité de retour à la moyenne peut varier avec le temps, nous faisons l'estimation sur des sous-intervalles plus courts. Nous choisissons 30 intervalles de longueur 1000 (4 années de données journalières), régulièrement espacées dans l'intervalle total. Sur chaque sous-intervalle, nous utilisons les premiers 800 jours pour estimer les paramètres, et les 200 restants pour mesurer à l'aide de ratios de vraisemblance la capacité du modèle à prédire les variations futures.

Pour les estimations uni-dimensionnelles, nous utilisons les valeurs maximisantes la vraisemblance (MLE). Pour les données multi-dimensionnelles, nous comparons le MLE au Lasso et au Lasso Adaptatif, ainsi qu'à un estimateur diagonal. Ce dernier consiste à construire une matrice diagonale avec les estimateurs uni-dimensionnels de chaque composante.

- **Retour à la moyenne des moyennes sectorielles**. Pour chaque secteur industriel, nous calculons une moyenne non-pondérée des EMA des actions. Une analyse uni-dimensionnelle de chaque secteur d'activité individuel indique l'existence de retours à la moyenne pour la plupart des secteurs. En effet, les valeurs typiques de a observées sont entre 0.11 et 0.14, par rapport à des valeurs de 0.105 ± 0.015 mesurées pour les retours indépendants. L'approche multi-dimensionnelle montre l'existence possible d'effets inter-sectoriels, l'estimateur Lasso non-diagonal battant l'estimateur diagonal. Cependant, les mesures de ratios de vraisemblance des estimateurs uni et multi-dimensionnels ne battent pas les valeurs de références de manière régulière.
- **Retour des actions à la moyenne de leur secteur**. Nous soustrayons ici la moyenne sectorielle de chaque processus correspondant aux actions individuelles. Les paramètres de retour à la moyenne prennent ici des valeurs typiques entre 0.11 et 0.15, ce qui indique l'existence de retours à la moyenne d'intensité dépassant ce qui est attendu pour des données indépendantes. Dans l'approche multi-dimensionnelle, le Lasso obtient de meilleurs résultats que le MLE et l'estimateur diagonal, indiquant la valeur des approches pénalisées, mais ne bat pas les valeurs de référence en termes de ratios de vraisemblance.

L'observation régulière de paramètres de retour à la moyenne au-dessus de ce qui est attendu de données non-corrélées dans le temps est contredite par l'impossibilité de la vérifier avec un ratio de vraisemblance mesuré sur un intervalle indépendant. Cette contradiction mérite une explication. Pour ceci, nous construisons un processus auto-regressif possédant par sa définition des propriétés de retour à la moyenne, auquel nous appliquons le même traitement qu'aux données réelles. Nous observons alors que la mesure de a est souvent significativement au-dessus de ce qui est attendu de données non-corrélées, mais cependant le ratio de vraisemblance ne permet généralement pas de confirmer le retour à la moyenne. Ceci nous incite à conclure que la mesure de ratio de vraisemblance n'est pas adaptée pour valider l'existence de retours à la moyenne dans les données.

3.2.2 Retours à la moyenne pour les EURO STOXX 50 Dividend Futures

Avec une approche par flux de trésorerie escomptées, nous pouvons voir une action comme l'équivalent du flux des futures dividendes. Avec quelques approximations sur les taux d'ecompte, nous avons

$$\tilde{S}_t \simeq \sum_{n \geq 0} D(t,n) F_t^{(n)}$$

où $\tilde{S}$ est le prix de l'action ajusté pour le paiement de dividendes, les D sont les paramètres d'escompte et les F sont les prix des futures sur dividendes. Les futures

sur dividendes sont des contrats construits sur un sous-jacent mesurant le total de dividendes payées pendant une année civile. Ainsi, si le prix de l'action est fixé, les variations des prix de dividendes sont liées par une relation linéaire qui peut créer des phénomènes de retour à la moyenne.

Nous prenons comme action à traiter l'index EURO STOXX 50, un index qui réplique la valeur des actions de 50 sociétés européennes, choisies par capitalisation totale. La bourse Eurex construit un index EURO STOXX 50 Dividend Points, qui compte le total des dividendes payées par les 50 entreprises. Cet index est alors utilisé comme sous-jacent pour les EURO STOXX 50 Dividend Futures. L'index peut alors être vu comme un action. Nous effectuons une régression des prix des Futures par rapport aux variations de l'index. Les coefficients de regression changent d'année en année, mais évoluent de manière similaire en fonction de la date d'échéance du Future. La partie restante peut être vue comme la partie des variations des prix de futures non-expliquée par les variations de l'index et est sujette de la suite de notre analyse.

Nous commençons par considérer les prix de futures comme composantes d'un processus d'Ornstein Uhlenbeck multi-dimensionnel, l'estimation donnant des résultats non concluants. Nous observons alors que les prix semblent se regrouper autour d'un processus moyen. Nous utilisons alors une approche en deux étapes. Premièrement, nous regardons le prix moyen comme un processus d'Ornstein-Uhlenbeck uni-dimensionnel. Le retour à la moyenne de celui-ci est significatif, mais il est important de ne pas faire d'estimation sur des intervalles trop longs. Deuxièmement, nous analysons les déviations de la moyenne, avec de nouveau des résultats non concluants. Nous pouvons conclure cependant que les prix des futures de dividendes, controllées par rapport aux variations de l'index, dévient peu d'une moyenne qui a tendance à retourner lentement vers sa moyenne.

Part I

How to infer the parameters of a multi-dimensional Ornstein-Uhlenbeck process from binary observations?

CHAPTER I

Parameter estimation of Ornstein-Uhlenbeck process generating a stochastic graph

Abstract

Given Y a graph process defined by an incomplete information observation of a multivariate Ornstein-Uhlenbeck process X, we investigate whether we can estimate the parameters of X. We define two statistics of Y. We prove convergence properties and show how these can be used for parameter inference. Finally, numerical tests illustrate our results and indicate possible extensions and applications.

Keywords: stochastic graph process, inference for stochastic process, incomplete information, asymptotic properties of estimators.

MSC 2010: 62Mxx; 05C80; 62F12.

1 Introduction

1.1 Statement of the problem

Take an Ornstein-Uhlenbeck process $X = (X_t : t \geq 0)$ with values in $\mathbb{R}^d$ ($d \in \mathbb{N}^+$), solution to the stochastic differential equation:

$$\mathrm{d}X_t = -\boldsymbol{A}_0 X_t \mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t, \quad X_0 \text{ given.} \tag{1.1.1}$$

We consider the model of stochastic graphs generated as follows: the adjacency value between vertices i and j is

$$Y_t^{ij} = \mathbb{1}_{X_t \in S^{ij}}$$

where $(S^{ij})_{i,j}$ are subsets of $\mathbb{R}^d$.

The topic of random graphs is a well-developed research area. Since the Erdős–Rényi model, many other ways of generating a random graph have been proposed, most notably the preferential attachment model, the Chung-Lu model or the Kronecker graph model [Bol01, MX07]. Most models have the goal to create a single instance of a random graph. Some proceed by "growing" the graph, i.e. by successively adding nodes and edges, as in the preferential attachment model. Other models also enable deleting nodes and edges.

In contrast, in our model the nodes are fixed, hence it is not a "growing" graph, but the edges are evolving continuously in time. One could in principle fix a T and consider the random graph Y_T, but the real richness of our model resides in the evolution of the graph in continuous time. For instance, it adds correlation between graphs at arbitrary time-scales.

Y gives only partial information about X. Hence, usual results on inference for stochastic processes can not be applied. We therefore aim at extending these results to our setting and ask then the question of finding $\boldsymbol{A}_0, \boldsymbol{\Sigma}$ from Equation (1.1.1), given the sole observation of Y.

We will consider that we have access to one realization of the process Y observed at discrete times $(k\Delta_n)_{0\leq k\leq n}$, $n \geq 1$. Therefore, we hope to get results in the long time limit $n\Delta_n \to +\infty$ as $n \to +\infty$, in which we can expect to use ergodic properties of the process X. Intuitively, doing so we will estimate parameters arising in the stationary distribution. Also, to estimate parameters related to local fluctuations (i.e. $\boldsymbol{\Sigma}$), we are interested in the high frequency limit $\Delta_n \to 0$.

1.2 Applications in systemic risk modeling

In [CFS15, FI13], authors present a model for interbank lending in which d bank reserves are modeled through real-valued random processes X^i. Whenever bank i has more reserves than bank j, i lends money to j, thus reducing reserve X^i and increasing X^j. Gaussian noise is added in order to model random variations of the reserves. We will refer to this model through Equation (1.2.1) (which is exactly the model from [CFS15] without a central bank):

$$\mathrm{d}X_t^i = -\frac{a}{d}\sum_{j=1}^{d}\left(X_t^i - X_t^j\right)\mathrm{d}t + \sigma \mathrm{d}W_t^i. \tag{1.2.1}$$

In this model, authors define a systemic event when the mean reserve falls below some predetermined value. The analysis of the model shows that the probability of the systemic event can be computed and the result depends explicitly on the values of the parameters, especially on the correlation of the Brownian motions with a common noise.

Therefore, it is crucial to know how to estimate the values of the parameters of the equation, even in the realistic situation where one wouldn't have complete information of the banks' reserves. We consider for instance that the regulator, whose perspective we are analysing here, would fix a regulatory threshold r and would observe all variables of the form

$$Y_t^{ij} := \mathbb{1}_{X_t^i - X_t^j > r}.$$

1.3 Summary of results

In this article, we define two statistics and show the following convergences.

- An "occupation time" statistic OT_n, which counts the number of times the process is present in a given set. Normalized by n, this number converges to the stationary measure of that set, in the following ways:
  - in L^2, with a speed of convergence bounded by $\sqrt{n\Delta_n}$, which is the square root of the time horizon of the estimation;
  - with the right normalization, convergence in law to a Gaussian variable, for one-dimensional processes.
- A "crossings" statistic $\mathcal{C}_n$, which counts the number of times the process X goes in or out of a given set S^{ij}, i.e. the number of changes of the $Y^{i,j}$-value. We show a convergence in L^2 of this - suitably normalized - statistic.

We show then how to use this in order to estimate the parameters of model (1.1.1).

1.4 Related work

The question that we investigate in this article is on the recovery of the parameters of an Ornstein-Uhlenbeck equation given the n observations of a single realization of the process. It relates therefore to the largely developed field of inference for stochastic processes. Many results exist on this subject: for instance, see [Kut04] for continuous-time observations and [KLS12] for discrete-time ones.

In this work, we are specifically interested in discrete-time observation schemes. We observe three distinct discrete-time settings. First, the low-frequency long-time (LF-LT) setting consists in fixing a time step Δ and observing at times $(i\Delta)_{i\leq n}$ with $n \to +\infty$ [Yos92]. Second, the high-frequency fixed-time (HF-FT) setting, where a time horizon T is fixed and observations are taken at $(i\Delta_n)_{i\leq n}$ with $\Delta_n = T/n \to 0$ [GCJ93]. Third, the high-frequency long-time (HF-LT) setting where one assumes observations at $(i\Delta_n)_{i\leq n}$ with the time step $\Delta_n \to 0$ and the time horizon $n\Delta_n \to +\infty$ [Kes97, Gob02, ASM04]. Our work is placed in the latter HF-LT setting.

Some results already exist on problems with observation of crossings of a given threshold. For instance, [Flo87] considers the estimation using only the observation of the sign of the process. However this is done in the LF-LT setting and in dimension 1. The same remark applies to [Flo89, Flo91]. We will extend her CLT results to the HF limit.

The HF-LT setting is combined with partial information observation in [IUY09]. The authors consider, for $\epsilon > 0$, the observation of $\mathbb{1}_{|X_t| \geq \epsilon} X_t$. Our assumption of a binary observation leaves us with even scarcer information, thus making the inference problem more delicate.

To sum up, the closest work to ours is seemingly [IUY09] and [Flo87] but our main original contribution concerns the multi-dimensional scope and the case of binary observation in the HF-LT setting.

Organisation of the paper. In the next subsection, we define the notations and assumptions used throughout this work. Then in Section 2, we study the first statistic based on occupation time. We first prove general convergence results useful to analyse the convergence of all estimators (Theorem I.1). L^2 and CLT results are proved (Theorems I.2 and I.3). Section 3 is devoted to the study of the second estimator based on crossings (Theorem I.4). Last in Section 4, we report numerical experiments. Applications to parameters inference are discussed along the sections. Technical results are postponed to Appendix.

1.5 Notations and assumptions

1.5.1 Notations

For $d \in \mathbb{N}^+$, take $m \in \mathbb{R}^d$ and V a symmetric positive definite $d \times d$ matrix. We call $\mathcal{N}(m, V)$ the law of a Gaussian r.v. with mean m and covariance matrix V. For $m = 0$, this centered Gaussian distribution is denoted by ν_V and its density by μ_V:

$$\mu_V(x) = (2\pi)^{-d/2} \det(V)^{-1/2} \exp\left(-\frac{1}{2} x^* V^{-1} x\right), \ x \in \mathbb{R}^d,$$

where x^* is the transpose of x. In dimension $d = 1$, we introduce additionally the CDF of $\mathcal{N}(0, 1)$:

$$N(x) = \int_{-\infty}^{x} \mu_1(s) \mathrm{d}s, \ x \in \mathbb{R}.$$

Given a measurable function $f : \mathbb{R}^d \to \mathbb{R}$ and a probability measure ν on $\mathbb{R}^d$, we denote $\nu(f) = \int f(x) \nu(\mathrm{d}x)$. For a measurable set $S \subset \mathbb{R}^d$, we write $\nu(S) = \nu(\mathbb{1}_S)$ by a slight abuse of notation.

1.5.2 Restatement of the model and standing assumptions

Consider two matrices $\boldsymbol{A}_0 \in \mathcal{M}_{d,d}(\mathbb{R})$ and $\boldsymbol{\Sigma} \in \mathcal{M}_{d,q}(\mathbb{R})$ where $d, q \in \mathbb{N}^+$, which serves to model (1.1.1). The standing assumptions on $\boldsymbol{A}_0$ and $\boldsymbol{\Sigma}$ are the following.

(H) The matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*$ is invertible and the spectrum of $\boldsymbol{A}_0$ has strictly positive real parts:

$$a_0 := \min_{\lambda \in \mathrm{Sp}(\boldsymbol{A}_0)} \mathcal{R}e(\lambda) > 0. \tag{1.5.1}$$

We define an important class of covariance matrices:

$$\boldsymbol{C}_t = \int_0^t e^{-\boldsymbol{A}_0 u} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^* e^{-\boldsymbol{A}_0^* u} \mathrm{d}u, \qquad \boldsymbol{C}_\infty = \int_0^{+\infty} e^{-\boldsymbol{A}_0 u} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^* e^{-\boldsymbol{A}_0^* u} \mathrm{d}u.$$

We easily check that $\boldsymbol{C}_\infty$ is well defined, symmetric positive definite. For one-dimensional processes, we simply have:

$$v_t = \frac{\sigma^2}{2a}\left(1 - e^{-2at}\right), \qquad v_\infty = \frac{\sigma^2}{2a}.$$

Let $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P}\right)$ be a filtered space and $(W_t)_{t\in\mathbb{R}^+}$ a q-dimensional Brownian motion with respect to $\mathcal{F}$. In this setting, we consider the multi-dimensional Ornstein-Uhlenbeck equation for X as introduced in (1.1.1):

$$\mathrm{d}X_t = -\boldsymbol{A}_0 X_t + \boldsymbol{\Sigma}\mathrm{d}W_t, \quad X_0 \overset{\mathrm{d}}{=} \mathcal{N}\left(0, \boldsymbol{C}_\infty\right), \tag{1.5.2}$$

where X_0 is a r.v. independent of W. In the following $X = (X_t : t \geq 0)$ stands for the $\mathbb{R}^d$-valued solution of (1.5.2). We recall some properties from [KS91, Chapter 5.6]. First X is stationary:

$$\forall t \in \mathbb{R}^+, X_t \overset{\mathrm{d}}{=} \mathcal{N}\left(0, \boldsymbol{C}_\infty\right). \tag{1.5.3}$$

To simplify we denote by ν_∞ the Gaussian distribution $\mathcal{N}\left(0, \boldsymbol{C}_\infty\right)$ and by μ_∞ its density. In the subsequent analysis, the initial distribution could be different from ν_∞, it would not change significantly the analysis since the OU-process converges exponentially fast to its stationary regime.

Second, X is Markovian and ergodic. Take $t > s$, we can write:

$$X_t = e^{-\boldsymbol{A}_0(t-s)} X_s + \int_s^t e^{-\boldsymbol{A}_0(t-u)} \boldsymbol{\Sigma}\mathrm{d}W_u, \tag{1.5.4}$$

from which we deduce

$$X_t \mid X_s \overset{\mathrm{d}}{=} \mathcal{N}\left(e^{-\boldsymbol{A}_0(t-s)} X_s, \boldsymbol{C}_{t-s}\right). \tag{1.5.5}$$

Equality (1.5.4) gives also an important insight on decorrelation of the process:

$$\mathbb{C}\mathrm{ov}\left(X_t, X_s\right) = e^{-\boldsymbol{A}_0(t-s)}\mathbb{V}\mathrm{ar}\left(X_s\right) = e^{-\boldsymbol{A}_0(t-s)}\boldsymbol{C}_\infty, \quad t \geq s. \tag{1.5.6}$$

In the following, all the limits will be considered as $n \to +\infty$, under the asymptotics of high frequency data ($\Delta_n \to 0$) on a long-time interval ($n\Delta_n \to +\infty$). Also, for simplicity, we assume $\Delta_n \leq 1$.

Remark I.1. *In Equations* (1.5.2) *and* (1.5.5)*, we see that the distribution of* X_0 *and* $X_t \mid X_s$ *depend on* $\boldsymbol{\Sigma}$ *only through* $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*$*. Hence we shall restate our inference problem as the estimation of* $(\boldsymbol{A}_0, \boldsymbol{\Sigma}\boldsymbol{\Sigma}^*)$.

2 Occupation time statistic

Consider the d-dimensional process governed by Equation (1.5.2). We define the first statistic:

Definition I.1. *Let* S *be a measurable subset of* $\mathbb{R}^d$*. Define:*

$$Y_t^S = \mathbb{1}_{X_t \in S}.$$

The occupation time statistic is defined as:

$$\mathrm{OT}_n^S = \frac{1}{n}\sum_{k=0}^{n-1} Y_{k\Delta_n}^S = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X_{k\Delta_n} \in S}.$$

This statistic gives the frequency of occupation of S by the process X, hence the name.

2.1 Preliminary tools

The study of convergence of OT_n^S and further statistics will be made possible by using some tight controls related to the mixing properties of X at different times. Note that we cannot directly invoke general mixing properties of Markov chains since here the Markov chain $(X_{k\Delta n} : k \geq 0)$ depends on n through Δ_n: this is the main difficulty. The current estimates are made possible using the Gebelein inequality (a.k.a. Lancaster inequality) about maximal correlation between Gaussian spaces.

Theorem I.1 (Mixing properties). *Assume that* X *solves the Ornstein-Uhlenbeck Equation* (1.5.2)*, and recall the definition of* a_0 *in* (1.5.1)*. There exists a finite constant* $C_{(2.1.1)}$*, depending only on the stationary distribution covariance matrix* $\boldsymbol{C}_\infty$*, such that for any*

$t \geq s \geq 0$ and for any functions $\varphi : \mathcal{C}^0([0,s],\mathbb{R}^d) \to \mathbb{R}, \phi : \mathcal{C}^0([t,+\infty),\mathbb{R}^d) \to \mathbb{R}$ such that φ, ϕ are square-integrable w.r.t. the law of X, we have

$$\left|\mathbb{C}\mathrm{ov}\left(\varphi\left((X_u)_{u\leq s}\right), \phi\left((X_v)_{v\geq t}\right)\right)\right| \leq C_{(2.1.1)} e^{-a_0|t-s|} \sqrt{\mathbb{V}\mathrm{ar}\left(\varphi\left((X_u)_{u\leq s}\right)\right) \mathbb{V}\mathrm{ar}\left(\phi\left((X_v)_{v\geq t}\right)\right)}. \tag{2.1.1}$$

The proof is done in Appendix I.B. A very useful corollary is related to the convergence study of sum of general local functionals of X.

Corollary I.1. *Consider a measurable function $g : \mathbb{N}\times\mathbb{N}\times\mathcal{C}^0([0,1],\mathbb{R}^d) \to \mathbb{R}$ such that $\mathbb{E}\left[g(k,n,(X_s)_{k\Delta_n\leq s\leq(k+1)\Delta_n})^2\right] < +\infty$ for any $k,n \in \mathbb{N}$. For $n \in \mathbb{N}$ define*

$$v_n^2 = \sup_{k<n} \mathbb{V}\mathrm{ar}\left(g(k,n,(X_s)_{k\Delta_n\leq s\leq(k+1)\Delta_n})\right),$$
$$\xi_k^{(n)} = \sqrt{\frac{\Delta_n}{n}} g(k,n,(X_s)_{k\Delta_n\leq s\leq(k+1)\Delta_n}).$$

Then, there is a finite constant $C_{(2.1.2)}$, dependent only on the parameters $\boldsymbol{A}_0, \boldsymbol{\Sigma}$ of the model, such that:

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=0}^{n-1} \xi_k^{(n)}\right) \leq C_{(2.1.2)} v_n^2. \tag{2.1.2}$$

Remark I.2. *If $\mathbb{E}\left[\sum_{k=0}^{n-1}\xi_k^{(n)}\right] \to l$ for some $l \in \mathbb{R}$, then $v_n \to 0$ implies*

$$\sum_{k=0}^{n-1} \xi_k^{(n)} \xrightarrow{L^2} l.$$

Proof of Corollary I.1. Denote $g_k = g(k,n,(X_s)_{k\Delta_n\leq s\leq(k+1)\Delta_n})$; without loss of generality, we can assume that $\mathbb{E}[g_k] = 0$. We have

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=0}^{n-1}\xi_k^{(n)}\right) = \frac{\Delta_n}{n}\sum_{k=0}^{n-1}\mathbb{V}\mathrm{ar}(g_k) + \frac{2\Delta_n}{n}\sum_{k=0}^{n-1}\sum_{l=k+1}^{n-1}\mathbb{C}\mathrm{ov}(g_k,g_l).$$

For $l > k$, we have $[k\Delta_n,(k+1)\Delta_n] \subset [0,(k+1)\Delta_n]$ and $[l\Delta_n,(l+1)\Delta_n] \subset [l\Delta_n,+\infty[$. Apply Theorem I.1:

$$\mathbb{C}\mathrm{ov}(g_k,g_l) \leq C_{(2.1.1)} e^{-a_0|k+1-l|\Delta_n} \sqrt{\mathbb{V}\mathrm{ar}(g_k)\,\mathbb{V}\mathrm{ar}(g_l)}.$$

Then we deduce

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=0}^{n-1}\xi_k^{(n)}\right) \leq \frac{\Delta_n}{n} n v_n^2 + \frac{2\Delta_n}{n} n \sum_{m\geq 0} C_{(2.1.1)} v_n^2 e^{-a_0 m\Delta_n}$$

$$\leq v_n^2 \left(\Delta_n + 2C_{(2.1.1)} \frac{\Delta_n}{1 - e^{-a_0 \Delta_n}} \right)$$
$$\leq C_{(2.1.2)} v_n^2,$$

where we set $C_{(2.1.2)} = \sup_{x \in [0,1]} \left(x + 2C_{(2.1.1)} \frac{x}{1-e^{-a_0 x}} \right)$. □

2.2 L^2 convergence of occupation time statistics

Theorem I.2. *For any measurable set $S \subset \mathbb{R}^d$, OT_n^S converges to $\nu_\infty(S)$ in L^2 and*

$$\mathbb{E}\left[(\mathrm{OT}_n^S - \nu_\infty(S))^2 \right] = O\left(\frac{1}{n\Delta_n} \right).$$

Proof. As the process is stationary, $\mathbb{E}\left[\mathrm{OT}_n^S\right] = \mathbb{E}\left[\mathbb{1}_{X_0 \in S}\right] = \nu_\infty(S)$. Next, we apply Corollary I.1 to $\mathrm{OT}_n^S = \sum_{k=0}^{n-1} \xi_k^{(n)}$ with

$$\xi_k^{(n)} = \frac{1}{n} \mathbb{1}_{X_{k\Delta_n} \in S} = \sqrt{\frac{\Delta_n}{n}} g(k, n, X_{k\Delta_n}),$$
$$g(k, n, X_{k\Delta_n}) = \frac{1}{\sqrt{n\Delta_n}} \mathbb{1}_{X_{k\Delta_n} \in S},$$
$$\mathbb{V}\mathrm{ar}\left(g(k, n, X_{k\Delta_n}) \right) = \frac{1}{n\Delta_n} \mathbb{V}\mathrm{ar}\left(\mathbb{1}_{X_{k\Delta_n} \in S} \right) = \frac{\nu_\infty(S)\left(1 - \nu_\infty(S)\right)}{n\Delta_n}.$$

Therefore, we get $\mathbb{E}\left[(\mathrm{OT}_n^S - \nu_\infty(S))^2 \right] = \mathbb{V}\mathrm{ar}\left(\mathrm{OT}_n^S \right) \leq \frac{C_{(2.1.2)}}{n\Delta_n}$. □

2.3 Central Limit Theorem for one-dimensional processes

Here we restrict the study to the one-dimensional situation. There are two technical reasons for this: we solve explicitly the Poisson equation (see Lemma I.4) and derive tractable bounds on it. Additionally, we take advantage of the one-dimensional situation to handle explicit computations. The validity of a Central Limit Theorem in the multi-dimensional setting remains an open question to us.

For $d = 1$, the model becomes

$$\mathrm{d}X_t = -aX_t \mathrm{d}t + \sigma \mathrm{d}W_t. \tag{2.3.1}$$

Assumption **(H)** reads $a > 0$ and $\sigma \neq 0$. We consider the case $S = [1, +\infty[$. The extension of the following results to the case where S is a finite union of intervals is straightforward, and it is left to the reader.

Theorem I.3. *As $n \to +\infty$, we have*

$$\sqrt{n\Delta_n}\left(\mathrm{OT}_n^{[1,+\infty[} - \nu_\infty\left([1,+\infty[\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \nu_\infty\left(\sigma^2 F'^2\right)\right)$$

where F is defined in (I.C.2.2) *and is such that*

$$F'(x) = \frac{2}{\sigma^2}\frac{N\left(\frac{x\wedge 1}{\sqrt{v_\infty}}\right) - N\left(\frac{1}{\sqrt{v_\infty}}\right)N\left(\frac{x}{\sqrt{v_\infty}}\right)}{\mu_\infty(x)} \in \left[0, 2\sqrt{\frac{\pi}{a\sigma^2}}\right].$$

Proof. A simple inspection on F' shows that it is non-negative. The upper bound is proved in Lemma I.5 (see inequality (I.C.2.7)). This proves the inclusion of $F'(x)$.

We now prove the Central Limit Theorem. We follow the approach by [Flo84]. The main difference is that the function $x \mapsto \mathbb{1}_{x\geq 1}$ is non continuous, which raises technical issues.

Consider first the continuous time extension of $\mathrm{OT}_n^{[1,+\infty[}$, i.e.

$$\mathrm{OT}_t^c = \frac{1}{t}\int_0^t \mathbb{1}_{X_s\geq 1}\mathrm{d}s.$$

Denote in this proof $f(x) = \mathbb{1}_{x\geq 1}$ and $\hat{f}(x) = f(x) - \nu_\infty\left([1,+\infty[\right)$, so that

$$\int_0^t \hat{f}(X_s)\mathrm{d}s = t\left(\mathrm{OT}_t^c - \nu_\infty\left([1,+\infty[\right)\right).$$

Introduce then $L = -ax\frac{\partial}{\partial x} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}$ the infinitesimal generator of X: Lemma I.4 in Appendix I.C ensures that F defined in (I.C.2.2) verifies the Poisson equation

$$LF = -\hat{f}.$$

Introduce $M_t = F(X_t) - F(X_0) + \int_0^t \hat{f}(X_s)\mathrm{d}s$. F is twice differentiable but F'' has a single point of discontinuity at 1. However, we can still apply Itô's formula in that case (see Lemma I.3). We get:

$$M_t = \int_0^t \sigma F'(X_s)\mathrm{d}W_s, \qquad \langle M\rangle_t = \int_0^t \sigma^2 F'(X_s)^2\mathrm{d}s.$$

F' being bounded, M is a martingale. As we have $t^{-1}\langle M\rangle_t \to \nu_\infty\left(\sigma^2 F'^2\right)$ in probability (ergodic theorem) as $t \to +\infty$, we can use a CLT for martingales (see Lemma III.10 with $K_t = t^{-1/2}$) to get

$$\frac{M_t}{\sqrt{t}} = \frac{F(X_t) - F(X_0) + \left(\mathrm{OT}_t^c - \nu_\infty\left([1,+\infty[\right)\right)t}{\sqrt{t}} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nu_\infty\left(\sigma^2 F'^2\right)\right).$$

Finally, F is sublinear (F' bounded), thus $\frac{1}{\sqrt{t}}\left(F(X_t) - F(X_0)\right) \xrightarrow{L^2} 0$. Consequently we have proved

$$\sqrt{t}\left(\mathrm{OT}_t^c - \nu_\infty\left([1,+\infty[\right)\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nu_\infty\left(\sigma^2 F'^2\right)\right).$$

We now aim at proving that the above result extends to the discrete version $\mathrm{OT}_n^{[1,+\infty[}$. For this, define

$$\begin{aligned} D_n &:= \sqrt{n\Delta_n}\left(\mathrm{OT}_n^{[1,+\infty[} - \mathrm{OT}_{n\Delta_n}^c\right) \\ &= \sqrt{\frac{\Delta_n}{n}} \sum_{k=0}^{n-1} \int_0^{\Delta_n} \frac{f\left(X_{k\Delta_n}\right) - f\left(X_{k\Delta_n+u}\right)}{\Delta_n} \mathrm{d}u \\ &:= \sqrt{\frac{\Delta_n}{n}} \sum_{k=0}^{n-1} g(k, n, (X_s)_{k\Delta_n \le s \le (k+1)\Delta_n}). \end{aligned}$$

Observe that it remains to prove that $D_n \xrightarrow{\mathbb{P}} 0$. In view of Corollary I.1 and since

$$\mathbb{E}\left[g(k, n, (X_s)_{k\Delta_n \le s \le (k+1)\Delta_n})\right] = 0,$$

it is enough to prove that

$$v_n^2 := \sup_{k<n} \mathbb{E}\left[g(k, n, (X_s)_{k\Delta_n \le s \le (k+1)\Delta_n})^2\right] \to 0.$$

Actually, by Jensen inequality, the stationarity property and since f takes values in $\{0,1\}$, we have

$$v_n^2 \le \frac{1}{\Delta_n} \int_0^{\Delta_n} \mathbb{E}\left[|f(X_0) - f(X_u)|^2\right] \mathrm{d}u = \int_0^1 \mathbb{E}\left[|f(X_0) - f(X_{t\Delta_n})|\right] \mathrm{d}t.$$

With probability 1, $f(X_{t\Delta_n}) \to f(X_0)$, since f is continuous except on a set of zero ν_∞-measure: by the dominated convergence theorem, we obtain $v_n \to 0$, then $D_n \xrightarrow{\mathbb{P}} 0$. From this we have:

$$\sqrt{n\Delta_n}\left(\mathrm{OT}_n^{[1,+\infty[} - \nu_\infty\left([1,+\infty[\right)\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nu_\infty\left(\sigma^2 F'^2\right)\right).$$

□

2.4 Application to parameter inference

Lemma I.1. *Fix $S \subset \mathbb{R}^d$ and recall **(H)**. Then $\nu_\infty(S)$ is a continuous function of $\boldsymbol{C}_\infty$.*

Proof. Write

$$\nu_\infty(S) = \int_S \mu_\infty(x)\mathrm{d}x = \int_S (2\pi)^{-d/2} \det(\boldsymbol{C}_\infty)^{-1/2} \exp\left(-\frac{1}{2}x^* \boldsymbol{C}_\infty^{-1} x\right) \mathrm{d}x.$$

As the determinant and the inverse are continuous functions, $\mu_\infty(x)$ is continuous in $\boldsymbol{C}_\infty$ for any x. We also have:

$$\mu_\infty(x) \leq (2\pi)^{-d/2} v_m^{-d/2} \exp\left(-\frac{v_M^{-1}}{2}|x|^2\right)$$

where $v_M = \max_{\lambda \in \mathrm{Sp}(\boldsymbol{C}_\infty)} \lambda, v_m = \min_{\lambda \in \mathrm{Sp}(\boldsymbol{C}_\infty)} \lambda$.

Applying the Hoffman-Wielandt theorem ([HJ86, Theorem 6.3.5]) we know that v_M and v_m are continuous functions of $\boldsymbol{C}_\infty$, which are also non-zero in the neighborhood of invertible $\boldsymbol{C}_\infty$. From this follows that there is a local bound (in the neighbourhood of every invertible $\boldsymbol{C}_\infty$) by an integrable function of the form:

$$\mu_\infty(x) \leq \mathrm{Cst} \exp\left(-\mathrm{Cst}|x|^2\right)$$

with a positive constant Cst. Conclude using the dominated convergence theorem. □

For one-dimensional processes. We consider here Equation (2.3.1). The limit value of OT_n depends on the stationary distribution of the process, which is a centered Gaussian r.v. with variance $v_\infty = \sigma^2/2a$ (see Section 1.5). If $\nu_\infty(S)$ is monotonous with respect to v_∞, then we can construct an estimator of v_∞.

For instance, if $S = [1, +\infty[$, then $\nu_\infty(S) = N\left(-1/\sqrt{v_\infty}\right)$ which is strictly increasing with v_∞.

However, v_∞ is not a one-to-one function of a and σ hence we need more information to find the parameters of the process, using for instance the crossings statistic of Section 3.

Assuming that the volatility parameter σ is known, we show a CLT-type convergence in law of the estimator of a constructed from OT_n using the Δ-method. If $S = [1, +\infty[$, then $a = h(\nu_\infty(S))$ where $h(x) := \frac{\sigma^2}{2} N^{-1}(x)^2$. Define then $\hat{a}_n := h(\mathrm{OT}_n)$. Given that $h'(\nu_\infty(S)) = \sqrt{2\pi}\sigma^2 \exp(2a^2/\sigma^4)$, we get

$$\sqrt{n\Delta_n}\left(\hat{a}_n - a\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 2\pi\sigma^4 \nu_\infty\left(\sigma^2 F'^2\right) \exp(4a^2/\sigma^4)\right).$$

For multi-dimensional processes. Here again the limit value of OT_n^S is the measure of S under the stationary distribution. This distribution depends on the value of the matrix V_∞. Without further information or assumptions, $\boldsymbol{C}_\infty$ is a symmetric $d \times d$

matrix, representing $d(d+1)/2$ unknowns. We can expect to be able to find these unknowns only if we consider more than one set S and the corresponding statistics.

In the following, we will use the fact that the covariance matrices of the marginals of a Gaussian variable are the restrictions of its covariance matrix to the relevant spaces.

Consider first for $i \leq d$ the set $S^i = \{x : x^i \geq 1\}$. Then $\nu_\infty(S^i)$ depends only on the value of $(\boldsymbol{C}_\infty)^{ii}$. Applying the result from the preceding paragraph, we can construct an estimator of that value.

Consider then for $i \neq j$ the set $S^{ij} = \{x : x^i \geq 1, x^j \geq 1\}$. Then $\nu_\infty(S^{ij})$ depends only on the values of $(\boldsymbol{C}_\infty)^{ii}, (\boldsymbol{C}_\infty)^{jj}, (\boldsymbol{C}_\infty)^{ij}$. From the previous point, we know we can construct estimators of $(\boldsymbol{C}_\infty)^{ii}, (\boldsymbol{C}_\infty)^{jj}$. For the last parameter, we use the following result.

Proposition I.1. *Take (G_1, G_2) a non-degenerate centered Gaussian vector. Denote ρ the correlation between G_1 and G_2. Denote $S = [1, +\infty[^2$ and μ_ρ the density of the distribution of (G_1, G_2). Then $\mu_\rho(S)$ is a strictly increasing function of ρ.*

Proof. Denote $\sigma_1 = \sqrt{\mathbb{V}\mathrm{ar}(G_1)}$, $\sigma_2 = \sqrt{\mathbb{V}\mathrm{ar}(G_2)}$. By symmetry between G_1 and G_2, we can safely assume $\sigma_1 \geq \sigma_2$. Introduce a standard centered Gaussian G. We can write

$$\begin{aligned}
(G_1, G_2) &\overset{\mathrm{d}}{=} \left(\rho\frac{\sigma_1}{\sigma_2}G_2 + \sigma_1\sqrt{1-\rho^2}G, G_2\right), \\
\mu_\rho(S) &= \mathbb{E}\left[\mathbb{1}_{G_2\geq 1}\mathbb{E}\left[\mathbb{1}_{\rho\frac{\sigma_1}{\sigma_2}G_2+\sigma_1\sqrt{1-\rho^2}G\geq 1} \mid G_2\right]\right] \\
&= \int_{\mathbb{R}} \mathbb{1}_{y\geq 1} N\left(\frac{\rho\frac{\sigma_1}{\sigma_2}y - 1}{\sigma_1\sqrt{1-\rho^2}}\right)\mu_{\sigma_2^2}(y)\mathrm{d}y.
\end{aligned}$$

Set $g(\rho, y) = \frac{\rho\frac{\sigma_1}{\sigma_2}y-1}{\sigma_1\sqrt{1-\rho^2}}$: then $\frac{dg}{d\rho}(\rho, y) = \frac{\frac{\sigma_1}{\sigma_2}y-\rho}{\sigma_1(1-\rho^2)^{3/2}}$. This is strictly positive for $y \geq 1$ since $\frac{\sigma_1}{\sigma_2} \geq 1$ and $\rho < 1$. Therefore $N(g(\rho, y))$ is strictly increasing in ρ, and so is $\mu_\rho(S)$ on $]-1, 1[$.

□

This shows that we can construct an estimator of $(\boldsymbol{C}_\infty)^{ij}$ given the knowledge of $(\boldsymbol{C}_\infty)^{ii}, (\boldsymbol{C}_\infty)^{jj}$, which we have as noted before. Therefore,using $d(d+1)/2$ estimators, we can recover the whole matrix $\boldsymbol{C}_\infty$.

3 Crossings statistic

Given a binary observation Y_t, a function of X_t, we will count how many times Y goes from 0 to 1 and vice-versa. The following defines a statistic counting the number of jumps between 0 and 1 of the discretization of Y.

Definition I.2. *We define the crossings statistic by:*

$$\mathcal{C}_n^S = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{Y^S_{k\Delta_n} \neq Y^S_{(k+1)\Delta_n}}.$$

In the following, we restrict the convergence analysis to sets which are half-spaces $S^i = \{x : x^i \geq 1\}$; by symmetry, we assume that $i = 1$. Therefore we drop the superscript S and

$$Y_t = \mathbb{1}_{X^1_t \geq 1}.$$

$\mathcal{C}_n$ counts the number of times the discretised projection of X on the first coordinate crosses 1.

3.1 L^2 convergence

Theorem I.4. *Assume that $n\Delta_n^{3/2} \to +\infty$. We have the following convergence:*

$$\frac{\mathcal{C}_n}{\sqrt{\Delta_n}} \xrightarrow{L^2} 2\sqrt{\frac{\left(\mathbf{\Sigma}\mathbf{\Sigma}^*\right)^{11}}{2\pi}}\mu_{\boldsymbol{C}^{11}_\infty}(1),$$
$$\mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right) = O\left(\frac{1}{n\Delta_n^{3/2}}\right).$$

Proof. For ease of writing, introduce a new notation:

$$Z_k^{(n)} = \mathbb{1}_{X^1_{k\Delta_n}<1}\mathbb{1}_{X^1_{(k+1)\Delta_n}\geq 1}.$$

Now, divide the sum in two similar parts:

$$\mathcal{C}_n = \mathcal{C}_n^{+-} + \mathcal{C}_n^{-+},$$
$$\mathcal{C}_n^{+-} = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X^1_{k\Delta_n}\geq 1}\mathbb{1}_{X^1_{(k+1)\Delta_n}<1}, \qquad \mathcal{C}_n^{-+} = \frac{1}{n}\sum_{k=0}^{n-1} Z_k^{(n)}.$$

Although the two sums aren't perfectly symmetric, we will show that our reasoning will apply to both sums. Concentrate then on the second sum. In order to apply Corollary I.1, we introduce the following notations:

$$\frac{\mathcal{C}_n^{-+}}{\sqrt{\Delta_n}} := \sum_{k=0}^{n-1} \xi_k^{(n)},$$

$$\xi_k^{(n)} = \frac{1}{n\sqrt{\Delta_n}} Z_k^{(n)} := \sqrt{\frac{\Delta_n}{n}} g\left(k, n, (X_s)_{k\Delta_n \le s \le (k+1)\Delta_n}\right),$$
$$v_n^2 := \sup_{k<n} \mathbb{V}\mathrm{ar}\left(g\left(k, n, (X_s)_{k\Delta_n \le s \le (k+1)\Delta_n}\right)\right).$$

We obviously have $v_n^2 = \frac{1}{n\Delta_n^2} \sup_{k<n} \mathbb{V}\mathrm{ar}\left(Z_k^{(n)}\right)$ and Corollary I.2 we have $\mathbb{V}\mathrm{ar}\left(Z_k^{(n)}\right) = O\left(\sqrt{\Delta_n}\right)$. Therefore $v_n^2 = O\left(\frac{1}{n\Delta_n^{3/2}}\right) = o(1)$ by assumption. Thus, Corollary I.1 gives us the convergence to 0 of the variance.

Using again Corollary I.2 we get

$$\mathbb{E}\left[\frac{\mathcal{C}_n^{-+}}{\sqrt{\Delta_n}}\right] = \frac{\mathbb{E}\left[Z_k^{(n)}\right]}{\sqrt{\Delta_n}} \sim \sqrt{\frac{\left(\mathbf{\Sigma}\mathbf{\Sigma}^*\right)^{11}}{2\pi}} \mu_{\boldsymbol{C}_\infty^{11}}(1).$$

Therefore, recalling Remark I.2 gives

$$\frac{\mathcal{C}_n^{-+}}{\sqrt{\Delta_n}} \xrightarrow{L^2} \sqrt{\frac{\left(\mathbf{\Sigma}\mathbf{\Sigma}^*\right)^{11}}{2\pi}} \mu_{\boldsymbol{C}_\infty^{11}}(1).$$

The same reasoning would apply to $\mathcal{C}_n^{+-}$ if only we replaced $Z_k^{(n)}$ by $\tilde{Z}_k^{(n)} = \mathbb{1}_{X^1_{k\Delta_n} \ge 1} \mathbb{1}_{X^1_{(k+1)\Delta_n} < 1}$. Observe

$$\begin{aligned}
\mathbb{E}\left[\tilde{Z}_{\Delta_n}\right] &= \mathbb{E}\left[\left(1 - \mathbb{1}_{X_0^1 < 1}\right)\left(1 - \mathbb{1}_{X_{\Delta_n}^1 \ge 1}\right)\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{X_0^1<1}\right] - \mathbb{E}\left[\mathbb{1}_{X_{\Delta_n}^1 \ge 1}\right] + \mathbb{E}\left[Z_{\Delta_n}\right] \\
&= \mathbb{E}\left[Z_{\Delta_n}\right]
\end{aligned}$$

using stationarity. Thus, we can transfer the estimate on $Z_k^{(n)}$ to $\tilde{Z}_k^{(n)}$, i.e. on $\mathcal{C}_n^{-+}$ to $\mathcal{C}_n^{+-}$, which gives the final result. □

3.2 Application to parameter inference

For one-dimensional processes. For $d = 1$, Theorem I.4 simplifies to

$$\frac{\mathcal{C}_n}{\sqrt{\Delta_n}} \xrightarrow{L^2} \frac{2\sigma}{\sqrt{2\pi}} \mu_\infty(1).$$

Thus the renormalized $\mathcal{C}_n$ converges to a value that depends on σ and the density $\mu_\infty(1)$, which itself depends on $v_\infty = \sigma^2/2a$. In Section 2.4, we show how to estimate the value of v_∞. Using this estimate, we can compute $\mu_\infty(1)$ and construct simply an estimator of σ. Finally, as we estimate both σ and μ_∞, we also estimate a. Therefore, the two statistics OT_n and $\mathcal{C}_n$ are sufficient to estimate the parameters of the model (2.3.1).

For multi-dimensional processes. As we show above, we can estimate parameters of one-dimensional processes. The extension to multi-dimensional processes is not obvious. However, in specific cases, we can leverage Theorem I.4. Consider specifically that $\boldsymbol{A}_0$ is a diagonal matrix with diagonal terms $(a_1, \dots, a_d)$. Then

$$(\boldsymbol{C}_\infty)^{ij} = \int_0^{+\infty} e^{-a_i u} \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\right)^{ij} e^{-a_j u} \mathrm{d}u = \frac{\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\right)^{ij}}{a_i + a_j}.$$

Assuming we know $\boldsymbol{C}_\infty$, which we can estimate using results from Section 2.4, and as we can estimate $\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\right)^{ii}$ using Theorem I.4, we can estimate the values of a_i:

$$a_i = \frac{\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\right)^{ii}}{2\left(\boldsymbol{C}_\infty\right)^{ii}}.$$

We complete the inference by using the relation $\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\right)^{ij} = (a_i + a_j)\left(\boldsymbol{C}_\infty\right)^{ij}$.

Moreover, if we consider a rotation $\boldsymbol{L}$, then $\boldsymbol{L}X$ is also an Ornstein-Uhlenbeck process. Its stationary distribution's covariance matrix is $\boldsymbol{L}\boldsymbol{C}_\infty\boldsymbol{L}^\top$ and the diffusion parameter becomes $\boldsymbol{L}\boldsymbol{\Sigma}$. Hence, using half-spaces $S^{\boldsymbol{L},i} = \{x : (\boldsymbol{L}x)^i \geq 1\}$, Theorem I.4 shows we can estimate $(\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{L}^\top)^{ii}$. A suitable choice of $\boldsymbol{L}$ gives then estimates of non-diagonal terms of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$.

Assuming $\boldsymbol{C}_\infty$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ are known, we can find the value of $\boldsymbol{A}$. To show this, we use the vectorization relation $\mathrm{vec}\ \boldsymbol{M}\boldsymbol{N}\boldsymbol{O} = \left(\boldsymbol{O}^\top \otimes \boldsymbol{M}\right)\ \mathrm{vec}\ \boldsymbol{N}$ where $\otimes$ is the Kronecker product:

$$\begin{aligned}
\boldsymbol{C}_\infty &= \int_0^{+\infty} e^{-\boldsymbol{A}u}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{-\boldsymbol{A}^\top u}\mathrm{d}u \\
\mathrm{vec}\ \boldsymbol{C}_\infty &= \int_0^{+\infty} \left(e^{-\boldsymbol{A}u} \otimes e^{-\boldsymbol{A}^\top u}\right)\ \mathrm{vec}\ \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathrm{d}u \\
\mathrm{vec}\ \boldsymbol{C}_\infty &= (\boldsymbol{A} \oplus \boldsymbol{A})^{-1}\ \mathrm{vec}\ \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\
(\boldsymbol{A} \otimes \mathrm{Id} + \mathrm{Id} \otimes \boldsymbol{A})\ \mathrm{vec}\ \boldsymbol{C}_\infty &= \mathrm{vec}\ \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\
\boldsymbol{C}_\infty\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{C}_\infty &= \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \qquad (3.2.1) \\
(\mathrm{Id} \otimes \boldsymbol{C}_\infty)\ \mathrm{vec}\ \boldsymbol{A}^\top + (\boldsymbol{C}_\infty \otimes \mathrm{Id})\ \mathrm{vec}\ \boldsymbol{A} &= \mathrm{vec}\ \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\
\mathrm{vec}\ \boldsymbol{A} &= \left((\mathrm{Id} \otimes \boldsymbol{C}_\infty)\,\boldsymbol{T} + (\boldsymbol{C}_\infty \otimes \mathrm{Id})\right)^{-1} \mathrm{vec}\ \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top
\end{aligned}$$

where we used the relation $e^{\boldsymbol{M}} \otimes e^{\boldsymbol{N}} = e^{\boldsymbol{M}\oplus\boldsymbol{N}}$ with $\boldsymbol{M} \oplus \boldsymbol{N} = \boldsymbol{M} \otimes \mathrm{Id} + \mathrm{Id} \otimes \boldsymbol{N}$ and introduced the permutation matrix $\boldsymbol{T}$ such that for any matrix $\boldsymbol{M}$, $\boldsymbol{T}\ \mathrm{vec}\ \boldsymbol{M} = \mathrm{vec}\ \boldsymbol{M}^\top$. Finally, the existence of the inverse in the last line is equivalent to the existence of A that solves the problem. If we replace $\boldsymbol{C}_\infty$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ by their

estimates, the inverse might not exist. Then we can define an estimator of $\boldsymbol{A}$ by a matrix that minimizes the difference between the LHS and RHS of Equation (3.2.1); we achieve unicity of the definition by requiring additionally the estimate to minimize for example the Frobenius norm:

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_F \quad \text{s.t. } \boldsymbol{A} \in \arg\min_{\boldsymbol{M}} \|\widehat{\boldsymbol{C}_\infty}\boldsymbol{M}^\top + \boldsymbol{M}\widehat{\boldsymbol{C}_\infty} - \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top}\|_F.$$

4 Numerical tests

In the following, we present some inference results in the case of one-dimensional Ornstein-Uhlenbeck process. The observations are obtained via simulation and we consider simulation lengths n between 10 and $10 \cdot 2^{14}$ and discretization time-steps Δ between 0.1 and $0.1 \cdot 2^{-7}$.

We simulate trajectories using $\sigma = a = 1$:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \mathrm{d}W_t.$$

For each data point characterized by (n, Δ), we compute the expectation and standard deviation of OT_n and $\mathcal{C}_n$. These are empirically computed using 50000 simulated trajectories, independently initialized in the stationary distribution. Each figure will show these values plotted against either n or Δ in blue, and regression lines are added to each series in red. The figures are plotted in log-log scale, in order to show dependence of the results as powers of n and Δ.

For some plots, the regressions don't give a clear answer on the power dependence. Because we know many of our results come from the long-time limit, we privilege data series that maximize $n\Delta_n$, the horizon of the simulation.

Plots. For programming and plotting purposes, it is clearer to define statistics OT_n and $\mathcal{C}_n$ omitting normalization. Therefore, only in this paragraph, we use

$$\widetilde{\mathrm{OT}_n} = \sum_{k=0}^{n-1} \mathbb{1}_{X_{k\Delta_n} > 1}, \qquad \widetilde{\mathcal{C}_n} = \sum_{k=0}^{n-1} \mathbb{1}_{X_{k\Delta_n} \neq X_{(k+1)\Delta_n}}.$$

Additionally, we use ts as a notation for Δ_n.

Figures I.1 and I.2 show a linear dependence in n and no dependence in Δ. This is in agreement with the stationarity of the process.

The lines fitted to the scatter in Figure I.3 are somewhat misleading. Their slopes vary from from 0.52 to 0.75. However, given our analysis in the paper, it is $n\Delta$, the inferring horizon, that has the largest impact on the quality of the convergence. For

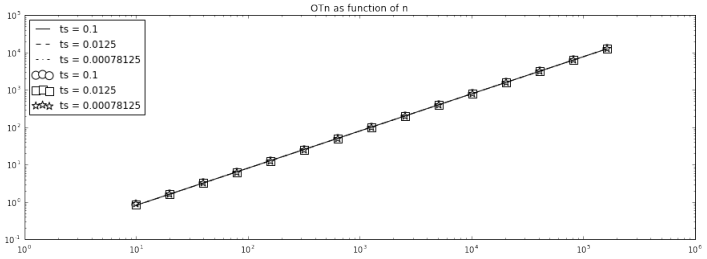


Figure I.1 – Expectation of $\widetilde{\mathrm{OT}_n}$ versus n for different values of time steps Δ_n (ts)

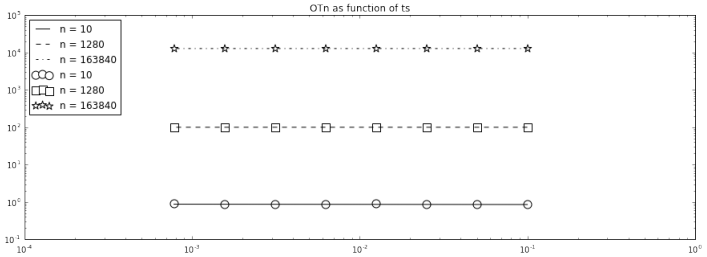


Figure I.2 – Expectation of $\widetilde{\mathrm{OT}_n}$ versus Δ_n (ts) for different values of n

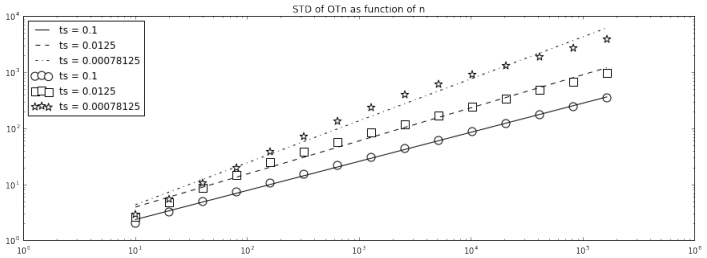


Figure I.3 – Standard deviation of $\widetilde{\mathrm{OT}_n}$ versus n for different values of time steps Δ_n (ts)

Figure I.4 – Standard deviation of $\widetilde{\mathrm{OT}_n}$ versus Δ_n (ts) for different values of n

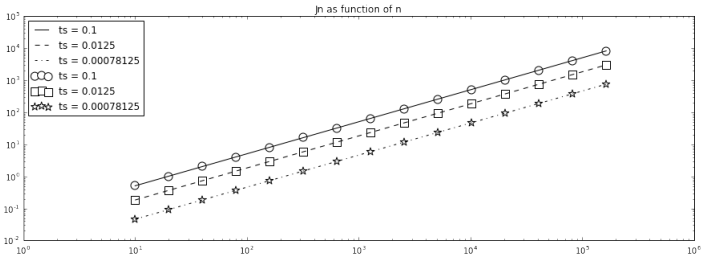


Figure I.5 – Expectation of $\widetilde{\mathcal{C}_n}$ versus n for different values of time steps Δ_n (ts)

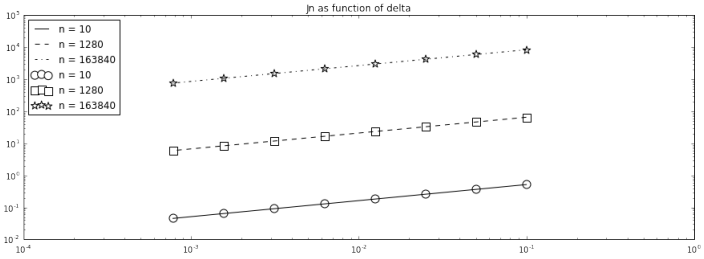


Figure I.6 – Expectation of $\widetilde{\mathcal{C}_n}$ versus Δ_n (ts) for different values of n

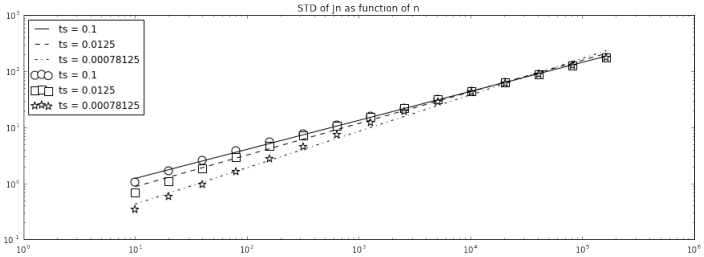


Figure I.7 – Standard deviation of $\widetilde{\mathcal{C}_n}$ versus n for different values of time steps Δ_n (ts)

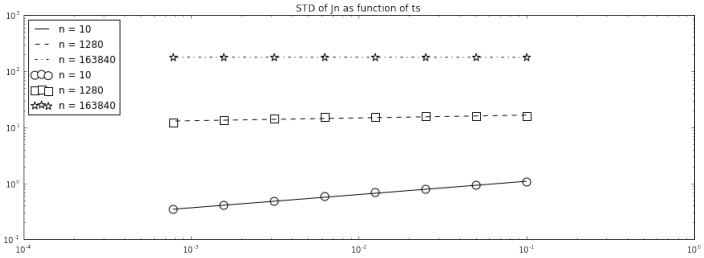


Figure I.8 – Standard deviation of $\widetilde{\mathcal{C}_n}$ versus Δ_n (ts) for different values of n

this reason, we should concentrate on high values of n. Further work on the last 5 points of the scatters concludes with a consistent slope of 0.5. The same applies to Figure I.4: although slopes vary from -0.07 to -0.5, we have the most confidence in the line corresponding to the highest value of n. We retain then the value of -0.5.

Thus, we numerically observe (up to constants)

$$\mathbb{E}\left[\widetilde{\mathrm{OT}_n}\right] \sim n, \qquad \sqrt{\mathbb{V}\mathrm{ar}\left(\widetilde{\mathrm{OT}_n}\right)} \sim n^{1/2}\Delta^{-1/2}.$$

Regarding $\widetilde{\mathcal{C}_n}$, Figures I.5 and I.6 show clear power dependencies of 1 and 0.5 with respect to n and Δ. Again, Figures I.7 and I.8 have to be observed only at the largest values of n. We conclude with slopes of 0.5 and 0.

$$\mathbb{E}\left[\widetilde{\mathcal{C}_n}\right] \sim n\Delta^{1/2}, \qquad \sqrt{\mathbb{V}\mathrm{ar}\left(\widetilde{\mathcal{C}_n}\right)} \sim n^{1/2}.$$

Summary. Taking the results from the preceding paragraph and rewriting them using our regular expressions of OT_n and $\mathcal{C}_n$, as in Definitions I.1 and I.2, we get from these numerical tests:

- For OT_n:
  - $\mathbb{E}\left[\mathrm{OT}_n\right] \propto 1$,
  - $\mathbb{V}\mathrm{ar}\left(\mathrm{OT}_n\right) \propto \frac{1}{n\Delta_n}$.

  This is in agreement with Theorem I.2.

- For $\mathcal{C}_n$:
  - $\mathbb{E}\left[\mathcal{C}_n\right] \propto \Delta_n^{1/2}$,
  - $\mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right) \propto \frac{1}{n\Delta_n}$.

  The expectation estimate is in agreement with Theorem I.4. However, our variance estimate is seemingly not optimal: the missing factor $\Delta_n^{1/2}$ may come from subtle cancellations in small time, in conjonction with the low regularity of indicator function. This issue is left to future research.

In Table I.1 we can compare our theoretical limits with the estimates from simulation.

	Simulation result	Theoretical value
$\mathbb{E}\left[\mathrm{OT}_n\right]$	0.07977	0.07865
$\mathbb{E}\left[\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right]$	0.16418	0.16560
$n\Delta_n \mathbb{V}\mathrm{ar}\left(\mathrm{OT}_n\right)$	0.25902	N/A
$n\Delta_n \mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right)$	0.43324	N/A

Table I.1 – Observed versus theoretical values of the limits of the expressions on the left

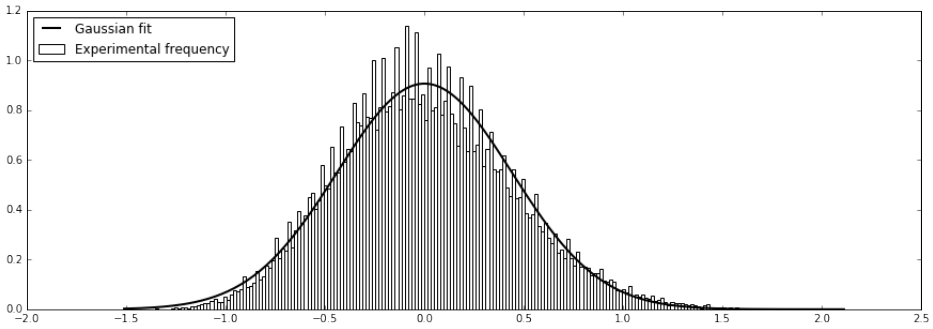


Figure I.9 – Normalized histogram of $\mathcal{C}_n$, for $n = 10 \cdot 2^{14}$ and $\Delta_n = 0.1 \cdot 2^{-7}$ and plot of PDF of Gaussian fit

Numerical investigation regarding a central limit theorem for $\mathcal{C}_n$. Our experimental observation of $\mathbb{V}\mathrm{ar}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}}\right) \propto \frac{1}{n\Delta_n}$ suggests we may expect a central limit theorem for $\sqrt{n\Delta_n}\left(\frac{\mathcal{C}_n}{\sqrt{\Delta_n}} - m\right)$ where $m = \lim_{n\to+\infty} \mathbb{E}\left[\frac{\mathcal{C}_n}{\Delta_n}\right]$. However, this result is out of the scope of the present paper.

Nonetheless we use the results of our simulations in order to see whether this conjecture is likely. In Figure I.9, we compare the normalized histogram of $\mathcal{C}_n$ to the probability density function of a fitted Gaussian. The agreement of the two seems to show that the validity of a central limit theorem is likely.

I.A Expectation of threshold crossing for OU processes in small time

In this section, X is the d-dimensional Ornstein-Uhlenbeck process of Equation (1.5.2). Define the indicator of the crossing of the threshold 1 by the first coordinate of X:

$$Z_\Delta = \mathbb{1}_{X_0^1<1}\mathbb{1}_{X_\Delta^1\geq 1}.$$

From Section 1.5.2, the first coordinate of $X_\Delta|X_0$ is a Gaussian vector centered at $\left(e^{-A\Delta}X_0\right)^1$ with variance $\boldsymbol{C}_\Delta^{11}$. Therefore $\mathbb{E}\left[\mathbb{1}_{X_\Delta^1\geq 1}\mid X_0\right] = N\left(\frac{\left(e^{-A\Delta}X_0\right)^1-1}{\sqrt{\boldsymbol{C}_\Delta^{11}}}\right)$ and

$$\mathbb{E}\left[Z_\Delta\right] = \int_{\mathbb{R}^d}\mathbb{1}_{x^1<1}N\left(\frac{\left(e^{-A\Delta}x\right)^1-1}{\sqrt{\boldsymbol{C}_\Delta^{11}}}\right)\mu_\infty(x)\mathrm{d}x. \tag{I.A.0.1}$$

Lemma I.2. *Let h be a bounded continuous increasing non-negative function, integrable in $-\infty$ and define:*

$$I_h(\Delta) := \int_{\mathbb{R}^d}\mathbb{1}_{x^1<1}h\left(\frac{\left(e^{-A\Delta}x\right)^1-1}{\sqrt{\boldsymbol{C}_\Delta^{11}}}\right)\mu_\infty(x)\mathrm{d}x.$$

Then when $\Delta\to 0$,

$$I_h(\Delta)\sim\sqrt{\Delta}\sqrt{\left(\Sigma\Sigma^*\right)^{11}}\mu_{\boldsymbol{C}_\infty^{11}}(1)\int_{-\infty}^0 h(x)\mathrm{d}x.$$

Proof. We use the following change of variables:

$$\begin{aligned}(y^1,y^2,\dots y^d) &= \left(\frac{x^1-1}{\sqrt{\boldsymbol{C}_\Delta^{11}}},x^2,\dots x^d\right),\\ \frac{I_h(\Delta)}{\sqrt{V_\Delta^{11}}} &= \int_{\mathbb{R}^d}\mathbb{1}_{y^1<0}h\left(\left(e^{-A\Delta}\right)^{11}y^1+\frac{\left(e^{-A\Delta}\right)^{11}-1+\sum_{i\geq 2}\left(e^{-A\Delta}\right)^{1i}y^i}{\sqrt{\boldsymbol{C}_\Delta^{11}}}\right)\\ &\quad\times\mu_\infty\left(1+\sqrt{V_\Delta^{11}}y^1,y^2,\dots,y^d\right)\mathrm{d}y.\end{aligned} \tag{I.A.0.2}$$

Using the power expansion of the matrix exponential $e^{-A\Delta}=\mathrm{Id}-\Delta A+O\left(\Delta^2\right)$, we have:

$$\begin{aligned}\left(e^{-A\Delta}\right)^{11} &= 1+O\left(\Delta\right),\\ \left(e^{-A\Delta}\right)^{1i} &= \Delta A^{1i}+O\left(\Delta^2\right),\quad i\geq 2.\end{aligned}$$

Simultaneously, we know that $\boldsymbol{C}_\Delta\sim\Delta\left(\Sigma\Sigma^*\right)$ as $\Delta\to 0$. Hence in the integral (I.A.0.2), $h(\dots)$ and $\mu_\infty(\dots)$ converge pointwise to $h(y^1)$ and $\mu_\infty(1,y^2,\dots,y^2)$ respectively, for any y. To pass to the limit for (I.A.0.2), it remains to dominate $h(\dots)\mu_\infty(\dots)$ uniformly in Δ, by an integrable function on $\mathbb{R}^d$.

- On the one hand, using the monotone and non-negative properties of h, we observe that $h(\dots)$ is bounded by $h(\frac{1}{2}y^1+2\sum_{j\geq 2}|y^j|)$ for any y such that $y^1<0$, provided that Δ is small enough.

- On the other hand, since $\boldsymbol{C}_\infty^{-1}$ is symmetric definite positive, $\mu_\infty(\dots)$ is bounded by $\mathrm{Cst}\exp(-\sum_{j\geq 2}|y^j|^2/\mathrm{Cst})$ for some positive constant Cst.

Finally, we easily check that the product of the two bounds is integrable, using in particular that h is integrable on $\mathbb{R}^-$ and bounded over $\mathbb{R}$. Thus the dominated convergence theorem yields

$$\frac{I_h(\Delta)}{\sqrt{\boldsymbol{C}_\Delta^{11}}} \to \int_{-\infty}^0 \int_{\mathbb{R}^{d-1}} h(y^1)\mu_\infty(1, y^2, \dots y^d)\mathrm{d}y,$$
$$I_h(\Delta) \sim \sqrt{\Delta}\sqrt{(\Sigma\Sigma^*)^{11}} \int_{-\infty}^0 h(x)\mathrm{d}x \int_{\mathbb{R}^{d-1}} \mu_\infty(1, y^2, \dots y^d)\mathrm{d}y.$$

We can simply rewrite the last integral. Let $G \overset{\mathrm{d}}{=} \mathcal{N}(0, \boldsymbol{C}_\infty)$:

$$\int_{-\infty}^x \int_{\mathbb{R}^{d-1}} \mu_\infty(y^1, z)\mathrm{d}y^1\mathrm{d}z = \mathbb{P}\left[G^1 \leq x\right] = \int_{-\infty}^x \mu_{\boldsymbol{C}_\infty^{11}}(y^1)\mathrm{d}y^1$$

where we have used that $G^1 \overset{\mathrm{d}}{=} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{11}\right)$ at the second equality. It gives, for any $x \in \mathbb{R}$,

$$\int_{\mathbb{R}^{d-1}} \mu_\infty(x, y^2, \dots y^d)\mathrm{d}y = \mu_{\boldsymbol{C}_\infty^{11}}(x),$$

and therefore the announced result. □

Corollary I.2. *In the limit $\Delta \to 0$, we have*

$$\mathbb{E}\left[Z_\Delta\right] \sim \mathbb{V}\mathrm{ar}\left(Z_\Delta\right) \sim \sqrt{\Delta}\sqrt{\frac{(\Sigma\Sigma^*)^{11}}{2\pi}}\mu_{\boldsymbol{C}_\infty^{11}}(1).$$

Proof. Since Z_Δ takes values in $\{0, 1\}$, if $\mathbb{E}\left[Z_\Delta\right] \to 0$, we have $\mathbb{V}\mathrm{ar}\left(Z_\Delta\right) = \mathbb{E}\left[Z_\Delta\right](1 - \mathbb{E}\left[Z_\Delta\right]) \sim \mathbb{E}\left[Z_\Delta\right]$. Thus, it remains to show the estimate on $\mathbb{E}\left[Z_\Delta\right]$.

Start from (I.A.0.1) and apply Lemma I.2 with $h(\cdot) = N(\cdot)$. We have $\int_{-\infty}^0 N(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}$, hence the result. □

I.B Maximal correlation inequality

In this section, we aim at proving the very useful Theorem I.1.

I.B.1 Gebelein's inequality

We start by an abstract version.

Theorem I.5 ([Jan97, Theorem 10.11]). *Take H, K two closed subspaces of some Gaussian Hilbert space. Define P_{HK} the restriction to H of the orthogonal projection onto K. Define the maximal correlation coefficient between variables A, B respectively measurable w.r.t. the sigma field generated by H and K:*

$$\rho(H,K) = \sup_{A\in L^2(H), B\in L^2(K)} |\mathbb{C}\mathrm{or}\,(A,B)|\,.$$

Then we have:

$$\rho(H,K) = \|P_{HK}\|$$

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

We now restate in a more convenient way the above result in a finite dimensional case and for given Gaussian vectors: we believe such a statement may exist in the literature but we could not find a ready reference. Consider a Gaussian Hilbert space and two Gaussian vectors X, Y in this space. Define then H_X and H_Y the subspaces spanned respectively by X and Y.

Denoting P_{XY} the orthogonal projection on H_Y restricted to H_X, from Theorem I.5 we have that $\rho(H_X, H_Y) = \|P_{XY}\|$. This value is independent of the enclosing Gaussian Hilbert Space.

Using the notations stated above, set

$$\rho\,(X,Y) = \rho(H_X, H_Y) = \|P_{XY}\|.$$

Corollary I.3. *Take X,Y two Gaussian vectors in the same Gaussian Hilbert space and f, g two functions such that $f(X)$, $g(Y)$ are square-integrable. Then we have:*

$$|\mathbb{C}\mathrm{ov}\,(f(X), g(Y))\,| \leq |\rho(X,Y)|\sqrt{\mathbb{V}\mathrm{ar}\,(f(X))\,\mathbb{V}\mathrm{ar}\,(g(Y))}.$$

In preparation of Theorem I.1, we now aim at making more explicit the coefficients $\rho(X,Y)$ in terms of the correlations of the components of X and Y, or of their transforms.

I.B.2 Finite-dimensional Gaussian vectors

Notation. We denote $\mathbb{C}\mathrm{ov}\,(X)$ the covariance matrix of X and $\mathbb{C}\mathrm{ov}\,(X,Y)$ the covariance matrix of X and Y, which is also the upper-right part of the covariance matrix of the vector (X,Y). We reserve the notation $\mathbb{V}\mathrm{ar}\,(X)$ for the variance of a real-valued X.

We set $K_{XX} = \mathbb{C}\mathrm{ov}\,(X)$, $K_{YY} = \mathbb{C}\mathrm{ov}\,(Y)$ and $K_{XY} = \mathbb{C}\mathrm{ov}\,(X, Y)$. As K_{XX} and K_{YY} are symmetricnon-negative definite matrices, there exist O_X, O_Y orthogonal matrices and D_X, D_Y diagonal non-negative definite matrices such that we have:

$$K_{XX} = O_X^* D_X O_X = \left(O_X^* D_X^{1/2} O_X\right)\left(O_X^* D_X^{1/2} O_X\right)^*,$$
$$K_{YY} = O_Y^* D_Y O_Y = \left(O_Y^* D_Y^{1/2} O_Y\right)\left(O_Y^* D_Y^{1/2} O_Y\right)^*.$$

Note that $\left(O_X^* D_X^{1/2} O_X\right)$ and $\left(O_Y^* D_Y^{1/2} O_Y\right)$ are symmetric.

Take now X' and Y' orthonormal basis of respectively H_X and H_Y. It is easy to see that the covariance matrix R of X' and Y' is the projection matrix from H_X on H_Y written in the basis X' and Y' and therefore $\rho(X, Y) = \|R\|$.

I.B.2.1 Non-degenerate case

Assume, in this paragraph, D_X and D_Y are non-degenerate or, equivalently, that K_{XX} and K_{YY} are non-degenerate.

We can choose $X' = \left(O_X^* D_X^{1/2} O_X\right)^{-1} X$ and $Y' = \left(O_Y^* D_Y^{1/2} O_Y\right)^{-1} Y$. Our calculation shows then that

$$R = \left(O_X^* D_X^{1/2} O_X\right)^{-1} K_{XY} \left(O_Y^* D_Y^{1/2} O_Y\right)^{-1}.$$

From the other side, $O_X^* D_X^{1/2} O_X$ and $O_Y^* D_Y^{1/2} O_Y$ are symmetric matrices which square to K_{XX} and K_{YY}. Therefore we have

$$(K_{XX})^{1/2} = O_X^* D_X^{1/2} O_X, \qquad (K_{YY})^{1/2} = O_Y^* D_Y^{1/2} O_Y.$$

From this we have the following proposition.

Proposition I.2. *Take (X, Y) a Gaussian vector. Assume that $\mathbb{C}\mathrm{ov}\,(X)$, $\mathbb{C}\mathrm{ov}\,(Y)$ are non-degenerate. Then we have*

$$\rho(X, Y) = \|\mathbb{C}\mathrm{ov}\,(X)^{-1/2}\, \mathbb{C}\mathrm{ov}\,(X, Y)\, \mathbb{C}\mathrm{ov}\,(Y)^{-1/2}\,\|.$$

Corollary I.4. *If X and Y are orthogonal with non degenerate components ($\mathbb{V}\mathrm{ar}\,(X^i) \neq 0$, $\mathbb{V}\mathrm{ar}\,(Y^j) \neq 0$, for any i, j), then we have $\rho(X, Y) = \|R\|$ with*

$$R^{ij} = \mathbb{C}\mathrm{or}\left(X^i, Y^j\right) = \frac{\mathbb{C}\mathrm{ov}\,(X^i, Y^j)}{\sqrt{\mathbb{V}\mathrm{ar}\,(X^i)\,\mathbb{V}\mathrm{ar}\,(Y^j)}}.$$

I.B.2.2 Degenerate case

In this part, we consider that X or Y is degenerate (or both). We know there are orthonormal matrices O_X and O_Y such that $O_X X$ and $O_Y Y$ both have diagonal covariances. These diagonals can have zero values; taking only the variables of $O_X X$ and $O_Y Y$ that have non-zero variances, we get a couple of orthonormal families. Applying results from previous paragraph and using the fact that extending a matrix with zeroes doesn't change its operator norm, we have the following.

Proposition I.3. *Let (X, Y) be a Gaussian vector and O_X, O_Y two orthogonal matrices such that $X_O := O_X X$ and $Y_O := O_Y Y$ are respectively orthogonal families. Then the maximal correlation coefficient verifies $\rho(X, Y) = \|R\|$ with*

$$R^{ij} = \mathbb{C}\mathrm{or}\left(X_O^i, Y_O^j\right) = \begin{cases} \frac{\mathbb{C}\mathrm{ov}\left(X_O^i, Y_O^j\right)}{\sqrt{\mathbb{V}\mathrm{ar}\left(X_O^i\right)\mathbb{V}\mathrm{ar}\left(Y_O^j\right)}} & \text{if } \mathbb{V}\mathrm{ar}\left(X_O^i\right)\mathbb{V}\mathrm{ar}\left(Y_O^j\right) \neq 0, \\ 0 & \text{if } \mathbb{V}\mathrm{ar}\left(X_O^i\right)\mathbb{V}\mathrm{ar}\left(Y_O^j\right) = 0. \end{cases}$$

I.B.3 Application to functions of Gaussian processes

We are now in a position to give the maximal correlation between X_s and X_t for $t, s \in \mathbb{R}^+$, in terms of the OU parameters (A, Σ) (Assumption **(H)**).

Proposition I.4. *Using the previous notation, we have for any $s, t \in \mathbb{R}^+$*

$$\rho(X_s, X_t) \leq \sqrt{\frac{v_M}{v_m}} e^{-a_0|t-s|},$$

where $a_0 := \min_{\lambda \in \mathrm{Sp}(A)} \mathcal{R}e(\lambda), v_M = \max_{\lambda \in \mathrm{Sp}(\boldsymbol{C}_\infty)} \lambda, v_m = \min_{\lambda \in \mathrm{Sp}(\boldsymbol{C}_\infty)} \lambda$.

Proof. Let $t \geq s \geq 0$. From (1.5.3) and (1.5.6), we have

$$\mathbb{C}\mathrm{ov}\left(X_s\right) = \mathbb{C}\mathrm{ov}\left(X_t\right) = \boldsymbol{C}_\infty, \qquad \mathbb{C}\mathrm{ov}\left(X_t, X_s\right) = e^{-A(t-s)}\boldsymbol{C}_\infty.$$

Since $\boldsymbol{C}_\infty$ is non degenerate (owing to Assumption **(H)**), we can apply proposition I.2, to get

$$\rho(X_s, X_t) = \|\boldsymbol{C}_\infty^{-1/2} e^{-A(t-s)} \boldsymbol{C}_\infty^{1/2}\|.$$

The bound on $\rho(X_s, X_t)$ is a consequence of sub-multiplicativity of the operator norm. □

We immediately deduce the following.

Corollary I.5. *For any $\varphi, \phi : \mathbb{R}^d \to \mathbb{R}$ square-integrable w.r.t. ν_∞, we have*

$$\left|\mathbb{C}\mathrm{ov}\left(\varphi(X_s), \phi(X_t)\right)\right| \leq \sqrt{\frac{v_M}{v_m}} e^{-a_0|t-s|} \sqrt{\mathbb{V}\mathrm{ar}\left(\varphi(X_s)\right)\mathbb{V}\mathrm{ar}\left(\phi(X_t)\right)}.$$

We can proceed to the proof of Theorem I.1. In its setting we have $t > s \geq 0$ and φ, ϕ take as arguments respectively $(X_u)_{0\leq u\leq s}$ and $(X_v)_{v\geq t}$.

Proof. For ease of writing, denote $\varphi_s = \varphi\left((X_u)_{0\leq u\leq s}\right)$ and $\phi_t = \phi\left((X_v)_{v\geq t}\right)$. Without loss of generality, we can assume $\mathbb{E}[\varphi_s] = \mathbb{E}[\phi_t] = 0$.We repeatedly use the Markov property and the tower property of conditional expectation to write

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\varphi_s, \phi_t) &= \mathbb{E}[\varphi_s\phi_t] \\
&= \mathbb{E}[\varphi_s\mathbb{E}[\phi_t \mid \mathcal{F}_s]] \\
&= \mathbb{E}[\varphi_s\mathbb{E}[\phi_t \mid X_s]] \\
&= \mathbb{E}[\mathbb{E}[\varphi_s \mid X_s]\,\mathbb{E}[\phi_t \mid X_s]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[\varphi_s \mid X_s]\,\phi_t \mid X_s]] \\
&= \mathbb{E}[\mathbb{E}[\varphi_s \mid X_s]\,\phi_t] \\
&= \mathbb{E}[\mathbb{E}[\varphi_s \mid X_s]\,\mathbb{E}[\phi_t \mid X_t]] \\
&= \mathbb{C}\mathrm{ov}(\mathbb{E}[\varphi_s \mid X_s], \mathbb{E}[\phi_t \mid X_t]).
\end{aligned}$$

We define now $f(X_s) = \mathbb{E}[\varphi_s \mid X_s]$, $g(X_t) = \mathbb{E}[\phi_t \mid X_t]$ and apply Corollary I.5:

$$\mathbb{C}\mathrm{ov}(f(X_s), g(X_t)) \leq \sqrt{\frac{v_M}{v_m}}e^{-a_0|t-s|}\sqrt{\mathbb{V}\mathrm{ar}(f(X_s))\,\mathbb{V}\mathrm{ar}(g(X_t))}.$$

Then, the announced inequality of Theorem I.1 stems from the standard decomposition $\mathbb{V}\mathrm{ar}(h) = \mathbb{V}\mathrm{ar}(\mathbb{E}[h \mid \mathcal{G}]) + \mathbb{E}[\mathbb{V}\mathrm{ar}(h \mid \mathcal{G})] \geq \mathbb{V}\mathrm{ar}(\mathbb{E}[h \mid \mathcal{G}])$ for any sigma-field $\mathcal{G}$ and any square-integrable variable h. □

I.C Central limit theorem for discontinuous functions of OU processes

In this section, we broaden the domain of application of the properties from [Flo84], precisely extending it to the case of a bounded function with a single point of irregularity, as $x \mapsto \mathbb{1}_{x\geq 1}$. In this section, the process X is one-dimensional.

I.C.1 Itô formula for piecewise C^2 function

First, we recall a generalization of Itô's lemma to functions that are not C^2.

Lemma I.3. *Let g be a function $g : \mathbb{R} \to \mathbb{R}$, twice differentiable and g'' is continuous except at a single point z. Assume also that $\forall x \neq z, |g''(x)| \leq K$. Then Itô's formula applies to g, i.e.:*

$$\mathrm{d}g(X_t) = g'(X_t)\mathrm{d}X_t + \frac{\sigma^2}{2}g''(X_t)\mathrm{d}t.$$

Proof. As we have $g \in C^1$ and g'' is integrable on any interval, we can apply [RW87, Lemma 45.9]. □

I.C.2 Solution to Poisson equation $LF = -f$

Consider now a function f smooth with the exception of a single point. Consider also L the infinitesimal generator associated to X verifying (2.3.1):

$$L\phi(x) = -ax\frac{\partial \phi}{\partial x}(x) + \frac{\sigma^2}{2}\frac{\partial^2 \phi}{\partial x^2}(x).$$

The next lemma gives a solution to the Poisson equation $LF = -f$.

Lemma I.4. *Let f be a bounded function. Then*

$$F(x) := -\frac{2}{\sigma^2}\int_0^x \frac{\mathrm{d}y}{\mu_\infty(y)}\int_{-\infty}^y f(u)\mu_\infty(u)\mathrm{d}u \tag{I.C.2.1}$$

is a solution to $LF = -f$.

With $f = \mathbb{1}_{[1,+\infty[} - N\left(-1/\sqrt{v_\infty}\right)$, we have:

$$F(x) = \frac{2}{\sigma^2}\int_0^x \frac{N\left(\frac{u\wedge 1}{\sqrt{v_\infty}}\right) - N\left(\frac{1}{\sqrt{v_\infty}}\right)N\left(\frac{u}{\sqrt{v_\infty}}\right)}{\mu_\infty(u)}\mathrm{d}u. \tag{I.C.2.2}$$

Proof. Simple computations give:

$$\begin{aligned}
\frac{\partial F}{\partial x}(x) &= -\frac{2}{\sigma^2\mu_\infty(x)}\int_{-\infty}^x f(u)\mu_\infty(u)\mathrm{d}u,\\
\frac{\partial^2 F}{\partial x^2}(x) &= -2\frac{2ax}{\sigma^2\mu_\infty(x)}\int_{-\infty}^x \frac{f(u)}{\sigma^2}\mu_\infty(u)\mathrm{d}u - 2\frac{f(x)}{\sigma^2}.
\end{aligned}$$

We then deduce $LF = -f$.

Choosing $f = \mathbb{1}_{[1,+\infty[} - N\left(-1/\sqrt{v_\infty}\right) = 1 - \mathbb{1}_{]-\infty,1[} - N\left(-1/\sqrt{v_\infty}\right) = -\mathbb{1}_{]-\infty,1[} + N\left(1/\sqrt{v_\infty}\right)$, we get:

$$\begin{aligned}
F(x) &= \frac{2}{\sigma^2}\left(\int_0^x \frac{\mathrm{d}y}{\mu_\infty(y)}\int_{-\infty}^{y\wedge 1}\mu_\infty(u)\mathrm{d}u - \int_0^x \frac{N\left(\frac{1}{\sqrt{v_\infty}}\right)}{\mu_\infty(y)}\mathrm{d}y\int_{-\infty}^y \mu_\infty(u)\mathrm{d}u\right)\\
&= \frac{2}{\sigma^2}\int_0^x \frac{N\left(\frac{u\wedge 1}{\sqrt{v_\infty}}\right) - N\left(\frac{1}{\sqrt{v_\infty}}\right)N\left(\frac{u}{\sqrt{v_\infty}}\right)}{\mu_\infty(u)}\mathrm{d}u.
\end{aligned}$$

□

We now establish bounds on F and its derivatives.

Lemma I.5. *Assume that f is bounded and such that $\nu_\infty(f) = 0$. Define F as in* (I.C.2.1). *Then there exist finite constants $C_{\text{(I.C.2.3)}}, C_{\text{(I.C.2.4)}}, C_{\text{(I.C.2.5)}}$ (depending only on the model) such that, for any $x \in \mathbb{R}$,*

$$|F(x)| \leq C_{\text{(I.C.2.3)}}|f|_\infty|x|, \tag{I.C.2.3}$$

$$|F'(x)| \leq C_{\text{(I.C.2.4)}}|f|_\infty \left(1 \wedge \frac{1}{|x|}\right), \tag{I.C.2.4}$$

$$|F''(x)| \leq C_{\text{(I.C.2.5)}}|f|_\infty. \tag{I.C.2.5}$$

Proof. Using the assumption that $\nu_\infty(f) = \int_{\mathbb{R}} f(x)\mu_\infty(x)\mathrm{d}x = 0$, we can write:

$$\begin{aligned}
F'(x) &= -\frac{2}{\sigma^2\mu_\infty(x)}\int_{-\infty}^{x} f(u)\mu_\infty(u)du \\
&= \frac{2}{\sigma^2\mu_\infty(x)}\int_{x}^{+\infty} f(u)\mu_\infty(u)du, \\
|F'(x)| &\leq \frac{2}{\sigma^2\mu_\infty(x)}\left(\int_{-\infty}^{x}|f(u)|\mu_\infty(u)du \wedge \int_{x}^{+\infty}|f(u)|\mu_\infty(u)du\right).
\end{aligned}$$

Applying now the assumption that f is bounded, we have:

$$|F'(x)| \leq \frac{2|f|_\infty}{\sigma^2}\frac{N\left(-\frac{|x|}{\sqrt{v_\infty}}\right)}{\mu_\infty(x)}. \tag{I.C.2.6}$$

Using the classic inequality $N(-|x|) \leq e^{-x^2/2}$ yields

$$|F'(x)| \leq \frac{2|f|_\infty}{\sigma^2}\frac{\exp\left(-\frac{x^2}{2v_\infty}\right)}{\frac{1}{\sqrt{2\pi v_\infty}}\exp\left(-\frac{x^2}{2v_\infty}\right)} \leq 2\sqrt{\frac{\pi}{a\sigma^2}}|f|_\infty. \tag{I.C.2.7}$$

By integrating, we complete the proof of Equation (I.C.2.3). Next, we use the Mills inequality,

$$N\left(-|x|\right) \leq \frac{e^{-x^2/2}}{\sqrt{2\pi}|x|},$$

which combined with (I.C.2.6) gives

$$|x||F'(x)| \leq \frac{2|f|_\infty}{\sigma^2}|x|\frac{N\left(-\frac{|x|}{\sqrt{v_\infty}}\right)}{\mu_\infty(x)} \leq \frac{2|f|_\infty}{\sigma^2}|x|\frac{\frac{1}{\sqrt{2\pi}\frac{|x|}{\sqrt{v_\infty}}}}{\frac{1}{\sqrt{2\pi v_\infty}}} \leq \frac{2|f|_\infty}{\sigma^2}v_\infty.$$

Thus, joined with (I.C.2.7), Inequality (I.C.2.4) is proved. Last, as $LF = -f$,

$$|F''(x)| = \frac{2}{\sigma^2}\left|axF'(x) - f(x)\right| \leq \frac{2}{\sigma^2}\left(a\frac{2v_\infty}{\sigma^2} + 1\right)|f|_\infty = \frac{4}{\sigma^2}|f|_\infty,$$

which proves (I.C.2.5). □

I.C.3 CLT for multi-dimensional continuous-time martingales

Lemma I.6 ([vZ00, Theorem 4.1]). *Let $(M_t; \mathcal{F}_t : t \geq 0)$ be a d-dimensional continuous local martingale. If there exist invertible, non-random $d \times d$-matrices $(K_t : t \geq 0)$ such that as $t \to \infty$*

- $K_t \langle M \rangle_t K_t^* \xrightarrow{\mathbb{P}} \eta\eta^*$ *where η is a random $d \times d$-matrix;*
- $|K_t| \to 0$*;*

then, for each $\mathbb{R}^k$-valued random vector X defined on the same probability space as M, we have

$$(K_t M_t, X) \xrightarrow{d} (\eta Z, X) \qquad \text{as} \quad t \to \infty,$$

where $Z \stackrel{\mathrm{d}}{=} \mathcal{N}(0, \mathrm{Id})$ and Z is independent of (η, X).

CHAPTER II

Properties of the stochastic graph in a model of interbank lending

In this Chapter, we will analyze the characteristics of the Ornstein-Uhlenbeck model of interbank lending. First, we develop an intuitive approach to the model and restate it in order to eliminate an issue related to the spectrum of $\boldsymbol{A}_0$. Second, we make a qualitative analysis of the stochastic graph resulting from thresholding differences in reserves, and compare the graph to the network that structure encoded in $\boldsymbol{A}_0$. Third, we analyze quantitatively the stochastic graph using an approach designed for evolving graphs.

1 Inter-bank lending model restatement

In Chapter I, we present the Ornstein-Uhlenbeck process as a candidate for modeling interbank lending. Indeed, the multi-dimensional linear structure is natural to model interactions between institutions. However, a model of interbank lending should verify for instance that every outflow of reserves from one institution is an inflow for another.

1.1 Intuitive approach to modeling interbank lending

Recall the model of interbank lending:

$$\mathrm{d}X_t^i = -\boldsymbol{A}_0^{ii} X_t^i \mathrm{d}t - \sum_{j \neq i} \boldsymbol{A}_0^{ij} X_t^j \mathrm{d}t + (\boldsymbol{\Sigma} \mathrm{d}W_t)^i. \tag{1.1.1}$$

We interpret the $\boldsymbol{A}_0^{ij} X_t^j$ as lending from i directed to j. Lending should happen when X_t^j becomes negative, hence we will require $\boldsymbol{A}_0^{ij}$ to be negative. As each outflow of

money for one bank is an inflow of money for another, bank i will receive flows equal from each bank j equal to $\boldsymbol{A}_0^{ji} X_t^i$. For this reason we identify

$$\boldsymbol{A}_0^{ii} = -\sum_{j \neq i} \boldsymbol{A}_0^{ji}.$$

The SDE becomes then:

$$\mathrm{d}X_t^i = \sum_{j \neq i} (\boldsymbol{A}_0^{ji} X_t^i - \boldsymbol{A}_0^{ij} X_t^j)\mathrm{d}t + (\boldsymbol{\Sigma}\mathrm{d}W_t)^i$$

which we interpret, in terms of the relation of banks i and j, as i borrowing in net from j proportionally to $\boldsymbol{A}_0^{ji} X_t^i - \boldsymbol{A}_0^{ij} X_t^j$. When that number becomes negative, borrowing becomes lending.

An interesting simplification comes from assuming that $\boldsymbol{A}_0$ is a symmetric matrix. Then borrowing is proportional to $\boldsymbol{A}_0^{ij}(X_t^i - X_t^j)$, a linear function of the difference in reserves. An additional advantage of assuming that $\boldsymbol{A}_0$ is symmetric is that its spectrum is real. For these reasons, we will assume in the rest of the chapter:

(H6) The matrix $\boldsymbol{A}_0$ is symmetric, for any $i \neq j$, $\boldsymbol{A}_0^{ij} \leq 0$ and $\boldsymbol{A}_0^{ii} = -\sum_{j \neq i} \boldsymbol{A}_0^{ji}$.

1.2 Linear reconfiguration

The formulation of the lending model above used that each outflow is equal to an outflow. In other terms, the total amount of money has no deterministic drift. We have

$$\mathrm{d}(\mathbb{1}^\top X_t) = -\mathbb{1}^\top \boldsymbol{A}_0 X_t + \mathbb{1}^\top \boldsymbol{\Sigma}\mathrm{d}W_t = \mathbb{1}^\top \boldsymbol{\Sigma}\mathrm{d}W_t \tag{1.2.1}$$

as we can verify $\mathbb{1}^\top \boldsymbol{A}_0 = 0$ when $\mathbb{1}$ is a vector of ones of the suitable size. This means for instance that $\boldsymbol{A}_0$ doesn't verify Assumption (H1). To remedy this, we observe that Equation (1.2.1) states that the average reserve has no deterministic drift. For this reason, we will choose to see the individual reserves as moving around the the average which is completely random. We introduce then then for all $i < d$, the reserve deviations $\tilde{X}_t^i = X_t^i - \bar{X}_t$ where $\bar{X}_t = d^{-1}\mathbb{1}^\top X_t$ is the average reserve. Then we can write the linear transformation

$$X_t = \boldsymbol{L}\begin{pmatrix} \tilde{X}_t \\ \bar{X}_t \end{pmatrix}, \qquad \boldsymbol{L} = \begin{pmatrix} \mathrm{Id} & \mathbb{1} \\ -\mathbb{1}^\top & 1 \end{pmatrix}, \boldsymbol{L}^{-1} = \begin{pmatrix} \mathrm{Id} - d^{-1}\mathbb{1}\mathbb{1}^\top & -d^{-1}\mathbb{1} \\ d^{-1}\mathbb{1}^\top & d^{-1} \end{pmatrix}.$$

Form this, we have the dynamic for the deviations and the average:

$$\mathrm{d}\begin{pmatrix} \tilde{X}_t \\ \bar{X}_t \end{pmatrix} = -\boldsymbol{L}^{-1}\boldsymbol{A}_0\boldsymbol{L}\begin{pmatrix} \tilde{X}_t \\ \bar{X}_t \end{pmatrix}\mathrm{d}t + \boldsymbol{L}^{-1}\boldsymbol{\Sigma}\mathrm{d}W_t$$

and we compute that

$$\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\boldsymbol{L}^{-1})^\top = \begin{pmatrix} \mathrm{Id} - d^{-1}\mathbb{1}\mathbb{1}^\top & 0 \\ 0 & d^{-1} \end{pmatrix} \text{ when } \boldsymbol{\Sigma} = \mathrm{Id},$$
$$\boldsymbol{L}^{-1}\boldsymbol{A}_0\boldsymbol{L} = \begin{pmatrix} \tilde{\boldsymbol{A}}_0 & 0 \\ 0 & 0 \end{pmatrix} \text{ with } \tilde{\boldsymbol{A}}_0 = \boldsymbol{A}_0^{|d-1}(\mathrm{Id} + \mathbb{1}\mathbb{1}^\top),$$

where $\boldsymbol{A}_0^{|d-1}$ denotes the upper-left $(d-1)\times(d-1)$ sub-matrix of $\boldsymbol{A}_0$. We see here that the average $\bar{X}$ has no deterministic part and that the deviations $\tilde{X}$ are independent of $\bar{X}$.

The volatility parameter becomes $\boldsymbol{L}^{-1}\boldsymbol{\Sigma}$ and will verify assumption (H2), which means that $\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\boldsymbol{L}^{-1})^\top$ will be invertible. Having eliminated the 0 eigenvalue that followed Equation (1.2.1), we will also assum that $\tilde{\boldsymbol{A}}_0$ verifies assumption (H1), which states that its spectrum is positive.

1.3 Application to the Carmona-Fouque model

The model of interbank lending studied in [CFS15, FI13] is stated similarly to Equation (1.1.1) with some specific values for the matrix $\boldsymbol{A_0}$. The model states that each pair of banks is actively lending and borrowing from each other, with the same intensity ad^{-1}, proportionally to the difference in reserves and in the direction of the one that has less reserves. This gives $\boldsymbol{A}_0 = a\mathrm{Id} - ad^{-1}\mathbb{1}\mathbb{1}^\top$. One can easily verify that $\mathbb{1}^\top\boldsymbol{A}_0 = 0$, hence that the average reserves are a martingale. Applying our method from the preceding Section 1.2, we get a dynamic for deviations from the average with parameter $\tilde{\boldsymbol{A}}_0 = a\mathrm{Id}$.

2 Examples of small interbank lending structures

In a interbank lending market governed by a stationary dynamic as in Equation (1.1.1), one is interested in the underlying structure $\boldsymbol{A}_0$. However, in practice one can observe only the lending activity, which depends on the reserve values for the banks. In this Section, we compare structure $\boldsymbol{A}_0$ to the observed evolving graph, as defined in Chapter I. We take here a real parameter r, and define a graph Y_t at any time t by

$$Y_t^{ij} = \mathbb{1}_{|X_t^i - X_t^j| \geq r}. \tag{2.0.1}$$

In the Carmona-Fouque model, this is equivalent to a thresholded observation of the lending activity, as it is proportional to the differences in reserves $|X_t^i - X_t^j|$. The choice of r will obviously impact the observed graphs. If r is close to 0, we will observe usually a complete graph, while if r is chosen very large then the graph will have no edges. We will choose r in a way that will give a reasonable value for the density, far enough from 0 and 1 that will lets us see variations.

2.1 Simulating a Carmona-Fouque lending model

In the Carmona-Fouque model, each bank is equivalent to any other, yielding a set of similar trajectories surrounding the mean. We present an example trajectory with $d = 10$ banks in Figure II.1, where $a = 1$ and $\mathbf{\Sigma} = \mathrm{Id}$.

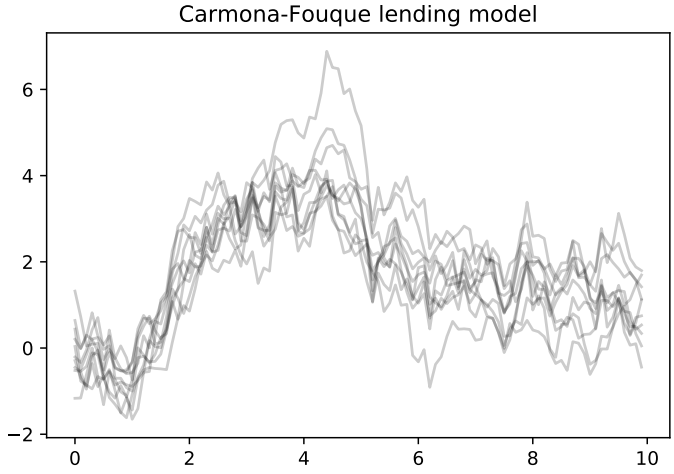
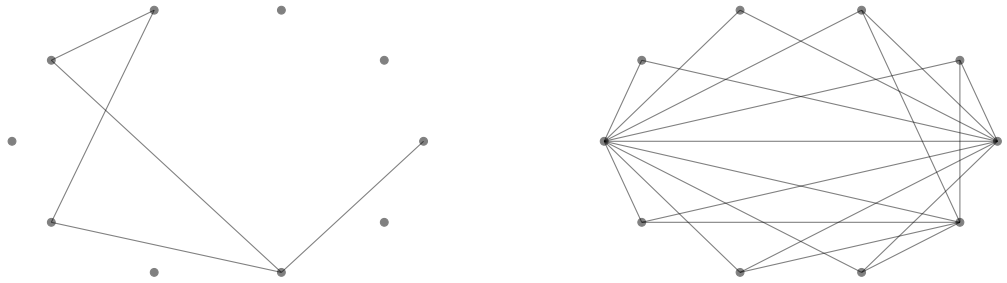


Figure II.1 – Example trajectory for a Carmona-Fouque lending model in dimension 10. The simulation uses a time-step equal to 0.1.

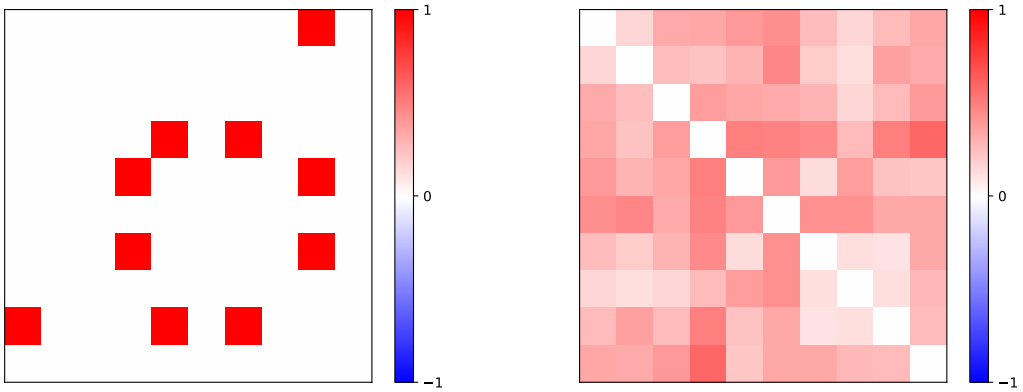
We then define the stochastic graph using Equation (2.0.1) and choosing $r = 1$. This gives an average density in our experiments of around 0.31. Depending on the situation, the observed graph can be quite different. We present in Figure II.2 two examples of graphs obtained from the trajectory presented in Figure II.1. The much larger density at time $t = 4.5$, equal to 0.47 compared to 0.11 at time $t = 1$, is due to a higher observed variance at $t = 4.5$. This feature, completely natural for Ornstein-Uhlenbeck processes, makes also sense in a real lending market. Indeed, decreased lending activity is not necessarily happening because of decreased willingness to lend, but could also happen if for random reasons the banks don't need lending at a given time.

The observed graphs show a network structure quite different from the homogeneity and complete connectedness of the Carmona-Fouque model. Not surprisingly, one shouldn't use a single snapshot of the lending activity to understand the lending structure, see Figure II.3a below. However, even after we average 100 snapshots, we observe a structure that appears to have some structure, see Figure II.3b below, that is absent in the Carmona-Fouque model.



(a) Time $t = 1$. (b) Time $t = 4.5$.

Figure II.2 – Observed graphs Y at two different times.



(a) At time $t = 1$. (b) Average over the interval.

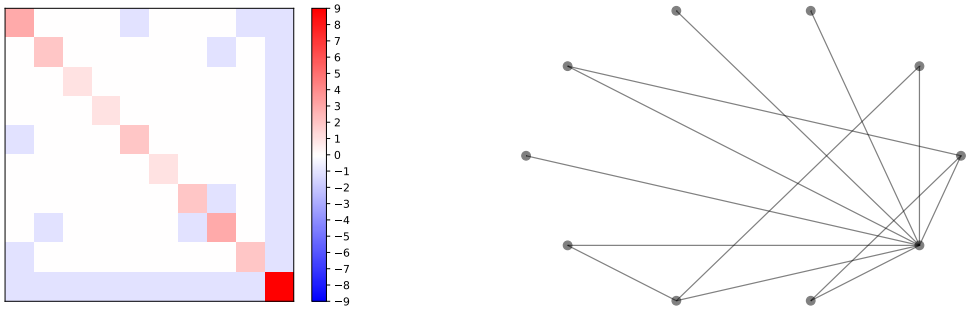
Figure II.3 – Observed adjacency matrix (left) and average of 100 adjacency matrices (right).

2.2 Simulating a sparse lending model

We consider here an alternative model which allows for heterogeneity among banks and respects the observed sparsity of real interbank lending models. We also add a central institution, connected to all other institutions. It may represent a central bank, or any other institution that has a highly central position in the network. The existence of such a central institution in our model allows us to be sure to get a connected network. Indeed, we add all other edges in the network randomly. In this Section, we use a small network with 10 banks interacting according to the mattrix presented in Figure II.4a. The resulting network representation is given in Figure II.4b. Most of the lending links involve the central institution.

The resulting trajectory is not very different on a first view from those observed for the Carmona-Fouque model. We chose to precise which component corresponds to the central institution, but one could hardly identify it otherwise.

We continue with defining the graph using Equation (2.0.1), with $r = 1$ as before.



(a) Adjacency matrix.

(b) Graph representation.

Figure II.4 – Representation of the lending model as an adjacency matrix (left) and as a graph (right). The last line and column of the matrix represent the central institution, which is visibly highly connected in the lending network.

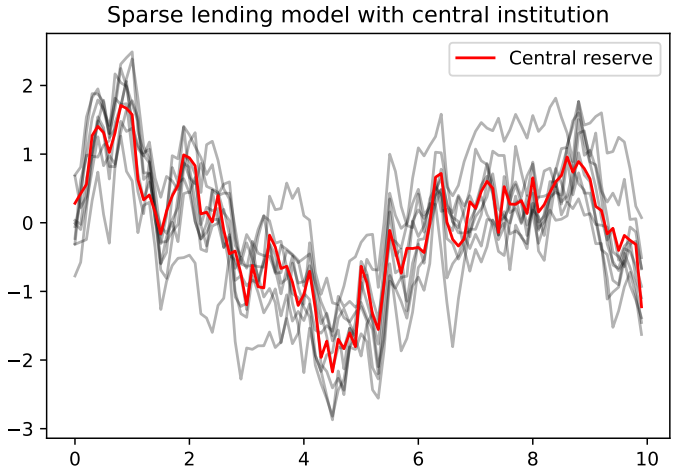
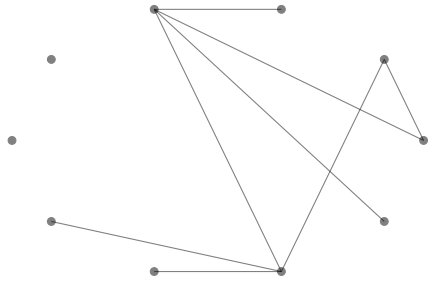
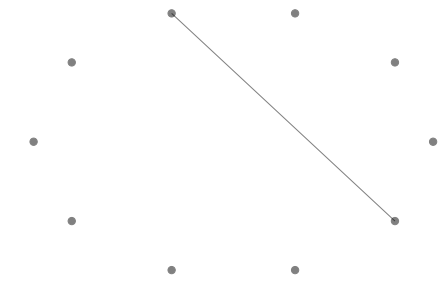


Figure II.5 – Example trajectory for the sparse lending model in dimension 10. The simulation uses a time-step equal to 0.1.

The resulting graphs are highly diverse in terms of structure, and we provide an example in Figure II.6a. However, in our model there is lending only if the banks are connected in the lending network from Figure II.4b. Conditioning on the existence of a link, we obtain an observed lending structure with barely any links. We interpret it in the following way. The central institution is highly connected and hence tends to be close to all other institutions in terms of reserves, which means its lending doesn't exceed often the threshold. The periphery institutions tend to be farther apart, but this doesn't lead to higher lending activity as there are not many links between them. As a result, we see that the observation of lending does not necessarily lead to an easy observation of the lending market. Indeed, a highly central institution can be hidden by the thresholded observation exactly because if its high centrality.



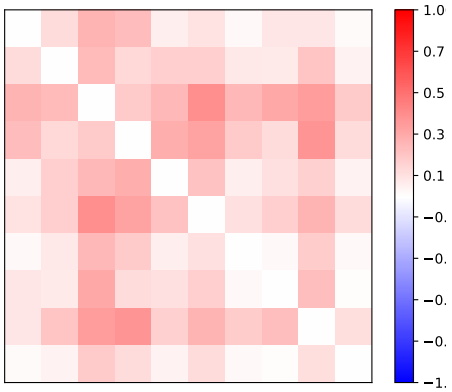
(a) Stochastic graph.



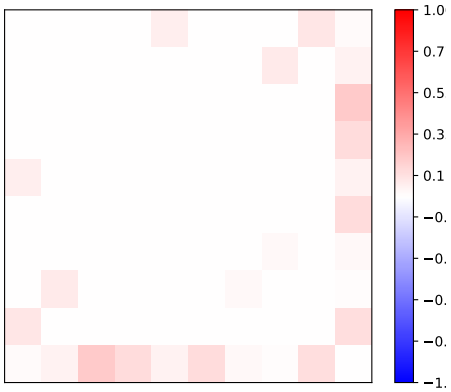
(b) Observed lending.

Figure II.6 – Graph Y_t obtained at $t = 3$ for the trajectory from Figure II.5 (left), and the resulting observed thresholded lending activity, conditioned on the existence of links, see Figure II.4b (right).

One would then naturally use several observations. After averaging, we get the results from Figure . The matrix in Figure II.7a, which is the average adjacency matrix for the graphs Y, doesn't have any specific structure that can help us in identifying the lending links and the key institutions. Obviously, when we average the observed thresholded lending activity, see Figure II.7b, we can clearly identify the links as the non-zero values. Indeed, over an interval it becomes unlikely that two connected components never exceed the threshold that makes the link observable.



(a) Average adjacency of Y.



(b) Average over the interval.

Figure II.7 – Averages of the adjacency matrices for the stochastic graph Y, see Figure II.6a and the thresholded lending activity, see Figure II.6b. The average are done over 100 points.

To conclude, it is not straightforward to estimate the lending structure from a thresholded observation such as the graph Y. If one has access to the lending activity over an interval, one can easily infer the structure of the model, but this isn't

surprising as in the model any lending activity is possible only if a connection exists in the network. This doesn't address issues such as noisy or false information and can't be used directly to estimate the lending intensity - which we assumed in this Section to be binary.

3 Evolving graph properties of the stochastic graphs

In this Section, we are interested in analyzing the stochastic graph defined in Equation (2.0.1) from the viewpoint of evolving graphs. We use here the approach outlined in [BFG+08]. The authors propose a three-pronged approach to analyzing evolving graphs. First, one can look at the evolution of statistics that are typically computed for static graphs. For a review of typical statistics compute on graphs, we refer to [AlB02]. We will be interested in particular in:

- the density of the graph,
- its degree distribution,
- its connectedness,
- its diameter and average shortest path length,
- its clustering.

Second, one can look at parameters that can't be defined on static graphs. We will be interested for instance in the distribution of durations of links. Third, one can look at the properties of some specific nodes. In the case of the Carmona-Fouque model, each node plays the same role, making this question irrelevant. To the contrary, in a sparse model we will observe the way central nodes connect to the rest of the network.

3.1 Analysis of Carmona-Fouque stochastic graphs

We take here a Carmona-Fouque interbank lending model in dimension 100, for which we compute the stochastic graph defined by Equation (2.0.1), where r takes the value 1 and 2. The value 1 is a natural value that we use in order to get a relatively large density and which is comparable to the setting from Section 2. However, as we will see, this creates graphs that are sometimes too dense for other metrics, which is why we will compare the results to those with $r = 2$. In addition to this, we will be able to see how some of the parameters evolve when one changes the value of r.

In Figure II.8, we present the variations of the density through time. For $r = 1$, the average density is observed at 31.5%, and 4.5% for $r = 2$. This decrease is

unsurprising given that the differences between the individual reserves are Gaussian. We also observe that the curve seems to be mean-reverting. Indeed, very large and very small densities are not stable, as they correspond to very dispersed and very concentrated reserves, respectively. Although we are not asserting that the density is an Ornstein-Uhlenbeck process, we perform an estimation of one-dimensional OU parameters estimation on the data in order to quantify the variations of the density. For $r = 1$, the volatility is estimated at $\hat{\sigma}_{r=1} \simeq 0.0238$ and the mean-return parameter at $\hat{a}_{r=1} \simeq 2.125$. For $r = 2$, we have $\hat{\sigma}_{r=2} \simeq 0.0101$ and $\hat{a}_{r=2} \simeq 2.172$. Interestingly, the mean-reversion parameter is very similar, while the volatility decreases, although not as fast as the mean value.

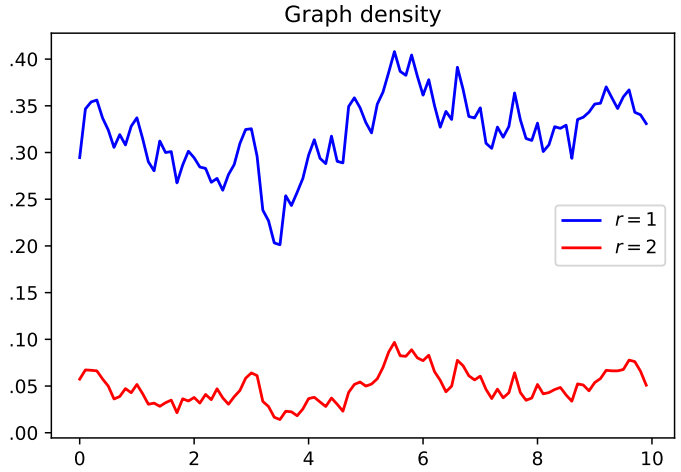
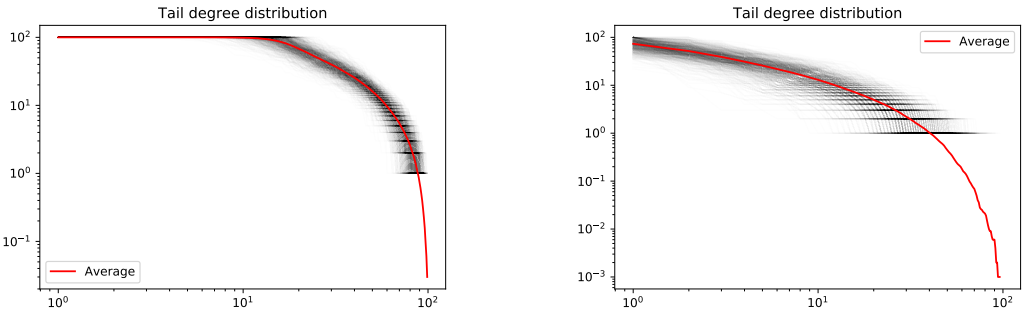


Figure II.8 – Density of the stochastic graphs defined for $r = 1$ and $r = 2$. The average over a longer interval for $r = 1$ is 0.315 and 0.045 for $r = 2$.

In order to analyse the degree distribution, we use the method of tail densities. In this method, instead of producing the degree histogram, we estimate the probabilities of a node to exceed a give degree. For this, we compute for each graph the number of nodes of degree k, where k is varying between 0 and the maximum degree. For real-world graphs, the distributions are shown to follow a power law, see [AlB02], which means the metric plotted in log-log scale is similar to a down-sloping line. In Figure II.9, we show all the tail distributions for a set of 1000 graphs obtained for $r = 1$ and $r = 2$. We can conclude that the stochastic graph does not have a power-law tail distribution. Indeed, the number of high-degree nodes is decreasing far too fast. However, the distributions are flatter when compared to ones obtained for Erodös-Renyi graphs with the same expected density. For these graphs, there are almost no nodes with very high degrees.

The degree distributions show that for $r = 2$, one gets sometimes isolated nodes. In that case, the graph is not connected. For $r = 1$, we haven't observed any isolated

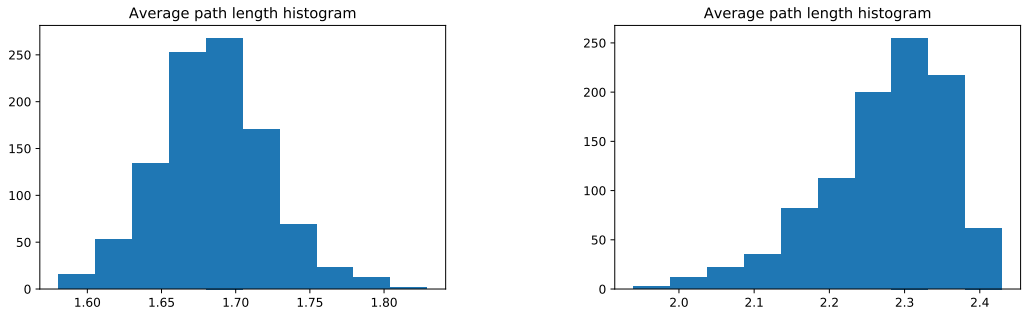


(a) Threshold $r = 1$.

(b) Threshold $r = 2$.

Figure II.9 – Tail degree distributions for graphs constructed with threshold $r = 1$ (left) and $r = 2$ (right). The red line is the average value for each degree, computed over 1000 graphs.

nodes and the graph was always connected. For $r = 2$, we observed up to 80 isolated nodes. However, once they were excluded, we were always left with a single connected component. We conclude that the model creates typically a graph that has one connected component and sometimes isolated nodes. On the connected component, we can compute statistics related to distances between nodes. For instance, for each pair of nodes, we can compute the path length, which is the minimal number of steps along edges one needs to make in order to get from one node to the other. Then, the largest of these values is the diameter; we compute also the average path length. We observe small values for both statistics. For $r = 1$, the diameter was 2 in 94% of the graphs, and 3 in the remaining 6%, and always equal to 3 for graphs computed with $r = 2$. For graphs with diameter 2, nodes that are not directly connected have a common neighbor. We show the average shortest path lengths histograms in Figure II.10. The small values confirm the small distances inside of the connected components of the stochastic graphs.



(a) Threshold $r = 1$.

(b) Threshold $r = 2$.

Figure II.10 – Histograms of the average path length statistic.

In the definition of the stochastic graph from Equation (2.0.1), we have an edge when two reserves are sufficiently far away. Assuming we have a triple of connected nodes, the distance between the largest reserve and the lowest is then at least equal to $2r$. We can expect this to be a rare event compared to the wedges, where two nodes have a common neighbor. Hence e expect the clustering of the resulting to be small. The corresponding statistic called transitivity, formally defined as 3 times the proportion of the number of triangles to the number of wedges, measures the clustering of the graph. For $r = 2$, almost 90% of the graphs had a transitivity of less than 0.01. For $r = 1$, we have an average transitivity of 0.165, which we can attribute to the relatively high density of the graphs. We conclude that the stochastic graph we analyze has typically low clustering, except when it is dense enough.

We end this analysis with the statistic edge duration. For this, we simply count the length between the moment a link is created and the moment when it disappears. We observe that most links disappear after a few time-steps at most. The tail distribution of the edge durations seems to follow an exponential distribution, see Figure II.11. The duration of edges for $r = 2$ is not only shorter than for $r = 1$, but also decreases faster.

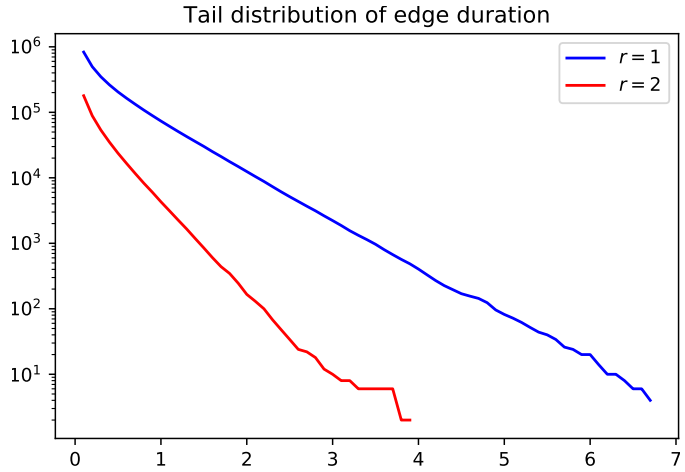
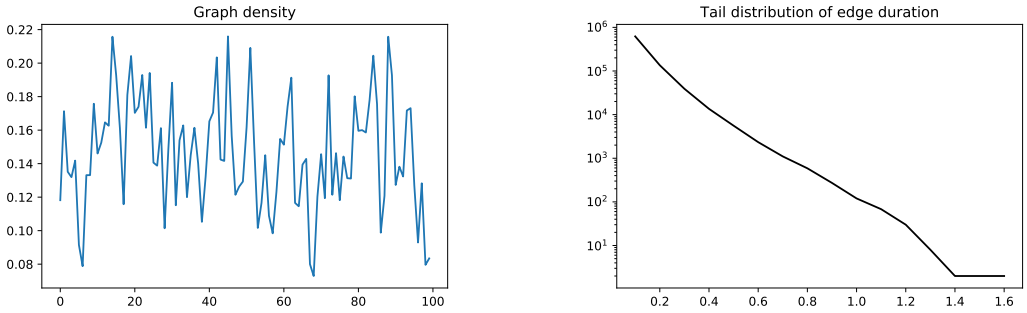


Figure II.11 – Tail distributions of edge durations for stochastic graphs defined with $r = 1$ and $r = 2$.

3.2 Analysis of sparse model stochastic graphs

In this Section, we take a sparse model of interbank lending, as introduced in Section 2.2, in dimension 100 and compute the resulting stochastic graph using the Equation (2.0.1) where $r = 0.5$. Choosing larger values gives a graph that is too sparse to be interesting. The density of the graph is presented in Figure II.12a and averages in our

simulations at 0.145. As in the preceding Section, we fit the data to a one-dimensional Ornstein-Uhlenbeck process, getting an estimated volatility parameter $\hat{\sigma}_{sp} \simeq 0.0385$ and a mean-reversion parameter $\hat{a}_{sp} \simeq 7.336$. The observed volatility is higher than for the Carmona-Fouque graphs, and so is the mean-reversion parameter. As a result, the edges are very short-lived, with an average edge duration of 0.13 and a tail distribution with fast exponential falloff. We have hence a fast rate of creation and disparition of edges.



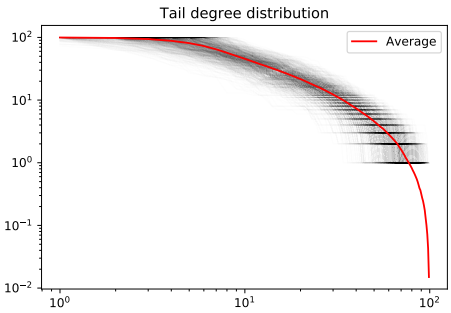
(a) Density.

(b) Edge duration distribution.

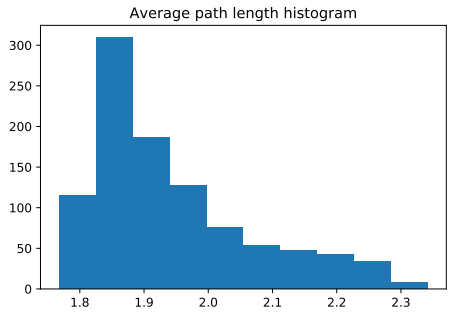
Figure II.12 – Density and edge duration distribution of the stochastic graphs defined for $r = 0.5$ for the sparse model. The density average over a longer interval is 0.145. The average edge duration is 0.13.

We follow with the analysis of the degree-distribution. We present the tail distribution in Figure II.13a. We observe that just as the density sits between the two densities computed for the Carmona-Fouque model, the tail distribution is placed between the two degree distributions from Figure II.9. Again, we get some isolated nodes, but excluding them we have a single connected component. That component has a diameter of 3 in 78% of the time, and a diameter of 2 otherwise. In that sense, the graph also sits between the two cases considered in Section 3.1. The average shortest path lengths fit also globally between the two cases, but we observe that although the diameter is usually equal to 3, the average path length is usually below 2. The mean value for the average shortest path length in our sample is equal to 1.94. The transitivity of the graphs averages at 0.046, which suggests we have again a weak clustering of the nodes.

We finish this analysis by comparing the central node to other nodes in terms of how it connects to the network. We compute for this the degree of the nodes and plot them as a function of time in Figure II.14. We observe that the central node of the lending network ends up very weakly connected to the network. While the average degree of the network is above 14, the central node averages only 5 connections. From one side, it is to be expected. The central node is borrowing and lending with all other nodes, hence is close to the average value and can connect only to the rare



(a) Tail degree distribution.



(b) Average shortest path length histogram.

Figure II.13 – Degree and average shortest path length distributions for the sparse model.

outliers. From the other side, it means that the central institution engages in mostly small but numerous lending activities and could the be hidden by a thresholded observation.

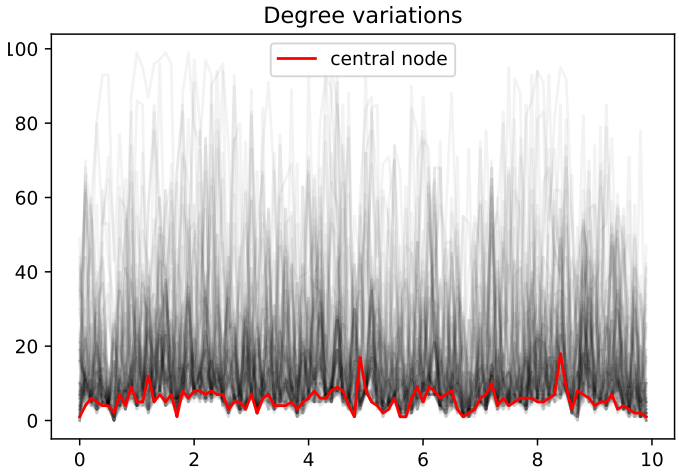


Figure II.14 – Density of the stochastic graphs defined for $r = 0.5$ for the sparse model. The average over a longer interval for 0.145.

4 Conclusion

Starting from an interbank lending model defined as a degenerate Ornstein-Uhlenbeck process, we describe how to corretly simulate corresponding trajectories. We apply this method to simulate some small interbank-lending models. We see that the

stochastic graphs obtained from thresholded imbalances can't be straightforwardly be used to estimate the structure of lending, which is a motivation for Chapter I. We follow with an analysis of the stochastic graphs from the viewpoint of evolving networks. The resulting graphs, both for the homogeneous Carmona-Fouque model and the sparse model, are very concentrated, with small diameter and average shortest path length, but usually without clustering. In that way, they are similar to Erdös-Renyi graphs. However, the degree distributions for the stochastic graph allow for much higher node degrees. Interestingly, the central node that we include in the sparse setting has a very low degree in the stochastic graph.

Part II

Can we improve the inference of Ornstein-Uhlenbeck parameters under a sparsity assumption?

CHAPTER III

Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process

Abstract

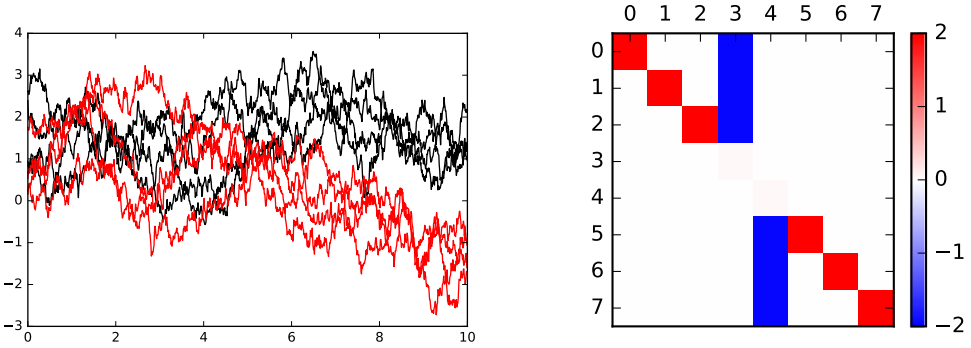
Given the observation of a high-dimensional Ornstein-Uhlenbeck (OU) process in continuous time, we proceed to the inference of the drift parameter under a row-sparsity assumption. Towards that aim, we consider the negative log-likelihood of the process, penalized by an ℓ^1-penalization (Lasso and Adaptive Lasso). We provide both non-asymptotic and asymptotic results for this procedure, by means of a sharp oracle inequality, and a limit theorem in the long-time asymptotics, including asymptotic consistency for variable selection. As a by-product, we point out the fact that for the Ornstein-Uhlenbeck process, one does not need an assumption of restricted eigenvalue type in order to derive fast rates for the Lasso, while it is well-known to be mandatory for linear regression for instance. Numerical results illustrate the benefits of this penalized procedure compared to standard maximum likelihood approaches both on simulations and real-world financial data.

Keywords. Ornstein-Uhlenbeck process; High-dimensional statistics; Sparse estimation; Lasso

MSC 2010. 60G15; 62H12; 62M99

1 Introduction

The Ornstein-Uhlenbeck, also called mean-reverting diffusion process, describes a process which evolves following a deterministic linear part with an added Gaussian



(a) Sample trajectory in 8 dimensions.

(b) Matrix parameter $\boldsymbol{A}_0$.

Figure III.1 – On the right, heat-map representation of a sparse matrix $\boldsymbol{A}_0$. In this particular example, the matrix is chosen in order to have two groups, 0 to 3 and 4 to 7, that are independent and tend to stay close within each group. On the left, plot of the 8 coordinates of the sample trajectory, each group being attributed a different color. Our estimation procedure can be applied to find this kind of hidden structure from non-obvious trajectories.

noise, similarly to a vector-autoregressive process in discrete time. This model is ubiquitous in quantitative finance, for instance the one-dimensional version is used for modeling rates and is called the Vasicek model [Hul09]. In a multi-dimensional setting, it can be therefore used to describe systems with linear interactions perturbed by Gaussian noise, see Figure III.1 below. Among many others, an example of application is inter-bank lending [CFS15, FI13], where lending is a flux of reserves and is proportional to the difference in reserves. A natural question is therefore how to estimate the interaction structure from the observation of the process. Unfortunately, the optimal solution based on the maximum likelihood estimator (MLE) is typically quite inaccurate in high-dimensional settings, because of the well-known curse of dimensionality, see for instance [BvdG11]. However, in real-world applications, the interaction structure is sparse: in the example mentioned above, banks have typically only a few lending partners [GG14, GSV15, BBvL15], as the lending arrangements are typically done on a personal level.

In this paper, we exploit this property by using a sparsity-inducing penalization. Sparse inference using convex penalization has known a strong development in the past decade [BvdG11, Gir14b, FBO12], mostly for linear supervised learning. Quite surprisingly, there is only a single previous attempt to this work to use these techniques in the setting of diffusion processes, in particular for the Ornstein-Uhlenbeck diffusion considered here, see [Sok13], with no theoretical guarantees nor applications on real-world data. The aim of this paper is therefore to fill this gap, and to give a

complete theory for this case, by developing both non-asymptotic results by means of a sharp oracle inequality, see Section 2, and asymptotic results (in the length of observation interval), see Section 3, where we notably establish asymptotic consistency for selection of the support of $\boldsymbol{A}_0$. We also prove a minimax lower-bound for the problem of sparse inference in this model in Section 2. As a by-product, we exhibit a surprising fact in our analysis that for the Ornstein-Uhlenbeck process, one does not need to assume the restricted eigenvalue assumption [BvdG11], which is known to be mandatory for the linear regression model, see for instance [ZWJ14].

1.1 Related work

We investigate in this article the question of recovery of the drift parameter of an Ornstein-Uhlenbeck process from the continuous observation of a single multidimensional trajectory on the interval $[0, T]$. This relates to the much developed area of inference for stochastic processes in continuous time, see [Kut04] for a survey on this topic. This work is also related to the field of high-dimensional statistics, in particular sparse inference, since we use a sparsity assumption on the parameter matrix, we refer to [BvdG11, Gir14a] for surveys on the topic. Indeed, in this paper we study the Lasso [Tib96] and Adaptive Lasso [Zou06] penalizations, applied to the multivariate Ornstein-Uhlenbeck process.

Note that, however, references that propose sparse inference techniques to stochastic processes are quite scarce. A Vector Auto-Regressive (VAR) process can be seen as a discretization in time of an Ornstein-Uhlenbeck process, where $\boldsymbol{A}_0$ is analogous to the VAR transition matrix.The sparse estimation of a VAR process using a Lasso is the subject of [BM15]. However, our work differs on two fundamental points. The first relates to the graph structure implied by $\boldsymbol{A}_0$. While [BM15] assumes sparsity of the whole graph, we place the sparsity on a node level, restricting the maximum degree of the graph, since we work under a row-sparsity assumption, see Assumption (H3) below. This prescribes for instance the existence of nodes which concentrate most connections, in line with observations of the interbank lending system, which note high connectedness only in the core of the network [GSV15]. The second relates to the continuous nature of the considered model. Since the VAR model has finite dimension both in time and space [BM15], it is possible to analyze them jointly in a space of finite dimension equal to the product of the two dimensions. In this paper, we work in continuous time, which forces us to treat time and dimensionality in a fundamentally different way. Another reference is [Sok13], where the Lasso is considered as a strategy to estimate Ornstein-Uhlenbeck parameters in a sparse setting, but no theoretical results nor numerical experiments are provided for this problem. Finally, we consider the particular notion of row-sparsity, which was considered previously for matrix estimation (with additive noise) in [KT15a], instead of the full sparsity of $\boldsymbol{A}_0$.

1.2 The model, main assumptions and tools

Throughout the article we consider a d-dimensional Ornstein-Uhlenbeck process $X = (X_t)_{t\geq 0}$, where $X_t \in \mathbb{R}^d$ for any $t \geq 0$ is solution to the following stochastic differential equation

$$\mathrm{d}X_t = -\boldsymbol{A}_0 X_t \mathrm{d}t + \mathrm{d}W_t, \quad \text{for any } t \geq 0, \tag{1.2.1}$$

where the initial value X_0 is given, $\boldsymbol{A}_0$ is an unknown $d \times d$ matrix to be inferred, and $(W_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$ defined on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$.

We observe the process on an interval $[0, T]$ with $T > 0$. Based on the observation $(X_t)_{t\in[0,T]}$, we want to estimate $\boldsymbol{A}_0$, under the assumption that $\boldsymbol{A}_0$ has sparse rows, namely that a large number of their entries are zeros. Note that $\boldsymbol{A}_0^{ij} \neq 0$ encodes the fact that the trajectory of process j influences the dynamic of process i, which is a property of particular interest for instance in interbank-lending as it implies lending activity from j to i. Row-sparsity implies that each institution borrows from a limited number of institutions. More generally, in time-series analysis, it means that each trajectory is impacted by a limited number of other trajectories.

Throughout the paper, we work under the following assumptions.

(H1) The spectrum of $\boldsymbol{A}_0$ has strictly positive real parts.

(H3) X_0 follows the stationary distribution of the process.

These are standard assumptions for Ornstein-Uhlenbeck processes: Assumption (H1) guarantees ergodicity of $(X_t)_{t\geq 0}$ and existence of a stationary distribution, and is necessary to ensure mean-reversion of the process, the real-world phenomenon that we want to capture and exploit in our modeling. Under (H3) the process is stationary, which is interesting for two reasons. First, it simplifies the results as the initial position doesn't have to be treated differently from the rest of the trajectory. Second, in typical applications one assumes an equilibrium, hence stationarity. For example, in interbank lending there is no reason to assume that the first day of observation is any different from days that precede and follow.

Under these assumptions, the Ornstein-Uhlenbeck verifies interesting properties. For instance, we have

$$X_t \sim \mathcal{N}(0, \boldsymbol{C}_\infty) \quad \text{with} \quad \boldsymbol{C}_\infty = \boldsymbol{C}_\infty(\boldsymbol{A}) := \int_0^\infty e^{-\boldsymbol{A}t} e^{-\boldsymbol{A}^\top t} \mathrm{d}t,$$

for all $t \geq 0$. For this and other classical properties we refer to [KS91], see Section 5.6 herein. In this model, the log-likelihood writes up to a constant as:

$$\mathcal{L}^* := -\frac{1}{2}\log|\boldsymbol{C}_\infty| - \frac{1}{2}X_0^\top \boldsymbol{C}_\infty X_0 - \int_0^T (\boldsymbol{A}X_t)^\top \mathrm{d}X_t - \frac{1}{2}\int_0^T (\boldsymbol{A}X_t)^\top \boldsymbol{A}X_t \mathrm{d}t,$$

where the first two terms become negligible when $T \to +\infty$. We define therefore a maximum-likelihood estimator (MLE) type estimator by minimizing the following normalized negative approximate log-likelihood:

$$\mathcal{L}_T(\boldsymbol{A}) := \frac{1}{T}\int_0^T (\boldsymbol{A}X_t)^\top \mathrm{d}X_t + \frac{1}{2T}\int_0^T (\boldsymbol{A}X_t)^\top \boldsymbol{A}X_t \mathrm{d}t,$$

and it can be written explicitly as

$$\widehat{\boldsymbol{A}}_{MLE} := -\left(\int_0^T \mathrm{d}X_t X_t^\top\right)\left(\int_0^T X_t X_t^\top \mathrm{d}t\right)^{-1}. \tag{1.2.2}$$

For simplicity, we call henceforth MLE this MLE-type estimator when there is no confusion. The inverse in Equation (1.2.2) above exists almost surely as the integral is almost surely a symmetric positive definite matrix (see Section 2 for more details). The asymptotic normality of the MLE-type estimator comes from the asymptotic normality of the MLE which is well-known and from the fact the extra terms in $\mathcal{L}^*$ are of order $O(1/T)$. Indeed we have

$$\sqrt{T}\left(\mathrm{vec}\ \widehat{\boldsymbol{A}}_{MLE} - \mathrm{vec}\ \boldsymbol{A}_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{-1} \otimes \mathrm{Id}\right), \tag{1.2.3}$$

see [Jac01], where $\xrightarrow{\mathcal{L}}$ stands for convergence in distribution, Id stands for the identity matrix in $\mathbb{R}^{d\times d}$, $\otimes$ is the matrix-Kronecker product and vec stands for the vectorization operator, that stacks rows of a $d \times d$ matrix into a flat vector with d^2 entries.

When d is large, the performance of the MLE deteriorates, because of the curse of dimensionality problem, see [BvdG11] and our numerical results in Section 4.2. So, as motivated above, we will reduce dimensionality using a sparsity-inducing penalization on this estimator, see Sections 2 and 3 below. Our analysis relies on the following two central quantities:

$$\widehat{\boldsymbol{C}}_T = \frac{1}{T}\int_0^T X_t X_t^\top \mathrm{d}t \quad \text{and} \quad \boldsymbol{\varepsilon}_T = \frac{1}{T}\int_0^T \mathrm{d}W_t X_t^\top.$$

The matrix $\widehat{\boldsymbol{C}}_T$ is the empirical covariance which satisfies $\mathbb{E}[\widehat{\boldsymbol{C}}_T] = \boldsymbol{C}_\infty$. It is analogous to the Gram matrix in linear regression. The matrix $\boldsymbol{\varepsilon}_T$ is a noise term, note that $(t\boldsymbol{\varepsilon}_t)_{t\geq 0}$ is a martingale with quadratic variation given by $\langle \mathrm{vec}\ t\boldsymbol{\varepsilon}_t\rangle = t\widehat{\boldsymbol{C}}_t \otimes \mathrm{Id}$. Using this matrix notation, we have for instance $\widehat{\boldsymbol{A}}_{MLE} = \boldsymbol{A}_0 - \boldsymbol{\varepsilon}_T\widehat{\boldsymbol{C}}_T^{-1}$ and the matrix formulation of the problem

$$\mathcal{L}_T(\boldsymbol{A}) = \mathrm{tr}\,\boldsymbol{A}^\top\boldsymbol{\varepsilon}_T + \frac{1}{2}(\boldsymbol{A}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T(\boldsymbol{A}-\boldsymbol{A}_0)^\top - \frac{1}{2}\boldsymbol{A}_0\widehat{\boldsymbol{C}}_T\boldsymbol{A}_0^\top. \tag{1.2.4}$$

Notation. For a matrix or a vector x, we denote by $\|x\|_q$ the entrywise ℓ_q norm for any $q \in [1, +\infty]$. The notation $\|x\|_0$ stands for the number of non-zero entries of x, $\|x\|_F = \|x\|_2$ for the Frobenius norm of x when it is a matrix; we consider also the Euclidean inner product $\langle \boldsymbol{U}, \boldsymbol{V} \rangle_F = \operatorname{tr} \boldsymbol{U}^\top \boldsymbol{V}$, where $\operatorname{tr} \boldsymbol{M}$ is the trace of a matrix $\boldsymbol{M}$ and define $\|\boldsymbol{M}\|_{\mathrm{op}}$ as the operator norm of $\boldsymbol{M}$. We also denote by $\sigma_{\min}(\boldsymbol{A})$ the smallest eigenvalue of a symmetric $\boldsymbol{A}$, and $\operatorname{diag}(\boldsymbol{A})$ stands for the vector formed by the diagonal of $\boldsymbol{A}$. We also denote by $\operatorname{supp}(x)$ the support of x, i.e. the set of indices of the non-null coordinates of x, where x is a matrix or a vector. Given a set of indices $\mathcal{I}$, we denote by $x_{|\mathcal{I}}$ the restriction of x to the indices in $\mathcal{I}$.

Moreover, $\operatorname{Sp} \boldsymbol{A}$ is the spectrum of $\boldsymbol{A}$ and $\operatorname{diag} \boldsymbol{A}$ is the extraction of the diagonal of $\boldsymbol{A}$. Additionally, we define

$$\|X\|_{L^2}^2 = \frac{1}{T} \int_0^T |X_t|_2^2 \mathrm{d}t \quad \text{and} \quad \langle X, Y \rangle_{L^2} = \frac{1}{T} \int_0^T X_t^\top Y_t \mathrm{d}t,$$

that correspond to the empirical norm and inner products along the observed trajectory of $(X_t)_{t \geq 0}$.

1.3 Main results and organization of the paper

In Section 2 we introduce the Lasso estimator of $\boldsymbol{A}_0$. Our main result, concerning non-asymptotic error bounds, is Theorem III.1. We show that this upper bound is asymptotically of the same order, up to logarithmic terms, as the lower bound we have in Theorem III.2. We conclude the section with Theorem III.3 which is an interesting by-product of the proof of Theorem III.1 and which states that a Restricted Eigenvalue condition is valid in our setting, when $\boldsymbol{A}_0$ is symmetric. In Section 3 we introduce the Adaptive Lasso estimator and prove in Theorem III.4 its asymptotic normality and support recovery properties. Numerical experiments are provided in Section 4, where we illustrate the benefits of sparse inference over direct maximum likelihood estimation. In Section 6 we provide the proofs of the properties from the preceding Sections.

2 Non-asymptotic error bounds for Lasso

Given a regularization parameter $\lambda > 0$, we define the Lasso estimator by:

$$\widehat{\boldsymbol{A}} = \operatorname*{arg\,min}_{\boldsymbol{A} \in \mathbb{R}^{d \times d}} \mathcal{L}_T(\boldsymbol{A}) + \lambda \|\boldsymbol{A}\|_1.$$

The uniqueness of $\widehat{\boldsymbol{A}}$ derives from the strict convexity of $\mathcal{L}_T$, which comes from the fact $\widehat{\boldsymbol{C}}_T$ is a.s. a positive definite matrix, see Equation (1.2.4). Indeed, observe that for any $u \in \mathbb{R}^d$, $\|u\|_2 = 1$, we have $u^\top \widehat{\boldsymbol{C}}_T u = T^{-1} \int_0^T (u^\top X_t)^2 \mathrm{d}t$ which can be zero only

if the trajectory is included in a hyperplane of $\mathbb{R}^d$. The observation length $T > 0$ is fixed in the whole Section. We also fix an integer $1 \leq s \leq d$ and express the sparsity of $\boldsymbol{A}_0$ in the following assumption:

(H4) The true parameter is row-s-sparse, i.e.

$$\|\boldsymbol{A}_0^{i,\bullet}\|_0 \leq s \quad \text{for all } i = 1, \dots, d,$$

where $\boldsymbol{A}^{i,\bullet}$ stands for the vector such that for any $j \leq d$, $(\boldsymbol{A}^{i,\bullet})^j = \boldsymbol{A}^{ij}$ for any matrix $\boldsymbol{A}$.

This assumption notably differs from a sparsity assumption on the whole matrix parameter, but has already been used in matrix estimation, for instance in [KT15b] for additive noise. We also need to introduce a technical hypothesis on the deviation of $\widehat{\boldsymbol{C}}_T$ from $\boldsymbol{C}_\infty$.

(H5) There exists a non-decreasing function H, positive on $\mathbb{R}^+$, such that for any vector u verifying $\|u\|_2 \leq 1$, we have:

$$\mathbb{P}\left[|u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u| \geq R\right] \leq 2\exp(-TH(R)).$$

We actually prove this assumption in the case where $\boldsymbol{A}_0$ is symmetric, see Theorem III.5 in Section 6.1. The proof is based on a concentration inequality for integrals of functionals of a stochastic process from [CG07]. Furthermore, the convergence of $\widehat{\boldsymbol{C}}_T$ to $\boldsymbol{C}_\infty$ is constrained by the speed of decorrelation of the process, which is the slowest precisely for symmetric parameters $\boldsymbol{A}_0$, see [HHMS93]. We therefore conjecture Assumption (H4) to hold also for a non-symmetric $\boldsymbol{A}_0$.

The set of Assumptions (H1) – (H4) are relatively unrestrictive. As already explained, Assumption (H1) is necessary for stationarity while Assumption (H3) could be possibly eliminated, since the exponentially decreasing autocorrelation means that the distribution of X is rapidly approaching the stationary distribution, but this would unnecessarily clutter our results. Assumption (H4) is not very restrictive: as mentioned above it is proved for a symmetric $\boldsymbol{A}_0$, and we conjecture it to be true in general (but were unable to prove the general case yet). Finally, Assumption (H3) is the sparsity assumption assumed throughout the paper on $\boldsymbol{A}_0$.

Theorem III.1. *Assume (H1), (H3) – (H5). Set* $\gamma > 1$, $0 \leq \tau < \gamma - 1$, $\epsilon_0 \in (0, 1)$ *and define*

$$\lambda_T := \gamma\sqrt{\frac{4e|\hat{\delta}_T|_\infty}{T}\left(\frac{1}{2}\log\frac{2\pi^2 d^2}{3\epsilon_0} + \log(2 + |\log(T\hat{\delta}_T)|_\infty)\right)} \tag{2.0.1}$$

where $\hat{\delta}_T := \operatorname{diag} \widehat{\boldsymbol{C}}_T$ *and* $\log$ *is applied entrywise on* $T\hat{\delta}_T$. *Set* $c_0 := \frac{\gamma+\tau+1}{\gamma-\tau-1}$, $\kappa := \sqrt{\frac{\min \mathrm{Sp}(\boldsymbol{C}_\infty)}{2}}$ *and assume that*

$$T \geq T_0 := H \left(\frac{\kappa^2}{9(c_0+2)^2} \right)^{-1} \left(s \log \left(21d \wedge \frac{21ed}{s} \right) + \log \left(\frac{4}{\epsilon_0} \right) \right).$$

Then, for any row-s*-sparse matrix* $\boldsymbol{A}$*, the lasso estimator* $\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{A}}_{\lambda_T}$ *verifies*

$$2\tau\gamma^{-1}\lambda_T \|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 \leq \|(\boldsymbol{A} - \boldsymbol{A}_0)X\|_{L^2}^2 + \left(\frac{1+\gamma+\tau}{\gamma\kappa} \right)^2 \lambda_T^2 ds \tag{2.0.2}$$

with probability at least $1 - \epsilon_0$.

The proof of Theorem III.1 is detailed in Section 6.2 below. It relies on a Restricted Eigenvalue property, see Theorem III.3 below, which we prove using Assumptions (H1)–(H4), as well as on a deviation property, see Theorem III.8 from Section 6.6 below. Theorem III.1 provides a sharp oracle inequality, with leading constant 1 in front of the bias term $\|(\boldsymbol{A} - \boldsymbol{A}_0)X\|_{L^2}^2$. The penalization parameter λ is a function of the observations through $\widehat{\boldsymbol{C}}_T$. However, the proof of Theorem III.1 uses Equation (6.2.1) which states that in the same set of events of probability at least $1 - \epsilon_0$, we have $\kappa^2 \leq \hat{\delta}_T^i \leq \boldsymbol{C}_\infty^{ii} + \kappa^2$ for any $i = 1, \ldots, n$. We can therefore safely bound $\hat{\delta}_T$ from below and above by deterministic constants in the statement of Theorem III.1.

The convergence rate obtained in Theorem III.1 almost matches the minimax lower bound provided in Theorem III.2 below. Indeed, the rate is $\lambda^2 ds$, up to numerical constants, and using the upper bound for $\hat{\delta}_T$ given above, we end up with a convergence rate of order

$$\frac{ds(\log d + \log\log T)}{T}.$$

Let us recall that ds is the sparsity of $\boldsymbol{A}_0$, under the row-sparsity (H3). The minimax lower bound from Theorem III.2 is of order $ds \log(d/s)/T$. The only main difference is between the terms d and d/s within the logarithm, and the negligible poly-logarithmic term $\log\log T$. We conjecture that an exact match (up to constants) between the upper and the minimax lower bound is possible, by considering ordered-ℓ_1 penalization, also called SLOPE, see [SC+16, BLT16], where such results are provided for linear regression only. However, such a development is way beyond the current focus of this paper: the choice of the weights involved in SLOPE is a difficult task in the setting considered here.

The next corollary provides errors bounds on the parameter $\boldsymbol{A}_0$ using different norms.

Corollary III.1. *With the same assumptions and notation as in Theorem III.1, the following holds with a probability larger than* $1 - \epsilon_0$*:*

1. for the empirical norm:

$$\|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2} \leq \frac{1+\gamma}{\gamma\kappa}\lambda_T\sqrt{ds} \tag{2.0.3}$$

2. for the ℓ^1 norm, with $\tau > 0$:

$$\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_1 \leq \frac{(1+\tau+\gamma)^2}{2\gamma\tau\kappa^2}\lambda_T ds \tag{2.0.4}$$

3. for the Frobenius norm:

$$\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F \leq \frac{1+\gamma}{\gamma\kappa^2}\lambda_T\sqrt{ds} \tag{2.0.5}$$

4. for the ℓ^q norm, with $q \in [1,2]$ and $\tau > 0$:

$$|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0|_q \leq (1+\tau+\gamma)^{4/q-2}(1+\gamma)^{2-2/q}(2\tau)^{1-2/q}\gamma^{-1}\kappa^{-2}\lambda_T(ds)^{1/q}.$$

All these inequalities are consequences of Equation (2.0.2), and are proved in Section 6.2. The next Theorem is a minimax lower bound over row-sparse matrices, for the considered model.

Theorem III.2. *For some constants $c > 0$ and $c' > 0$, we have:*

$$\inf_{\hat{\boldsymbol{A}}} \sup_{\boldsymbol{A}\in\Gamma_s} \mathbb{E}_{\boldsymbol{A}}\left[\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\right] \geq \frac{c'ds\log(cd/s)}{T},$$

where Γ_s is the set of row-s-sparse matrices and the infimum is taken over all possible estimators.

The proof of Theorem III.2 above is in Section 6.3. It uses the approach from [Tsy08], where we construct a set of matrices that are separated enough in Frobenius norm but close enough in terms of the resulting probability densities. For this, we need a set of row-s-sparse matrices that are invertible, that we create using regular graph adjacency matrices. The complexity of this set is controlled thanks to precise combinatorial results, such as the ones from [MW91].

Finally, we present an interesting by-product of the proof of Theorem III.1. Theorem III.3 below expresses that a Restricted Eigenvalue condition, see [BRT09], is, quite surprisingly, satisfied in the case of the Ornstein-Uhlenbeck drift estimation, while it is well-known to be a mandatory assumption for the linear regression model, see [ZWJ14], when one wants to prove optimal convergence rates for polynomial-time sparsity inducing algorithms, such as ℓ_1 penalization.

Theorem III.3. *Assume (H1) – (H4). Set* $s \leq d$ *and* $c_0 > 0$. *Define* $C(s, c_0) := \{u \in \mathbb{R}^d : \|u\|_1 \leq (1 + c_0)\|u_{|\mathcal{I}_s(u)}\|_1\}$ *where* $\mathcal{I}_s(u)$ *stands for the set of indices of the* s *largest entries of* $|u|$. *Consider* $\epsilon_0 \in (0, 1)$ *and* T_0 *given in Theorem III.1. Then, for any* $T \geq T_0$, *we have*

$$\mathbb{P}\left[\inf_{u \in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa\right] \geq 1 - \frac{\epsilon_0}{2}. \tag{2.0.6}$$

The proof of Theorem III.3 is given in Section 6.4 and uses explicitly Assumption (H4), which is proved in Theorem III.5, see Section 6.1, for a symmetric $\boldsymbol{A}_0$. We can interpret it equivalently as a lower bound on $\mathrm{tr}\,(\boldsymbol{A}\widehat{\boldsymbol{C}}_T\boldsymbol{A}^\top)$ (see Lemma III.9 from Section 6.4), hence as a RE property for $\widehat{\boldsymbol{C}}_T$ over row-s-sparse matrices $\boldsymbol{A}$. Observe that the values of κ and ϵ_0 are independent on s and c_0 and the validity of Equation (2.0.6) depends on s, c_0 solely through the condition $T \geq T_0(s, c_0)$. In other words, the validity of a Restricted Eigenvalue property in our model is achieved as long as T is large enough.

3 Asymptotic oracle properties for Adaptive Lasso

The MLE is asymptotically optimal, as observed with the asymptotic normality property from Equation (1.2.3). In this Section we derive similar properties for the ℓ^1-penalized estimator. Furthermore, another desirable property from a sparsity-inducing estimator is consistency in variable selection [BvdG11]. We define it by the property that the support of a $\mathrm{supp}(\widehat{\boldsymbol{A}})$ converges to the support of the true parameter $\mathrm{supp}(\boldsymbol{A}_0)$. It is known in the context of Gaussian linear regression that the Lasso cannot satisfy both properties with the same parameter λ, see [Zou06] while the Adaptive Lasso does. The Adaptive Lasso in our context is defined as

$$\widehat{\boldsymbol{A}}_{ad.} = \underset{\boldsymbol{A} \in \mathbb{R}^{d \times d}}{\arg\min}\, \mathcal{L}_T(\boldsymbol{A}) + \lambda \|\boldsymbol{A} \circ |\widehat{\boldsymbol{A}}_{MLE}|^{-\gamma}\|_1, \tag{3.0.1}$$

for fixed positive parameters λ and γ, where $\circ$ stands for the Hadamard product, and $|\widehat{\boldsymbol{A}}_{MLE}|^{-\gamma}$ stands for the matrix obtained by computing entrywise the absolute value, and exponentiation by $-\gamma$ of the MLE estimator (1.2.2). The idea of the Adaptive Lasso, involving a penalization level proportional to the entries of $|\widehat{\boldsymbol{A}}_{MLE}|^{-\gamma}$ (any $\sqrt{T}$-consistent estimator can be used theoretically), is to penalize more the entries expected to be actually zeros (trusting the MLE) and to penalize less those expected to be non-zero. Note that the MLE entries are non-zero almost surely.

Note that many quantities, such as λ and estimators $\widehat{\boldsymbol{A}}_{MLE}$ and $\widehat{\boldsymbol{A}}_{ad.}$, implicitly depend on T, and that we consider in this section asymptotics $T \to +\infty$.

Theorem III.4. *Assume (H1) – (H2). Fix* $\gamma > 0$ *and assume that* λ *verifies* $\lambda T^{1/2} \to 0$ *and* $\lambda T^{(\gamma+1)/2} \to +\infty$ *when* $T \to +\infty$. *Then, we have the following properties.*

1. *Consistency of the variable selection: as* $T \to +\infty$*, we have*

$$\mathbb{P}\left[\mathrm{supp}(\widehat{\boldsymbol{A}}_{ad.}) = \mathcal{A}_0\right] \to 1.$$

2. *Asymptotic normality: as* $T \to +\infty$*, we have*

$$\sqrt{T}\left((\widehat{\boldsymbol{A}}_{ad.})_{|\mathcal{A}_0} - (\boldsymbol{A}_0)_{|\mathcal{A}_0}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0 \times \mathcal{A}_0}\right)^{-1}\right),$$

where $\mathcal{A}_0 = \mathrm{supp}(\boldsymbol{A}_0)$ *is the support of the parameter* $\boldsymbol{A}_0$*.*

The proof of Theorem III.4 is in Section 6.5. It expresses two crucial asymptotic behaviors of the Adaptive Lasso for the Ornstein-Uhlenbeck drift estimation. The first point shows that Adaptive Lasso can be reasonably used for support recovery of the drift parameter, whenever T is large enough. The second point proves that the Adaptive Lasso shares the property of asymptotic efficiency with the MLE, over the support of the true parameter.

4 Numerical results

This Section proposes numerical experiments, both on simulated and real datasets, that confirm our theoretical findings. We start in Figure III.2 with an illustration of estimation results using MLE, Lasso and Adaptive Lasso, where the advantage of penalized methods can be seen at a first glance. The penalization level λ of all estimators are tuned using a cross-validation procedure described in Section 4.1 below.

In the next sections we verify this observation using repeated experiments in different settings, in order to see the impact on estimation performance of the observation length T and the dimension of the process d (Section 4.2). The support recovery ability of Lasso and Adaptive Lasso are illustrated in Section 4.3 and a brief analysis of the issue of trajectory discretization is discussed in Section 4.4. An application to real-world financial data is proposed in Section 4.5. In all our experiments, we use a time-step equal to 10^{-2} for the discretization of the continuous trajectories, see Section 4.4 for details.

4.1 Cross-validation for selection of λ

The Lasso and the Adaptive Lasso use a parameter λ that must be fixed by the user. Our theoretical results suggest a value for λ for the Lasso, see Equation (2.0.1). However, theoretical penalization parameters are known to be very pessimistic, in the

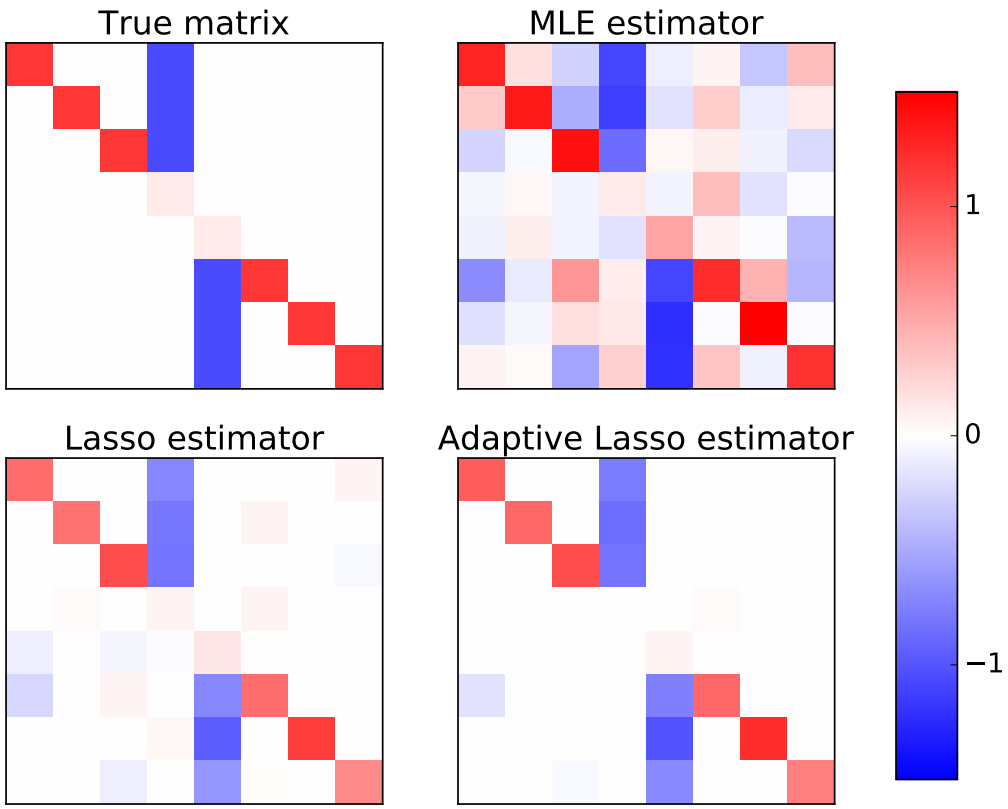


Figure III.2 – MLE, Lasso and Adaptive Lasso estimates compared to the ground truth matrix. The Lasso shows a significant improvement over the MLE, and the Adaptive Lasso further improves on the Lasso, especially in terms of support recovery. All penalization parameters are obtained through cross-validation.

sense that they are typically too large in most situations, see for instance [BvdG11]. We propose instead to tune λ through cross-validation.

In our setting, we implement cross-validation by using the first 80% of the trajectory as the training set and the remaining 20% as the validation set, in the following way.

$$\widehat{\boldsymbol{A}}_\lambda = \operatorname*{arg\,min}_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}_{.8T}(\boldsymbol{A}) + \lambda\|\boldsymbol{A}\|_1,$$
$$\hat{\lambda} = \operatorname*{arg\,min}_{\lambda\geq 0} \mathcal{L}_{[.8T,T]}(\widehat{\boldsymbol{A}}_\lambda),$$

where $\mathcal{L}_{[.8T,T]}$ is the negative log-likelihood constructed from the interval $[.8T, T]$. The cross-validated Lasso is then $\widehat{\boldsymbol{A}}_{\hat{\lambda}}$, and we define similarly the cross-validated adaptive Lasso. In the next sections, referring to Lasso and Adaptive Lasso will always correspond to the Lasso and Adaptive Lasso with cross-validated λ. Note that

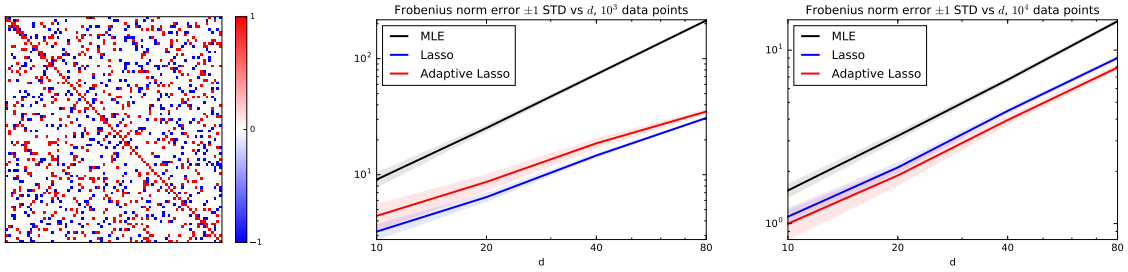
the selection of λ is performed in a grid on a logarithmic scale between 10^{-2} and 10^3, in all our experiments.

4.2 Influence of the observation length and of the dimension

In Figures III.3 and III.4, we plot the average Frobenius norm estimation error of MLE, Lasso and Adaptive Lasso, respectively as a function of the dimension d, and as a function of the sample size T. The bold lines and the shaded areas correspond respectively to the means and standard deviations of the errors obtained over 100 independent simulated trajectories. The ground truth parameter $\boldsymbol{A}_0$ is chosen as a random matrix with sparsity equal to $0.2d$, with non-zero entries equal to ± 1. Such a matrix is displayed for $d = 80$ on the left-hand size of Figure III.3. Note that all y-axis are on a logarithmic scale.

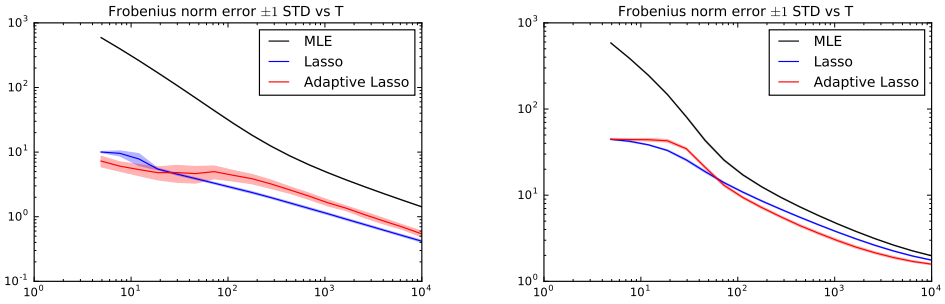
In Figure III.3, we observe the deterioration of the estimation error with an increasing dimension d. We observe also that penalized procedures perform much better than the MLE, but that slopes are very close: this comes from the fact that the row sparsity is fixed to $0.2 \times d$ leading to a $0.2 \times d^2$ overall sparsity, which is not much smaller than the dense case d^2 for the range of values of d considered in the experiment.

In Figure III.4, we observe the improvement of the estimation error with an increasing sample size T. We observe that the curves are consistent with a common convergence rate of order $\sqrt{T}$, but that penalized estimates constantly outperform the MLE.



(a) Ground truth $\boldsymbol{A}_0$ (b) Errors with $T = 10$ (c) Errors with $T = 100$

Figure III.3 – (a): Example of a simulated ground truth matrix, with row sparsity equal to $0.2d$; (b): estimation errors measured by the Frobenius norm for Lasso, Adaptive Lasso and MLE, as a function of d, with a sample size $T = 10$; (c): same as (b) with $T = 100$. Bold lines and shaded areas correspond respectively to the means and standard deviations of the error obtained over 100 independent simulated trajectories.



(a) Estimation error for $d = 10$ (b) Estimation error for $d = 100$

Figure III.4 – Estimation errors measured by the Frobenius norm for Lasso, Adaptive Lasso and MLE, as a function of T for (a) $d = 10$ and (b) $d = 100$. Bold lines and shaded areas correspond respectively to the means and standard deviations of the error obtained over 100 independent simulated trajectories.

4.3 Support recovery

Penalization methods such as Lasso and Adaptive Lasso can be used for variable selection, because of their sparsity-inducing property. Indeed, we proved in Theorem III.4 from Section 3 that the Adaptive Lasso is consistent for variable selection of the drift parameter $\boldsymbol{A}_0$. In Figure III.5, we consider the estimation problem of a 80×80 matrix $\boldsymbol{A}_0$ with sparsity $0.1 \times d$, and with a sample size $T = 100$. We simulate 100 trajectories, and compute the F_1-score obtained for support selection achieved by the MLE, Lasso and Adaptive Lasso. Figure III.5 then displays the box-plots of these F_1-scores. The F_1-score obtained by each estimator is computed as follows: first, we binarize the entries of the estimators and of the ground-truth matrix: zero entries are kept equal to zero, while non-zero entries are replaced by ones. Then, we count the true positives, false positives and false negatives in order to compute the precision and recall values.

The MLE does not lead to a sparse solution, so that its F_1-score is constant and is computed from the average row-sparsity s using the fromula $2s/(1+s)$, which is around 0.2 in our case as observed in Figure III.5. Indeed, the MLE always classifies all entries as non-zero, hence the corresponding true positive, false positive and false negative values are always equal respectively to sd, d^2 and 0. The strong improvement of the Adaptive Lasso over Lasso is clearly illustrated on this example: its F_1-score is almost equal to 1, while the Lasso achieves an F_1-score slightly below 0.5.

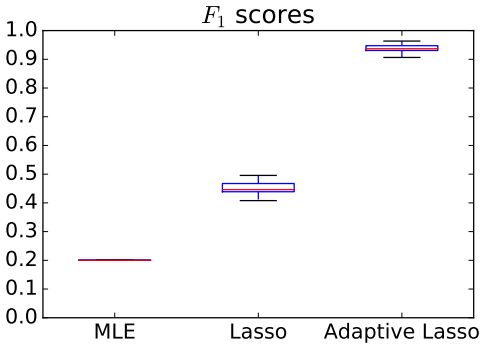


Figure III.5 – Accuracy scores for different estimation methods, using a 80×80 matrix $\boldsymbol{A}_0$ and an observation length $T = 100$. We classify as positive detection all non-zero entries of the estimators. The plot illustrates a clear advantage of Adaptive Lasso over Lasso in terms of support recovery. The MLE accuracy is provided here for convenience, as a lower-bound for any procedure, that corresponds to the sparsity of $\boldsymbol{A}_0$.

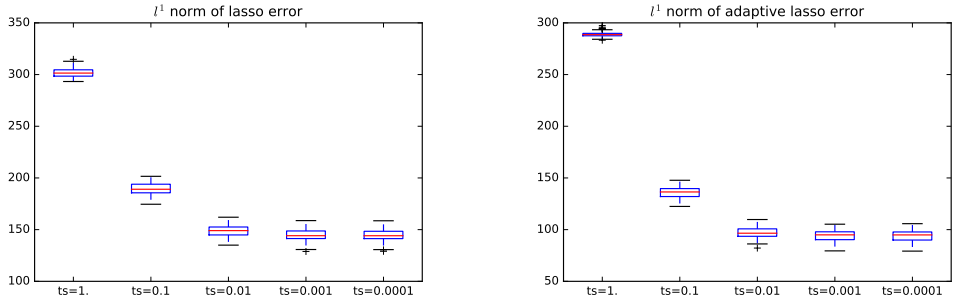
4.4 Influence of the time-step

The theoretical results proposed in this paper assumes a continuous observation of the trajectory of the data. However, in practice, any simulation or real-data analysis will have to use some discretization method. In our simulations we use a constant time-step $\delta t = 10^{-2}$. This value has been decided in hindsight, after a study of the impact of the time-step on the quality of the estimators. This study is illustrated in Figure III.6, where we display box-plots of the estimation errors obtained with varying discretization time steps $\delta t \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We observe in Figure III.6 that results improved with a decreasing δt, which is to be expected, since a smaller time step means more data points, but we observed no improvement for $\delta t \leq 10^{-2}$.

4.5 Application to financial data

The Ornstein-Uhlenbeck model is a popular method in finance to model mean-reverting processes, for instance for pairs trading [Hul09, CFS15, FI13]. In a typical setting, one chooses two related financial assets (for example the stocks of two companies in the same sector). Upon verification in the data, it is assumed that some linear combination of the stocks reverts to some "normal" value, often chosen as 0. This combination, denoted X, can be modeled by a one-dimensional Ornstein-Uhlenbeck process.

However, this method does not address two issues. First, one needs a separate method to find the relevant pairs. Second, instead of pairs, one could be interested



(a) ℓ^1 error for the Lasso.

(b) ℓ^1 error for the Adaptive Lasso.

Figure III.6 – Estimation errors for Lasso and Adaptive Lasso as a function of the discretization step δt. Box-plots are computed from the estimation errors obtained from 100 simulations of the same process, with a decreasing time step for discretization.

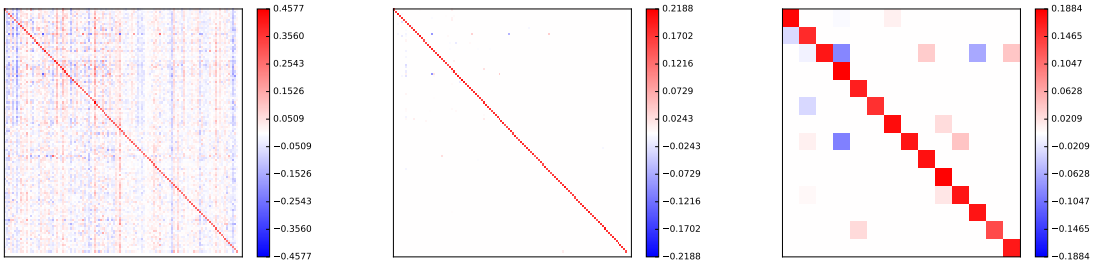
in more general linear combinations or in situations where the evolution of one price impacts another, when the pair does not fit a mean-reverting process. The multi-dimensional Ornstein-Uhlenbeck process is a way to address both problems, as it allows to involve an unrestricted number of assets. Moreover, thanks to the sparsity-inducing penalization considered in this paper, a sparse estimator might help in finding relevant combinations.

To illustrate this, we take daily close data of SP500 stocks, for companies in the financial and IT sectors with long enough history in the SP500 index. Our choice of sectors is arbitrary and is motivated by simplicity. We take the log-returns, then compute the exponential moving average (EMA), which will be the data we want to model using an Ornstein-Uhlenbeck process. By design, the EMA has a mean-reverting property and hence is a good candidate for fitting an Ornstein-Uhlenbeck process. We denote that process R and assume the model:

$$\mathrm{d}R_t = -\boldsymbol{A}(R_t - m)\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t \tag{4.5.1}$$

where $\boldsymbol{\Sigma}$ is typically not the identity because of high correlations between certain stocks. We estimate m and $\boldsymbol{\Sigma}$ using the mean and the squared variations. Because of $\boldsymbol{\Sigma}$, in order to estimate $\boldsymbol{A}$, we need to maximize a slightly modified log-likelihood which takes $\boldsymbol{\Sigma}$ into account, which is done easily using a proximal gradient descent algorithm for instance. The resulting MLE and cross-validated Adaptive Lasso give the matrices in Figure III.7. The heavy diagonals are explained by the fact that the data is an exponential moving average. However, the non-zero values away from the diagonal are non-trivial, since they can't be explained by covariances, that were already captured by the estimation of $\boldsymbol{\Sigma}$.

The sparse estimation selects the most significant stock prices that influence other stock prices. This can be an indication to find interesting stock pairs. The highest



(a) MLE (b) Adaptive Lasso (c) Adaptive Lasso (zoom)

Figure III.7 – (a) MLE estimator; (b) Adaptive Lasso estimator; (c) Zoom of (b) for stocks with the highest non-diagonal values; all for the estimation of $\boldsymbol{A}$ in the model of Equation (4.5.1). The diagonal values are expected from the design of the EMA. The MLE gives a very noisy estimate, while the Adaptive Lasso is highly sparse.

value, in absolute value, that we find outside of the diagonal is at tick-coordinates ('PRU','FITB'), and takes a value roughly equal to -0.1. This means in practice that given an above-average value for the exponential moving average of log-returns of FITB, the model predicts an increase of the exponential moving average of log-returns of PRU, all else being controlled: by controlling the correlation between the two stocks, we get the plot from Figure III.8. In layman terms, recent above-average returns of FITB predict above-average returns of PRU. As a disclaimer, we should point out that this study has been conducted with a very simple approach, and shouldn't be considered as trading advice.

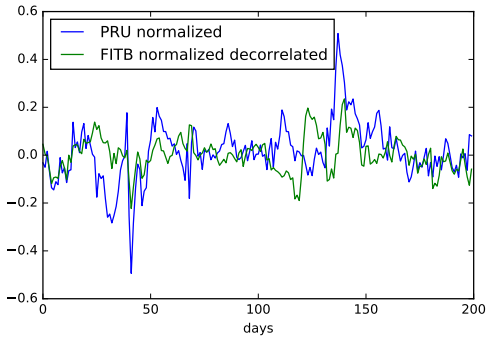


Figure III.8 – Plot of normalized EMA for PRU and FITB, the latter subtracted a fraction of the PRU data for decorrelation.

5 Conclusion

This paper provides a complete theory for the estimation of the drift parameter of a Ornstein-Uhlenbeck processes under a row-sparsity assumption. This is, to the best of our knowledge, the first paper to provide such results, either in a non-asymptotic or asymptotic framework, for Lasso and Adaptive Lasso. This paper is therefore a first attempt towards the use of sparsity-inducing penalization, widely used in the context of generalized linear models, to high-dimensional diffusion processes.

A natural extension of our work consists in assuming a correlated Brownian noise, modeled by a non-diagonal parameter $\boldsymbol{\Sigma}$ in front of $\mathrm{d}W_t$ in Equation (1.2.1). This parameter is exactly computable in the continuous observation setting. However, in a high-dimensional setting, $\boldsymbol{\Sigma}$ should be considered sparse as well, and one could therefore consider a joint estimation procedure for $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$, with dedicated sparsity-inducing penalizations. However, it turns out to be a much more difficult task, since the negative log-likelihood is not jointly convex with respect to $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. Such a development is therefore way beyond the scope of the present paper, and might actually involve a very different approach than the one considered here.

Another natural extension is to consider matrices $\boldsymbol{A}_0$ with non-positive spectra. A very interesting property is zero eigenvalues, which leads to a reduced rank and hence to co-integrated processes. A method to reduce rank is to penalize it, see [BSW11, BSW12] for application to Gaussian regression, or to use the so-called trace norm or nuclear norm penalization, which corresponds to a convex relaxation of the rank.

6 Proofs

In this Section, we provide proofs of the theorems and other statements from Sections 2 and 3.

6.1 Proof of Assumption (H4) in the reversible case

Theorem III.5 below expresses that assumption (H4) is true when $\boldsymbol{A}_0$ is symmetric. This condition is equivalent in our case to the reversibility of the process, see [GS16].

Theorem III.5. *Assume that $\boldsymbol{A}_0$ is symmetric. Then there exists a non-decreasing, non-negative function H such that for any vector u, $\|u\|_2 \leq 1$, we have:*

$$\mathbb{P}\left[|u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u| \geq R\right] \leq 2\exp(-TH(R)).$$

Theorem III.5 above follows from Lemmas III.1 and III.2 below, after taking $H(R) = H_1(R) \wedge H_2(R)$. The proof of the Lemmas is based on Theorem III.6

below, which shows a deviation inequality for the integral of a functional of an ergodic process from its long-time limit.

Theorem III.6 ([CG07], Theorem 2.1). *Let L be the infinitesimal generator associated to an ergodic diffusion X with stationary distribution μ. If μ satisfies the* $\log$*-Sobolev inequality:*

$$c \int f^2 \log f^2 \mathrm{d}\mu \leq - < Lf, f >_\mu \tag{6.1.1}$$

for some $c > 0$ and for all functions f in the domain of definition of L such that $\int_{\mathbb{R}^d} f^2 \mathrm{d}\mu = 1$, then for all $Q \in \mathbb{L}^1(\mu)$ and $R > 0$:

$$\mathbb{P}\left[\frac{1}{T}\int_0^T Q(X_t)\mathrm{d}t - \int Q \mathrm{d}\mu \geq R\right] \leq \exp\left(-tH^*\left(R\right)\right) \tag{6.1.2}$$

where

$$H^*(R) := \sup_{0 \leq \rho < \rho_{max}} \left\{ \rho R - c \log \int \exp\left(\frac{\rho}{c}(Q - \int Q \mathrm{d}\mu)\right) \mathrm{d}\mu \right\}$$

and ρ_{max} is such that the integral above is finite for any $0 \leq \rho < \rho_{max}$.

Remark III.1. *When $\boldsymbol{A}_0$ is symmetric, a simple integration by parts shows that $< Lf, f >_\mu := \int_{\mathbb{R}^d} Lf \cdot f \mathrm{d}\mu = -\frac{1}{2}\int_{\mathbb{R}^d} \|\nabla f\|_F^2 \mathrm{d}\mu$ and hence that Equation* (6.1.1) *holds due to the classical log-Sobolev inequality [Gro75], with $c = 1/4$.*

Observe that Equation (6.1.2) applies to a one-sided inequality. Therefore, in order to get Theorem III.5, we will have to work with two inequalities. We deal with the first one in Lemma III.1 below.

Lemma III.1. *Assume $\boldsymbol{A}_0$ is symmetric. Then for any vectors u such that $\|u\|_2 \leq 1$:*

$$\mathbb{P}\left[u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u > R\right] \leq \exp(-TH_1(R))$$

where $H_1(R) = \frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} - \log\det\left(\mathrm{Id} + R\frac{\boldsymbol{C}_\infty u u^\top}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right)$.

Proof. Observe first that it suffices to prove the Lemma for $\|u\|_2 = 1$. In the following, $u \in \mathbb{R}^d$ verifies that condition. We apply Theorem III.6, which applies with $c = 1/4$ as explained in Remark III.1, to the function $Q(X) = u^\top X X^\top u = (u^\top X)^2$. Then $\int Q \mathrm{d}\mu = u^\top \boldsymbol{C}_\infty u$. It remains to write explicitly $H^*(R)$. We have $H^*(R) = \sup_{0 \leq \rho < \rho_{max}} \rho(R + u^\top \boldsymbol{C}_\infty u) - \frac{1}{4}\log I_\rho$ where:

$$I_\rho := \int \exp\left(4\rho u^\top X X^\top u\right) \mathrm{d}\mu$$

$$= (2\pi)^{-d/2}(\det \boldsymbol{C}_\infty)^{-1/2} \int \exp\left(-\frac{1}{2} X^\top \left(\boldsymbol{C}_\infty^{-1} - 8\rho u u^\top\right) X\right) \mathrm{d}X$$
$$= (\det \boldsymbol{C}_\infty)^{-1/2}(\det \boldsymbol{\Sigma}_\rho)^{1/2} \tag{6.1.3}$$

and $\boldsymbol{\Sigma}_\rho^{-1} := \boldsymbol{C}_\infty^{-1} - 8\rho u u^\top$. The product $uu^\top$ is a symmetric matrix of rank 1 and its only non-zero eigenvalue is 1, as $uu^\top u = u$.

The integral I_ρ is defined and Equation (6.1.3) is valid if and only if $\boldsymbol{\Sigma}_\rho^{-1}$ is indeed a matrix with positive spectrum. We have

$$\min \mathrm{Sp}(\boldsymbol{\Sigma}_\rho^{-1}) \geq \min \mathrm{Sp}(\boldsymbol{C}_\infty^{-1}) - 8\rho \max \mathrm{Sp}(uu^\top) = (\max \mathrm{Sp}(\boldsymbol{C}_\infty))^{-1} - 8\rho.$$

Hence we choose $\rho_{max} := \frac{1}{8}(\max \mathrm{Sp}(\boldsymbol{C}_\infty))^{-1}$ and I_ρ is well defined for $\rho < \rho_{max}$. Note also for later that $\|8\rho \boldsymbol{C}_\infty u u^\top\|_{\mathrm{op}} \leq 8\rho \max \mathrm{Sp}(\boldsymbol{C}_\infty) < 1$.

To find $H^*(R)$, we differentiate the argument of the supremum. For this, we need $\frac{\mathrm{d} \det \boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho}$. We have $\frac{\mathrm{d}\boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho} = -\boldsymbol{\Sigma}_\rho \frac{\mathrm{d}\boldsymbol{\Sigma}_\rho^{-1}}{\mathrm{d}\rho} \boldsymbol{\Sigma}_\rho = 8\boldsymbol{\Sigma}_\rho u u^\top \boldsymbol{\Sigma}_\rho$. Hence

$$\frac{\mathrm{d} \det \boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho} = \mathrm{tr}\left(\mathrm{adj}(\boldsymbol{\Sigma}_\rho)\frac{\mathrm{d}\boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho}\right) = 8\mathrm{tr}\left(\mathrm{adj}(\boldsymbol{\Sigma}_\rho)\boldsymbol{\Sigma}_\rho u u^\top \boldsymbol{\Sigma}_\rho\right) = 8(\det \boldsymbol{\Sigma}_\rho) u^\top \boldsymbol{\Sigma}_\rho u.$$

Therefore, to find $H^*(R)$, we solve in ρ the equation $R + u^\top \boldsymbol{C}_\infty u - u^\top \boldsymbol{\Sigma}_\rho u = 0$. We can actually compute $u^\top \boldsymbol{\Sigma}_\rho u$ using a geometric series, recalling that $\|8\rho \boldsymbol{C}_\infty u u^\top\|_{\mathrm{op}} < 1$:

$$\begin{aligned}
\boldsymbol{\Sigma}_\rho &= \left(\mathrm{Id} - 8\rho \boldsymbol{C}_\infty u u^\top\right)^{-1} \boldsymbol{C}_\infty = \sum_{k\geq 0} \left(8\rho \boldsymbol{C}_\infty u u^\top\right)^k \boldsymbol{C}_\infty \\
&= \boldsymbol{C}_\infty + 8\rho \boldsymbol{C}_\infty u \sum_{k\geq 0} \left(8\rho u^\top \boldsymbol{C}_\infty u\right)^k u^\top \boldsymbol{C}_\infty \\
&= \boldsymbol{C}_\infty + 8\rho \frac{\boldsymbol{C}_\infty u u^\top \boldsymbol{C}_\infty}{1 - 8\rho u^\top \boldsymbol{C}_\infty u} \\
u^\top \boldsymbol{\Sigma}_\rho u &= \frac{u^\top \boldsymbol{C}_\infty u}{1 - 8\rho u^\top \boldsymbol{C}_\infty u}.
\end{aligned}$$

We get $H^*(R)$ for $\rho = \rho^* := \frac{1}{8}\frac{R}{u^\top \boldsymbol{C}_\infty u(R + u^\top \boldsymbol{C}_\infty u)}$. Then:

$$\begin{aligned}
\Sigma_{\rho^*} &= \boldsymbol{C}_\infty \left(\mathrm{Id} + R\frac{u u^\top \boldsymbol{C}_\infty}{(u^\top \boldsymbol{C}_\infty u)^2}\right) \\
H^*(R) &= \rho^*(R + u^\top \boldsymbol{C}_\infty u) + \frac{1}{8}\log(\det \boldsymbol{C}_\infty) - \frac{1}{8}\log(\det \boldsymbol{\Sigma}_{\rho^*}) \\
&= \frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} - \log\det\left(\mathrm{Id} + R\frac{u u^\top \boldsymbol{C}_\infty}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right).
\end{aligned}$$

□

We deal with the second inequality in Lemma III.2 below.

Lemma III.2. *Assume $\boldsymbol{A}_0$ is symmetric. Then for any vectors u such that $\|u\|_2 \leq 1$:*

$$\mathbb{P}\left[u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u \leq -R\right] \leq \exp(-TH_2(R)) \tag{6.1.4}$$

$$\textit{where } H_2(R) = \begin{cases} -\frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} + \log\det\left(\mathrm{Id} - R\frac{\boldsymbol{C}_\infty uu^\top}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right) & \textit{if } R < u^\top \boldsymbol{C}_\infty u, \\ +\infty & \textit{else.} \end{cases}.$$

Proof. Observe first that if $R \geq u^\top \boldsymbol{C}_\infty u$, the probability is zero, as $\widehat{\boldsymbol{C}}_T$ is a.s. a positive definite matrix. We assume henceforth that $R < u^\top \boldsymbol{C}_\infty u$. The same reasoning as in the proof of Theorem III.1 to $Q(X) = -u^\top XX^\top u$ gives:

$$\begin{aligned} H^*(R) &:= \sup_{\rho \geq 0} \rho R - \rho u^\top \boldsymbol{C}_\infty u + \frac{1}{8}\log\det \boldsymbol{C}_\infty - \frac{1}{8}\log\det \boldsymbol{\Sigma}_\rho \\ \boldsymbol{\Sigma}_\rho^{-1} &:= \boldsymbol{C}_\infty^{-1} + 8\rho uu^\top. \end{aligned}$$

We restrict the supremum to $\rho \leq \rho_{max} = \frac{1}{8}(\max \mathrm{Sp}(\boldsymbol{C}_\infty))^{-1}$ as in the proof of Lemma III.1, as it is sufficient to get the upper bound (6.1.4). This enables the geometric series calculation as in the proof of Lemma III.1. We get:

$$\begin{aligned} u^\top \boldsymbol{\Sigma}_\rho u &= \frac{u^\top \boldsymbol{C}_\infty u}{1 + 8\rho u^\top \boldsymbol{C}_\infty u} \\ \rho^* &:= \frac{1}{8}\frac{R}{u^\top \boldsymbol{C}_\infty u(u^\top \boldsymbol{C}_\infty u - R)} \\ \Sigma_{\rho^*} &= \boldsymbol{C}_\infty\left(\mathrm{Id} - R\frac{SS^\top \boldsymbol{C}_\infty}{(S^\top \boldsymbol{C}_\infty S)^2}\right) \\ H^*(R) &= -\frac{1}{8}\left(\frac{R}{S^\top \boldsymbol{C}_\infty S} + \log\det\left(\mathrm{Id} - R\frac{SS^\top \boldsymbol{C}_\infty}{(S^\top \boldsymbol{C}_\infty S)^2}\right)\right). \end{aligned}$$

□

For completeness, we state an interesting corollary.

Corollary III.2. *For any vectors S_1, S_2 such that $\|u_1\|_2 \leq 1$,$\|u_2\|_2 \leq 1$, and for any $i, j \leq d$:*

$$\begin{aligned} \mathbb{P}\left[|u_1^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u_2| > 3R\right] &\leq 6\exp(-TH(R)) \\ \mathbb{P}\left[|\widehat{\boldsymbol{C}}_T^{ij} - \boldsymbol{C}_\infty^{ij}| > 3R\right] &\leq 6\exp(-TH(R)). \end{aligned}$$

Proof. Denote $\Delta\boldsymbol{C} := \widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty$. We have

$$|u_1^\top \Delta\boldsymbol{C} u_2| \leq \frac{1}{2}\left|(u_1+u_2)^\top \Delta\boldsymbol{C}(u_1+u_2) + u_1^\top \Delta\boldsymbol{C}u_1 + u_2^\top \Delta\boldsymbol{C}u_2\right|.$$

Each time with probability at least $1 - \exp(-TH^*(R))$, we have $|u_1^\top \Delta \boldsymbol{C} u_1| \leq R$, $|u_2^\top \Delta \boldsymbol{C} u_2| \leq R$ and $|(u_1+u_2)^\top \Delta \boldsymbol{C}(u_1+u_2)| \leq 4R$.

For the second inequality, apply the first with u_1 and u_2 set respectively as the i-th and j-th vectors of the canonical basis.

□

6.2 Proof of Theorem III.1 (Lasso error bound)

Lemma III.3. *For any matrix $\boldsymbol{A}$ and any $\lambda > 0$, we have:*

$$\|(\widehat{\boldsymbol{A}}-\boldsymbol{A}_0)X\|_{L^2}^2 - \|(\boldsymbol{A}-\boldsymbol{A}_0)X\|_{L^2}^2 \leq 2\langle \boldsymbol{\varepsilon}_T, \boldsymbol{A}-\widehat{\boldsymbol{A}}\rangle_F - \|(\boldsymbol{A}-\widehat{\boldsymbol{A}})X\|_{L^2}^2 + 2\lambda(\|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1).$$

Proof. As $\mathcal{L}_T(\boldsymbol{A}) = \operatorname{tr} \boldsymbol{A}^\top \boldsymbol{\varepsilon}_T - \frac{1}{2}(\boldsymbol{A}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T(\boldsymbol{A}-\boldsymbol{A}_0)^\top + \frac{1}{2}\boldsymbol{A}_0\widehat{\boldsymbol{C}}_T\boldsymbol{A}_0^\top$, the gradient is $\boldsymbol{\varepsilon}_T + (\boldsymbol{A}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T$. The optimality condition applied to $\widehat{\boldsymbol{A}}$ gives that there exists a $\boldsymbol{B}$ in the sub-derivative of the ℓ^1 norm computed at $\widehat{\boldsymbol{A}}$ such that $\boldsymbol{\varepsilon}_T + (\widehat{\boldsymbol{A}}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T + \lambda \boldsymbol{B}$. $\boldsymbol{B}$ being in the sub-derivative, we have $\langle \boldsymbol{B}, \boldsymbol{A} - \widehat{\boldsymbol{A}}\rangle_F \leq \|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1$.

Applying this to the following formula and observing that for any matrix $\boldsymbol{M}$, $\|\boldsymbol{M}X\|_{L^2}^2 = \langle \boldsymbol{M}^\top \boldsymbol{M}, \widehat{\boldsymbol{C}}_T\rangle_{L^2}$, we get:

$$\begin{aligned}
S &:= \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 - \|(\boldsymbol{A}-\boldsymbol{A}_0)X\|_{L^2}^2 + \|(\widehat{\boldsymbol{A}}-\boldsymbol{A})X\|_{L^2}^2 \\
&= \langle \widehat{\boldsymbol{C}}_T, \left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right)^\top (\widehat{\boldsymbol{A}} - \boldsymbol{A}_0) - (\boldsymbol{A}-\boldsymbol{A}_0)^\top(\boldsymbol{A}-\boldsymbol{A}_0) + (\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top(\widehat{\boldsymbol{A}}-\boldsymbol{A})\rangle_F \\
&= 2\langle \widehat{\boldsymbol{C}}_T, (\boldsymbol{A}_0 - \widehat{\boldsymbol{A}})^\top(\boldsymbol{A}-\widehat{\boldsymbol{A}})\rangle_F \\
&= 2\langle \boldsymbol{\varepsilon}_T + \lambda \boldsymbol{B}, \boldsymbol{A} - \widehat{\boldsymbol{A}}\rangle_F \\
&\leq 2\langle \boldsymbol{\varepsilon}_T, \boldsymbol{A} - \widehat{\boldsymbol{A}}\rangle_F + 2\lambda(\|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1).
\end{aligned}$$

□

Observe the preceding is true for any value of λ. Choose then λ as in (2.0.1). We have for instance $\gamma^{-1}\lambda = \theta(x, X)$ with $x = \frac{1}{2}\log\frac{2\pi^2 d^2}{3\epsilon_0}$ and θ as in Equation (6.6.1). Second, we assume $T \geq T_1 := H\left(\frac{\kappa^2}{9(c_0+2)^2}\right)^{-1}(s\log(21d \wedge 21ed/s) + \log 4\epsilon_0^{-1})$. Therefore, using Theorems III.8 and Corollary III.4, we have for any matrix $\boldsymbol{U}$:

$$\mathbb{P}\left[\inf_{u \in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa \cap \langle \boldsymbol{U}, \boldsymbol{\varepsilon}_T\rangle_F \leq \gamma^{-1}\lambda \|\boldsymbol{U}\|_1 \cap \forall i, \kappa^2 \leq \hat{\delta}_T^{ii} \leq \boldsymbol{C}_\infty^{ii} + \kappa^2\right] \geq 1 - \epsilon_0. \quad (6.2.1)$$

We proceed to the proof of Theorem III.1.

Proof. We assume for the all what follows that the observation falls in the set of events defined by Equation (6.2.1) where we take $c_0 := \frac{\gamma+\tau+1}{\gamma-\tau-1}$. Therefore the inequalities we prove hold with probability at least $1-\epsilon_0$. Denote $\boldsymbol{U} = \boldsymbol{A} - \widehat{\boldsymbol{A}}$. From Lemma III.3, we have

$$\begin{aligned}
S &:= 2\tau\gamma^{-1}\lambda\|\boldsymbol{U}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 - \|(\boldsymbol{A} - \boldsymbol{A}_0)\,X\|_{L^2}^2 + \|\boldsymbol{U}X\|_{L^2}^2 \\
&\leq 2\tau\gamma^{-1}\lambda\|\boldsymbol{U}\|_1 + 2\langle\boldsymbol{\varepsilon}_T, \boldsymbol{U}\rangle_F + 2\lambda(\|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1) \\
&\leq 2\lambda\left((1+\tau)\gamma^{-1}\|\boldsymbol{U}\|_1 + \|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1\right) \\
&\leq 2\lambda\sum_{i=1}^{d}(1+\tau)\gamma^{-1}\|\boldsymbol{U}^{i,\bullet}\|_1 + \|\boldsymbol{A}^{i,\bullet}\|_1 - \|\widehat{\boldsymbol{A}}_\lambda^{i,\bullet}\|_1 \\
&\leq 2\lambda\sum_{\Delta^i>0}\Delta^i
\end{aligned} \tag{6.2.2}$$

where $\Delta^i := (1+(1+\tau)\gamma^{-1})\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i,\bullet}}\|_1 - (1-(1+\tau)\gamma^{-1})\|\boldsymbol{U}^{i,\bullet}_{|\bar{\mathcal{A}}^{i,\bullet}}\|_1$ and $\mathcal{A}^{i,\bullet} = \operatorname{supp}\boldsymbol{A}^{i,\bullet}$. This last inequality comes from Lemma III.4 where we use the fact that $(1+\tau)\gamma^{-1} < 1$. We only need to consider the indices i such that $\Delta^i > 0$, for which

$$\begin{aligned}
\|\boldsymbol{U}^{i,\bullet}_{|\bar{\mathcal{A}}^{i,\bullet}}\|_1 < \frac{1+(1+\tau)\gamma^{-1}}{1-(1+\tau)\gamma^{-1}}\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i,\bullet}}\|_1 = c_0\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i,\bullet}}\|_1 \\
\|\boldsymbol{U}^{i,\bullet}\|_1 < (1+c_0)\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i,\bullet}}\|_1 \leq (1+c_0)\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{U}^{i,\bullet}}\|_1
\end{aligned}$$

where $\mathcal{U}^{i,\bullet} = \operatorname{supp}\boldsymbol{U}^{i,\bullet}$.

We have then $\boldsymbol{U}^{i,\bullet} \in C(s, c_0)$. We apply the condition from Equation (6.2.1) and get $\Delta^i \leq \gamma^{-1}(\gamma+\tau+1)\sqrt{s}\|\boldsymbol{U}^{i,\bullet}\|_2 \leq \gamma^{-1}(\gamma+\tau+1)\sqrt{s}\kappa^{-1}\|(\boldsymbol{U}^{i,\bullet})^\top X\|_{L^2}$. Observing that

$$\sum_{\Delta^i>0}\|(\boldsymbol{U}^{i,\bullet})^\top X\|_{L^2} \leq \sum_{i=1}^{d}\|(\boldsymbol{U}^{i,\bullet})^\top X\|_{L^2} \leq \sqrt{d}\|\boldsymbol{U}X\|_{L^2},$$

we get $S \leq 2\lambda\gamma^{-1}(\gamma+\tau+1)\sqrt{s}\kappa^{-1}\|\boldsymbol{U}X\|_{L^2}$. Using

$$2\lambda\gamma^{-1}(\gamma+\tau+1)\sqrt{s}\kappa^{-1}\|\boldsymbol{U}X\|_{L^2} - \|\boldsymbol{U}X\|_{L^2}^2 \leq \left(\frac{\gamma+\tau+1}{\gamma\kappa}\right)^2\lambda^2 ds,$$

we conclude with Equation (2.0.2).

□

The proof of Corollary III.1 consists in using Theorem III.1 with specific values of the parameters. We explicit it in the following proof.

Proof. 1. It suffices to take $\tau = 0$ and $\boldsymbol{A} = \boldsymbol{A}_0$ in Equation (2.0.2).

2. We take $\boldsymbol{A} = \boldsymbol{A}_0$ and $\tau > 0$ in Equation (2.0.2). We bound The L^2 norm from below by 0 and get the result.

3. It suffices to apply Equation (2.0.3) with the Restricted Eigenvalue condition which states that $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F \leq \kappa^{-1}\|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}$ as long as each line of $\widehat{\boldsymbol{A}} - \boldsymbol{A}_0$ is in $C(s, c_0)$ (see Lemma III.9). To prove that last point, we fix an index $i \leq d$ and continue the proof from Equation (6.2.2). Choose $\boldsymbol{A}$ the matrix equal to $\widehat{\boldsymbol{A}}$ except on the i-th line, where we assume it is equal to $\boldsymbol{A}_0$. Then $\boldsymbol{U}$ is null except on the i-th line, where it is equal to the i-th line of $\boldsymbol{A}_0 - \widehat{\boldsymbol{A}}$. As we have already assumed $\tau = 0$, we get:

$$\begin{aligned}
2\lambda\Delta^i &\geq \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right) X\|_{L^2}^2 - \|\left(\boldsymbol{A} - \boldsymbol{A}_0\right) X\|_{L^2}^2 + \|\boldsymbol{U}X\|_{L^2}^2 \\
&= \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right) X\|_{L^2}^2 - \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0 + e_i(\boldsymbol{U}^{i,\bullet})^\top\right) X\|_{L^2}^2 + \|(\boldsymbol{U}^{i,\bullet})^\top X\|_{L^2}^2 \\
&\geq 2\langle(\boldsymbol{A}_0 - \widehat{\boldsymbol{A}})X, e_i(\boldsymbol{U}^{i,\bullet})^\top X\rangle_{L^2} = 2\|(\boldsymbol{U}^{i,\bullet})^\top X\|_{L^2}^2 \geq 0.
\end{aligned}$$

   Where e_i is the i-th element of the canonical basis of $\mathbb{R}^d$. Using the same argument as in the proof of III.1, we conclude that for each i, $(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)^{i,\bullet} \in C(s, c_0)$ which is what we wanted.

4. We apply the norm interpolation inequality: $\|\boldsymbol{U}\|_q^q \leq \|\boldsymbol{U}\|_1^{2-q}\|\boldsymbol{U}\|_F^{2q-2}$ to equations (2.0.4) and (2.0.5).

□

We finish this Section with the statement of a few useful inequalities

Lemma III.4. *Take $\boldsymbol{A}, \boldsymbol{B}$ two $d \times d$ matrices, $\gamma \in (0, 1)$ and denote $\boldsymbol{U} = \boldsymbol{A} - \boldsymbol{B}$, $\mathcal{A} := \operatorname{supp} \boldsymbol{A}$. Then we have the following inequalities:*

$$\begin{aligned}
\gamma\|\boldsymbol{U}\|_1 + \|\boldsymbol{A}\|_1 - \|\boldsymbol{B}\|_1 &= \gamma\|\boldsymbol{U}_{|\mathcal{A}}\|_1 + \gamma\|\boldsymbol{U}_{|\bar{\mathcal{A}}}\|_1 + \|\boldsymbol{A}_{|\mathcal{A}}\|_1 - \|\boldsymbol{B}_{|\mathcal{A}}\|_1 - \|\boldsymbol{B}_{|\bar{\mathcal{A}}}\|_1 \\
&= \gamma\|\boldsymbol{U}_{|\mathcal{A}}\|_1 + \gamma\|\boldsymbol{U}_{|\bar{\mathcal{A}}}\|_1 + \|\boldsymbol{A}_{|\mathcal{A}}\|_1 - \|\boldsymbol{B}_{|\mathcal{A}}\|_1 - \|\boldsymbol{U}_{|\bar{\mathcal{A}}}\|_1 \\
&\leq (1+\gamma)\|\boldsymbol{U}_{|\mathcal{A}}\|_1 - (1-\gamma)\|\boldsymbol{U}_{|\bar{\mathcal{A}}}\|_1 \\
&\leq (1+\gamma)\sqrt{\|\boldsymbol{A}\|_0}\|\boldsymbol{U}_{|\mathcal{A}}\|_F - (1-\gamma)\|\boldsymbol{U}_{|\bar{\mathcal{A}}}\|_1.
\end{aligned}$$

6.3 Proof of Theorem III.2 (lower bound of the estimation error)

Lemma III.5. *For some constant $c > 0$, there exists a set $\Omega_a \subset \{-1, 0, 1\}^{d\times d}$ such that for any $\boldsymbol{B} \neq \boldsymbol{B}' \in \Omega_a$:*

1. $\boldsymbol{B}$ *is antisymmetric and its upper triangular section has non-negative entries*
2. $\boldsymbol{B}$ *has at most* $s-1$ *non-zero entries by row*
3. $\sum_{ij} \mathbb{1}_{\boldsymbol{B}^{ij} \neq \boldsymbol{B}'^{ij}} \geq cds$

and $\log |\Omega_a| \geq cds \log \frac{ced}{s}$.

Proof. Consider the sets of matrices $\Omega'_{\leq r}, \Omega'_r$ of antisymmetric matrices in $\{-1,0,1\}^{d\times d}$ such that the entries in the upper triangular section are all non-negative and with at most r non-zero entries in each row for $\Omega'_{\leq r}$ and exactly r non-zero entries in each row for Ω'_r. For any $r \leq s-1$, $\Omega'_r \subset \Omega'_{\leq s-1}$. Fix then r to be the largest even number smaller or equal to $s-1$: we have $s-2 \leq r \leq s-1$. Ω'_r is one-to-one with the set $\mathrm{reg}(r,d)$, the set of r-regular labeled graphs with d vertices. To see this, observe that applying the absolute value to a matrix in Ω'_r gives an adjacency matrix of a regular graph. The relation is one-to-one given assumption 1.

We know the asymptotic of $|\mathrm{reg}(r,d)|$. Take for example [MW91] and apply $k! = k^k e^{-k}\sqrt{2\pi k}\Psi(k)$ where $(12j+1)^{-1} < \log \Psi(k) < (12j)^{-1}$. We keep track only of the highest order on d, the relevant variable that is considered high.

$$\begin{aligned}\log |\mathrm{reg}(r,d)| &\sim -\frac{r^2-1}{4} - \frac{r^3}{12d} + \log(dr)! - \log(dr/2)! - \frac{dr}{2}\log 2 - d\log r! \\ &\sim -\frac{r^2-1}{4} - \frac{r^3}{12d} + \frac{dr}{2}\log\frac{ed}{r} \\ &\quad - \frac{d}{2}\log r - d\log\sqrt{2\pi} + \frac{1}{2}\log 2 - \frac{1}{12dr} - \frac{d}{12r}.\end{aligned}$$

Keeping only the highest order in d gives $\log|\mathrm{reg}(r,d)| \sim \frac{dr}{2}\log\frac{ed}{r}$. We know due to the Erdős-Gallai theorem [EG60] that the number of r-regular graphs is non-zero for $d > r$ as r is chosen to be even. Hence there exists a constant $2c > 0$ such that $|\mathrm{reg}(r,d)| \geq 2cd(r+2)\log ed/r$.

We continue by applying the Gilbert–Varshamov bound. For this, we need to compute the maximum volume of a Hamming ball of radius K, where K is an integer: we fix some $\boldsymbol{A}$ and will count the number of $\boldsymbol{A}'$ that differ from $\boldsymbol{A}$ on a maximum of K entries. That volume is bounded by the one in the larger space of matrices with at most dr non-zero entries. We have thus:

$$V \leq \sum_{i=1}^{K}\binom{d^2}{i} \leq \sum_{i=1}^{K}\frac{K^i}{i!}\left(\frac{d^2}{K}\right)^i \leq \left(\frac{ed^2}{K}\right)^K.$$

Choose then $K = \lfloor cd(r+2)\rfloor \geq \lfloor cds\rfloor$. As $x \mapsto (ed^2/x)^x$ is increasing for $x \leq d^2$, we have $V \leq \left(\frac{ed}{c(r+2)}\right)^{cd(r+2)} \leq \left(\frac{ed}{cr}\right)^{cd(r+2)}$. The Gilbert-Varshamov bound gives us the existence of a set $\Omega_a \subset \Omega'_r$, verifying condition 3 and its size is at least:

$$\log|\Omega_a| \geq \log|\Omega'_r| - \log V = cd(r+2)\log\frac{ced}{r} \geq cds\log\frac{ced}{s}.$$

□

Lemma III.6 gives a way to construct a family of drift parameters with corresponding diagonal stationary covariance from a family of antisymmetric matrices.

Lemma III.6. *Take $\boldsymbol{A} = \alpha\mathrm{Id} + \boldsymbol{B}$ for $\alpha > 0$ and $\boldsymbol{B}$ an antisymmetric matrix. Then we have:*

$$\boldsymbol{C}_\infty(\boldsymbol{A}) = \int_0^\infty e^{-\boldsymbol{A}t}e^{-\boldsymbol{A}^\top t}\mathrm{d}t = \frac{1}{2\alpha}\mathrm{Id}.$$

Proof. We have $\boldsymbol{B}^\top = -\boldsymbol{B}$ hence $i\boldsymbol{B}$ is Hermitian and therefore unitarily diagonalizable. There exists an unitary matrix $\boldsymbol{U}$ such that $\boldsymbol{B} = i\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ where $\boldsymbol{D}$ is a diagonal real matrix. Then $e^{-(\alpha\mathrm{Id}+\boldsymbol{B})t} = e^{-\alpha t}\boldsymbol{U}e^{-i\boldsymbol{D}t}\boldsymbol{U}^*$ and $e^{-(\alpha\mathrm{Id}+\boldsymbol{B})^\top t} = e^{-(\alpha\mathrm{Id}-\boldsymbol{B})t} = e^{-\alpha t}\boldsymbol{U}e^{i\boldsymbol{D}t}\boldsymbol{U}^*$ hence $\boldsymbol{C}_\infty(\boldsymbol{A}) = \int_0^\infty e^{-2\alpha t}\mathrm{d}t\ \mathrm{Id} = \mathrm{Id}/(2\alpha)$. □

Corollary III.3. *For some constant $c > 0$ and $0 < \alpha < 1/8$, there exists a set Ω with $\log|\Omega| \geq cds\log\frac{cd}{es}$ such that for any $\boldsymbol{A} \neq \boldsymbol{A}' \in \Omega$:*

1. *$\boldsymbol{A}$ is row-s-sparse*
2. *$\boldsymbol{C}_\infty(\boldsymbol{A}) = \mathrm{Id}$*
3. *$KL(\mathbb{P}_{\boldsymbol{A}}, \mathbb{P}_{\boldsymbol{A}'}) \leq \alpha\log|\Omega|$*
4. *$\|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 \geq \alpha c^2 T^{-1} ds\log\frac{ced}{s}$.*

Proof. Define $\Omega = \{\frac{1}{2}\mathrm{Id} + w\boldsymbol{B} : \boldsymbol{B} \in \Omega_a\}$ for some $w > 0$ and Ω_a as defined in Lemma III.5. Then $|\Omega| = |\Omega_a|$ and hence $\log|\Omega| \geq cds\log\frac{ced}{s}$. Condition 1 is verified trivially and Lemma III.6 gives condition 2. Further, from Lemma III.5 point 3. $\|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 = w^2\|\boldsymbol{B} - \boldsymbol{B}'\|_F^2 \geq w^2 cds$. Also, the maximum Hamming distance being $2ds$, we get $\|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 \leq 2w^2 ds$.

The Kullback-Leibler divergence writes

$$\begin{aligned} KL(\mathbb{P}_{\boldsymbol{A}}, \mathbb{P}_{\boldsymbol{A}'}) &= \mathbb{E}_{\boldsymbol{A}}\left[\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{A}'}}\right] = \mathbb{E}_{\boldsymbol{A}}\left[\mathcal{L}^*(\boldsymbol{A}) - \mathcal{L}^*(\boldsymbol{A}')\right] \\ &= \frac{T}{2}\|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 \leq w^2 Tds. \end{aligned}$$

Indeed, the terms from $\mathcal{L}^*$ that we neglect in the definition of $\mathcal{L}_T$ are cancelling each other due to condition 2.

Choose $0 < \alpha < 1/8$ and $w^2 = \alpha c T^{-1}\log\frac{ced}{s}$ such that $KL(\mathbb{P}_{\boldsymbol{A}}, \mathbb{P}\boldsymbol{A}') \leq \alpha\log|\Omega|$. Then we also have $\|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 \geq \alpha c^2 T^{-1} ds\log\frac{ced}{s}$. □

Theorem III.2 is the corollary of the preceding and of Theorem 2.7 from [Tsy08].

6.4 Proof of Theorem III.3 (Restricted Eigenvalue property)

Lemma III.7. *Take a random symmetric matrix $\boldsymbol{C}$. Define $K(s) := \{u : \|u\|_0 \leq s\}$. Assume that for any vector $u \in K(s)$ such that $\|u\|_2 \leq 1$, and any $R \geq 0$, we have $\mathbb{P}\left[|u^\top \boldsymbol{C} u| \geq R\right] \leq p(R)$. Then:*

$$\mathbb{P}\left[\sup_{u \in K(s), \|u\|_2 \leq 1} |u^\top \boldsymbol{C} u| \geq 3R\right] \leq 21^s (d^s \wedge (ed/s)^s) p(R).$$

We omit the proof as it follows exactly the same steps as the one of Lemma F.2 from the supplement to [BM15]. Recall now the definition $C(s, c_0) = \{u : \|u\|_1 \leq (1 + c_0)\|u_{\mathcal{I}_s(u))}\|_1\}$, where $\mathcal{I}_s(u))$ is the set of indices of the s largest values of $|u|$. Using Lemmas F.1 and F.3 from the supplement to [BM15], we get that:

$$\sup_{u \in C(s,c_0), \|u\|_2 \leq 1} u^\top \widehat{\boldsymbol{C}}_T u \leq 3(c_0 + 2)^2 \sup_{u \in K(2s), \|u\|_2 \leq 1} u^\top \widehat{\boldsymbol{C}}_T u.$$

Taking this combined with Lemma III.7 applied using hypothesis (H4), we get the following Lemma.

Lemma III.8. *For any $R \geq 0$, we have:*

$$\mathbb{P}\left[\sup_{\substack{u \in C(s,c_0) \\ \|u\|_2 \leq 1}} |u^\top (\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty) u| \geq R\right] \leq 2 \exp\left(-TH\left(\frac{R}{9(c_0+2)^2}\right) + s \log(21d \wedge 21ed/s)\right).$$

We conclude by proving the Restricted Eigenvalue inequality.

Theorem III.7 (Restricted Eigenvalue).

$$\mathbb{P}\left[\inf_{\substack{u \in C(s,c_0) \\ \|u\|_2 \leq 1}} |u^\top \widehat{\boldsymbol{C}}_T u| \leq \kappa^2\right] \leq 2 \exp\left(-TH\left(\frac{\kappa^2}{9(c_0+2)^2}\right) + s \log(21d \wedge 21ed/s)\right).$$

Proof. We apply the lemma with $R = \kappa^2 = \min \mathrm{Sp}(\boldsymbol{C}_\infty)/2$ and use the fact that

$$\inf_{u \in C(s,c_0), \|u\|_2 = 1} u^\top \boldsymbol{C}_\infty u \geq \min \mathrm{Sp}(\boldsymbol{C}_\infty). \qquad \square$$

We can actually have, with the same probability, an additional upper bound on $u^\top \widehat{\boldsymbol{C}}_T u$, from which we get a bound on the diagonal elements of $\widehat{\boldsymbol{C}}_T$, as stated in

Corollary III.4. *Set* $\epsilon_0 \in (0,1)$. *For* $T \geq T_0 := H\left(\frac{\kappa^2}{9(c_0+2)^2}\right)^{-1}(s\log(21d \wedge 21ed/s) + \log 4\epsilon_0^{-1})$, *we have:*

$$\mathbb{P}\left[\inf_{u\in C(s,c_0),\|u\|_2\leq 1} |u^\top \widehat{\boldsymbol{C}}_T u| \geq \kappa^2, \|\operatorname{diag}\widehat{\boldsymbol{C}}_T\|_\infty \leq \|\operatorname{diag}\boldsymbol{C}_\infty\|_\infty + \kappa^2\right] \geq 1 - \frac{\epsilon_0}{2}.$$

Proof. From Lemma III.8 we get a set of events of probability $1 - \frac{\epsilon_0}{2}$ which verifies $\sup_{\substack{u\in C(s,c_0)\\ \|u\|_2\leq 1}} |u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u| \leq R$. From this follows the infimum condition as in Theorem III.7. Further, by taking for some i, $u = e_i \in C(s,c_0)$, we get $|\widehat{\boldsymbol{C}}_T^{ii} - \boldsymbol{C}_\infty^{ii}| \leq \kappa^2$ and the result follows. □

We finish this Section by showing different ways the Restricted Eigenvalue property can be expressed, using matrices, vectors, $\widehat{\boldsymbol{C}}_T$ or L^2 norm. Using this we get for instance Theorem III.3.

Lemma III.9. *For any subset* $E \subset \mathbb{R}^d$, *we have:*

$$\begin{aligned}\sup_{\boldsymbol{A}:\forall i\leq d, \boldsymbol{A}^{i,\bullet}\in E} \frac{\|\boldsymbol{A}X\|_{L^2}}{\|\boldsymbol{A}\|_F} &= \left(\sup_{\boldsymbol{A}:\forall i\leq d, \boldsymbol{A}^{i,\bullet}\in E, \|\boldsymbol{A}\|_F\leq 1} \operatorname{tr}\boldsymbol{A}\widehat{\boldsymbol{C}}_T\boldsymbol{A}^\top\right)^{1/2}\\ &= \left(\sup_{u\in E,\|u\|_2\leq 1} u^\top\widehat{\boldsymbol{C}}_T u\right)^{1/2}\\ &= \sup_{u\in E}\frac{\|u^\top X\|_{L^2}}{\|u\|_2}.\end{aligned}$$

$$\begin{aligned}\inf_{\boldsymbol{A}:\forall i\leq d, \boldsymbol{A}^{i,\bullet}\in E} \frac{\|\boldsymbol{A}X\|_{L^2}}{\|\boldsymbol{A}\|_F} &= \left(\inf_{\boldsymbol{A}:\forall i\leq d, \boldsymbol{A}^{i,\bullet}\in E, \|\boldsymbol{A}\|_F\leq 1} \operatorname{tr}\boldsymbol{A}\widehat{\boldsymbol{C}}_T\boldsymbol{A}^\top\right)^{1/2}\\ &= \left(\inf_{u\in E,\|u\|_2\leq 1} u^\top\widehat{\boldsymbol{C}}_T u\right)^{1/2}\\ &= \inf_{u\in E}\frac{\|u^\top X\|_{L^2}}{\|u\|_2}.\end{aligned}$$

Proof. We have the following relations for any matrix $\boldsymbol{A}$:

$$\|\boldsymbol{A}X\|_{L^2}^2 = \operatorname{tr}\boldsymbol{A}\widehat{\boldsymbol{C}}_T\boldsymbol{A}^\top = \sum_{i=1}^d (\boldsymbol{A}^{i,\bullet})^\top\widehat{\boldsymbol{C}}_T(\boldsymbol{A}^{i,\bullet}) = \sum_{i=1}^d \|\boldsymbol{A}^{i,\bullet}\|_2^2 \left(\frac{\boldsymbol{A}^{i,\bullet}}{\|\boldsymbol{A}^{i,\bullet}\|_2}\right)^\top \widehat{\boldsymbol{C}}_T \left(\frac{\boldsymbol{A}^{i,\bullet}}{\|\boldsymbol{A}^{i,\bullet}\|_2}\right)$$

Similarly, for a vector u, $\|u^\top X\|_{L^2}^2 = u^\top \widehat{\boldsymbol{C}}_T u$.

Assume that for any $i \leq d$, $\boldsymbol{A}^{i,\bullet} \in E$. We immediately get $\|\boldsymbol{A}X\|_{L^2}^2 \leq \|\boldsymbol{A}\|_F^2 \sup_{u\in E} \frac{\|u^\top X\|_{L^2}^2}{\|u\|_2^2}$. Hence we have $\sup_{\boldsymbol{A}:\forall i\leq d, \boldsymbol{A}^{i,\bullet}\in E} \frac{\|\boldsymbol{A}X\|_{L^2}}{\|\boldsymbol{A}\|_F} \leq \sup_{u\in E} \frac{\|u^\top X\|_{L^2}}{\|u\|_2}$. Choose now a vector u that realizes the supremum on the RHS. By choosing $\boldsymbol{A} = \mathbb{1}u^\top$, we get the equality in the inequality.

The proof for the infimums is exactly analogous.

□

6.5 Proof of Theorem III.4 (asymptotic properties of the Adaptive Lasso)

$\widehat{\boldsymbol{A}}_{ad.}$ is defined as the minimizer of the penalized log-likelihood, see Equation (3.0.1). The penalization includes the MLE and we denote $\boldsymbol{\Gamma} = 1/|\widehat{\boldsymbol{A}}_{MLE}|^\gamma$. We start by re-centering and changing the normalization of the objective function, then separating the log-likelihood from the penalization:

$$\begin{aligned}\sqrt{T}(\widehat{\boldsymbol{A}}_{ad.} - \boldsymbol{A}_0) = \hat{\boldsymbol{U}} &= \arg\min_{\boldsymbol{U}} \phi_1(\boldsymbol{U}) + \phi_2(\boldsymbol{U}) \\ \phi_1(\boldsymbol{U}) &:= T\mathcal{L}_T(\boldsymbol{A}_0 + \boldsymbol{U}/\sqrt{T}) - T\mathcal{L}_T(\boldsymbol{A}_0) \\ \phi_2(\boldsymbol{U}) &:= \lambda T\|(\boldsymbol{A}_0 + \boldsymbol{U}/\sqrt{T}) \odot \boldsymbol{\Gamma}\|_1 - \lambda T\|\boldsymbol{A}_0 \odot \boldsymbol{\Gamma}\|_1.\end{aligned}$$

Using equation (1.2.4), we characterize the limit structure of ϕ_1.

$$\begin{aligned}\phi_1(\boldsymbol{U}) &= \sqrt{T}\operatorname{tr}\boldsymbol{U}^\top\boldsymbol{\varepsilon}_T + \frac{1}{2}\operatorname{tr}\boldsymbol{U}\widehat{\boldsymbol{C}}_T\boldsymbol{U}^\top \\ &= \frac{1}{\sqrt{T}}\int_0^T (\boldsymbol{U}X_t)^\top \mathrm{d}W_t + \frac{1}{2T}\int_0^T \|\boldsymbol{U}X_t\|_2^2 \mathrm{d}t.\end{aligned}$$

1. From assumption (H1), X is ergodic and therefore we can apply the ergodic theorem for the classical integral:

$$\frac{1}{T}\int_0^T \|\boldsymbol{U}X_t\|_2 \mathrm{d}t \xrightarrow{\mathbb{P}} \mathbb{E}\left[\|\boldsymbol{U}X_0\|_2^2\right] = \operatorname{tr}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top.$$

2. The stochastic integral $M_T = \int_0^T (\boldsymbol{U}X_t)^\top \mathrm{d}W_t$ is a martingale, for which we apply the central limit theorem for martingales, recalled below in Lemma III.10.

   Lemma III.10 ([vZ00, Theorem 4.1]). *Let $(M_t; \mathcal{F}_t : t \geq 0)$ be a d-dimensional continuous local martingale. If there exist invertible, non-random $d \times d$-matrices $(K_t : t \geq 0)$ such that as $t \to \infty$*

- $K_t \langle M \rangle_t K_t^\top \xrightarrow{\mathbb{P}} \eta\eta^\top$ *where* η *is a random* $d \times d$*-matrix;*
- $|K_t| \to 0$*;*

then, for each $\mathbb{R}^k$*-valued random vector* X *defined on the same probability space as* M*, we have*

$$(K_t M_t, X) \xrightarrow{\mathrm{d}} (\eta Z, X) \qquad \textit{as} \quad t \to \infty,$$

where $Z \stackrel{\mathrm{d}}{=} \mathcal{N}(0, \mathrm{Id})$ *and* Z *is independent of* (η, X).

We have:

$$\langle M \rangle_T = \int_0^T \|\boldsymbol{U} X_t\|_2^2 \mathrm{d}t$$
$$\left(\frac{1}{\sqrt{T}}\right)^2 \langle M \rangle_T \xrightarrow{\mathbb{P}} \operatorname{tr} \boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top.$$

Hence

$$\frac{1}{\sqrt{T}} \int_0^T (\boldsymbol{U} X_t)^\top \mathrm{d}W_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \operatorname{tr} \boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top\right).$$

Introduce then a centered Gaussian $d \times d$ matrix $\boldsymbol{G}$ such that $\mathbb{C}\mathrm{ov}\left(\boldsymbol{G}^{ij}, \boldsymbol{G}^{kl}\right) = \mathbb{1}_{j=l} \boldsymbol{C}_\infty^{ik}$. Then, for any matrix $\boldsymbol{U}$:

- $\operatorname{tr} \boldsymbol{U}\boldsymbol{G}$ is a Gaussian variable
- $\mathbb{E}\left[\operatorname{tr} \boldsymbol{U}\boldsymbol{G}\right] = 0$
- $\mathbb{V}\mathrm{ar}\left(\operatorname{tr} \boldsymbol{U}\boldsymbol{G}\right) = \sum_{ijkl} \boldsymbol{U}^{ji}\boldsymbol{U}^{lk} \mathbb{C}\mathrm{ov}\left(\boldsymbol{G}^{ij}\boldsymbol{G}^{kl}\right) = \sum_{ijk} \boldsymbol{U}^{jk}\boldsymbol{U}^{ji}\boldsymbol{C}_\infty^{ik} = \operatorname{tr} \boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top$

From there,

$$\frac{1}{\sqrt{T}} \int_0^T (\boldsymbol{U} X_t)^\top \mathrm{d}W_t \xrightarrow{\mathcal{L}} \operatorname{tr} \boldsymbol{U}\boldsymbol{G}.$$

From the two preceding points, we conclude:

$$\phi_1(\boldsymbol{U}) \xrightarrow{\mathcal{L}} \operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}\boldsymbol{G}\right). \tag{6.5.1}$$

Second, we observe the limit structure of the penalization ϕ_2. Denote $\mathcal{A}_0 = \operatorname{supp} \boldsymbol{A}_0$. We have $\phi_2(\boldsymbol{U}) = \lambda T \sum_{ij} \boldsymbol{\Gamma}^{ij} \left(\left|\boldsymbol{A}_0^{ij} + \boldsymbol{U}^{ij}/\sqrt{T}\right| - \left|\boldsymbol{A}_0^{ij}\right|\right)$.

1. If $(i,j) \in \mathcal{A}_0$, for high enough T, $\sqrt{T}\left|\boldsymbol{A}_0^{ij} + \boldsymbol{U}^{ij}/\sqrt{T}\right| - \left|\boldsymbol{A}_0^{ij}\right| = \operatorname{sign}(\boldsymbol{A}_0^{ij})|\boldsymbol{U}^{ij}|$, and $\boldsymbol{\Gamma}^{ij} \xrightarrow{\mathbb{P}} |\boldsymbol{A}_0^{ij}|^{-\gamma}$, a positive constant. From our assumption, $\lambda\sqrt{T} \to 0$, hence
$$\lambda T \boldsymbol{\Gamma}^{ij}\left(\left|\boldsymbol{A}_0^{ij} + \boldsymbol{U}^{ij}/\sqrt{T}\right| - \left|\boldsymbol{A}_0^{ij}\right|\right) \xrightarrow{\mathbb{P}} 0.$$
2. Else, $(i,j) \in \bar{\mathcal{A}}_0$ and then $\lambda T \boldsymbol{\Gamma}^{ij}\left(\left|\boldsymbol{A}_0^{ij} + \boldsymbol{U}^{ij}/\sqrt{T}\right| - \left|\boldsymbol{A}_0^{ij}\right|\right) = \lambda T^{(\gamma+1)/2}|\sqrt{T}\widehat{\boldsymbol{A}}_{MLE}^{ij}|^{-\gamma}\left|\boldsymbol{U}^{ij}\right|$. We know the MLE is root-t consistent, hence $\sqrt{T}\widehat{\boldsymbol{A}}_{MLE}^{ij} = O(1)$ and by assumption $\lambda T^{(\gamma+1)/2} \to +\infty$. Hence, the expression diverges to $+\infty$.

For high T, ϕ_2 becomes flat 0 on the support and infinite outside. Combining with the result from Equation (6.5.1), we have:

$$\phi_1(\boldsymbol{U}) + \phi_2(\boldsymbol{U}) \xrightarrow{\mathcal{L}} \begin{cases} +\infty & \text{if } \boldsymbol{U}_{\bar{\mathcal{A}}_0} = 0, \\ \operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}G\right) & \text{else.} \end{cases}$$

We finally need to compute the minimum of that function. Take $\boldsymbol{U}$ such that $\boldsymbol{U}_{\bar{\mathcal{A}}_0} = 0$. Recall that we treat a matrix restricted to a set of indices as a vector. Then:

$$\begin{aligned} \operatorname{tr}\boldsymbol{U}\boldsymbol{G} &= \sum_{ij} \boldsymbol{U}^{ij}\boldsymbol{G}^{ji} \\ &= \left((\boldsymbol{G}^\top)_{|\mathcal{A}_0}\right)^\top \boldsymbol{U}_{|\mathcal{A}_0} \\ \operatorname{tr}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top &= \sum_{ijkl} \boldsymbol{U}^{ij}\boldsymbol{U}^{kl}(\mathbb{1}_{i=k}\boldsymbol{C}_\infty^{jl}) \\ &= (\boldsymbol{U}_{|\mathcal{A}_0})^\top(\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0^2}\boldsymbol{U}_{|\mathcal{A}_0}. \end{aligned}$$

$(\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0^2}$ is the restriction of $\boldsymbol{C}_\infty \otimes \mathrm{Id}$ to the indices in $\mathcal{A}_0^2 := \mathcal{A}_0 \times \mathcal{A}_0$ and $\boldsymbol{C}_\infty \otimes \mathrm{Id}$ is invertible and symmetric, with inverse $\boldsymbol{C}_\infty^{-1} \otimes \mathrm{Id}$, hence $(\boldsymbol{C}_\infty \otimes \mathrm{Id})_{\mathcal{A}_0^2}$ is invertible and symmetric. Hence, $\operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}\boldsymbol{G}\right) = \frac{1}{2}(\boldsymbol{U}_{|\mathcal{A}_0})^\top(\mathrm{Id} \otimes \boldsymbol{C}_\infty)_{|\mathcal{A}_0^2}\boldsymbol{U}_{\mathcal{A}_0} + (\boldsymbol{U}_{|\mathcal{A}_0})^\top(\boldsymbol{G}^\top)_{|\mathcal{A}_0}$, which is a quadratic function of $\boldsymbol{U}_{|\mathcal{A}_0}$ and we compute easily the minimum, which shows that $\widehat{\boldsymbol{U}}_{|\mathcal{A}_0}$ is a centered Gaussian and completes the proof of point 2:

$$\widehat{\boldsymbol{U}}_{|\mathcal{A}_0} \xrightarrow{\mathcal{L}} -(\boldsymbol{G}^\top)_{|\mathcal{A}_0}\left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0^2}\right)^{-1}, \quad \widehat{\boldsymbol{U}}_{\bar{\mathcal{A}}_0} \xrightarrow{\mathcal{L}} 0$$

and we find that $(\boldsymbol{G}^\top)_{|\mathcal{A}_0}\left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0^2}\right)^{-1} \sim \mathcal{N}(0,\mathcal{V})$ with $\mathcal{V} := \left((\boldsymbol{C}_\infty \otimes \mathrm{Id})_{|\mathcal{A}_0^2}\right)^{-1}$.

Proceed now with point 1. We proved in the preceding the asymptotic normality of the convergence on $\mathcal{A}_0$, from which we deduce $\forall (i,j) \in \mathcal{A}_0, \mathbb{P}\left[\widehat{\boldsymbol{A}}_{ad.}^{ij} \neq 0\right] \to 1$.

Take now $(i,j) \in \bar{\mathcal{A}}_0$ and assume the event $\widehat{\boldsymbol{A}}^{ij}_{ad.} \neq 0$. We write the optimality conditions, multiplied by $\sqrt{T}$, and apply the absolute value: $\left|\sqrt{T}S^{ij} + (\sqrt{T}\widehat{\boldsymbol{A}}_{ad.}\widehat{\boldsymbol{C}}_T)^{ij}\right| = \lambda \boldsymbol{\Gamma}^{ij}\sqrt{T}$. When $T \to +\infty$:

$$\begin{aligned}
\sqrt{T}S^{ij} = \frac{1}{\sqrt{T}}\int_0^T X_t^j \mathrm{d}W_t^i &\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{C}^{jj}_\infty\right) \\
\sqrt{T}\widehat{\boldsymbol{A}}_{ad.} &\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathcal{V}\right) \\
\widehat{\boldsymbol{C}}_T &\xrightarrow{\mathbb{P}} \boldsymbol{C}_\infty \\
\lambda \boldsymbol{\Gamma}^{ij}\sqrt{T} = \lambda T^{(\gamma+1)/2}(\sqrt{T}|\widehat{\boldsymbol{A}}^{ij}_{MLE}|)^{-\gamma} &\xrightarrow{\mathbb{P}} +\infty.
\end{aligned}$$

When $T \to +\infty$, we can therefore bound the probability of $\widehat{\boldsymbol{A}}^{ij}_{ad.} \neq 0$ by the probability that the sum of some two Gaussians is equal to a diverging number, in absolute value. This is clearly of a probability converging to zero. Therefore, $\forall (i,j) \in \mathcal{A}_0, \mathbb{P}\left[\widehat{\boldsymbol{A}}^{ij}_{ad.} = 0\right] \to 1$.

6.6 Deviation bound

Recall Bernstein's inequality, see Chapter 4, Exercise 3.16 in [RY99]:

Lemma III.11 (Bernstein's inequality). *Let M be a scalar continuous local martingale. For all $a > 0, b > 0$:*

$$\mathbb{P}\left[M_t \geq a, \langle M\rangle_t \leq b\right] \leq \exp\left(-\frac{a^2}{2b}\right).$$

Lemma III.12. *Let M be a scalar continuous local martingale. For any $x > 0$:*

$$\mathbb{P}\left[M_t \geq \sqrt{4e\langle M\rangle_t\left(x + \log\left(2 + |\log\langle M\rangle_t|\right)\right)}\right] \leq \frac{\pi^2}{3}\exp(-2x).$$

Proof. Observe that if $j \leq \log\langle M\rangle_t \leq j+1$ for some integer j, then $|\log\langle M\rangle_t| \geq |j| - 1$.

$$\begin{aligned}
P &= \mathbb{P}\left[M_t \geq \sqrt{4e\langle M\rangle_t\left(x + \log\left(2 + |\log\langle M\rangle_t|\right)\right)}\right] \\
&= \sum_{j\in\mathbb{Z}} \mathbb{P}\left[M_t \geq \sqrt{4e\langle M\rangle_t\left(x + \log\left(2 + |\log\langle M\rangle_t|\right)\right)}, e^j \leq \langle M\rangle_t < e^{j+1}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{j\in\mathbb{Z}} \mathbb{P}\left[M_t \geq \sqrt{4e^{j+1}\left(x+\log\left(1+|j|\right)\right)}, \langle M\rangle_t < e^{j+1}\right] \\
&\leq \sum_{j\in\mathbb{Z}} \exp\left(-2\left(x+\log\left(1+|j|\right)\right)\right) \\
&= \exp(-2x)\sum_{j\in\mathbb{Z}} \frac{1}{\left(1+|j|\right)^2} \\
&= 2\exp(-2x)\sum_{j\in\mathbb{N}^*} \frac{1}{j^2} \\
&= \frac{\pi^2}{3}\exp(-2x).
\end{aligned}
$$

□

Theorem III.8. *Define for $x > 0$:*

$$
\theta(x,(X_t)) := \sqrt{4eT^{-1}|\mathrm{diag}\left(\widehat{\boldsymbol{C}}_T\right)|_\infty \left(x+\log(2+|\log T\mathrm{diag}\left(\widehat{\boldsymbol{C}}_T\right)|_\infty)\right)}, \tag{6.6.1}
$$

where we denote diag *the extraction of the diagonal of a matrix and* log *applies naturally to each term (which are all positive).*

For any matrix $\boldsymbol{U}$, the set of events

$$
\mathbb{P}\left[\left\langle \boldsymbol{U}, T^{-1}\int_0^T \mathrm{d}W_t X_t^\top \right\rangle_F \leq \theta(x,(X_t))\|\boldsymbol{U}\|_1\right] \geq 1-\frac{\pi^2}{3}\exp(-2x+2\log d). \tag{6.6.2}
$$

Proof. Set $i,j \leq d$. Recall that $\int_0^T \mathrm{d}W_s^i X_t^j$ is a martingale, and its bracket is $\int_0^T (X_t^j)^2\mathrm{d}t = T\widehat{\boldsymbol{C}}_T^{jj}$. By applying Lemma III.12:

$$
\mathbb{P}\left[\int_0^T \mathrm{d}W_s^i X_t^j \geq \sqrt{4eT\widehat{\boldsymbol{C}}_T^{jj}\left(x+\log\left(2+\left|\log T\widehat{\boldsymbol{C}}_T^{jj}\right|\right)\right)}\right] \leq \frac{\pi^2}{3}\exp(-2x).
$$

We have $\sqrt{4eT\widehat{\boldsymbol{C}}_T^{jj}\left(x+\log\left(2+\left|\log T\widehat{\boldsymbol{C}}_T^{jj}\right|\right)\right)} \leq T\theta(x,(X_t))$, hence using an union bound:

$$
\mathbb{P}\left[\left\|\int_0^T \mathrm{d}W_s X_t^\top\right\|_\infty \geq T\theta(x,(X_t))\right] \leq \frac{\pi^2}{3}d^2\exp(-2x). \tag{6.6.3}
$$

Observe now that by homogeneity, it suffices to prove Equation (6.6.2) for any matrix $\boldsymbol{U}$ such that $\theta(x,(X_t))\|\boldsymbol{U}\|_1 \leq 1$. Then we have:

$$\left|\left\langle \boldsymbol{U}, T^{-1}\int_0^T \mathrm{d}W_t X_t^\top \right\rangle_F\right| = \left|\sum_{ij} \theta(x,(X_t))\boldsymbol{U}^{ij}\frac{\int_0^T \mathrm{d}W_t^i X_t^j}{T\theta(x,(X_t))}\right| \leq \frac{\|\int_0^T \mathrm{d}W_s X_t^\top\|_\infty}{T\theta(x,(X_t))}.$$

Equation (6.6.2) follows from Equation (6.6.3).

□

CHAPTER IV

Sparse estimation applied to finance: stock returns, dividend futures

In this Chapter, we apply the sparse estimation methods introduced in Chapter III to real-world financial datasets. In a first part, which closely follows Section 4.5 from Chapter III, we analyze the properties of daily returns of stocks present in the SP500 index. In a second part, we analyze daily prices of EURO STOXX 50 dividend futures.

1 Measuring mean-return properties of stock returns using the Ornstein-Uhlenbeck model

Under the efficient market hypothesis, the price of an asset reflects all available information on its value. As a consequence, prices are often modeled by random walks, which means future returns can't be predicted using he knowledge of past returns. In order to challenge this hypothesis, extensive research has been done to uncover mean-return properties of stock returns.

Mean-return properties are a broad term which describes the tendency of a time-series to have above-average performances followed by below-average performances and *vice-versa*. The average here can be computed over a variety of sets, the most notable example being the long-term average such as the average return over several years. Mean-return properties are notable with respect to the efficient market hypothesis for instance because they have a natural interpretation in behavioral economics. Indeed, if investors over-react to recent past information, the changes in price are then corrected on the longer term, from which follows mean-reversion

[DBT85]. Also, mean-reversion can be used for a possibly market-beating contrarian strategy, consisting in selling past over-performers and buying past under-performers [SSS05, BW06]. However, exhibiting mean-reversion cannot be directly used as evidence of a failure of the efficient market hypothesis. Indeed, an alternative to the behavioral interpretation is that mean-reversion can be the result of an efficient market with an additional time-varying variable, as compared to the CAPM, such as varying risk premia [GR96]. Also, the contrarian strategy must be shown to work under real-market conditions, including for instance transaction costs [CFS05], and not to be a false positive in a data-mining approach. In the end, the efficient market hypothesis holds a strong position in the literature and many arguments conciliate mean-reversion with market efficiency [Fam98].

The overreaction hypothesis for explaining mean-reversion has been formulated in [DBT85], where the authors provide first evidence of the existence of mean-reversion in the US stock market. Negative autocorrelation is found in [FF88] on a time-scale of 3-5 years. Further evidence is given in [PS88], where the authors show mean-reversion of monthly returns in the period 1926 to 1985 on a time-scale of several years, but fail to reach high significance levels. The authors use for this the variance ratio approach, which consists in computing statistics that are equal to 1 in expectation under the random walk hypothesis. If the series is negatively auto-correlate, a sign of mean-reversion, the statistic falls below 1. For a detailed explanation, see Chapter 2 in [CLM97], which also does a review of empirical results. A similar analysis but with stocks sorted by industries is done in [Gro04]. Mean-reverting properties have also been reported in an international setting, see [SSS05, BW06], and more specifically in emerging markets [AS13] and in China [KLN02]. For relatively recent results, the period 1967-2007 for US stocks is analyzed in [Muk11], and the period 1900-2009 for 18 OECD countries is analyzed in [SBvdH12]. Both articles point to the existence of mean-reversion in stock-returns, but with time-scales of the order of a few years or more.

Our work differs from the results presented above in several major points. First, we are looking for mean-reversion at a much shorter time-scale, of the order of a few days instead of years. Mean-reversion at that time-scale is reported in [Leh90], which analyzes weakly data, but on a different period than the one we consider. Second, we use more frequent data, as we will be using daily Close prices. Third, our approach consisting in fitting an Ornstein-Uhlenbeck model to our data is parametric and is therefore very much different from the results above which typically are non-parametric, for instance when one computes an auto-correlation function. Fourth, we will present results for sector averages as well as for individual stocks, and for both we will present a one-dimensional and a multi-dimensional analysis.

1.1 Analysis method

Our objective here is to investigate whether recent performances of the stock market provide information on future performances of the stock market. It is difficult to define objectively what are recent performances. One naturally considers the average return on a few past days, which leads to using moving averages, but it isn't clear how many days one should consider nor whether the weighting should be uniform or not. Instead, we decide to use here the exponential moving average, which relies on weighting past values with an exponential of the lag. For a local semimartingale M defined on $\mathbb{R}^+$, we can define its EMA E with parameter α by

$$E_t := \alpha \int_0^t e^{-\alpha(t-s)} \mathrm{d}M_s.$$

The exponential moving average puts a large weight on the most recent values of the time-series, which coincides the most with our intuition of what a recent performance is. It is also unaffected by the threshold effect that comes with simple moving averages, which can have large jumps because of new data, but also because some past data leaves the lag over which the moving average is computed. Also, we believe that the EMA has a more intuitive structure in terms of its evolution in time. Indeed, Ito's Lemma gives

$$\mathrm{d}E_t = -\alpha E_t \mathrm{d}t + \alpha \mathrm{d}M_t. \tag{1.1.1}$$

For these reasons, we use throughout this section the exponential moving averages of logarithmic single stock returns. For practical applications, we have to specify the discretisation setting that we use. We will assume observations at discrete times with constant time-step. Then linear approximations with $\alpha\delta t << 1$ give

$$E_{(k+1)\delta t} \simeq (1 - \alpha\delta t) E_{k\delta t} + \alpha\delta t \frac{\delta M_{(k+1)\delta t}}{\delta t}.$$

The EMA at time $(k+1)\delta t$ can therefore be seen as a weighted average of the past EMA value and the innovation $\frac{\delta M_{(k+1)\delta t}}{\delta t}$.

Let $S_t^{(i)}$ be the price of stock i at time t. The discretisation scheme means we observe them at times $(k\delta t)_{0 \leq k \leq n}$ for some fixed n. We will fix $\delta t = 1$, the time being counted in days, which is consistent with the use of close prices and interpreting the time variable as trading time. Then, define $\Delta \log S^{(i)}_{(k+1)\delta t} := \log S^{(i)}_{(k+1)\delta t} / S^{(i)}_{k\delta t}$, the logarithmic daily return. Using this, we set:

$$X_t^{(i)} := \alpha \int_0^t e^{-\alpha(t-s)} \mathrm{d} \log S_t^{(i)},$$

which converts in discrete time to

$$X_{k+1}^{(i)} \simeq (1-\alpha)X_k^{(i)} + \alpha\Delta \log S_{(k+1)\delta t}^{(i)}.$$

The process X_t is then a vector-valued stochastic process in $\mathbb{R}^d$, where d is the number of stocks that are considered. An industrial sector is then seen as a set of stocks, characterized by a set of indices $\mathcal{I}$. We define the sector average as the equal weighted average

$$Z_t^{\mathcal{I}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} X_t^{(i)}.$$

The choice of a uniform weighting is here arbitrary and a natural alternative is weighting by capitalization. In that comparison, the uniform weighting is putting more emphasis on smaller companies and past work has generally observed that this setting leads to stringer mean-reverting properties. Finally, we define sector-detrended returns. Industrial sectors define a partition of the stock market, hence for each stock i, there is a unique sector $\mathcal{I}$ such that $i \in \mathcal{I}$. We define then

$$Y_t^{(i)} := X_t^{(i)} - Z_t^{(\mathcal{I})}.$$

1.1.1 The data

As the source of the data, we use Yahoo Finance's historical adjusted close prices [Yah]. The prices are adjusted for stock splits and dividends. We use data spanning the period between January 3, 2000 and February 10, 2017, which corresponds to 4305 trading days. We consider all stocks that are components of the Standard & Poor's SP500 index at the final date of the dataset, which is reported by Thomson Reuters [Fin]. We restrict then the analysis to the stocks that have been part of the index since the start date of our dataset. This exposes us to some survival bias, but this is beyond the scope of this analysis. We classify each stock into its industrial sector using the GICS methodology [MSC].

1.1.2 The model

Given a process F that we wish to analyze we consider shorter sub-processes to which we fit an Ornstein-Uhlenbeck process

$$\mathrm{d}F_t = -\boldsymbol{A}(F_t - m)\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t. \tag{1.1.2}$$

For one-dimensional processes, we prefer to make clear that the parameters are real valued:

$$\mathrm{d}F_t = -a(F_t - m)\mathrm{d}t + \sigma\mathrm{d}W_t. \tag{1.1.3}$$

We compute the values of m and $\mathbf{\Sigma}$ using mean values and covariances. For our data which is computed using exponential moving averages, the value of m is not the average return of the underlying stock, as its computation weights more first returns than the last. However, if we assume that the returns are stationary, the m converges to the expected returns. The covariance matrix $\mathbf{\Sigma}$ is proportional to stock return covariances. Both the dynamics of the expected return and the dynamics of covariances and volatilities are important questions in finance but are beyond the scope of this analysis. We will simply assume that expected returns and volatilities are constant on the interval that we consider. We will also assume that the values of m and $\mathbf{\Sigma}$ are suitable estimations of those quantities in adjacent intervals.

For multi-dimensional processes, we will use several estimation methods and compare their performances. For one-dimensional processes, we will use the value of a which maximizes the log-likelihood:

$$\hat{a} := \frac{\int_0^T X_t \mathrm{d}X_t}{\int_0^T X_t^2 \mathrm{d}t}.$$

1.1.3 Benchmarking

As we are trying to show the existence of mean-reversion properties of the stock prices, our natural benchmark is the situation where all returns are uncorrelated. As we are not interested in the precision of our estimation of $\mathbf{\Sigma}$, we simulate data with that target covariance, but without temporal correlation. The result is a Brownian Motion log-price, from which we compute an exponential moving average, which we denote BMEMA (Brownian Motion Exponential Moving Average).

As seen in Equation (1.1.1), the BMEMA is mean-reverting with mean-reversion parameter α. Therefore, returns present mean-reverting properties if the fit to the resulting EMA has a mean-reversion parameter exceeding that value. In order to avoid false positives, we also compute a range of values of the mean-reversion parameter that can be obtained from a BMEMA using Monte-Carlo simulations.

Another important measure is the degree at which the results help in predicting future returns. Indeed, if the EMA verifies both Equation (1.1.1) and Equation (1.1.2), we have

$$\begin{aligned} -\alpha F_t \mathrm{d}t + \mathrm{d}M_t &= -\hat{\alpha}(F_t - m)\mathrm{d}t + \sigma \mathrm{d}W_t \\ \mathrm{d}M_t &= \alpha F_t \mathrm{d}t - \hat{\alpha}(F_t - m)\mathrm{d}t + \sigma \mathrm{d}W_t \end{aligned} \tag{1.1.4}$$

which lets us make a prediction on future increments of F. However, it is very difficult to beat in practice the prediction error of a constant, because of usually large volatilities compared to the predictable part. Instead, we will compute the log-likelihood values on adjacent intervals and compare them between those for the mean-reversion

parameter estimate from data and the one observed for the BMEMA. We believe this approach has theoretical value, but we also tested its power on simulated data, see Section 1.4, and conclude that it is very hard to reject the no-correlation hypothesis using this metric.

Finally, we wish to point out that in order to reject the efficient market hypothesis, we would need to show the possibility to consistently beat the market. For this, we would at least have to consider a trading strategy and compute its historical performance. In a valid approach, one needs to consider transaction costs and bid-ask spreads. This approach is beyond the scope of this work.

1.2 Sector averages

In this section, we analyze the mean-reverting properties of sector averages, denoted Z above. In the following work, we do the estimation on 30 sub-intervals of length 800, representing slightly above 3 years of daily close prices. We use for each interval the following 200 datapoints in order to evaluate the validity of our method by comparing its prediction to the benchmark method. On each interval, we estimate compute a set of statistics. Hence, we have for each statistic a set of 50 observations, one for each sub-interval. These statistics will evolve in time, but we will be only interested in their distribution, which we will usually characterise by its mean and standard deviation. Indeed, we expect these statistic to evolve continuously as at sub-interval has around 90% in common with the preceding sub-interval.

1.2.1 One-dimensional processes analysis

As expected from equation (1.1.1), the processes that we consider present clear mean-reversion properties. We also remark that they are visually similar to processes that are simulated based on parameters estimated from the real data, see Figure IV.1.

Let us first show results on the estimation of m and σ. We present the results in Figure IV.2, where each sector is indexed by its initials. We show both the values obtained from the estimation on the whole interval of study, in purple, as well as the distribution of the results obtained on the sub-intervals. The Figures provide the mean values for each statistic and the shaded are represents plus-minus one standard deviation of the distribution. We see that the distributions of m are typically more on the positive side, which is consistent with a long-term growth of the market, but the variations are large enough to have periods of negative growth. The values of σ vary more but are all of the same order of magnitude.

Let us proceed to the estimation of a. As said above, the estimation results vary with time, see Figure IV.3. We can attribute a part of the stationarity of the estimations to the fact that the sub-intervals we consider have non-void intersections.

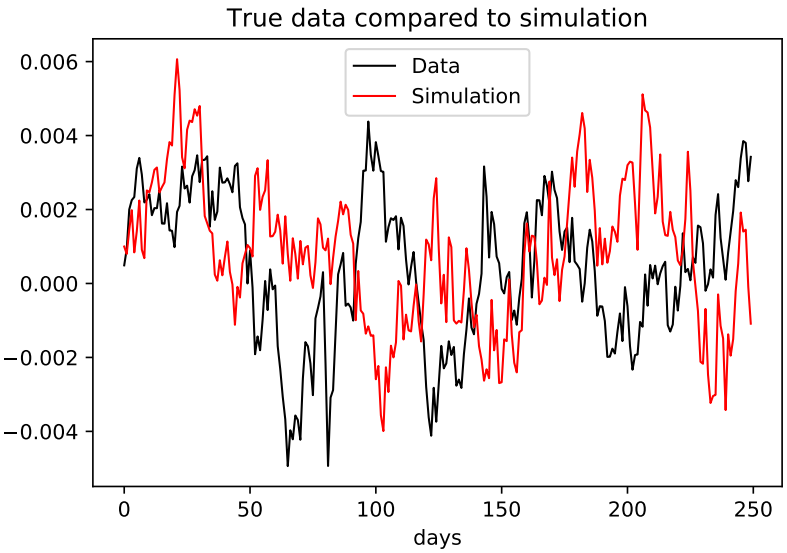
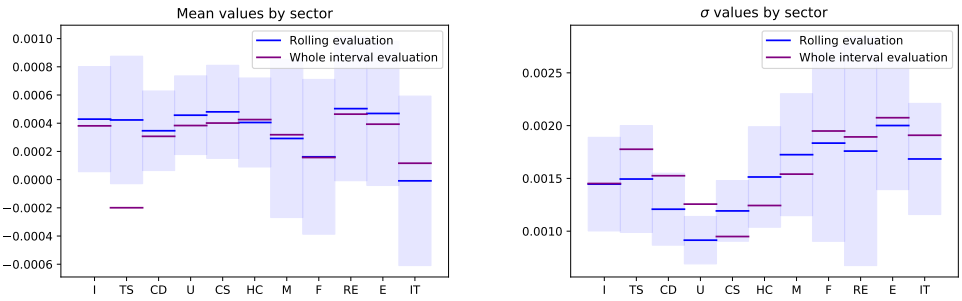


Figure IV.1 – Average process A for the Utilities sector, compared to a simulated trajectory. The simulation uses parameters fitted to the data; we show trajectories of length 250 days, while the estimation was done on a longer interval of length 800. We observe visually similarity between the two processes.



(a) Mean values of the EMA, by sector. (b) Volatilities of the EMA, by sector.

Figure IV.2 – Mean values and volatilities by sector. Whole interval evaluation corresponds to 17 years of data to which we fit a one-dimensional Ornstein-Uhlenbeck process as in Equation (1.1.3). We fit the model to 30 intervals of length 800 spanning the 17 years and report the mean value, the shaded areas representing $\pm$ one standard deviation computed from the 30-element sample. We see that the two values don't necessarily coincide and can be even quite different, but are usually relatively close.

Therefore, we will prefer to simply observe the distribution of the estimated values of a.

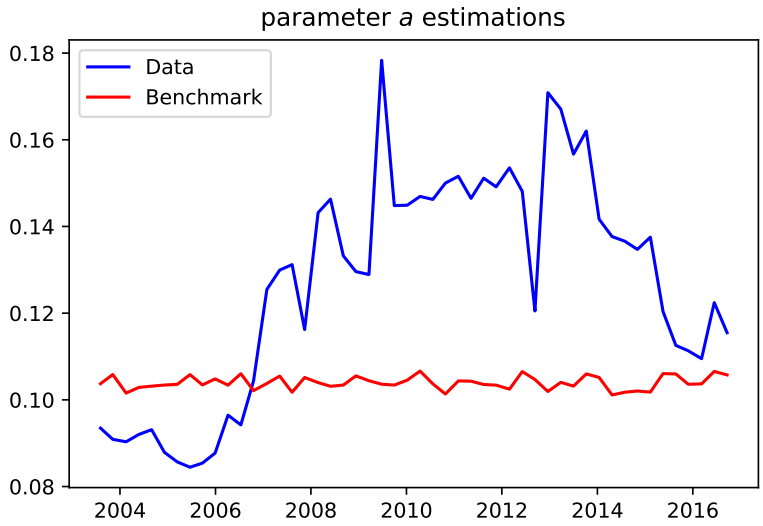


Figure IV.3 – Evolution of estimated values for a, for the Utilities sector average and for simulated benchmarks. The date points to the middles of the interval of length 200 on which the estimation is tested.

We present these results by sector in Figure IV.4. We compare there the results of estimation on real data to results of estimation on simulated BMEMA. For the benchmark, we mark the mean value, while the shaded area represents plus and minus one standard deviation observed in a thousand simulated trajectories of BMEMA. If the stock returns data was consistent with the BMEMA dynamic, which represents no time-correlation in returns, estimated values of a would be expected to fall in the shaded area. The blue marker shows the mean of the 30 estimations of a for each sector, while the shaded area show plus and minus one standard deviation of their distribution. We see that for 9 out of 11 sectors, the mean of the estimated values of a falls above the expected range consistent with BMEMA, while no mean value falls below the BMEMA mean. If we had time-correlation, we would expect roughly half of the 11 sector mean values to fall below the benchmark mean and most mean values inside the shaded area. We believe this is strong evidence of existence of a mean-reversion phenomenon for the returns of the sector averages. t also seems that some sectors are more concerned by this phenomenon than others - it is the strongest for Utilities and Financials, and weakest for Telecommunication Services and Information Technology. The excess mean-reversion varies between 0.005 and 0.04. Using Equation (1.1.4), and taking $m = 0$ for simplicity, we see that

an exponential moving average of returns at 1% indicates an expected return on the next day of negative 0.005% to 0.04%. These values are below the volatilities that we report in Figure IV.2b, hence our prediction will be of limited accuracy. However, for the Utilities sector, which has the highest mean-reversion mean parameter, we see that an EMA at 2.5% indicates an expected return of the same order as the volatility.

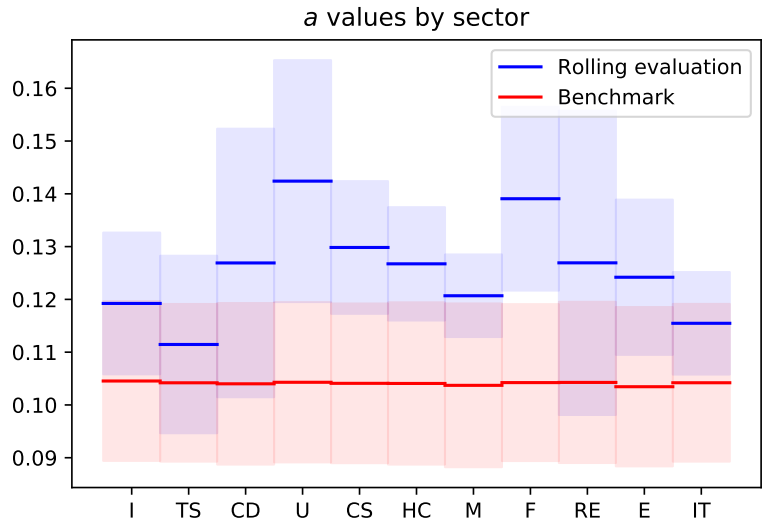
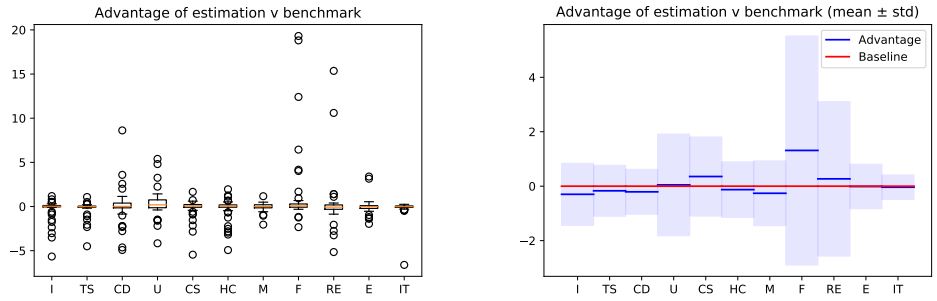


Figure IV.4 – Values of a computed for sector averages.

In order to verify the validity of our estimates, we use the results of the estimations on the sub-intervals of length 800 and compute their negative log-likelihoods on the following intervals of length 200. We also compute the value of the log-likelihodd on that same interval corresponding to the mean-reversion parameter obtained for the BMEMA. We compute the difference of the two scores, which is also the logarithm of their likelihood ratio, and show their distribution for each sector. We present the results in Figure IV.5. We see here that the estimate on real data doesn't have a consistently better score than the benchmark value. Looking at the detail, we see that most of the outliers, both positive and negative, are situated between 2008 and 2010. However, we will show in Section 1.4 that the log-likelihood test is not adapted to estimating whether data has mean-reversion properties.

1.2.2 Multi-dimensional process analysis

We further our analysis by considering the 11-dimensional vector process of sector averages. The multi-dimensional approach has the advantage of considering more



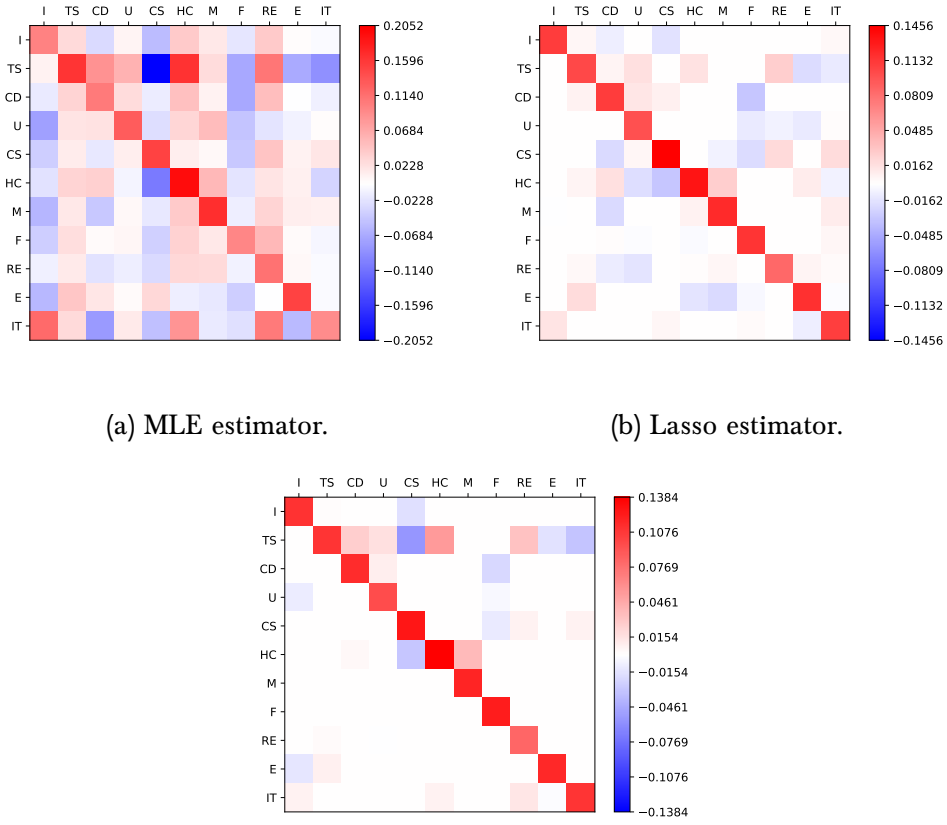
(a) Difference presented as a box-plot. (b) Means and standard deviations.

Figure IV.5 – Log-likelihood-ratios of parameters estimated on real data compared to benchmark values obtained using the BMEMA (positive values mean the real data estimation outperforms the benchmark). The distributions are heavily concentrated around 0, with some large deviations, see left figure, which influence greatly the mean and the standard deviation, see right figure.

complex mean-reversion properties, which involve multiple price processes simultaneously. For instance, through fitting the data to a multi-dimensional Ornstein-Uhlenbeck process, we could observe that some above-average recent returns of one process are followed by above or below-average returns for another process. Also, considering the vector process we can deal more easily with statistical issues coming with high-dimensional data, in this case that we can observe high values of mean-reversion parameters because of the large number of processes considered.

We will compare here the performances of the three estimators that we use in Chapter III: the Maximum Likelihood Estimator, the Lasso and the Adaptive Lasso. We present examples of estimation results in Figure IV.6. We see on a first glance that the penalized methods create stronger sparsity and remarkably, that they make diagonal entries relatively more relevant. We will also use a diagonal estimator, which uses the one-dimensional estimations from Section 1.2.1 above. Additionally, we will use a diagonal-matrix benchmark, based on a multi-dimensional BMEMA.

Finally, as in Section 1.2.1, we compute the log-likelihood score of the 5 estimators on the sub-interval following the estimation data. We present the differences compared to the benchmark in Figure IV.7. We see that the Maximum Likelihood Estimator has consistently the worse score. The reason behind it is the relatively high noise and the high number of matrix entries that we are estimating. This leads us to believe that the drift parameter matrix should be sparse. Restricting the analysis to only the diagonal gives more consistent results but on average has a worse result than the Lasso estimator, while not allowing to discover cross-sectorial effects. We note however that all methods fail to beat the benchmark in this analysis.



(a) MLE estimator.

(b) Lasso estimator.

(c) Adaptive Lasso estimator.

Figure IV.6 – Estimations of Ornstein-Uhlenbeck drift parameter fitted to the vectorial process of sector averages.

1.3 In-sector mean-return

In this section, we consider the single-stock processes detrended with respect to their sector average, denoted Y above. To the contrary of sector averages, where we couldn't assume that the returns are centered as the stock market is growing in the long-term, here we don't see any reason why would a stock price consistently outperform its sector. This leads us to setting $m = 0$ in Equations (1.1.2) and (1.1.3). We will follow the same methodology as for the sector average processes, first analysing single stock processes separately, then considering them as parts of vector valued processes.

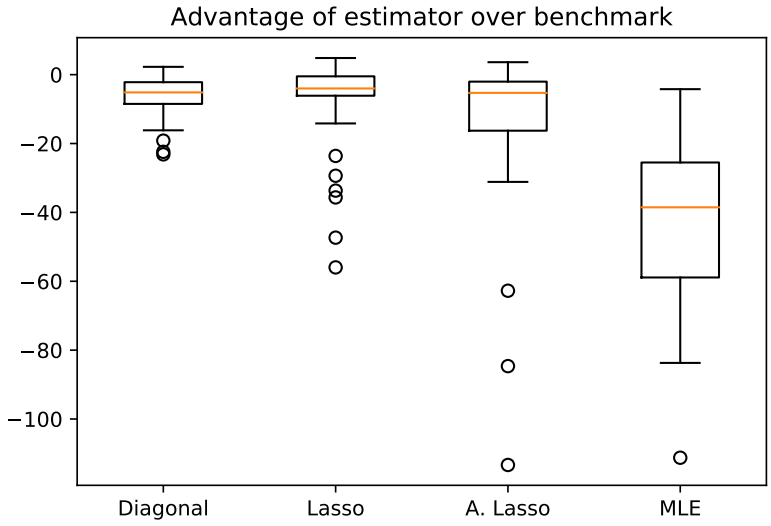


Figure IV.7 – Log-likelihood ratios of the 4 estimators and the benchmark. Values below zero mean that the estimators give worse results than the benchmark.

1.3.1 One-dimensional process analysis

Following the same methodology as in Section 1.2.1, we compute the values of a for each stock first. We provide the results for the 378 individual stocks that we consider in the analysis in Figure IV.8. For each sector and globally, the mean-reversion parameter values exceed the values expected from the benchmark. Around half of the parameters are one standard deviation above that mean. This provides evidence for the existence of mean-reversion in each individual process. We have computed the log-likelihood ratios for each process but similarly to the results in Figure IV.5, we get distributions highly concentrated around 0.

1.3.2 Multi-dimensional process analysis

We follow with a multi-dimensional analysis in each sector. As we have detrended each process using the sector average, we have created a linear dependency between the processes Y. For this reason, we exclude one stock process from each multi-dimensional analysis. That stock-process is simply the opposite of the sum of the other stock processes. This doesn't impact strongly our results, as if we show that all stocks except one mean-revert as a vector, the stock that was left apart is also mean-reverting. However, we won't be able to recover individual relations of that stock with the others.
The performance of the MLE in this experiment was not satisfying. The resulting

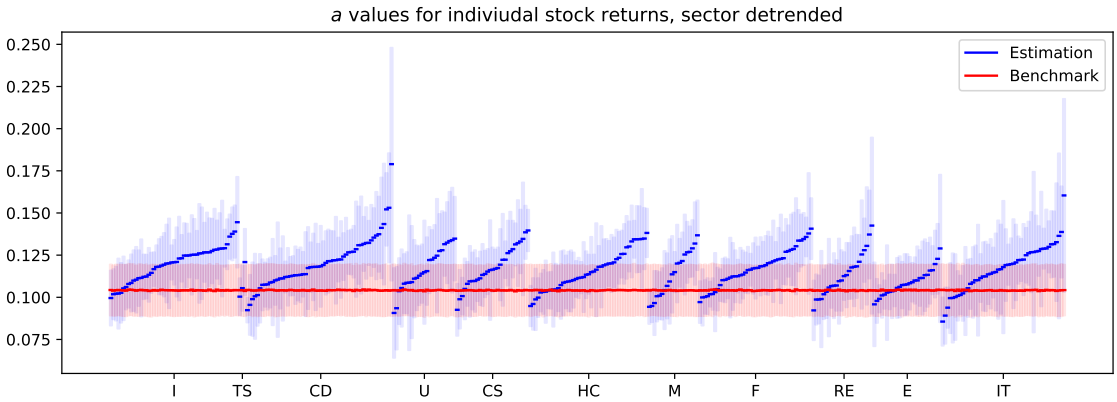


Figure IV.8 – Distributions of the mean-return parameter computed for individual stocks, compared to benchmark values for the BMEMA. The stocks are sorted by industrial sector, and in sector by growing values of mean a.

log-likelihood was of orders higher than for all other methods. The situation was similar to the one from Figure IV.7 hence we excluded it from the comparison that we present in Figure IV.9. To decrease clutter, we present here the mean values of the log-likelihood computed on 10 intervals without providing the standard deviations. We observe here that while the penalized methods usual perform better than the diagonal approach, they still get beaten by the benchmark which is the multi-dimensional BMEMA.

1.4 Power of log-likelihood ratio score

We have seen both in Section 1.2 and Section 1.3 that fitting the Ornstein-Uhlenbeck model to the exponential moving averages that we have defined gives mean-reversion parameters that are consistently above what is expected for data that does not present any mean-reversion properties. On the other side, comparing the log-likelihoods of the estimates to those of benchmark values gave results that did not show any significant improvement.

In this Section, we take data that we design to be mean-reverting and compare its estimation results to data that we design to have no mean-reversion. For data without mean-reversion, we take standard normal variables, of which we compute an exponential moving average. For data with mean-reversion, we use an auto-regressive process:

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon_t$$

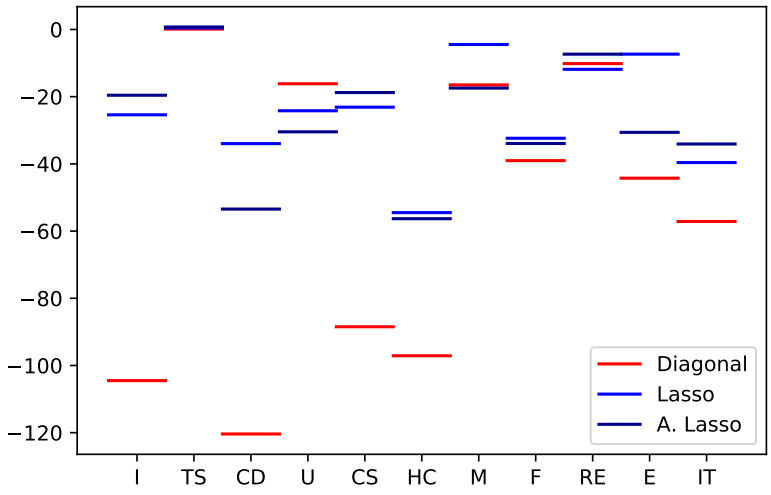
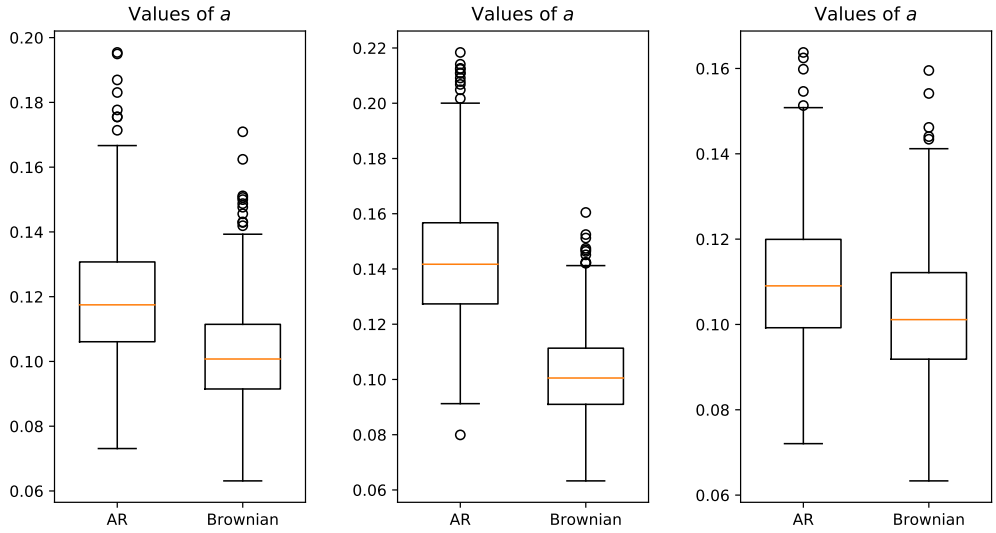


Figure IV.9 – Evaluation log-likelihood ratios of fitted multi-dimensional Ornstein-Uhlenbeck, by sector. The sectors are represented by their initials. The ratio is to the benchmark, and higher values are better. The markers represent the average value for a given estimator in a given sector, the average being over all intervals considered. The negative values for all estimators mean that no estimator beat the benchmark. However, we see that the penalized non-diagonal methods give a better result than the diagonal method as well as the non-penalized MLE (not shown, as its values are far below the scale). Comparisons between sectors have little sense, as the log-likelihood is not consistent when one changes datasets and dimensions.

where α is a vector of real values and for any t, $\epsilon_t \sim \mathcal{N}(0,1)$. By putting negative values in the vector α, we force mean-reversion on the data, as future returns will be negatively correlated with past returns.

In Figure IV.10 we show the distributions of a in the two scenarios, where we change the parameter α. In Figures IV.10a and IV.10b, we use an AR(1) process, and we put negative correlation at 0.1 and 0.2. For 10% negative correlation, our estimation is on average above what is expected from a BMEMA, but in many cases one could have difficulties to reject the uncorrelated model. For 20%, the difference becomes significant. In Figure IV.10c, we use a more complex AR(5) process, which puts positive correlation with the latest value and negative correlation with more distant data. We still observe a higher value of a, though a lower value than for the AR(1) processes.

In Figure IV.11 we show the distributions of the log-likelihood ratios for the same scenarios. Only for the AR(1) process with 20% negative correlation we observe a



(a) $\alpha = (-.1)$. (b) $\alpha = (-.2)$. (c) $\alpha = (.1, 0, 0, -.1, -.1)$.

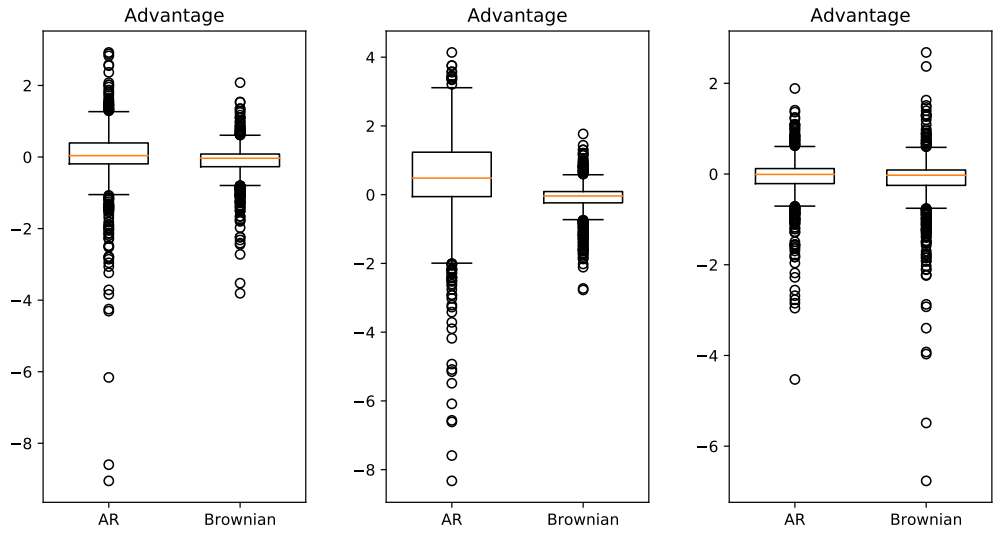
Figure IV.10 – Values of the mean-reversion parameter a for an EMA based on a AR process compared to an EMA based on a uncorrelated process. Positive values mean the estimator gives a value of higher likelihood than the benchmark, based on an independent set of data.

noticeable increase in mean log-likelihood over the benchmark. However, in all three cases, the typical values are small compared to the variations.

Based on our simulations, we assert that our method correctly gives higher values of a for data that is mean-reverting, but this is not followed with significant increases in log-likelihood obtained through cross-validation. In practice, we can correctly identify mean-reversion, but the cross-validation could still give a negative log-likelihood ratio. This shows that we can't rely on cross-validation log-likelihood ratios in order to decide on the existence of mean-reversion.

1.5 Conclusion

We have shown that the equally-weighted sector averages are consistent with a mean-reversion property in the short-term. The log-likelihood ratio validation couldn't confirm the significance of that result, but based on simulations involving mean-reverting processes we argue that the log-likelihood ratio criterium is exceedingly difficult to obtain. Multi-dimensional analysis... We have also shown that most stocks taken separately are behaving consistently with mean-reversion to the sector average in the short-term. In order to decide on the validity of the results that we present, we believe it would be best to define a trading strategy based on the predictions



(a) $\alpha = (-.1)$. (b) $\alpha = (-.2)$. (c) $\alpha = (.1, 0, 0, -.1, -.1)$.

Figure IV.11 – Values of log-likelihood ratios of the parameter estimated on the data and of the BMEMA benchmark. We expect 0 for the Brownian data.

that follow the model. Further work could also investigate varying parameters of the model. Changing the parameter of the exponential moving average would help in finding the time-scales on which mean-r version has the most impact. By varying the length of the sub-intervals one could see on which timescales are mean-reversion properties stationary. One could also consider data with shorter time-steps.

2 Dividend futures

If one equates a stock with the cash-flow of its future dividends, the price of a stock has to be equal to the expectation of discounted future dividends. Indeed, if the market expectations for future dividends didn't match the price of a stock, one could build based on this an arbitrage opportunity. The market expectations on dividend payoffs are observed in the prices of dividend futures. Based on the arbitrage argument, we believe that the dividend futures prices should revert to some equilibrium value, once we have controlled for stock price movements. Also, as time passes, the market's expectations on when will be dividends paid can vary, hence creating flows of futures prices between maturities. We wish to look at both phenomena and to see if we can find them in data. We take the example of the EURO STOXX 50 index, which isn't technically a stock but works as one in practice.

2.1 Presentation of the data

The EURO STOXX 50 is a composite index that tracks the total value of 50 Eurozone blue-chip companies. At present, the composition includes banks, energy companies, retail companies and others [Sto]. The composition of the index is updated, but in practical purposes the index can be seen as a stock. The prices that are used for the computation are adjusted for dividends, hence the resulting index should be seen as a stock price adjusted for dividends.

In a similar manner one can compute the weighted average of dividends paid by the components of the EURO STOXX 50 index. The resulting EURO STOXX 50 DVP Dividend Points (DVP) index, can be seen as the year-to-date total dividends paid by the EURO STOXX 50 index, when seen as a stock. One can then buy future contracts based on this underlying: the EURO STOXX 50 Index Dividend Futures [Eur], henceforth called simply dividend futures. These are traded at the Eurex Exchange, and are available for the next 10 yearly maturities.

For practical purposes, we can imagine that we have a stock, a dividend index and dividend futures contracts with several maturities. However, it is widely believed that dividend futures are not liquid for single stocks, which makes their analysis difficult in our framework. For this reason, we work only with index dividend futures, which are liquid enough at least for the shortest maturities.

2.2 Theoretical approach

Using a discounted cash flow approach, such as in the dividend discount model, the price of a stock is equal to the sum of the expected discounted future dividends. Let t be the time counted in years, S_t be the price of a dividend yielding stock and δ_t the instantaneous payout function of the stock, assumed integrable on any interval. The continuous payout of dividends is of course an approximation, as stocks pay out dividends at specific times. $D(t,s)$ the discount factor that is the value at time t of a unit of cash flow at time $s \geq t$. $D(t,s)$ is typically assumed decreasing in s and increasing in t. The discount factor corresponds to the risk-free rate between times t and s and is therefore typically computed from the yield curve. If $r_t^{(s-t)}$ is the yield at time t for a maturity of $s-t$, then we have $D(t,s) = e^{-r_t^{(s-t)}(s-t)}$. If one neglects the variations of the yield curve then $r_t^{(s-t)} = r^{(s-t)}$. Given all this, we write the stock price as a conditional expectation integral

$$S_t = \int_t^{+\infty} D(t,s)\mathbb{E}\left[\delta_s \mid \mathcal{F}_t\right] \mathrm{d}s$$

Assume that the discount factors are slowly varying, i.e. that the risk-free rates are small. Then we will write that for any integer n and any real $n \leq s \leq n+1$,

$D(t,s) \simeq D(t,n)$. From this:

$$S_t \simeq \int_t^{\lceil t \rceil} \mathbb{E}\left[\delta_s \mid \mathcal{F}_t\right] \mathrm{d}s + \sum_{n \geq \lceil t \rceil} D(t,n) \int_n^{n+1} \mathbb{E}\left[\delta_s \mid \mathcal{F}_t\right] \mathrm{d}s$$

where $F_t^{(n)} := \int_n^{n+1} \mathbb{E}\left[\delta_s \mid \mathcal{F}_t\right] \mathrm{d}s$ is the future price of the dividend paid between n and $n+1$, which is the strike such that the future has zero value at signing. Assuming $t \leq 1$, we continue with

$$S_t + \int_0^t \delta_s \mathrm{d}s \simeq \sum_{n \geq 0} D(t,n) F_t^{(n)}. \tag{2.2.1}$$

We use here that $F_t^{(0)}$, the future price of the dividend paid during the current year, can be written as $F_t^{(0)} = \int_0^t \delta_s \mathrm{d}s + \int_t^1 \mathbb{E}\left[\delta_s \mid \mathcal{F}_t\right] \mathrm{d}s$. Further, on the left-hand side we have a price that is adjusted for dividends. Indeed, one typically defines an adjusted price such that it coincides with the real price at the present time, hence subtracting dividends from past prices:

$$\tilde{S}_t^{(t_0)} := S_t - \int_t^{t_0} \delta_u \mathrm{d}u.$$

The adjusted price has therefore meaning up to a constant. By choosing in the formula above $t_0 = 0$, we get the LHS of Equation (2.2.1) and we see that in order to get an adjusted price we can instead add the dividend flow when rolling forward.

$$\tilde{S}_t := S_t + \int_0^t \delta_u \mathrm{d}u.$$

This gives a simple formula:

$$\tilde{S}_t \simeq \sum_{n \geq 0} D(t,n) F_t^{(n)}. \tag{2.2.2}$$

Let us explore now the implications of the equation above. First, the stock price is seen here as as a weighted sum of dividend futures. The variations of the market's expectations of dividend payouts in a future year have then a linear impact on the price of the stock. Conversely, the variations of the stock price should be accompanied with variations of dividend future prices. Each future can react differently, but one can predict linear responses to relatively small variations of the stock price. Second, the discount factors can change change with time, hence the stock price reflects beliefs on future dividends as well as the risk structure in the market. Third, taking constant stock price and discount factors, the equation shows that a variation of some dividend future price must be followed by at least one move of another dividend future in the

opposite direction - all dividend future prices can't move together or else we have theoretically an arbitrage opportunity. For that same reason, the variations of one future price are likely to have an impact on other future prices.

2.3 Analysis of dividend futures regressed on the stock price

In the following, we will be interested in analysing the properties of the dividend future prices. As mentioned above, we expect them to hold a linear erlation with the stock price, hence we will write:

$$F_t^{(i)} = \alpha^i S_t + \beta^i + X_t^i. \tag{2.3.1}$$

The process X is the object of our interest, as it corresponds to the variations of the dividend future price as if the stock price was constant. It is by definition a centered process and its variations are orthogonal to the market.

Before delving into the analysis of X, we present the results of the regression regarding the linearity coefficient α. We expect it to be positive, as increases in market price should be expected to be accompanied by increases rather than decreases of the dividend expectations. This is what we observe in data. Taking for example the 2014 prices of dividend futures with maturity 2016, and plotting them against the stock price, we obtain the result in Figure IV.12.

Figure IV.12 – 2014 prices of 2016 dividend futures, plotted against the index values.

In order for the model from Equation (2.3.1) to make sense with constant regression coefficients, we have to restrict the regressions to limited time-spans. Indeed,

we observe that the linearity breaks-down on time-spans of a few years. The futures having yearly maturities, the natural time-span to work with is therefore one year.

During one year, one future plays a very specific role, which is the future that is settled at the end of the year. Its value is less impacted by the beliefs of investors than by the observed dividend payouts throughout the year. With the end of the year approaching, its value converges to the settle price, with decreasing volatility. Its specificity would require a different approach, which is the reason we will systematically exclude it from our analysis.

Also, we have to add that we have restricted information on future prices. Only up to 10-year maturity futures are traded at the Eurex exchange, which means one can't compute in practice the sum in Equation (2.2.2). Even among the 10 futures, only the first are liquid enough to give sense to the quoted price. Finally, in practice we don't have access to all prices that have ever been quoted. Furthermore, our dataset has regularly missing prices. When our dataset doesn't cover at least 80% of the year, we exclude it. When some data points are missing, we complete the dataset with plausible values, through linear interpolation, where the linearity is with respect to trading days. This data completion has a possible impact on our result, hence we tested different completion methods and obtained similar results, which leads us to believe the sparse missing points have a limited impact on the results.

We start the presentation of regression results by showing the slopes values by year and by maturity. We present the results in Figure IV.13. We observe a common pattern of a rising curve with increased maturity, which leads to the conclusion that later maturity dividend futures are more impacted by stock prices. We interpret it as the market having a better idea of the dividends in the near future, hence being less inclined to change the previsions when stock market changes. Conversely, a change in stock price reflects more a change in dividend predictions in the long term than in the short term.

We continue then with the analysis of the future prices controlled by the stock price, denoted X above. We provide plots of future prices in Figure IV.14. We observe that the prices are typically moving together. The deviations from the mean are far smaller than the deviations of the mean from 0 (recall that through the regression each trajectory is centered). This is very visible in the prices from 2011, slightly less so in 2013. In the latter, we have a period of large variations, followed by a period when all prices cluster around the mean, itself followed by half a year of large variations. The second half of the data shows prices that are separated by approximately constant values. This suggests that working with shorter time-spans could lead to more relevant results, but on the other side this would prevent us from obtaining results such as for the 2011 prices.

We consider the series X^i as the components of a vector process, the dimension

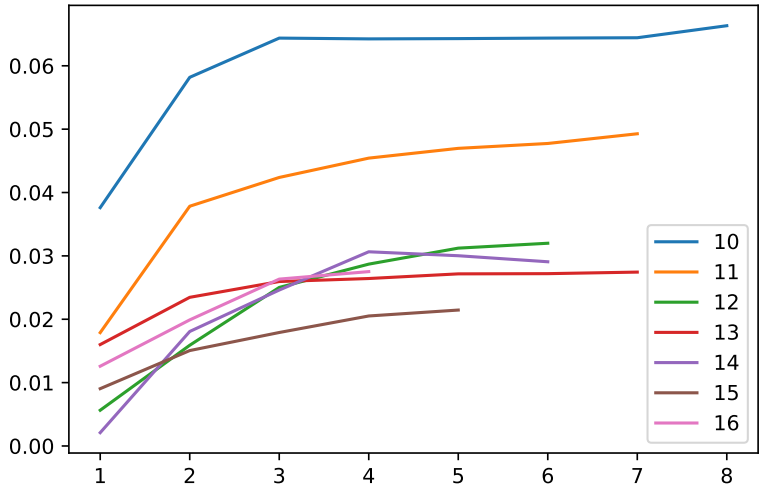
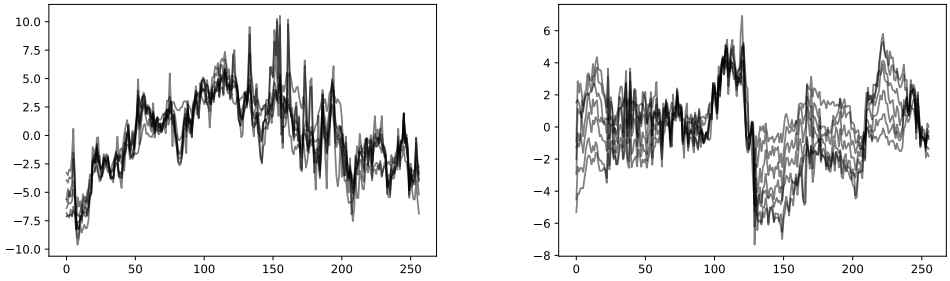


Figure IV.13 – Regression slopes α of dividend futures as a function of the stock price. Line k represents the slopes computed from the trading year $2000+k$. On the x-axis, are represented the maturities, in years calculated starting from the trading year. For example, Point 2 on line number 11 is the linear coefficient for the 2013 dividend future computed in 2011.



(a) Daily dividend future prices in 2011, controlled for the stock price, with maturities 2012-2018.

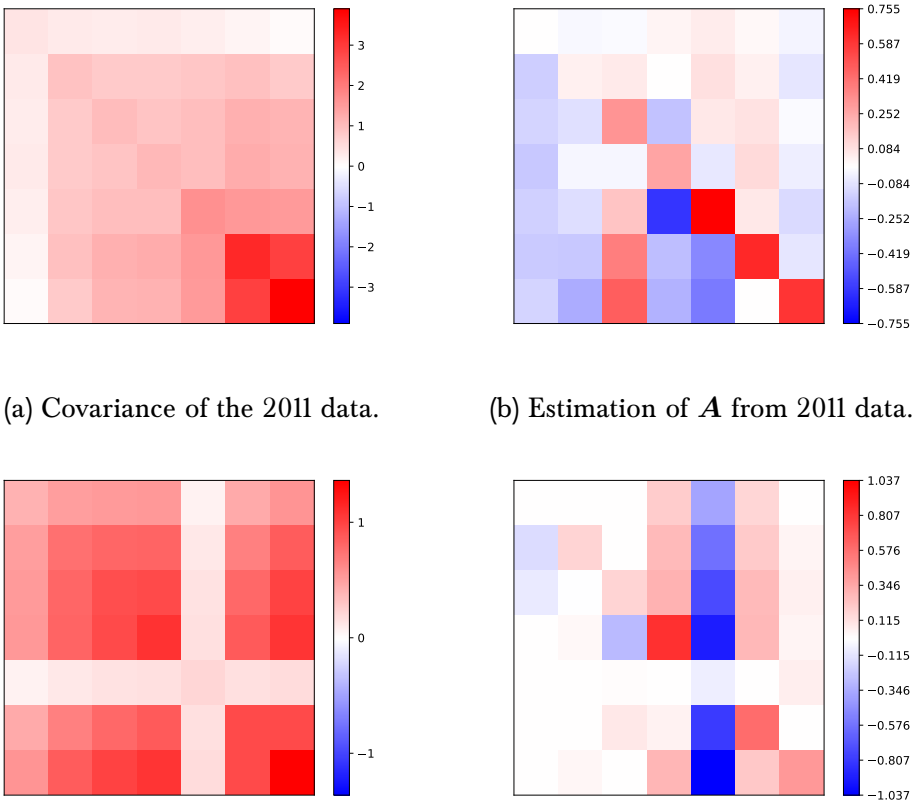
(b) Daily dividend future prices in 2013, controlled for the stock price, with maturities 2014-2020.

Figure IV.14 – Plots of the trajectories of the remainders X of the dividend future prices after regressing against the stock price. We observe a common movement around a mean that explains most of the movements, although this is more valid for 2011 data than for 2013. The x axis is expressed in trading days.

being the number of available maturities subtracted one, as we exclude the shortest maturity. We then fit to that trajectory an Ornstein-Uhlenbeck model, governed by the stochastic differential equation

$$\mathrm{d}X_t = -\boldsymbol{A}X_t\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t.$$

We estimate $\boldsymbol{\Sigma}$ using the quadratic variations, then we estimate $\boldsymbol{A}$ by minimizing the negative log-likelihood, penalized in the case of the Lasso and Adaptive Lasso estimators. We present some example results on Figure IV.15. We observe that although some diagonal terms are important, the results are much less diagonal-heavy than for instance what we had in Section 1, Figure IV.6.



(a) Covariance of the 2011 data.

(b) Estimation of $\boldsymbol{A}$ from 2011 data.

(c) Covariance of the 2013 data.

(d) Estimation of $\boldsymbol{A}$ from 2013 data.

Figure IV.15 – Covariances and drift estimations for data presented in Figure IV.14. The estimator here is the Adaptive Lasso. In both cases, the results were close to the MLE estimator, but the penalisation provides some moderate sparsity.

We continue by validating our approach by testing if our method can predict variations of the process in future dates. For this, we divide each year-long set into a first part, representing 80% of the data, as a training set that we use to estimate the parameters, and a second part, the validation set that we use to verify the ability to predict variations. Our natural benchmark is the trivial benchmark $\boldsymbol{A} = 0$. Our metric is the negative log-likelihood ratio of the estimator and the benchmark. Negative values mean we have beaten the benchmark. In Figure IV.16, we present the results. The best estimator in this experiment is the Lasso estimator, which beats the baseline in out of 7 years, and is equivalent to the benchmark in one more year. While this is not necessarily a very strong result, it leads us to believe that some fine-tuning may lead to more significant results.

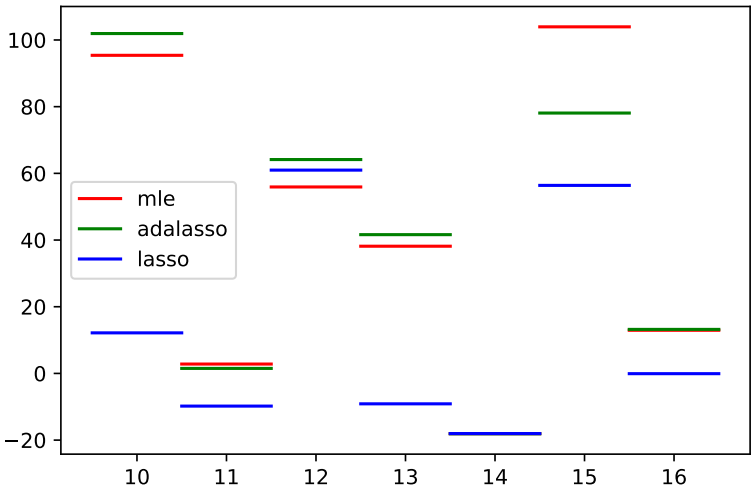


Figure IV.16 – Log-likelihood based evaluation of estimation methods.

2.4 Analysis of price movements around the average

As we have already pointed out above, the series X typically gather around their mean value. This suggests a two-step approach. First, we will analyze the mean-return properties of the mean. Second, we will look into the way prices move around that mean.

For the first point, we fit the mean series to Ornstein-Uhlenbeck models. In Figure IV.17 we plot an example of such a trajectory alongside simulated data that use the parameters fitted to the data, which gives a visual validation of our approach.

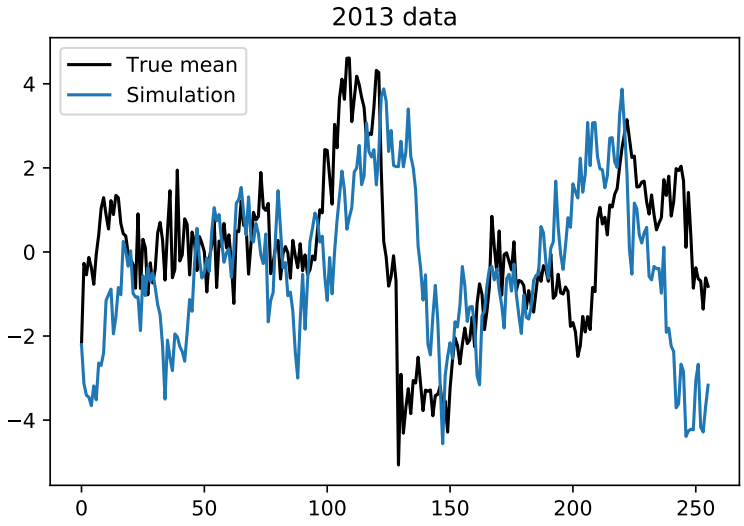
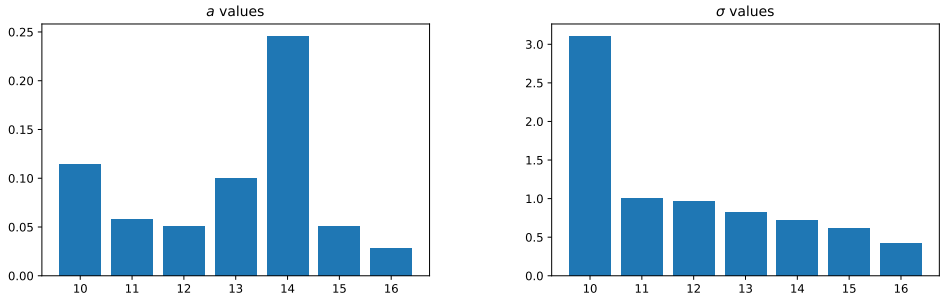


Figure IV.17 – Dividend future price mean compared to sample trajectory.

The resulting values of the one-dimensional Ornstein-Uhlenbeck parameters a and σ are presented in Figure IV.18 for each year. The values of σ are falling from year to year, with a big step down between 2010 and 2011. The shape is difficult to interpret, as the series is an average of remainders of a regression procedure. The values of a are consistently positive, with typical values in the $0.05 - 0.1$ range. The time-step in this analysis is one day, which suggests mean-return at the time-scale of the order of 10 to 20 days. 2014 is here an outlier, on which we will comment more broadly below.

Again, we test the validity of our approach through the same procedure as above. We provide our results in Figure IV.19a. We observe a negative value on all years except 2014 and 2016, hence beating the benchmark. The huge value for 2014 forces us to give an explanation. We plot the mean in Figure IV.19b, where we also mark the cutoff between the first 80 and the last 20 percent. The estimation of the parameters is done on the first part, while the evaluation on the second one. The fact that the data is close to 0 in the first part suggests a high value of a, which ewe observe in Figure IV.18, but this is incompatible with values away from 0 in the second part. This leads to very high values for the negative log-likelihood. This suggests again that one could work with shorter time-spans.

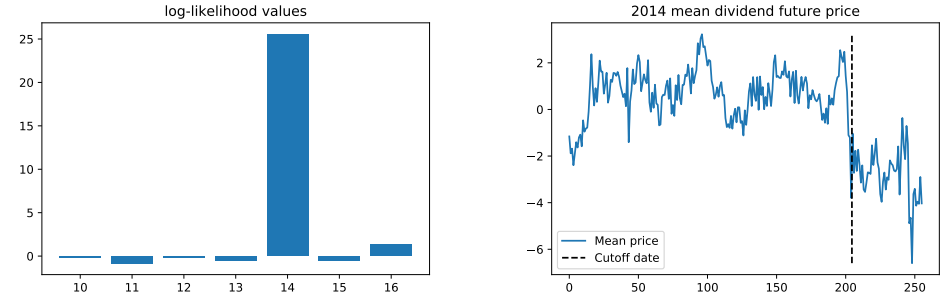
Going to the second step, we subtract the mean from the processes X. As in Section 1.3.2, we create a linear dependence, which we address again by excluding the longest maturity from the analysis. The resulting processes are shown in Figure IV.20. Visually, we observe processes that are closer to being independent.



(a) Mean-return drift coefficients for the mean price of dividend futures, by year.

(b) Volatility parameter of the mean price of dividend futures, by year.

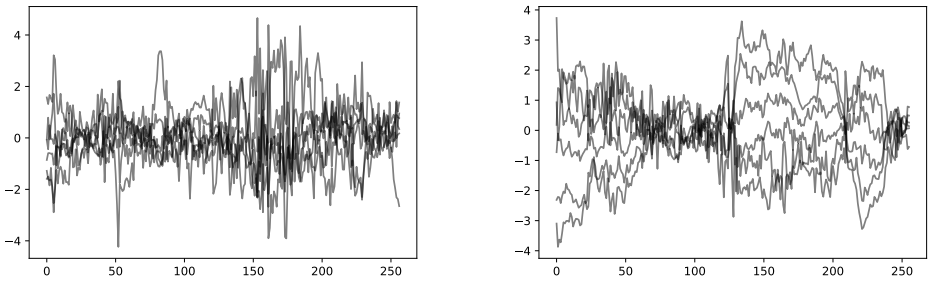
Figure IV.18 – Parameters of the one-dimensional Ornstein-Uhlenbeck processes fitted to the mean price processes, by year.



(a) Log-likelihoods obtained after independent evaluation, by observation year.

(b) Mean price of dividend futures in 2014 and training/evaluation cutoff.

Figure IV.19 – Evaluation of the validity of the estimation using negative log-likelihoods. The 2014 spike in log-likelihood is explained by the spurious coincidence of a drop in dividend values and the date chosen for separating the training and evaluation sets.



(a) Daily stock-controlled dividend future price deviations from the mean in 2011.

(b) Daily stock-controlled dividend future price deviations from the mean in 2013.

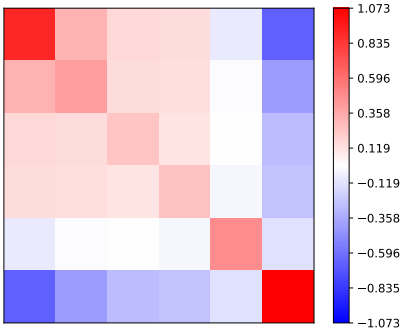
Figure IV.20 – Plots of the deviations from the mean of dividend future prices. The images agree better with a usual Ornstein-Uhlenbeck process.

We present example results, for 2011 and 2013, in Figure IV.21. The results are drastically different from those in Figure IV.15, which shows that our two-step approach is interesting. The matrix results we present are more diagonal-heavy, which suggests that every single price deviates from the mean independently of the others.

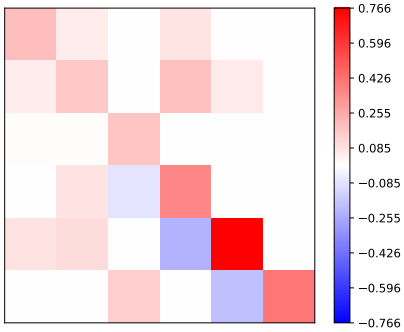
We conclude this analysis by presenting the validation using the log-likelihood ratios computed on independent intervals. The results in Figure IV.22 show that the Lasso has again the best performance. The Lasso beats the benchmark only in 2 out of 7 years and matches the benchmark in three years. This is again a relatively weak result, but it might be improved after fine-tuning. Importantly, we see that in 2014 we beat the benchmark despite the huge error for the mean. This suggests again that the mean-return properties of the average can be analyzed independently of the individual prices.

2.5 Conclusion

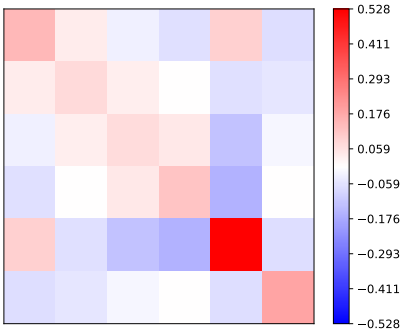
Using a discounted cash flow approach, we show that a stock price and its dividend futures are linked through a linear relation. The relation imposes that all maturity dividend future prices cannot move in the same direction, once controlled for stock price variations. However, this is exactly what we observe: dividend future prices for all maturities typically move together and are gathered around a slowly mean-reverting average. While market inefficiency may be an explanation, we formulate two others. First, we may be impacted by the limited number of maturities that are traded, as the variations of short maturities may be offset by opposite variations of non-observed long maturities. Second, other factors, such as varying discount functions, can be the reason for variation of the stock price that we mistakenly attribute to the dividends. The mean-reversion of the mean appears to be significant, but situations



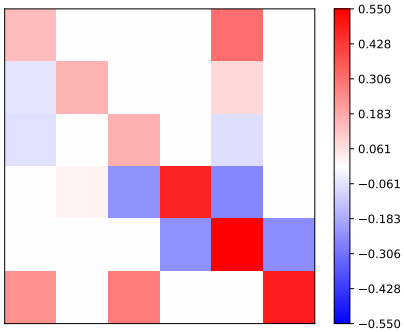
(a) Covariance of the 2011 data.



(b) Adaptive Lasso estimation of $\boldsymbol{A}$ from 2011 data.



(c) Covariance of the 2013 data.



(d) Adaptive Lasso estimation of $\boldsymbol{A}$ from 2013 data.

Figure IV.21 – Covariances and drift estimations for data presented in Figure IV.20. As in Figure IV.14, the results were close to the MLE estimators, but the penalisation provides some moderate sparsity.

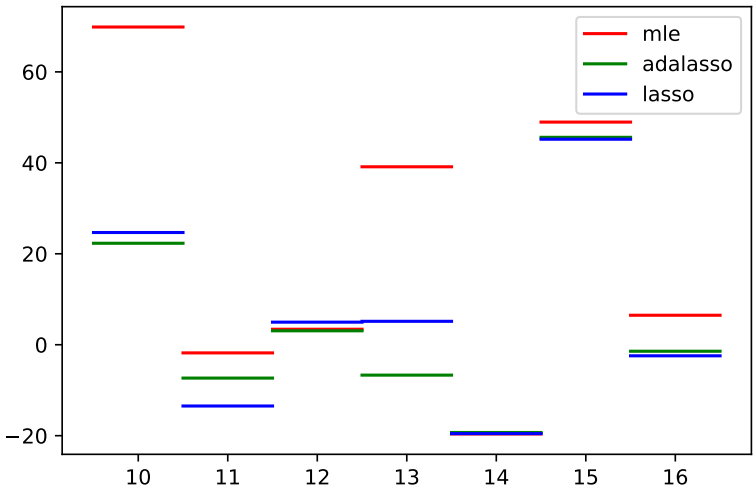


Figure IV.22 – Log-likelihood based evaluation of estimation methods.

such as for the 2014 data show that a robust approach would need to consider various time-spans for estimation. Then, variations around the mean of individual prices seem to fit a mean-reverting pattern, but our multi-dimensional approach gives mixed results in terms of variations prediction.

Our analysis notably excludes the shortest maturity, because of how different it is from other dividend futures. However, the shortest maturity could possibly bring interesting information that could help us better understand the other dividend futures. Including that shortest maturity future in the analysis is a direction for future research, as it could also help in providing a unified approach that could also be applied across different years.

Bibliography

[AC10] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[ACM12] H. Amini, R. Cont, and A. Minca. Stress testing the resilience of financial networks. *International Journal of Theoretical and Applied Finance*, 15(01):1250006, 2012.

[AHH+16] L. Arciero, R. Heijmans, R. Heuver, M. Massarenti, C. Picillo, and F. Vacirca. How to Measure the Unsecured Money Market: The Eurosystem's Implementation and Validation Using TARGET2 Data. *International Journal of Central Banking*, 12(1):247–280, March 2016.

[AKS11] G. Afonso, A. Kovner, and A. Schoar. Stressed, not frozen: The federal funds market in the financial crisis. *The Journal of Finance*, 66(4):1109–1139, 2011.

[AKS14] G. Afonso, A. Kovner, and A. Schoar. Trading partners in the interbank lending market. Staff Reports 620, Federal Reserve Bank of New York, 2014.

[AlB02] R. Albert and A. lászló Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys*, 2002.

[AŁL+16] A. Anagnostopoulos, J. Łącki, S. Lattanzi, S. Leonardi, and M. Mahdian. Community detection on evolving graphs. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3522–3530. Curran Associates, Inc., 2016.

[ANP09] P. Angelini, A. Nobili, and M. C. Picillo. The interbank market after August 2007: what has changed, and why? Temi di discussione (Economic working papers) 731, Bank of Italy, Economic Research and International Relations Area, October 2009.

[AS13] Y. D. Akarim and S. Sevim. The impact of mean reversion model on portfolio investment strategies: Empirical evidence from emerging markets. *Economic Modelling*, 31(C):453–459, 2013.

[ASM04] Y. Ait-Sahalia and P. A. Mykland. Estimators of diffusions with randomly spaced discrete observations: A general theory. *Ann. Statist.*, 32(5):2186–2222, 10 2004.

[Aza89] J.-M. Azaïs. Approximation des trajectoires et temps local des diffusions. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 25(2):175–194, 1989.

[Bar99] Barabási, Albert-László and Albert, Réka. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[BBvL15] F. Blasques, F. Bräuning, and I. van Lelyveld. A dynamic network model of the unsecured interbank lending market. BIS Working Papers 491, Bank for International Settlements, February 2015.

[BFG+08] P. Borgnat, E. Fleury, J.-L. Guillaume, C. Magnien, C. Robardet, and A. Scherrer. Evolving networks. In *Mining Massive Data Sets for Security*, volume 19, pages 198–203. IOS Press, 2008.

[BKMU12] B. Bahmani, R. Kumar, M. Mahdian, and E. Upfal. Pagerank on an evolving graph. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 24–32, New York, NY, USA, 2012. ACM.

[Bla14] P. G. Blackwell. *Ornstein–Uhlenbeck Process: Overview.* John Wiley & Sons, Ltd, 2014.

[BLT16] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. Submitted to the Annals of Statistics, 05 2016.

[BM15] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 08 2015.

[Bol01] B. Bollobás. *Random Graphs.* Cambridge Studies in Advanced Mathematics, 2nd edition, 2001.

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

[BSW11] F. Bunea, Y. She, and M. H. Wegkamp. Optimal selectionof reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[BSW12] F. Bunea, Y. She, and M. H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388, 10 2012.

[BTW07] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, 1:169–194, 2007.

[BvdG11] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[BW06] R. Balvers and Y. Wu. Momentum and mean reversion across national equity markets. *Journal of Empirical Finance*, 13(1):24–48, 2006.

[CF06] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), June 2006.

[CFS05] G. Carcano, P. Falbo, and S. Stefani. Speculative trading in mean reverting markets. *European Journal of Operational Research*, 163(1):132–144, 2005.

[CFS15] R. Carmona, J.-P. Fouque, and L.-H. Sun. Mean field games and systemic risk. *Communications in Mathematical Sciences*, 13(4):911–933, 2015.

[CG07] P. Cattiaux and A. Guillin. Deviation bounds for additive functionals of markov processes. *ESAIM: Probability and Statistics*, 12:12–29, 11 2007.

[CIR85] J. C. Cox, J. E. Ingersoll, and S. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.

[CL93] T. Q. Cook and R. K. LaRoche, editors. *Instruments of the money market*. Federal Reserve Bank of Richmond, 1993.

[CL02] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[CLM97] J. Campbell, A. Lo, and A. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[CZ16] J. Chen and W. Zhang. The right way to search evolving graphs. In *Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International*. IEEE, 2016.

[CZS+07] Y. Chi, S. Zhu, X. Song, J. Tatemura, and B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 163–172, New York, NY, USA, 2007. ACM.

[DBT85] W. F. M. De Bondt and R. Thaler. Does the stock market overreact? *Journal of Finance*, 40(3):793–805, 1985.

[Doh87] G. Dohnal. On estimating the diffusion coefficient. *Journal of Applied Probability*, 24(1):105–114, 1987.

[EG60] P. Erdős and T. Gallai. Gráfok előírt fokszámú pontokkal. *Matematikai Lapok*, 11:264–274, 1960.

[ER59] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959 1959.

[Eur] Eurex. Eurex dividend derivatives. https://www.eurexchange.com/blob/830906/5bb5239d4f7775c2a9d41f1c1391251c/data/presentation_dividend_derivatives.pdf.pdf.

[Fam98] E. F. Fama. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, 49(3):283 – 306, 1998.

[FBO12] J. M. Francis Bach, Rodolphe Jenatton and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[FF88] E. F. Fama and K. R. French. Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273, 1988.

[FI13] J. Fouque and T. Ichiba. Stability in a model of interbank lending. *SIAM J. Financial Math.*, 4(1):784–803, 2013.

[Fin] Financial Times. Index constituents. https://markets.ft.com/data/indices/tearsheet/constituents?s=INX:IOM.

[FL13] J. Fouque and J. Langsam, editors. *Handbook on Systemic Risk*. Cambridge University Press, 2013.

[Flo84] D. Florens-Zmirou. Statistics on crossings of discretized diffusions and local time. *Stochastic Processes and their Applications*, 39(1):139 – 151, 1984.

[Flo86] D. Florens-Zmirou. Estimation de paramètres d'une diffusion à partir des temps de passages en zéro. *Comptes Rendus de l'Académie des Sciences Paris*, 303(13):655 – 658, 1986.

[Flo87] D. Florens-Zmirou. Estimation du paramètre d'une diffusion par les changements de signe de sa discrétisée. *Comptes Rendus de l'Académie des Sciences Paris*, 305:661 – 664, 1987.

[Flo88] D. Florens-Zmirou. Statistiques de diffusions et temps local. *Annales de l'I.H.P. Probabilités et statistiques*, 24(1):99 – 130, 1988.

[Flo89] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.

[Flo91] D. Florens-Zmirou. Statistics on crossings of discretized diffusions and local time. *Stochastic Processes and their Applications*, 39(1):139 – 151, 1991.

[Flo93] D. Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability*, 30(4):pp. 790–804, 1993.

[GCJ93] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 29(1):119–151, 1993.

[GCJ94] V. Genon-Catalot and J. Jacod. Estimation of the diffusion coefficient for diffusion processes: Random sampling. *Scandinavian Journal of Statistics*, 21(3):193–221, 1994.

[GG14] S. Gabrieli and C.-P. Georg. A network view on interbank market freezes. Technical report, Banque de France, November 2014.

[Gir14a] C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, December 2014.

[Gir14b] C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[Gob02] E. Gobet. Propriété LAN pour les diffusions ergodiques avec observations discrètes. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(5):711–737, 2002.

[GR96] P. Gangopadhyay and M. R. Reinganum. Interpreting mean reversion in stock returns. *The Quarterly Review of Economics and Finance*, 36(3):377–394, 1996.

[Gro75] L. Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[Gro04] J. Gropp. Mean reversion of industry stock returns in the u.s., 1926-1998. *Journal of Empirical Finance*, 11(4):537–551, 2004.

[GS16] E. Gobet and Q. She. Perturbation of Ornstein-Uhlenbeck stationary distributions: expansion and simulation. working paper or preprint, July 2016.

[GSV15] S. Gabrieli, D. Salakhova, and G. Vuillemey. Cross-border interbank contagion in the european banking sector. Document de travail 545, Banque de France, March 2015.

[Har69] F. Harary. *Graph Theory*. Addison-Wesley Series in Mathematics. Addison Wesley, 1969.

[HHMS93] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating gaussian diffusions. *The Annals of Applied Probability*, 3(3):897–913, 1993.

[HJ86] R. A. Horn and C. R. Johnson, editors. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986.

[HKW16] T. Hartmann, A. Kappes, and D. Wagner. Clustering evolving networks. In *Algorithm Engineering*, 2016.

[Hof99] M. Hoffmann. Adaptive estimation in diffusion processes. *Stochastic Processes and their Applications*, 79(1):135 – 163, 1999.

[HS12] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97 – 125, 2012. Temporal Networks.

[HS13] P. Holme and J. Saramäki, editors. *Temporal Networks*. Springer-Verlag Berlin Heidelberg, 2013.

[Hul09] J. C. Hull. *Options, Futures and Other Derivatives*. Options, Futures and Other Derivatives. Pearson/Prentice Hall, 2009.

[IP17] J. Idier and T. Piquard. Pandemic crises in financial systems: a simulation-model to complement stress-testing frameworks. Working Paper 621, Banque de France, January 2017.

[IUY09] S. M. Iacus, M. Uchida, and N. Yoshida. Parametric estimation for partially hidden diffusion processes sampled at discrete times. *Stochastic Processes and their Applications*, 119(5):1580 – 1600, 2009.

[Jac01] J. Jacod. Inference for stochastic processes. Prépublications du laboratoire de probabilités et modèles aléatoires, 2001.

[Jan97] S. Janson. *Gaussian Hilbert Spaces*. Cambridge Tracts in Mathematics. Cambridge University Press, 1997.

[Kes97] M. Kessler. Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997.

[KLN02] J. Kang, M.-H. Liu, and S. X. Ni. Contrarian and momentum strategies in the china stock market: 1993-2000. *Pacific-Basin Finance Journal*, 10(3):243–265, 2002.

[KLS12] M. Kessler, A. Lindner, and M. Sorensen. *Statistical Methods for Stochastic Differential Equations*, volume 124 of *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Taylor & Francis, 2012.

[Kos09] V. Kostakos. Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007–1023, 2009.

[KS91] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991.

[KS99] M. Kessler and M. Sørensen. Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299–314, 04 1999.

[KT15a] O. Klopp and A. B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.

[KT15b] O. Klopp and A. B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.

[Kut04] Y. A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. Springer, 2004.

[LB76] A. Le Breton. On continuous and discrete sampling for parameter estimation in diffusion type processes. In R.-B. Wets, editor, *Stochastic Systems: Modeling, Identification and Optimization, I*, volume 5 of *Mathematical Programming Studies*, pages 124–144. Springer Berlin Heidelberg, 1976.

[LB77] A. Le Breton. *Parameter Estimation in a Linear Stochastic Differential Equation*, pages 353–366. Springer Netherlands, Dordrecht, 1977.

[LCK+10] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042, March 2010.

[Leh90] B. N. Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105(1):1–28, 1990.

[LL09] C. Laing and G. J. Lord, editors. *Stochastic Methods in Neuroscience*. Oxford University Press, 2009.

[LL15] T. Leung and X. Li. Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance*, 18(03):1550020, 2015.

[MSC] MSCI. Gics. https://www.msci.com/gics.

[Muk11] S. Mukherji. Are stock returns still mean-reverting? *Review of Financial Economics*, 20(1):22–27, 2011.

[MW91] B. D. McKay and N. C. Wormald. Asymptotic enumeration by degree sequence of graphs with degrees $o(\sqrt{n})$. *Combinatorica*, 11(4):369–382, 1991.

[MX07] M. Mahdian and Y. Xu. *Stochastic Kronecker Graphs*, pages 179–186. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[Ped95] A. R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71, 1995.

[PS88] J. Poterba and L. Summers. Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics*, 22(1):27–59, 1988.

[RW87] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes and Martingales, vol 2: Ito calculus*. John Wiley, 1987.

[RY99] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg, 1999.

[SBvdH12] L. Spierdijk, J. A. Bikker, and P. van den Hoek. Mean reversion in international stock markets: An empirical analysis of the 20th century. *Journal of International Money and Finance*, 31(2):228–249, 2012.

[SC10] E. Santos and R. Cont. The brazilian interbank network structure and systemic risk. Working Papers Series 219, Central Bank of Brazil, Research Department, 2010.

[SC+16] W. Su, E. Candes, et al. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[Sok13] A. Sokol. *On martingales, causality, identifiability and model selection.* PhD thesis, University of Copenhagen, November 2013.

[Sor97] M. Sorensen. *Estimating Functions for Discretely Observed Diffusions: A Review*, volume 32 of *Lecture Notes–Monograph Series*, pages 305–326. Institute of Mathematical Statistics, Hayward, CA, 1997.

[SSS05] Q. Shen, A. C. Szakmary, and S. Sharma. Momentum and contrarian strategies in international stock markets: Further evidence. *Journal of Multinational Financial Management*, 15(3):235–255, 2005.

[Sto] Stoxx. Euro stoxx 50 components. https://www.stoxx.com/index-details?symbol=sx5e.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York, 2008.

[UO30] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.

[Vas77] O. Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2):177–188, 1977.

[vL07] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007.

[vZ00] H. van Zanten. A multivariate central limit theorem for continuous local martingales. *Statistics and Probability Letters*, 50(3):229 - 235, 2000.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 06 1998.

[Yah] Yahoo. Yahoo finance. https://finance.yahoo.com.

[YL06] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[Yos92] N. Yoshida. Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis*, 41(2):220 – 242, 1992.

[Zou06] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[ZWJ14] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 921–948, 2014.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Inférence statistique de processus d'Ornstein-Uhlenbeck: génération de graphes stochastiques, sparsité, applications en finance

Mots clefs : Processus stochastiques, Inférence statistique, Sparsité, Réseaux évolutifs, Données financières

Résumé : Le sujet de cette thèse est l'inférence statistique de processus d'Ornstein-Uhlenbeck multi-dimensionnels. Dans une première partie, nous introduisons un modèle de graphes stochastiques définis comme observations binaires de trajectoires. Nous montrons alors qu'il est possible de déduire la dynamique de la trajectoire sous-jacente à partir des observations binaires. Pour ceci, nous construisons des statistiques à partir du graphe et montrons de nouvelles propriétés de convergence dans le cadre d'une observation en temps long et en haute fréquence. Nous analysons aussi les propriétés des graphes stochastiques du point de vue des réseaux évolutifs. Dans une deuxième partie, nous travaillons sous l'hypothèse d'information complète et en temps continu et ajoutons une hypothèse de sparsité concernant le paramètre de *drift* du processus d'Ornstein-Uhlenbeck. Nous montrons alors des propriétés d'oracle pointues de l'estimateur Lasso, prouvons une borne inférieure sur l'erreur d'estimation au sens minimax et démontrons des propriétés d'optimalité asymptotique de l'estimateur Lasso Adaptatif. Nous appliquons ensuite ces méthodes pour estimer la vitesse de retour à la moyenne des retours journaliers d'actions américaines ainsi que des prix de futures de dividendes pour l'indice EURO STOXX 50.

Title : Statistical inference of Ornstein-Uhlenbeck processes: generation of stochastic graphs, sparsity, applications in finance

Keywords : stochastic processes, statistical inference, sparsity, evolving networks, financial data

Abstract : The subject if this thesis is the statistical inference of multi-dimensional Ornstein-Uhlenbeck processes. In a first part, we introduce a model of stochastic graphs, defined as binary observations of a trajectory. We show then that it is possible to retrieve the dynamic of the underlying trajectory from the binary observations. For this, we build statistics of the stochastic graph and prove new results on their convergence in the long-time, high-frequency setting. We also analyse the properties of the stochastic graph from the point of view of evolving networks. In a second part, we work in the setting of complete information and continuous time. We add then a sparsity assumption applied to the drift matrix coefficient of the Ornstein-Uhlenbeck process. We prove sharp oracle inequalities for the Lasso estimator, construct a lower bound on the estimation error for sparse estimators and show optimality properties of the Adaptive Lasso estimator. Then, we apply the methods to estimate mean-return properties of real-world financial datasets: daily returns of SP500 components and EURO STOXX 50 Dividend Future prices.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France