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

Stéphane Gaïffas*, Gustaw Matulewicz†

June 26, 2017

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

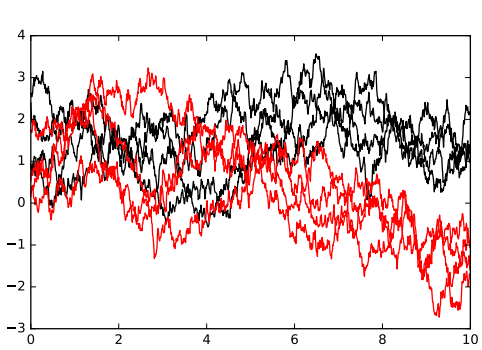
MSC 2010. 60G15; 62H12; 62M99

1 Introduction

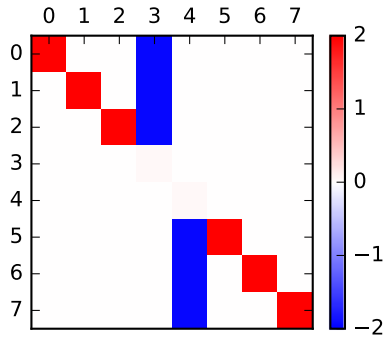
The Ornstein-Uhlenbeck, also called mean-reverting diffusion process, describes a process which evolves following a deterministic linear part with an added Gaussian noise, similarly to a vector-autoregressive process in discrete time. This model is ubiquitous in quantitative finance, for instance the 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 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.

*CMAP, Ecole Polytechnique and CNRS, Université Paris Saclay, Route de Saclay, 91128 Palaiseau cedex, France. Email: stephane.gaiffas@polytechnique.edu.

†CMAP, Ecole Polytechnique and CNRS, Université Paris Saclay, Route de Saclay, 91128 Palaiseau cedex, France. Email: gustaw.matulewicz@polytechnique.edu. This work was funded jointly by *Chaire Risques Financiers* of the *Risk Foundation*, the *Finance for Energy Market Research Centre*, the Natixis Foundation for Quantitative Research, and the Data Science Initiative of Ecole polytechnique



(a) Sample trajectory in 8 dimensions.



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

Figure 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.

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, Gir14a, 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, Gir14b] 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

$$\mathrm{d}X_t = -\boldsymbol{A}_0 X_t \mathrm{d}t + \mathrm{d}W_t, \quad \text{for any } t \geq 0, \tag{1.1}$$

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.

(H2) 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 (H2) 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 maximum-likelihood estimator (MLE) is given as the argument minimum of the following negative 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

$$\widehat{\boldsymbol{A}}_{MLE} := -\left(\int_0^T \mathrm{d}X_t X_t^\top\right)\left(\int_0^T X_t X_t^\top \mathrm{d}t\right)^{-1}. \tag{1.2}$$

The inverse exists almost surely as the integral is almost surely a symmetric positive definite matrix (see Section 2 for more details). The asymptotic normality of this MLE is well-know, indeed we have

$$\sqrt{T}\left(\mathrm{vec}\,\widehat{\boldsymbol{A}}_{MLE} - \mathrm{vec}\,\boldsymbol{A}_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{-1} \otimes \boldsymbol{I}\right), \tag{1.3}$$

see [Jac01], where $\xrightarrow{\mathcal{L}}$ stands for convergence in distribution, $\boldsymbol{I}$ 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 $\langle \mathrm{vec}\, t\boldsymbol{\varepsilon}_t\rangle = t\widehat{\boldsymbol{C}}_t \otimes \boldsymbol{I}$. 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

$$\mathcal{L}_T(\boldsymbol{A}) = \mathrm{tr}\,\boldsymbol{A}^\top \boldsymbol{\varepsilon}_T + \frac{1}{2}(\boldsymbol{A}-\boldsymbol{A}_0)\widehat{\boldsymbol{C}}_T(\boldsymbol{A}-\boldsymbol{A}_0)^\top - \frac{1}{2}\boldsymbol{A}_0\widehat{\boldsymbol{C}}_T\boldsymbol{A}_0^\top. \tag{1.4}$$

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 $\langle \boldsymbol{U}, \boldsymbol{V}\rangle_F = \mathrm{tr}\,\boldsymbol{U}^\top\boldsymbol{V}$, where $\mathrm{tr}\,\boldsymbol{M}$ 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 $\mathrm{diag}\,(\boldsymbol{A})$ stands for the vector formed by the diagonal of $\boldsymbol{A}$. We also denote by $\mathrm{supp}(x)$ 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, $\mathrm{Sp}\,\boldsymbol{A}$ is the spectrum of $\boldsymbol{A}$ and $\mathrm{diag}\,\boldsymbol{A}$ 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 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 2. We conclude the section with Theorem 3 which is an interesting by-product of the proof of Theorem 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 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:

$$\widehat{\boldsymbol{A}} = \underset{\boldsymbol{A}\in\mathbb{R}^{d\times d}}{\arg\min}\, \mathcal{L}_T(\boldsymbol{A}) + \lambda\|\boldsymbol{A}\|_1.$$

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 (1.4). 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 $1 \le s \le d$ and express the sparsity of $\boldsymbol{A}_0$ in the following assumption:

(H3) The true parameter is row-s-sparse, i.e.

$$\|\boldsymbol{A}_0^{i,\bullet}\|_0 \le s \;\text{ for all } i = 1,\dots,d,$$

where $\boldsymbol{A}^{i,\bullet}$ stands for the vector such that for any $j \le d$, $(\boldsymbol{A}^{i,\bullet})^j = \boldsymbol{A}^{ij}$ 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 $\boldsymbol{C}_\infty$.

(H4) There exists a non-decreasing function H, positive on $\mathbb{R}^+$, such that for any vector u verifying $\|u\|_2 \le 1$, we have:

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

We actually prove this assumption in the case where $\boldsymbol{A}_0$ is symmetric, see Theorem 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 (H2) 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 1. *Assume (H1) – (H4). Set $\gamma > 1$, $0 \le \tau < \gamma - 1$, $\epsilon_0 \in (0,1)$ and define*

$$\lambda_T := \gamma\sqrt{\frac{4e|\hat{\delta}_T|_\infty}{T}\left(\frac{1}{2}\log\frac{2\pi^2 d^2}{3\epsilon_0} + \log(2 + |\log(T\hat{\delta}_T)|_\infty)\right)} \tag{2.1}$$

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*

$$2\tau\gamma^{-1}\lambda_T \|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 \leq \|(\boldsymbol{A} - \boldsymbol{A}_0)X\|_{L^2}^2 + \left(\frac{1+\gamma+\tau}{\gamma\kappa} \right)^2 \lambda_T^2 ds \quad (2.2)$$

with probability at least $1 - \epsilon_0$.

The proof of Theorem 1 is detailed in Section 6.2 below. It relies on a Restricted Eigenvalue property, see Theorem 3 below, which we prove using Assumptions (H1)–(H4), as well as on a deviation property, see Theorem 8 from Section 6.6 below. Theorem 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 1 uses Equation (6.5) 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 1.

The convergence rate obtained in Theorem 1 almost matches the minimax lower bound provided in Theorem 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 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 1. *With the same assumptions and notation as in Theorem 1, the following holds with a probability larger than* $1 - \epsilon_0$*:*

1. *for the empirical norm:*

$$\|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2} \leq \frac{1+\gamma}{\gamma\kappa} \lambda_T \sqrt{ds} \quad (2.3)$$

2. *for the* ℓ^1 *norm, with* $\tau > 0$*:*

$$\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_1 \leq \frac{(1+\tau+\gamma)^2}{2\gamma\tau\kappa^2} \lambda_T ds \quad (2.4)$$

3. for the Frobenius norm:

$$\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F \leq \frac{1+\gamma}{\gamma\kappa^2}\lambda_T\sqrt{ds} \tag{2.5}$$

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 (2.2), and are proved in Section 6.2. The next Theorem is a minimax lower bound over row-sparse matrices, for the considered model.

Theorem 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 Theorem 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 1. Theorem 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 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 1. Then, for any $T \geq T_0$, we have*

$$\mathbb{P}\left[\inf_{u\in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa\right] \geq 1 - \frac{\epsilon_0}{2}. \tag{2.6}$$

The proof of Theorem 3 is given in Section 6.4 and uses explicitly Assumption (H4), which is proved in Theorem 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 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 (2.6) 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 (1.3). 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

$$\widehat{\boldsymbol{A}}_{ad.} = \underset{\boldsymbol{A}\in\mathbb{R}^{d\times d}}{\arg\min}\, \mathcal{L}_T(\boldsymbol{A}) + \lambda \|\boldsymbol{A} \circ |\widehat{\boldsymbol{A}}_{MLE}|^{-\gamma}\|_1, \tag{3.1}$$

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). 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 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*

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

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

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

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

The proof of Theorem 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 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.

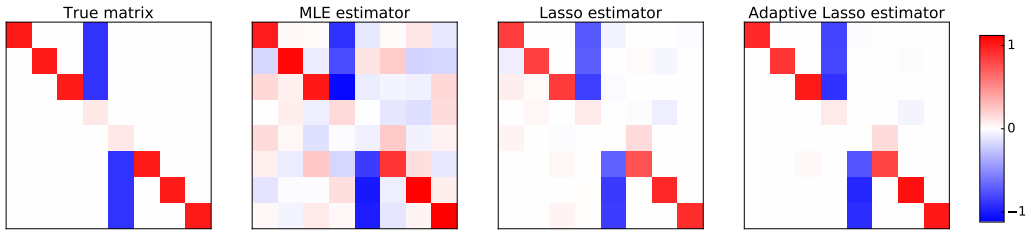


Figure 2: MLE, Lasso and Adaptive Lasso estimates compared to the ground truth matrix. The Lasso shows a significant improvement over the MLE, and the Adaptive Lasso further improves on the Lasso, especially in terms of support recovery. All penalization parameters are obtained through cross-validation.

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 (2.1). However, theoretical penalization parameters are known to be very pessimistic, in the 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.

$$\widehat{\boldsymbol{A}}_\lambda = \underset{\boldsymbol{A}\in\mathbb{R}^{d\times d}}{\arg\min}\, \mathcal{L}_{.8T}(\boldsymbol{A}) + \lambda\|\boldsymbol{A}\|_1,$$
$$\hat{\lambda} = \underset{\lambda\geq 0}{\arg\min}\, \mathcal{L}_{[.8T,T]}(\widehat{\boldsymbol{A}}_\lambda),$$

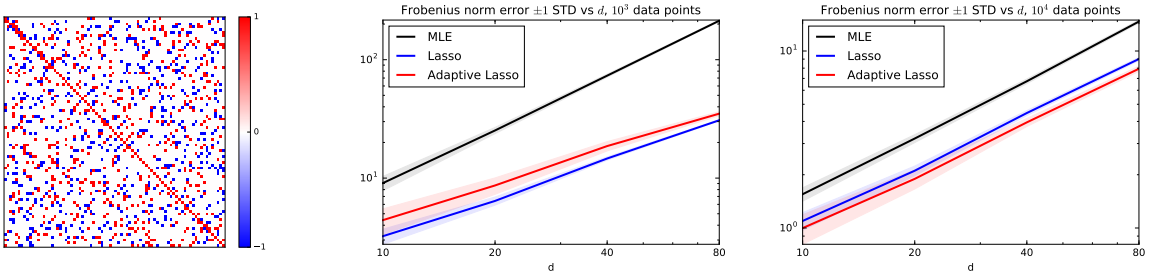
where $\mathcal{L}_{[.8T,T]}$ is the negative log-likelihood constructed from the interval $[.8T, T]$. The cross-validated Lasso is then $\widehat{\boldsymbol{A}}_{\hat{\lambda}}$, and we define similarly the cross-validated adaptive Lasso. In the next sections, referring to Lasso and Adaptive Lasso 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 T and of the dimension d

In Figures 3 and 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 3. Note that all y-axis are on a logarithmic scale.

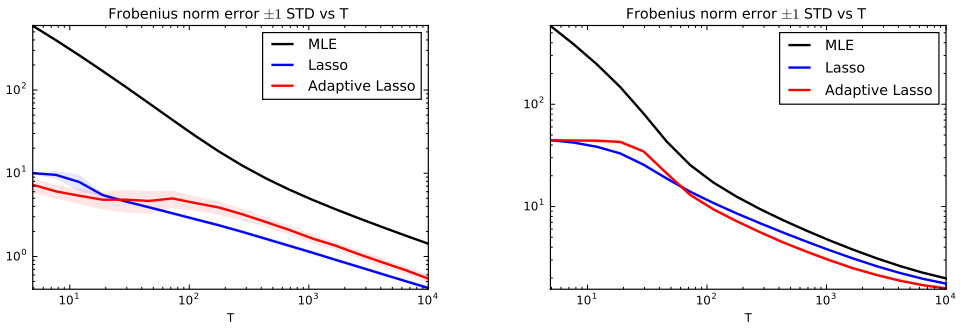
In Figure 3, we observe the deterioration of the estimation error with an increasing dimension d. We observe also that penalized procedures perform much better than the MLE, but that slopes are very close: this comes from the fact that the row sparsity is fixed to $0.2 \times d$ leading to a $0.2 \times d^2$ 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 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 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 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 4 from Section 3 that the Adaptive Lasso is consistent for variable selection of the drift parameter $\boldsymbol{A}_0$. In Figure 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 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 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.

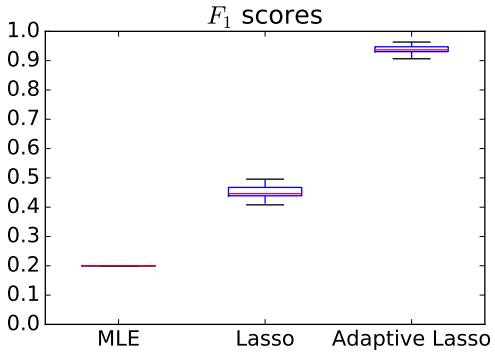
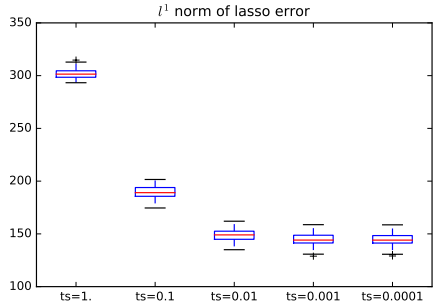


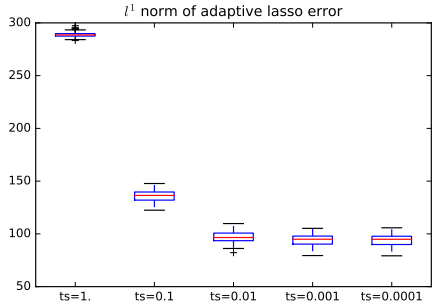
Figure 5: Accuracy scores for different estimation methods, using a 80×80 matrix $\boldsymbol{A}_0$ and an observation length $T = 100$. We classify as positive detection all non-zero entries of the estimators. The plot illustrates a clear advantage of Adaptive Lasso over Lasso in terms of support recovery. 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 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 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 $\delta t \leq 10^{-2}$.



(a) ℓ^1 error for the Lasso estimator.



(b) ℓ^1 error for the Adaptive Lasso estimator.

Figure 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.

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.

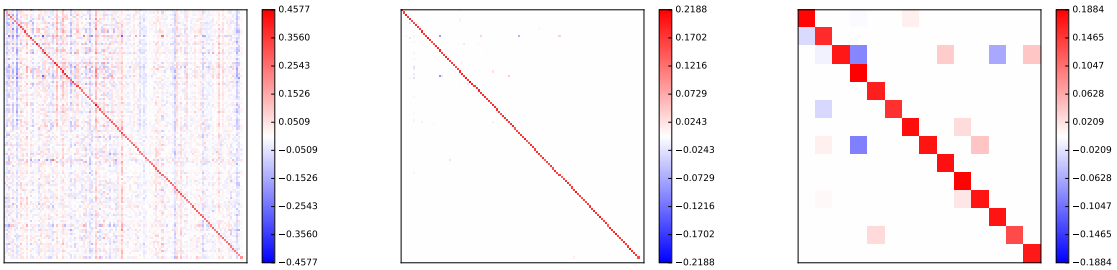
However, this method does not address two issues. First, one needs a separate method to find the relevant pairs. Second, instead of pairs, one could be interested in more general linear combinations or in situations where the evolution of one price impacts another, when the pair does not fit a mean-reverting 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:

$$\mathrm{d}R_t = -\boldsymbol{A}(R_t - m)\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}W_t \tag{4.1}$$

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 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 value, in absolute value, that we find outside of the diagonal is at tick-coordinates ('PRU','FITB'), and takes a value roughly



(a) Estimation using MLE (b) Estimation using Ad. Lasso (c) Zoom of Ad. Lasso

Figure 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 (4.1). 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.

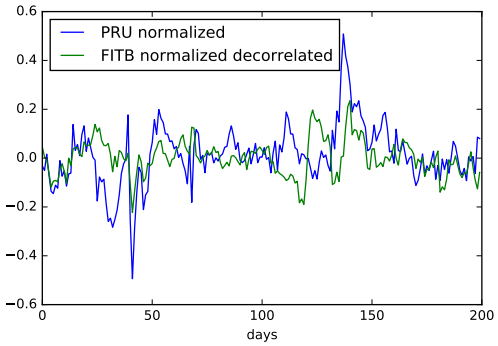


Figure 8: Plot of normalized EMA for PRU and FITB, the latter subtracted a fraction of the PRU data for decorrelation.

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 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.

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 (1.1). 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 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 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 5 above follows from Lemmas 1 and 2 below, after taking $H(R) = H_1(R) \wedge H_2(R)$. The proof of the Lemmas is based on Theorem 6 below, which shows a deviation inequality for the integral of a functional of an ergodic process from its long-time limit.

Theorem 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:*

$$c\int f^2 \log f^2 \mathrm{d}\mu \leq - < Lf, f >_\mu \tag{6.1}$$

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 $R > 0$:

$$\mathbb{P}\left[\frac{1}{T}\int_0^T Q(X_t)\mathrm{d}t - \int Q\mathrm{d}\mu \geq R\right] \leq \exp\left(-tH^*\left(R\right)\right) \tag{6.2}$$

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 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) *holds due to the classical log-Sobolev inequality [Gro75], with $c = 1/4$.*

Observe that Equation (6.2) applies to a one-sided inequality. Therefore, in order to get Theorem 5, we will have to work with two inequalities. We deal with the first one in Lemma 1 below.

Lemma 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 $H_1(R) = \frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} - \log\det\left(\boldsymbol{I} + R\frac{\boldsymbol{C}_\infty uu^\top}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right)$.

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 6, which applies with $c = 1/4$ as explained in Remark 1, to the function $Q(X) = u^\top XX^\top u = (u^\top X)^2$. 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:

$$\begin{aligned} I_\rho &:= \int \exp\left(4\rho u^\top XX^\top u\right)\mathrm{d}\mu \\ &= (2\pi)^{-d/2}(\det \boldsymbol{C}_\infty)^{-1/2}\int \exp\left(-\frac{1}{2}X^\top\left(\boldsymbol{C}_\infty^{-1} - 8\rho uu^\top\right)X\right)\mathrm{d}X \\ &= (\det \boldsymbol{C}_\infty)^{-1/2}(\det \boldsymbol{\Sigma}_\rho)^{1/2} \end{aligned} \tag{6.3}$$

and $\boldsymbol{\Sigma}_\rho^{-1} := \boldsymbol{C}_\infty^{-1} - 8\rho uu^\top$. 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 (6.3) 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 $\|8\rho \boldsymbol{C}_\infty uu^\top\|_{\mathrm{op}} \leq 8\rho \max \mathrm{Sp}(\boldsymbol{C}_\infty) < 1$.

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 $\frac{\mathrm{d}\boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho} = -\boldsymbol{\Sigma}_\rho \frac{\mathrm{d}\boldsymbol{\Sigma}_\rho^{-1}}{\mathrm{d}\rho}\boldsymbol{\Sigma}_\rho = 8\boldsymbol{\Sigma}_\rho uu^\top \boldsymbol{\Sigma}_\rho$. Hence

$$\frac{\mathrm{d}\det \boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho} = \mathrm{tr}\left(\mathrm{adj}(\boldsymbol{\Sigma}_\rho)\frac{\mathrm{d}\boldsymbol{\Sigma}_\rho}{\mathrm{d}\rho}\right) = 8\mathrm{tr}\left(\mathrm{adj}(\boldsymbol{\Sigma}_\rho)\boldsymbol{\Sigma}_\rho uu^\top \boldsymbol{\Sigma}_\rho\right) = 8(\det \boldsymbol{\Sigma}_\rho)u^\top \boldsymbol{\Sigma}_\rho u.$$

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 $\|8\rho \boldsymbol{C}_\infty uu^\top\|_{\mathrm{op}} < 1$:

$$\begin{aligned} \boldsymbol{\Sigma}_\rho &= \left(\boldsymbol{I} - 8\rho \boldsymbol{C}_\infty uu^\top\right)^{-1}\boldsymbol{C}_\infty = \sum_{k\geq 0}\left(8\rho \boldsymbol{C}_\infty uu^\top\right)^k \boldsymbol{C}_\infty \\ &= \boldsymbol{C}_\infty + 8\rho \boldsymbol{C}_\infty u\sum_{k\geq 0}\left(8\rho u^\top \boldsymbol{C}_\infty u\right)^k u^\top \boldsymbol{C}_\infty \\ &= \boldsymbol{C}_\infty + 8\rho\frac{\boldsymbol{C}_\infty uu^\top \boldsymbol{C}_\infty}{1 - 8\rho u^\top \boldsymbol{C}_\infty u} \\ u^\top \boldsymbol{\Sigma}_\rho u &= \frac{u^\top \boldsymbol{C}_\infty u}{1 - 8\rho u^\top \boldsymbol{C}_\infty u}. \end{aligned}$$

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:

$$\Sigma_{\rho^*} = \boldsymbol{C}_\infty\left(\boldsymbol{I} + R\frac{uu^\top \boldsymbol{C}_\infty}{(u^\top \boldsymbol{C}_\infty u)^2}\right)$$

$$H^*(R) = \rho^*(R + u^\top \boldsymbol{C}_\infty u) + \frac{1}{8}\log(\det \boldsymbol{C}_\infty) - \frac{1}{8}\log(\det \boldsymbol{\Sigma}_{\rho^*})$$
$$= \frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} - \log\det\left(\boldsymbol{I} + R\frac{uu^\top \boldsymbol{C}_\infty}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right). \qquad \square$$

We deal with the second inequality in Lemma 2 below.

Lemma 2. *Assume $\boldsymbol{A}_0$ is symmetric. Then for any vectors u such that $\|u\|_2 \leq 1$:*

$$\mathbb{P}\left[u^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u \leq -R\right] \leq \exp(-TH_2(R)) \tag{6.4}$$

where $H_2(R) = \begin{cases} -\frac{1}{8}\left(\frac{R}{u^\top \boldsymbol{C}_\infty u} + \log\det\left(\boldsymbol{I} - R\frac{\boldsymbol{C}_\infty uu^\top}{(u^\top \boldsymbol{C}_\infty u)^2}\right)\right) & \text{if } R < u^\top \boldsymbol{C}_\infty u, \\ +\infty & \text{else.} \end{cases}$.

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 1 to $Q(X) = -u^\top XX^\top u$ gives:

$$H^*(R) := \sup_{\rho \geq 0} \rho R - \rho u^\top \boldsymbol{C}_\infty u + \frac{1}{8}\log\det \boldsymbol{C}_\infty - \frac{1}{8}\log\det \boldsymbol{\Sigma}_\rho$$
$$\boldsymbol{\Sigma}_\rho^{-1} := \boldsymbol{C}_\infty^{-1} + 8\rho uu^\top.$$

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 1, as it is sufficient to get the upper bound (6.4). This enables the geometric series calculation as in the proof of Lemma 1. We get:

$$u^\top \boldsymbol{\Sigma}_\rho u = \frac{u^\top \boldsymbol{C}_\infty u}{1 + 8\rho u^\top \boldsymbol{C}_\infty u}$$
$$\rho^* := \frac{1}{8}\frac{R}{u^\top \boldsymbol{C}_\infty u(u^\top \boldsymbol{C}_\infty u - R)}$$
$$\Sigma_{\rho^*} = \boldsymbol{C}_\infty\left(\boldsymbol{I} - R\frac{SS^\top \boldsymbol{C}_\infty}{(S^\top \boldsymbol{C}_\infty S)^2}\right)$$
$$H^*(R) = -\frac{1}{8}\left(\frac{R}{S^\top \boldsymbol{C}_\infty S} + \log\det\left(\boldsymbol{I} - R\frac{SS^\top \boldsymbol{C}_\infty}{(S^\top \boldsymbol{C}_\infty S)^2}\right)\right). \qquad \square$$

For completeness, we state an interesting corollary.

Corollary 2. *For any vectors S_1, S_2 such that $\|u_1\|_2 \leq 1, \|u_2\|_2 \leq 1$, and for any $i, j \leq d$:*

$$\mathbb{P}\left[|u_1^\top(\widehat{\boldsymbol{C}}_T - \boldsymbol{C}_\infty)u_2| > 3R\right] \leq 6\exp(-TH(R))$$
$$\mathbb{P}\left[|\widehat{\boldsymbol{C}}_T^{ij} - \boldsymbol{C}_\infty^{ij}| > 3R\right] \leq 6\exp(-TH(R)).$$

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.

$\square$

6.2 Proof of Theorem 1 (Lasso error bound)

Lemma 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 (2.1). 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 (6.8). 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 8 and Corollary 4, we have for any matrix $\boldsymbol{U}$:

$$\mathbb{P}\left[\inf_{u\in C(s,c_0)} \frac{\|u^\top X\|_{L^2}}{\|u\|_2} \geq \kappa \cap \langle \boldsymbol{U}, \boldsymbol{\varepsilon}_T\rangle_F \leq \gamma^{-1}\lambda\|\boldsymbol{U}\|_1 \cap \forall i, \kappa^2 \leq \hat{\delta}_T^{ii} \leq \boldsymbol{C}_\infty^{ii} + \kappa^2\right] \geq 1 - \epsilon_0. \quad (6.5)$$

We proceed to the proof of Theorem 1.

Proof. We assume for the all what follows that the observation falls in the set of events defined by Equation (6.5) 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 3, we have

$$\begin{aligned}
S &:= 2\tau\gamma^{-1}\lambda\|\boldsymbol{U}\|_1 + \|(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0)X\|_{L^2}^2 - \|(\boldsymbol{A} - \boldsymbol{A}_0)\,X\|_{L^2}^2 + \|\boldsymbol{U}X\|_{L^2}^2 \\
&\leq 2\tau\gamma^{-1}\lambda\|\boldsymbol{U}\|_1 + 2\langle \boldsymbol{\varepsilon}_T, \boldsymbol{U}\rangle_F + 2\lambda(\|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1) \\
&\leq 2\lambda\left((1+\tau)\gamma^{-1}\|\boldsymbol{U}\|_1 + \|\boldsymbol{A}\|_1 - \|\widehat{\boldsymbol{A}}\|_1\right) \\
&\leq 2\lambda\sum_{i=1}^{d}(1+\tau)\gamma^{-1}\|\boldsymbol{U}^{i,\bullet}\|_1 + \|\boldsymbol{A}^{i,\bullet}\|_1 - \|\widehat{\boldsymbol{A}}_\lambda^{i,\bullet}\|_1 \\
&\leq 2\lambda\sum_{\Delta^i>0}\Delta^i \qquad (6.6)
\end{aligned}$$

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 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

$$\|\boldsymbol{U}^{i,\bullet}_{|\bar{\mathcal{A}}^{i},\bullet}\|_1 < \frac{1+(1+\tau)\gamma^{-1}}{1-(1+\tau)\gamma^{-1}}\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i},\bullet}\|_1 = c_0\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i},\bullet}\|_1$$
$$\|\boldsymbol{U}^{i,\bullet}\|_1 < (1+c_0)\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{A}^{i},\bullet}\|_1 \leq (1+c_0)\|\boldsymbol{U}^{i,\bullet}_{|\mathcal{U}^{i},\bullet}\|_1$$

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 (6.5) 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

$$2\lambda\gamma^{-1}(\gamma+\tau+1)\sqrt{s}\kappa^{-1}\|\boldsymbol{U}X\|_{L^2} - \|\boldsymbol{U}X\|^2_{L^2} \leq \left(\frac{\gamma+\tau+1}{\gamma\kappa}\right)^2\lambda^2 ds,$$

we conclude with Equation (2.2).

□

The proof of Corollary 1 consists in using Theorem 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 (2.2).

2. We take $\boldsymbol{A} = \boldsymbol{A}_0$ and $\tau > 0$ in Equation (2.2). We bound The L^2 norm from below by 0 and get the result.

3. It suffices to apply Equation (2.3) 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 9). To prove that last point, we fix an index $i \leq d$ and continue the proof from Equation (6.6). 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:

$$\begin{aligned}
2\lambda\Delta^i &\geq \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right)X\|^2_{L^2} - \|(\boldsymbol{A} - \boldsymbol{A}_0)\,X\|^2_{L^2} + \|\boldsymbol{U}X\|^2_{L^2} \\
&= \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right)X\|^2_{L^2} - \|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0 + e_i(\boldsymbol{U}^{i,\bullet})^\top\right)X\|^2_{L^2} + \|(\boldsymbol{U}^{i,\bullet})^\top X\|^2_{L^2} \\
&\geq 2\langle(\boldsymbol{A}_0 - \widehat{\boldsymbol{A}})X, e_i(\boldsymbol{U}^{i,\bullet})^\top X\rangle_{L^2} = 2\|(\boldsymbol{U}^{i,\bullet})^\top X\|^2_{L^2} \geq 0.
\end{aligned}$$

   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 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 (2.4) and (2.5).

□

We finish this Section with the statement of a few useful inequalities

Lemma 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 2 (lower bound of the estimation error)

Lemma 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 $\operatorname{reg}(r,d)$, 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 $|\operatorname{reg}(r,d)|$. 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.

$$\begin{aligned}
\log|\operatorname{reg}(r,d)| &\sim -\frac{r^2-1}{4} - \frac{r^3}{12d} + \log(dr)! - \log(dr/2)! - \frac{dr}{2}\log 2 - d\log r! \\
&\sim -\frac{r^2-1}{4} - \frac{r^3}{12d} + \frac{dr}{2}\log\frac{ed}{r} - \frac{d}{2}\log r - d\log\sqrt{2\pi} + \frac{1}{2}\log 2 - \frac{1}{12dr} - \frac{d}{12r}.
\end{aligned}$$

Keeping only the highest order in d gives $\log|\operatorname{reg}(r,d)| \sim \frac{dr}{2}\log\frac{ed}{r}$. 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 $|\operatorname{reg}(r,d)| \geq 2cd(r+2)\log ed/r$.

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 6 gives a way to construct a family of drift parameters with corresponding diagonal stationary covariance from a family of antisymmetric matrices.

Lemma 6. *Take $\boldsymbol{A} = \alpha \boldsymbol{I} + \boldsymbol{B}$ for $\alpha > 0$ and $\boldsymbol{B}$ an antisymmetric matrix. Then we have:*

$$\boldsymbol{C}_\infty(\boldsymbol{A}) = \int_0^\infty e^{-\boldsymbol{A}t} e^{-\boldsymbol{A}^\top t} \mathrm{d}t = \frac{1}{2\alpha} \boldsymbol{I}.$$

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 $e^{-(\alpha \boldsymbol{I} + \boldsymbol{B})t} = e^{-\alpha t} \boldsymbol{U} e^{-i\boldsymbol{D}t} \boldsymbol{U}^*$ and $e^{-(\alpha \boldsymbol{I} + \boldsymbol{B})^\top t} = e^{-(\alpha \boldsymbol{I} - \boldsymbol{B})t} = e^{-\alpha t} \boldsymbol{U} e^{i\boldsymbol{D}t} \boldsymbol{U}^*$ hence $\boldsymbol{C}_\infty(\boldsymbol{A}) = \int_0^\infty e^{-2\alpha t} \mathrm{d}t \; \boldsymbol{I} = \boldsymbol{I}/(2\alpha)$. □

Corollary 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. *$\boldsymbol{C}_\infty(\boldsymbol{A}) = \boldsymbol{I}$*
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 $\Omega = \{\frac{1}{2}\boldsymbol{I} + w\boldsymbol{B} : \boldsymbol{B} \in \Omega_a\}$ for some $w > 0$ and Ω_a as defined in Lemma 5. Then $|\Omega| = |\Omega_a|$ and hence $\log |\Omega| \geq cds \log \frac{ced}{s}$. Condition 1 is verified trivially and Lemma 6 gives condition 2. Further, from Lemma 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 $KL(\mathbb{P}_{\boldsymbol{A}}, \mathbb{P}_{\boldsymbol{A}'}) = \mathbb{E}_{\boldsymbol{A}} \left[\log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{A}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{A}'}}\right] = \frac{T}{2} \mathrm{tr}\,(\boldsymbol{A}' - \boldsymbol{A}) \boldsymbol{C}_\infty(\boldsymbol{A}) (\boldsymbol{A}' - \boldsymbol{A})^\top$. Using condition 2, $KL(\mathbb{P}_{\boldsymbol{A}}, \mathbb{P}_{\boldsymbol{A}'}) = \frac{T}{2} \|\boldsymbol{A} - \boldsymbol{A}'\|_F^2 \leq w^2 T ds$.

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 2 is the corollary of the preceding and of Theorem 2.7 from [Tsy08].

6.4 Proof of Theorem 3 (Restricted Eigenvalue property)

Lemma 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 7 applied using hypothesis (H4), we get the following Lemma.

Lemma 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 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 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:*

$$\mathbb{P}\left[\inf_{u \in C(s,c_0), \|u\|_2 \leq 1} |u^\top \widehat{\boldsymbol{C}}_T u| \geq \kappa^2, \|\mathrm{diag}\, \widehat{\boldsymbol{C}}_T\|_\infty \leq \|\mathrm{diag}\, \boldsymbol{C}_\infty\|_\infty + \kappa^2\right] \geq 1 - \frac{\epsilon_0}{2}.$$

Proof. From Lemma 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 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. $\square$

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 3.

Lemma 9. *For any subset $E \subset \mathbb{R}^d$, we have:*

$$\sup_{\boldsymbol{A}: \forall i \leq d, \boldsymbol{A}^{i,\bullet} \in E} \frac{\|\boldsymbol{A}X\|_{L^2}}{\|\boldsymbol{A}\|_F} = \left(\sup_{\boldsymbol{A}: \forall i \leq d, \boldsymbol{A}^{i,\bullet} \in E, \|\boldsymbol{A}\|_F \leq 1} \mathrm{tr}\, \boldsymbol{A} \widehat{\boldsymbol{C}}_T \boldsymbol{A}^\top\right)^{1/2}$$

$$= \left(\sup_{u \in E, \|u\|_2 \leq 1} u^\top \widehat{\boldsymbol{C}}_T u \right)^{1/2}$$
$$= \sup_{u \in E} \frac{\|u^\top X\|_{L^2}}{\|u\|_2}.$$

$$\inf_{\boldsymbol{A}: \forall i \leq d, \boldsymbol{A}^{i,\bullet} \in E} \frac{\|\boldsymbol{A}X\|_{L^2}}{\|\boldsymbol{A}\|_F} = \left(\inf_{\boldsymbol{A}: \forall i \leq d, \boldsymbol{A}^{i,\bullet} \in E, \|\boldsymbol{A}\|_F \leq 1} \operatorname{tr} \boldsymbol{A} \widehat{\boldsymbol{C}}_T \boldsymbol{A}^\top \right)^{1/2}$$
$$= \left(\inf_{u \in E, \|u\|_2 \leq 1} u^\top \widehat{\boldsymbol{C}}_T u \right)^{1/2}$$
$$= \inf_{u \in E} \frac{\|u^\top X\|_{L^2}}{\|u\|_2}.$$

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 4 (asymptotic properties of the Adaptive Lasso)

$\widehat{\boldsymbol{A}}_{ad.}$ is defined as the minimizer of the penalized log-likelihood, see Equation (3.1). 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:

$$\sqrt{T}(\widehat{\boldsymbol{A}}_{ad.} - \boldsymbol{A}_0) = \hat{\boldsymbol{U}} = \arg\min_{\boldsymbol{U}} \phi_1(\boldsymbol{U}) + \phi_2(\boldsymbol{U})$$
$$\phi_1(\boldsymbol{U}) := T\mathcal{L}_T(\boldsymbol{A}_0 + \boldsymbol{U}/\sqrt{T}) - T\mathcal{L}_T(\boldsymbol{A}_0)$$
$$\phi_2(\boldsymbol{U}) := \lambda T \|(\boldsymbol{A}_0 + \boldsymbol{U}/\sqrt{T}) \odot \boldsymbol{\Gamma}\|_1 - \lambda T \|\boldsymbol{A}_0 \odot \boldsymbol{\Gamma}\|_1.$$

Using equation (1.4), we characterize the limit structure of ϕ_1.

$$\phi_1(\boldsymbol{U}) = \sqrt{T} \operatorname{tr} \boldsymbol{U}^\top \boldsymbol{\varepsilon}_T + \frac{1}{2} \operatorname{tr} \boldsymbol{U} \widehat{\boldsymbol{C}}_T \boldsymbol{U}^\top$$
$$= \frac{1}{\sqrt{T}} \int_0^T (\boldsymbol{U}X_t)^\top \mathrm{d}W_t + \frac{1}{2T} \int_0^T \|\boldsymbol{U}X_t\|_2^2 \mathrm{d}t.$$

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 10.

Lemma 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

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

where $Z \overset{\mathrm{d}}{=} \mathcal{N}(0, \boldsymbol{I})$ 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:

$$\phi_1(\boldsymbol{U}) \xrightarrow{\mathcal{L}} \operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}\boldsymbol{G}\right). \tag{6.7}$$

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 (6.7), we have:

$$\phi_1(\boldsymbol{U}) + \phi_2(\boldsymbol{U}) \xrightarrow{\mathcal{L}} \begin{cases} +\infty & \text{if } \boldsymbol{U}_{\bar{\mathcal{A}}_0} = 0, \\ \operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}\boldsymbol{G}\right) & \text{else.} \end{cases}$$

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:

$$\begin{aligned} \operatorname{tr} \boldsymbol{U}\boldsymbol{G} &= \sum_{ij} \boldsymbol{U}^{ij}\boldsymbol{G}^{ji} \\ &= \left((\boldsymbol{G}^\top)_{|\mathcal{A}_0}\right)^\top \boldsymbol{U}_{|\mathcal{A}_0} \\ \operatorname{tr} \boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top &= \sum_{ijkl} \boldsymbol{U}^{ij}\boldsymbol{U}^{kl}(\mathbb{1}_{i=k}\boldsymbol{C}_\infty^{jl}) \\ &= (\boldsymbol{U}_{|\mathcal{A}_0})^\top (\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{|\mathcal{A}_0^2} \boldsymbol{U}_{|\mathcal{A}_0}. \end{aligned}$$

$(\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{|\mathcal{A}_0^2}$ is the restriction of $\boldsymbol{C}_\infty \otimes \boldsymbol{I}$ to the indices in $\mathcal{A}_0^2 := \mathcal{A}_0 \times \mathcal{A}_0$ and $\boldsymbol{C}_\infty \otimes \boldsymbol{I}$ is invertible and symmetric, with inverse $\boldsymbol{C}_\infty^{-1} \otimes \boldsymbol{I}$, hence $(\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{\mathcal{A}_0^2}$ is invertible and symmetric. Hence, $\operatorname{tr}\left(\frac{1}{2}\boldsymbol{U}\boldsymbol{C}_\infty\boldsymbol{U}^\top + \boldsymbol{U}\boldsymbol{G}\right) = \frac{1}{2}(\boldsymbol{U}_{|\mathcal{A}_0})^\top (\boldsymbol{I} \otimes \boldsymbol{C}_\infty)_{|\mathcal{A}_0^2} \boldsymbol{U}_{\mathcal{A}_0} + (\boldsymbol{U}_{|\mathcal{A}_0})^\top (\boldsymbol{G}^\top)_{|\mathcal{A}_0}$, 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:

$$\widehat{\boldsymbol{U}}_{|\mathcal{A}_0} \xrightarrow{\mathcal{L}} -(\boldsymbol{G}^\top)_{|\mathcal{A}_0} \left((\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{|\mathcal{A}_0^2}\right)^{-1}, \quad \widehat{\boldsymbol{U}}_{\bar{\mathcal{A}}_0} \xrightarrow{\mathcal{L}} 0$$

and we find that $(\boldsymbol{G}^\top)_{|\mathcal{A}_0} \left((\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{|\mathcal{A}_0^2}\right)^{-1} \sim \mathcal{N}(0, \mathcal{V})$ with $\mathcal{V} := \left((\boldsymbol{C}_\infty \otimes \boldsymbol{I})_{|\mathcal{A}_0^2}\right)^{-1}$.

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 $\widehat{\boldsymbol{A}}_{ad.}^{ij} \neq 0$. 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 $T \to +\infty$:

$$\begin{aligned} \sqrt{T}S^{ij} = \frac{1}{\sqrt{T}} \int_0^T X_t^j \mathrm{d}W_t^i &\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{C}_\infty^{jj}\right) \\ \sqrt{T}\widehat{\boldsymbol{A}}_{ad.} &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}) \\ \widehat{\boldsymbol{C}}_T &\xrightarrow{\mathbb{P}} \boldsymbol{C}_\infty \\ \lambda \boldsymbol{\Gamma}^{ij}\sqrt{T} = \lambda T^{(\gamma+1)/2} (\sqrt{T}|\widehat{\boldsymbol{A}}_{MLE}^{ij}|)^{-\gamma} &\xrightarrow{\mathbb{P}} +\infty. \end{aligned}$$

When $T \to +\infty$, we can therefore bound the probability of $\widehat{\boldsymbol{A}}_{ad.}^{ij} \neq 0$ 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, $\forall (i,j) \in \mathcal{A}_0, \mathbb{P}\left[\widehat{\boldsymbol{A}}_{ad.}^{ij} = 0\right] \to 1$.

6.6 Deviation bound

Recall Bernstein's inequality, see Chapter 4, Exercise 3.16 in [RY99]:

Lemma 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 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 $|\log\langle M\rangle_t| \geq |j| - 1$.

$$\begin{aligned}
P &= \mathbb{P}\left[M_t \geq \sqrt{4e\langle M\rangle_t \left(x + \log\left(2 + |\log\langle M\rangle_t|\right)\right)}\right] \\
&= \sum_{j\in\mathbb{Z}} \mathbb{P}\left[M_t \geq \sqrt{4e\langle M\rangle_t \left(x + \log\left(2 + |\log\langle M\rangle_t|\right)\right)}, e^j \leq \langle M\rangle_t < e^{j+1}\right] \\
&\leq \sum_{j\in\mathbb{Z}} \mathbb{P}\left[M_t \geq \sqrt{4e^{j+1}\left(x + \log\left(1 + |j|\right)\right)}, \langle M\rangle_t < e^{j+1}\right] \\
&\leq \sum_{j\in\mathbb{Z}} \exp\left(-2\left(x + \log\left(1 + |j|\right)\right)\right) \\
&= \exp(-2x)\sum_{j\in\mathbb{Z}} \frac{1}{\left(1 + |j|\right)^2} \\
&= 2\exp(-2x)\sum_{j\in\mathbb{N}^*} \frac{1}{j^2} \\
&= \frac{\pi^2}{3}\exp(-2x).
\end{aligned}$$

□

Theorem 8. *Define for $x > 0$:*

$$\theta(x, (X_t)) := \sqrt{4eT^{-1}|\mathrm{diag}\left(\widehat{\boldsymbol{C}}_T\right)|_\infty \left(x + \log(2 + |\log T\mathrm{diag}\left(\widehat{\boldsymbol{C}}_T\right)|_\infty)\right)}, \tag{6.8}$$

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

$$\mathbb{P}\left[\left\langle \boldsymbol{U}, T^{-1}\int_0^T \mathrm{d}W_t X_t^\top \right\rangle_F \leq \theta(x, (X_t))\|\boldsymbol{U}\|_1\right] \geq 1 - \frac{\pi^2}{3}\exp(-2x + 2\log d). \tag{6.9}$$

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 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:

$$\mathbb{P}\left[\left\|\int_0^T \mathrm{d}W_s X_t^\top\right\|_\infty \geq T\theta(x,(X_t))\right] \leq \frac{\pi^2}{3}d^2\exp(-2x). \tag{6.10}$$

Observe now that by homogeneity, it suffices to prove Equation (6.9) 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 (6.9) follows from Equation (6.10).

□

References

[BBvL15] Francisco Blasques, Falk Bräuning, and Iman van Lelyveld. A dynamic network model of the unsecured interbank lending market. BIS Working Papers 491, Bank for International Settlements, February 2015.

[BLT16] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. Submitted to the Annals of Statistics, 05 2016.

[BM15] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 08 2015.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

[BSW11] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selectionof reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[BSW12] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388, 10 2012.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[CFS15] Rene Carmona, Jean-Pierre Fouque, and Li-Hsien Sun. Mean field games and systemic risk. *Communications in Mathematical Sciences*, 13(4):911–933, 2015.

[CG07] Patrick Cattiaux and Arnaud Guillin. Deviation bounds for additive functionals of markov processes. *ESAIM: Probability and Statistics*, 12:12–29, 11 2007.

[EG60] Paul Erdős and Tibor Gallai. Gráfok előírt fokszámú pontokkal. *Matematikai Lapok*, 11:264–274, 1960.

[FBO12] Julien Mairal Francis Bach, Rodolphe Jenatton and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[FI13] Jean-Pierre Fouque and Tomoyuki Ichiba. Stability in a model of interbank lending. *SIAM J. Financial Math.*, 4(1):784–803, 2013.

[GG14] Silvia Gabrieli and Co-Pierre Georg. A network view on interbank market freezes. Technical report, Banque de France, November 2014.

[Gir14a] C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[Gir14b] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, December 2014.

[Gro75] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[GS16] Emmanuel Gobet and Qihao She. Perturbation of Ornstein-Uhlenbeck stationary distributions: expansion and simulation. working paper or preprint, July 2016.

[GSV15] Silvia Gabrieli, Dilyara Salakhova, and Guillaume Vuillemey. Cross-border interbank contagion in the european banking sector. Document de travail 545, Banque de France, March 2015.

[HHMS93] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating gaussian diffusions. *The Annals of Applied Probability*, 3(3):897–913, 1993.

[Hul09] John C. Hull. *Options, Futures and Other Derivatives*. Options, Futures and Other Derivatives. Pearson/Prentice Hall, 2009.

[Jac01] Jean Jacod. Inference for stochastic processes. Prépublications du laboratoire de probabilités et modèles aléatoires, 2001.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991.

[KT15a] Olga Klopp and Alexandre B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.

[KT15b] Olga Klopp and Alexandre B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.

[Kut04] Yury A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. Springer, 2004.

[MW91] Brendan D McKay and Nicholas C Wormald. Asymptotic enumeration by degree sequence of graphs with degrees $o(\sqrt{n})$. *Combinatorica*, 11(4):369–382, 1991.

[RY99] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg, 1999.

[SC+16] Weijie Su, Emmanuel Candes, et al. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[Sok13] Alexander Sokol. *On martingales, causality, identifiability and model selection*. PhD thesis, University of Copenhagen, November 2013.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

[vZ00] Harry van Zanten. A multivariate central limit theorem for continuous local martingales. *Statistics and Probability Letters*, 50(3):229 – 235, 2000.

[Zou06] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[ZWJ14] Yuchen Zhang, Martin J. Wainwright, and Michael 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.