# Minimax Estimation of Partially-Observed Vector AutoRegressions

Guillaume Dalle, Yohann de Castro

# Minimax Estimation of Partially-Observed Vector AutoRegressions

**Guillaume Dalle** GUILLAUME.DALLE@ENPC.FR
*CERMICS, École des Ponts*
*Marne-la-Vallée, France*

**Yohann De Castro** YOHANN.DE-CASTRO@EC-LYON.FR
*Institut Camille Jordan, École Centrale Lyon*
*Écully, France*

## Abstract

To understand the behavior of large dynamical systems like transportation networks, one must often rely on measurements transmitted by a set of sensors, for instance individual vehicles. Such measurements are likely to be incomplete and imprecise, which makes it hard to recover the underlying signal of interest.

Hoping to quantify this phenomenon, we study the properties of a partially-observed state-space model. In our setting, the latent state $X$ follows a high-dimensional Vector AutoRegressive process $X_t = \theta X_{t-1} + \varepsilon_t$. Meanwhile, the observations $Y$ are given by a noise-corrupted random sample from the state $Y_t = \Pi_t X_t + \eta_t$. Several random sampling mechanisms are studied, allowing us to investigate the effect of spatial and temporal correlations in the distribution of the sampling matrices $\Pi_t$.

We first prove a lower bound on the minimax estimation error for the transition matrix $\theta$. We then describe a sparse estimator based on the Dantzig selector and upper bound its non-asymptotic error, showing that it achieves the optimal convergence rate for most of our sampling mechanisms. Numerical experiments on simulated time series validate our theoretical findings, while an application to open railway data highlights the relevance of this model for public transport traffic analysis.

**Keywords:** vector autoregression, partial observation, sparsity, minimax lower bound, railway modeling.

## 1. Introduction

In this paper, we focus on the estimation of a *partially-observed Vector AutoRegression*, which is a kind of state-space model endowed with a randomized observation mechanism.

### 1.1 Context of the Study

Because they provide a natural representation for periodic measurements of a stochastic process, *time series* have long been a major focus of the statistics community (Douc et al., 2014). Among the many possible models, those defined by linear Gaussian recursions may be the most widely used and the easiest to study. This is the case for the well-known *AutoRegressive* (AR) process and its numerous extensions, such as the multivariate *Vector AutoRegressive* (VAR) process (Lütkepohl, 2005).

To model complex dynamic systems, time series can be generalized to encompass hidden data, giving rise to so-called *state-space models*. The general form of these models is

$$X_{t+1} = f_t(X_t, \varepsilon_t) \qquad \text{and} \qquad Y_t = g_t(X_t, \eta_t),$$

where $X_t$ is an unknown vector representing the *latent state* of the system, $Y_t$ is a vector of *observations* derived from $X_t$, $\varepsilon_t$ is often called the *innovation* process in the linear case, and $\eta_t$ is the *noise* process. State-space models have found numerous applications in engineering, control, maintenance, climatology, finance and many more fields (Douc et al., 2014).

Unfortunately, in the real world, data acquisition is not only noisy, but limited in size. Whether this limitation is due to costly sensors, physical constraints or random sampling phenomena, it is often impossible to measure every component of a multivariate stochastic process at all times. Learning a system's dynamics based on this kind of partial information certainly seems more difficult. It is therefore natural to ask: *how much harder does parameter estimation become when one only observes a fraction $p$ of the process values?*

This question is the topic of the present paper. Here we study a simple state-space model where the state $X$ follows a high-dimensional VAR process of order 1, while the observations $Y$ are generated by applying a random sampling matrix $\Pi$ (following a known distribution $\mathcal{D}$) to $X$, and then corrupting the result with noise:

$$X_t = \theta X_{t-1} + \varepsilon_t \qquad \text{and} \qquad \begin{cases} \Pi \sim \mathcal{D} \\ Y = \Pi X + \eta. \end{cases}$$

As we will explain in Section 1.2, this model approximates a variety of relevant real-life situations. Our running example will be transportation network congestion, which is often estimated indirectly by monitoring some of the network's users as they make their way through the edges of a graph. We will justify the relevance of our model in this case by applying it to historical data from the Zürich tramway network in Switzerland.

Analyzing the properties of this *partially-observed* VAR process will provide two complementary answers to our main question. On the one hand, we will obtain a lower bound for the minimax estimation error on the transition matrix $\theta$ of the VAR process. On the other hand, we will construct an estimator that provably achieves this optimal rate of convergence for most of the sampling mechanisms we consider, at least up to noise-related constants. A rough summary of our results is that an optimal estimator $\widehat{\theta}$ satisfies

$$\|\widehat{\theta} - \theta\|_\infty \propto \left(1 + \frac{\omega^2}{\sigma^2}\right) \frac{s}{p\sqrt{T}} \tag{1}$$

with high probability. In Equation (1), $\|\cdot\|_\infty$ denotes the operator $\ell_\infty$ norm, $\sigma$ is the standard deviation of the innovations $\varepsilon_t$, $\omega$ is the standard deviation of the noise $\eta_t$, $T$ is the duration of the observation period, $s$ is the number of non-zero coefficients in each row of $\theta$, and $p$ is the fraction of observed state components.

Novel features of our work include the first proof of a minimax lower bound in this setting (to the best of our knowledge), the investigation of random sampling mechanisms displaying temporal or spatial correlations, as well as detailed numerical experiments on both simulated and real data.

## 1.2 Partially-Observed VAR Processes for Railway Modeling

The partially-observed VAR model can be relevant in situations where the components of an underlying process are sampled and observed in a non-deterministic manner. Since the present work originated from a collaboration with the French railway operator SNCF, we describe the initial use case of this formulation: a simple *model for railway delay propagation.*

### 1.2.1 Hidden Congestion and Observed Arrival Times

Let us consider a train network represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of vertices and $\mathcal{E}$ is the set of edges. We, as a railway operator, want to study the network congestion $X = (X_{t,e})_{t\in\mathbb{N}, e\in\mathcal{E}}$, which measures how hard it is to cross edge $e$ at time $t$ due to the presence of other trains on the tracks. This information can be used to regulate traffic or adjust the timetable.

We make the assumption that the congestion evolves according to a linear Gaussian recursion of the form $X_t = \theta X_{t-1} + \varepsilon_t$. This is a way to account for delay propagation between neighboring regions of the network. Indeed, if many trains are gathered on edge $e$ at time $t$, it may create a traffic jam that forces trains on upstream edges $(e-1, e-2, ...)$ at future times $(t+1, t+2, ...)$ to slow down (here we used imprecise notations by assuming some kind of topological order on the edges). In this regard, the transition matrix $\theta$ will be closely related to the adjacency structure of the network graph $\mathcal{G}$.

Our problem is that the congestion process is hidden from us: indeed, the notion of congestion itself is more of an abstract quantity than a measurable metric. The only things we can observe are the arrival times of each train, which may also depend on factors unrelated to network congestion: driver decisions, passenger behavior, mechanical failures, etc. Therefore, the travel time of some train crossing edge $e$ at time $t$ is affected by the underlying congestion value $X_{t,e}$, but not only: in some sense, the observations are a noisy version of the congestion.

To make this more formal, let us denote by $A_{k,v}$ (resp $\bar{A}_{k,v}$) the actual (resp. planned) arrival time of train $k$ at station $v$. For an edge $e = (u, v)$, we are interested in the additional delay $Y_{k,e}$ suffered along this edge by train $k$:

$$Y_{k,e} = \underbrace{(A_{k,v} - \bar{A}_{k,v})}_{\text{delay at } v} - \underbrace{(A_{k,u} - \bar{A}_{k,u})}_{\text{delay at } u} = \underbrace{(A_{k,v} - A_{k,u})}_{\text{actual crossing time}} - \underbrace{(\bar{A}_{k,v} - \bar{A}_{k,u})}_{\text{planned crossing time}} \tag{2}$$

Equation (2) tells us how to deduce $Y$ from arrival times, but now we want to relate it to the latent process $X$. As announced, the additional delay $Y_{k,e}$ is caused partly by the congestion of edge $e$ when train $k$ reaches it, and partly by other individual factors which we will regroup under a noise term $\eta_{k,e}$:

$$Y_{k,e} = X_{t,e} + \eta_{k,e} \qquad \text{where} \qquad t = \left\lfloor \frac{A_{k,u}}{\Delta t} \right\rfloor \tag{3}$$

Here $\Delta t$ is the duration of the discretization interval, i.e. the actual time elapsed between $X_t$ and $X_{t+1}$. For dense suburban networks, it should be no more than a few minutes (because of how quick the congestion evolves).

### 1.2.2 The Random Sampling Complication

It is critical to notice that the congestion term in Equation (3) is not individual (relative to a single train) but collective (it applies on the network level). Indeed, it does not depend on a specific train number $k$, but only on *the time $t$ at which train $k$ reaches the first vertex $u$ of edge $e$*. This has two main consequences.

The first consequence is that our observations are limited in size. Indeed, the only congestion values $X_{t,e}$ that give rise to observations $Y_{k,e}$ are those for which one train (or more) reaches edge $e$ at time $t$. Subsequently, the number of "activated" couples $(t, e)$ is directly related to the number of trains circulating on the network, and will represent only a fraction $p$ of the complete set $[T] \times \mathcal{E}$ of possible couples.

This activation pattern can be represented as a sampling matrix $\Pi \sim \mathcal{D}$ with random binary entries such that $Y = \Pi X + \eta$. The sampling matrix contains one row per edge crossing, and if row number $o$ (for "observation") corresponds to train $k$ on edge $e$, only the coefficient in column $(t, e)$ will equal 1, while all others are 0. Note that the model is written $Y = \Pi X + \eta$ and not $Y = \Pi(X + \eta)$ because a single entry of $X$ can be activated multiple times by different trains, hence with different individual noise values.

The second consequence is that the selection of these activated couples $(t, e)$, that is, the generation of the sampling matrix $\Pi$, can be considered random. First, railway timetables are often large and complex, potentially varying from day to day and subject to last-minute modifications. Therefore, it may be simpler to assume that the spatio-temporal locations of selected couples $(t, e)$ are randomly sampled.

But more importantly, travel times themselves are not deterministic. If train number 1234 reaches edge $e$ slightly later than usual one day, it may face a different congestion value, say $X_{t+1,e}$ instead of $X_{t,e}$. And because of this, the distribution of $\Pi$ is even influenced by the values of $X$ itself, since train 1234's unusual delay at edge $e$ may have been caused by a traffic jam on edge $e - 1$. This dependence structure implies that the sampling matrix $\Pi$ can exhibit *spatial and temporal correlations*, which will be a crucial aspect of our work.

### 1.2.3 Estimating the Transition Matrix

In order to predict the future values of the congestion, it is necessary to learn the parameters governing its evolution. The most important parameter here is the *transition matrix* $\theta$, which quantifies spatial interactions between all edges of the network. Estimating it provides valuable insight into the dynamics of delays. Unfortunately, railway networks are often large and complex, while the available data is limited in size. Suppose we study a network of dimension $|\mathcal{E}| = D$, and we have access to $N$ days of observations, each day being split into $T$ time steps (so that $24h = T \times \Delta t$). Can we recover the transition matrix $\theta$ with sufficient accuracy in the general case?

In high dimension, without additional assumptions, the answer will often be no. Fortunately, we expect $\theta$ to have a very specific structure because it describes local interactions on our network graph $\mathcal{G}$. Therefore, a *sparsity* hypothesis on this transition matrix seems natural, and will help us recover its coefficients with much higher precision: the *number of nonzero coefficients $s$* in every row of $\theta$ is expected to play a critical role.

Beyond the sparsity level, we will see that the precision of estimation also depends on other critical parameters. The first one is the *fraction of observations* $p$, which compares the total information provided by passing trains to the overall size of the network. The second one is a kind of *signal-to-noise ratio* $\gamma \propto 1 + \frac{\omega^2}{\sigma^2}$, which in our case quantifies the relative importance of network congestion compared to all other sources of delay. These insights are critical to evaluate the reliability of our estimate $\widehat{\theta}$, and thus provide trustworthy information to passengers and traffic managers alike.

#### 1.2.4 Beyond Railway Applications

Of course, railway traffic analysis is not the only application of partially-observed state-space models. Here are a few other situations that could lend themselves to a similar approach:

- Road traffic analysis based on information transmitted by drivers, as is done by some popular smartphone apps;
- Patient load prediction in interlinked hospital services based on the individual lengths of stay;
- Interaction modeling within a social network based on a small sample of its contents or users.

In many cases, the underlying process of interest might not behave as a VAR process, but studying the linear Gaussian case hopefully gives interesting intuitions which extend to more general situations.

### 1.3 Related Works

The theory of VAR processes has been well-known for a long time: the book of Lütkepohl (2005) provides an extensive account. In particular, Chapter 3 describes three standard estimators for the VAR transition matrix:

- The *Maximum Likelihood Estimator* (MLE), based on maximizing the log-likelihood of the parameters;
- The *Conditional Least Squares* estimator, based on minimizing $\sum_t \|X_t - \theta X_{t-1}\|_2^2$;
- The *Method of Moments* estimator, based on the Yule-Walker equation recursively defining the covariance matrices (which is similar to the one we present in Section 2.2.1);

In the fully-observed Gaussian case, these three estimators are asymptotically equivalent. However, the equivalence no longer holds when observations become noisy or when some data goes missing. On the one hand, autoregressions with noisy or faulty measurements have been studied extensively for decades (Buonaccorsi, 2010, Section 12.3). On the other hand, much less research effort has been devoted to the situation we consider, namely the case of a vector-valued autoregression whose components are partially sampled in a random way.

### 1.3.1 Literature Review

In theory, the MLE is easily extended from fully-observed time series to partially- or noisily-observed time series, a.k.a. state-space models (Cappé et al., 2006). Most of the time, exact or approximate inference is achievable using some version of the Kalman filter (Kalman, 1960) or particle filter methods (Doucet et al., 2000), whereas parameter estimation typically involves the Expectation-Maximization (EM) algorithm (Shumway and Stoffer, 1982). Unfortunately, the EM algorithm is hard to analyze explicitly in terms of statistical error, which is why other methods are sometimes preferred in theoretical studies. In particular, plug-in methods using covariance estimates have been quite popular recently in the machine learning community. Our work is a direct continuation of this part of the literature, which we summarize below.

Since our model is designed to tackle real-world problems such as train delay prediction, a core challenge lies in the dimension $D$ of the VAR process $X_t$. On large networks, the transition matrix $\theta \in \mathbb{R}^{D\times D}$ may be impossible to estimate precisely without additional structural assumptions. This is why the latest approaches to high-dimensional VAR process estimation use sparsity penalties as a way to reduce data requirements and computational workload (see the book of Hastie et al., 2015).

In the literature on sparse learning, two closely related methods stand out. We describe them in the case of standard linear regression:

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{Z}.$$

When $\beta$ is known to be sparse, the most popular estimation procedure is the LASSO (Tibshirani, 1996), which requires solving the optimization problem

$$\min_{\beta} \quad \|\mathcal{Y} - \mathcal{X}\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \leq \lambda,$$

where the $\ell_1$ norm acts as a convex substitute to the $\ell_0$ "norm". During the last ten years, statisticians have begun extending this theory to random designs exhibiting correlations or missing data, starting with the seminal work of Loh and Wainwright (2012), who showed how to obtain statistical guarantees in spite of non-convexity. Other studies followed with a more precise focus on VAR processes, each one deriving slightly different non-asymptotic error bounds on the LASSO estimator (see for example Basu and Michailidis, 2015; Kock and Callot, 2015; Melnyk and Banerjee, 2016). Among the most recent works in this line of research, Jalali and Willett (2018) investigated a componentwise independent missing data scenario by using measure concentration for sub-Gaussian processes.

A few years after the LASSO emerged, another method for sparse estimation was introduced by Candes and Tao (2007): the Dantzig selector, which is the solution to

$$\min_{\beta} \quad \|\beta\|_1 \quad \text{s.t.} \quad \|\mathcal{X}'(\mathcal{Y} - \mathcal{X}\beta)\|_\infty \leq \lambda.$$

This time, the sparsity penalization is the objective, while data fidelity is enforced in the constraints. Although the LASSO and the Dantzig selector can be shown to exhibit similar behaviors (Bickel et al., 2009), the latter has some computational advantages. Not only is it the solution to a Linear Program, or LP (instead of a Quadratic Program for the LASSO), but this LP can even be parallelized across dimensions to speed up computations.

The most prominent application of the Dantzig selector to VAR estimation was proposed by Han et al. (2015). The error bounds they provided are much simpler than the ones obtained with LASSO-related methods, and they require less intricate assumptions. In the same line of research, Rao et al. (2017b) studied the more general scenario in which a hidden VAR process is randomly sampled or projected, and then corrupted with noise. They adapted the Dantzig selector method of Han et al. (2015) using custom concentration inequalities and obtained results that are quite similar to ours, although less generic in several aspects. In addition, it is our opinion that their proof is incomplete.[1]

A salient feature of our work is that we not only provide upper bounds on the error of a particular estimator, but we also investigate minimax lower bounds to prove optimality. To the best of our knowledge, this was only done in one previous study for (partially-observed) VAR processes. Rao et al. (2017b) presented a minimax lower bound on estimation error in a scenario very similar to ours, but we think that the proof of this result is incorrect as well.[2]

### 1.3.2 Contributions

When we compare it to the previously-surveyed state-of-the-art, our work brings the following novelties:

- We study random sampling mechanisms which are not independent across time or space, such as fixed-size sampling or Markov sampling. As we have seen in Section 1.2, such dependencies are sometimes necessary to approximate realistic situations.
- These assumptions force us to use slightly involved probabilistic results (combination of discrete and continuous concentration inequalities, recent results on Markov chain convergence) to provide upper bounds on the estimation error.
- We give a general minimax lower bound, which matches our upper bounds for non-Markov sampling mechanisms, proving the optimality of the sparse estimator for this task.
- Until the very end, our minimax proof is largely independent of the subset of admissible transition matrices, which makes it easy to handle many types of structured transitions: sparse, Toeplitz, banded, etc.
- We present extensive numerical experiments that empirically support our rates of convergence on simulated data, and we justify our model by applying it on real data.

1. Indeed, the combination of discrete and Gaussian concentration inequalities as performed on page 2 (middle of right column) of the supplementary material for Rao et al. (2017a) glosses over the fact that $L_F$ is itself a random variable. As we will discover during our own proof (Lemma 7), this introduces an additional difficulty and forces us to use a more complex discrete concentration inequality. See `https://web.stanford.edu/~milind/papers/system_id_icassp_proof.pdf` for the supplementary material in question.
2. In particular, the covariance matrix $\Phi_A$ presented on page 5 (top of right column) of the supplementary material for Rao et al. (2017b) does not seem suited to a VAR process initialized at $x_0 = 0$: its variance should increase with time as it converges to the stationary value we called $\Gamma_0(A)$. See `https://web.stanford.edu/~milind/papers/system_id_isit_proof.pdf` for the supplementary material in question.

### 1.4 Outline of the Paper

In Section 2, we define the generative procedure behind the partially-observed VAR process, and we present a sparse estimator $\widehat{\theta}$ of the transition matrix $\theta$. We then state both of our theoretical results in Section 3: a minimax lower bound on the error of any estimation procedure, and then an upper-bound on the error of our specific estimator. Section 4 contains numerical experiments demonstrating the influence of various parameters, as well as an application to a real-life railway data set. We conclude in Section 5.

Our statistical analysis heavily relies on preliminary groundwork laid out in Appendix A. Appendix B gives a detailed proof of the minimax lower bound, highlighting the role of Fano's method and Kullback-Leibler (KL) divergences. Appendix C on the other hand is dedicated to proving the convergence rate of the sparse estimator, by combining discrete and continuous concentration inequalities before analyzing the behavior of the Dantzig selector. A number of well-known results from linear algebra and probability are presented in Appendix D to make the present paper as self-contained as possible. Appendix E contains a glossary.

### 1.5 Notations

Assigning meaning to a symbol is done with $:=$. For any integer $n$, let $[n] := \{1, ..., n\}$. The symbol $\mathbf{1}_{\{...\}}$ stands for an indicator function. When dealing with random variables, we write $\mathbb{P}(X = x)$ for a probability density, $\mathbb{E}[X]$ for an expectation, $\mathrm{Var}[X]$ for a variance (scalar of vector) and $\mathrm{Cov}[X, Y]$ for a covariance (scalar or matrix). The symbols $\mathcal{B}(p)$, $\mathcal{B}(n, p)$ and $\mathcal{N}(\mu, \Sigma)$ denote a Bernoulli distribution, a binomial distribution and a (possibly multivariate) Gaussian distribution respectively. When we write $\log(x)$, we mean the natural (base-$e$) logarithm.

Given a real number $a$, we denote by $|a|$ its absolute value. Given a vector $x$, we denote by $\|x\|_2$ (resp. $\|x\|_1$, $\|x\|_\infty$, $\|x\|_0$) its Euclidean norm (resp. $\ell_1$ norm, $\ell_\infty$ norm, number of nonzero entries). The notation $\mathbf{1}_i$ stands for a vector with a single non-zero coordinate at position $i$, while $\mathbf{1} := \sum \mathbf{1}_i$.

A matrix can be defined by its coefficients $M = (M_{i,j})_{i,j}$ or by its blocks $M = (M_{[b_1,b_2]})_{b_1,b_2}$. We write $I$ for the identity matrix, and $J_r$ for the square matrix entirely filled with zeroes, except for the subdiagonal of rank $r$ which is filled with ones. The notation $\mathrm{diag}(\lambda)$ stands for the diagonal matrix with coefficients $\lambda_1, ..., \lambda_n$, while $\mathrm{bdiag}_T(M)$ stands for a block-diagonal matrix with $T$ copies of $M$ on the diagonal and zeroes elsewhere. We write $\mathrm{vec}(M)$ for the column-wise flattening of matrix $M$ into a vector. When we want to apply a function elementwise, we often use notation that is standard for real numbers but not for matrices: for instance, $\sqrt{M} := (\sqrt{M_{i,j}})_{i,j}$ and $1/M := (1/M_{i,j})_{i,j}$. Given a real matrix $M$, we denote by

- $M'$ its transposition, $M^+$ its Moore-Penrose pseudo-inverse and $M^{-1}$ its inverse;
- $\mathrm{Tr}(M)$ its trace and $\det(M)$ its determinant;
- $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$, $\lambda_i(M)$) its maximum (resp. minimum, $i$-th largest) eigenvalue, so that

$$\lambda_{\max}(M) = \lambda_1(M) \geq \lambda_2(M) \cdots \geq \lambda_n(M) = \lambda_{\min}(M)$$

- $\varsigma_{\max}(M)$ (resp. $\varsigma_{\min}(M)$, $\varsigma_i(M)$) its maximum (resp. minimum, $i$-th largest) singular value;
- $\|M\|_1 = \sup \frac{\|Mx\|_1}{\|x\|_1} = \max_j \sum_i |M_{i,j}|$ its operator $\ell_1$ norm, which is the maximum $\ell_1$ norm of a column of $M$;
- $\|M\|_2 = \sup \frac{\|Mx\|_2}{\|x\|_2} = |\varsigma_{\max}(M)| = \sqrt{\lambda_{\max}(M'M)}$ its operator $\ell_2$ norm, also known as the spectral norm;
- $\|M\|_\infty = \sup \frac{\|Mx\|_\infty}{\|x\|_\infty} = \max_i \sum_j |M_{i,j}|$ its operator $\ell_\infty$ norm, which is the maximum $\ell_1$ norm of a row of $M$;
- $\|M\|_F = \|\mathrm{vec}(M)\|_2 = \mathrm{Tr}(M'M)$ its Frobenius norm;
- $\|M\|_{\max} = \|\mathrm{vec}(M)\|_\infty = \max_{i,j} |M_{i,j}|$ the maximum absolute value of its entries;
- $\rho(M)$ its spectral radius.

See Petersen and Pedersen (2012) for a collection of inequalities relating all of these matrix quantities. Given two real matrices $A$ and $B$, we denote by

- $A \otimes B$ their Kronecker product;
- $A \odot B$ Hadamard (elementwise) product;
- $A \succeq B$ or $A \preceq B$ the (partial) Loewner order on symmetric matrices.

In all our derivations, the letter $c$ (or $c_1$, $c_2$, etc.) will denote a universal positive constant, which may change from one line to the next but never depends on any *varying* problem parameters. More specifically, statements involving it should always be understood as "there exists $c > 0$ such that"... As for other letters used throughout the paper, the most frequent ones are listed in Appendix E with their usual meanings.

## 2. The Partially-Observed VAR Process and its Sparse Estimator

Before stating our theoretical results, we first introduce our statistical model and the estimator we use.

### 2.1 Model Definition

The model we study can be described by the following generative procedure.

#### 2.1.1 Underlying State Process

We start by drawing $X = (X_{t,d})_{t\in[T],d\in[D]} \in \mathbb{R}^{TD}$ according to a stationary VAR process of order 1. This process has dimension $D$ and the following recursive definition:

$$X_t = \theta X_{t-1} + \varepsilon_t. \tag{4}$$

Here $\theta \in \mathbb{R}^{D\times D}$ is the transition matrix, taken from a row-sparse parameter set

$$\Theta_s = \{\theta \in \mathbb{R}^{D\times D} : \|\theta\|_2 \leq \vartheta < 1 \quad \text{and} \quad \forall i, \|\theta_{i,\cdot}\|_0 \leq s\}. \tag{5}$$

The innovations $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$ are Gaussian vectors of size $D$, independent across time. To ensure stationarity of the VAR process, we must impose $\rho(\theta) < 1$. Throughout the paper, we actually make the following (stronger) assumption, reflected in the definition of $\Theta_s$: there exists $\vartheta \in [0, 1[$ such that for all the values of $\theta$ we consider, $\|\theta\|_2 \leq \vartheta < 1$. Furthermore, we denote by $\sigma^2_{\min} := \lambda_{\min}(\Sigma)$ and $\sigma^2_{\max} := \lambda_{\max}(\Sigma)$ the minimum and maximum eigenvalues of $\Sigma$.

### 2.1.2 Generating the Observations

As announced, we do not have access to the latent process $X$ itself. To construct the observations $Y$, we first apply a sampling operation $X \mapsto \Pi X$, where $\Pi$ is a *random* binary matrix with one 1 per row. This constraint accounts for the fact that in our railway model, each observation depends on a single component of the hidden congestion process. Then, a vector $\eta$ of i.i.d. Gaussians with variance $\omega^2$ is added, and we observe the result. The formula reads

$$Y = \Pi X + \eta \qquad \text{or} \qquad Y_t = \Pi_t X_t + \eta_t \tag{6}$$

Note that the number of observations at $t$, which is equal to the number of rows of $\Pi_t$, is stochastic. In this sense, the generation of $\Pi_t$ also determines the number of components in $\eta_t$.

An essential hypothesis we make is the *mutual independence* between our three sources of randomness: the innovations $\varepsilon$, the sampling matrix $\Pi$ and the observation noise $\eta$ (at least once its dimension is known). Although this independence property is not satisfied in our initial use case (for the railway model, $\Pi$ can be influenced by $X$), it simplifies the analysis while retaining qualitative features of the model's behavior.

Our data set is built from $N$ independent realizations of this process, indexed by $n$. We denote by $X^{1:N}$ the collection $(X^n)_{n\in[N]}$, and the same goes for $\Pi^{1:N}$ and $Y^{1:N}$. The observed data at our disposal is the couple $(\Pi^{1:N}, Y^{1:N})$, while $X^{1:N}$ remains hidden. Note that the sampling matrix $\Pi^n$ may differ for every $n$.

For the sake of simplicity however, we will prove all convergence theorems in the case $N = 1$: extending those results to the general case will generally amount to replacing $T$ with $NT$ in the resulting error bounds.

### 2.1.3 Sampling Mechanisms

As stated in the beginning, the major feature of the present work is the nondeterministic selection of observed state components, that is, the fact that $\Pi$ is not a constant but a random variable following a known distribution $\mathcal{D}$. In order to sum up the amount of information available using only one parameter $p \in ]0, 1[$, we want $\mathcal{D}$ to satisfy the following condition: each component $X_{t,d}$ of the latent state must be activated (i.e. involved in an observation) on average $p$ times.

We now present three examples of sampling mechanisms satisfying this condition, which we will study in the rest of this paper:

1. *Independent sampling* $\mathcal{D}_{\mathsf{indep}}$: each index $(t, d)$ is selected with probability $p$, independently of all others.

2. *Fixed-size sampling* $\mathcal{D}_{\mathsf{fixed}}$: at each time step $t$, a number $pD$ of indices are drawn from $[D]$ with replacement (assuming $pD$ is an integer).

3. *Markov sampling* $\mathcal{D}_{\mathsf{Markov}}$: independently along each dimension $d$, time indices $t$ are selected according to a binary Markov chain with transition matrix $\mathcal{Q} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$ such that the chain is stationary with invariant measure $(\frac{b}{a+b}, \frac{a}{a+b}) = (1-p, p)$. We also assume there exists a universal constant $\chi$ such that $0 < \chi \leq a, b \leq 1 - \chi \leq 1$.

Although they may seem arbitrary, these three sampling mechanisms are each interesting for different reasons. Independent sampling is the simplest and most intuitive approach. Markov sampling provides a framework to study the effect of temporal dependencies, and it reduces to independent sampling when $a = 1 - b = p$. Finally, fixed-size sampling with replacement exhibits spatial dependency at each time step, and it is the only mechanism for which a single state component can be activated by multiple observations, which is crucial in practice when studying traffic data generated by dense transportation networks. In railway applications, the fixed size of the sample would be equal to the cumulated journey lengths of all trains during the observation period.

### 2.2 Sparse Transition Estimator

We now present the specific estimator we study.

#### 2.2.1 Estimator Construction

Our construction is a straightforward generalization of the one used by Rao et al. (2017a). As we will see in Lemma 11, the lag-$h$ covariance matrices of the VAR process $X$ are given by a simple recursion (see Lemma 11):

$$\Gamma_h(\theta) = \mathrm{Cov}_\theta[X_{t+h}, X_t] = \theta^h \Gamma_0(\theta) \tag{7}$$

Equation (7) is often called the Yule-Walker equation (for $h = 0$), and we can use it to define a simple estimation procedure:

1. For a given $h_0$, build estimators $\widehat{\Gamma}_{h_0}$ and $\widehat{\Gamma}_{h_0+1}$ of the covariances $\Gamma_{h_0}$ and $\Gamma_{h_0+1}$.
2. Use them to approximate the transition matrix, for example with $\widehat{\theta} = \widehat{\Gamma}_{h_0+1}\widehat{\Gamma}^+_{h_0}$.

The problem with this procedure is that is does not guarantee sparsity of $\widehat{\theta}$. To obtain a sparse result, Han et al. (2015) suggest casting the Yule-Walker Equation (7) as a soft constraint enforcing proximity between $\widehat{\Gamma}_{h_0+1}$ and $\widehat{\theta}\widehat{\Gamma}_{h_0}$. The algorithm they propose is a variant of the Dantzig selector (Candes and Tao, 2007) designed specifically for VAR processes. It requires solving the following constrained optimization problem:

$$\widehat{\theta} \in \operatorname*{argmin}_{M \in \mathbb{R}^{D \times D}} \|\mathrm{vec}(M)\|_1 \quad \text{s.t.} \quad \|M\widehat{\Gamma}_{h_0} - \widehat{\Gamma}_{h_0+1}\|_{\max} \leq \lambda_0 \tag{LP}$$

This problem can be reformulated as a linear program and decomposed along each dimension, which allows for an efficient and parallel solution procedure. The only thing left to do is decide how to estimate the covariance matrices.

#### 2.2.2 Moment-Based Covariance Estimation

Since $Y_t = \Pi_t X_t + \eta_t$ where $\eta_t$ is zero-mean, a natural proxy for $X_t$ is given by $\widehat{X}_t = \Pi_t^+ Y_t$. It would therefore seem logical to build an estimator of $\Gamma_h$ by plugging this proxy into the empirical covariance between $X_{t+h}$ and $X_t$. However, if we want it to work, we must make two adjustments to this idea:

1. To account for the random sampling, this plug-in empirical covariance must be scaled by a factor $S(h)$. Intuitively, since $\widehat{X}_{t+h}\widehat{X}_t'$ has a fraction $p^2$ of nonzero coefficients, we need to divide it by something close to $p^2$ to get an unbiased covariance estimator.

2. To account for the observation noise, we must incorporate an additive correction $C(h)$ multiplied by $\omega^2$. This correction becomes unnecessary for $h \geq 1$ since the observation noise $\eta_t$ is independent across time.

We obtain the following estimator:

$$\widehat{\Gamma}_h := \left[\frac{1}{S(h)} \odot \frac{1}{T-h} \sum_{t=1}^{T-h} \left(\Pi_{t+h}^+ Y_{t+h}\right)\left(\Pi_t^+ Y_t\right)'\right] - C(h). \tag{8}$$

The scaling matrix $S(h)$ and the noise correction $C(h)$ both depend on the sampling mechanism. Their expressions are given in Lemma 18.

## 3. Lower and Upper Bound on the Estimation Error

We now have all the necessary background to state our theoretical results.

### 3.1 Minimax Lower Bound

We start with a lower bound on the minimax estimation error for partially-observed Vector AutoRegressions. This lower bound is estimator-independent, and quantifies the intrinsic difficulty of our statistical problem. The term *minimax* means that we study the *worst-case probability of error* for the *best possible estimator*. One can picture it this way:

1. First, we pick an estimation algorithm $\widehat{\theta}$ among all measurable functions of the observations.

2. Then, the universe replies by choosing the worst possible true value $\theta \in \Theta_s$ in terms of estimator performance (see below).

3. To measure the performance of our estimator, we draw observations from $\mathbb{P}_\theta(Y, \Pi)$. For a given threshold $\zeta$, we then compute the probability that the $\ell_\infty$ operator distance between $\widehat{\theta}$ and $\theta$ exceeds $\zeta$.

In not so many words, the quantity of interest is

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_s} \mathbb{P}_\theta \left[\|\widehat{\theta} - \theta\|_\infty \geq \zeta\right],$$

and we want to find a threshold $\zeta$ such that the probability of exceeding it is non-negligible, for instance equal to $1/2$. The evolution of this threshold will tell us how the error behaves with respect to the various problem parameters.

**Theorem 1 (Minimax lower bound)** *Consider the partially-observed VAR model defined in Section 2.1 and suppose that $T$ is "large enough", as specified by Equations* (10) *and* (11)*. We define*

$$\gamma_\ell(\mathcal{D}) = (1-\vartheta)^{3/2}\frac{\mathbf{1}_{\mathcal{D}\neq\mathcal{D}_{\mathsf{fixed}}}\sigma_{\min}^2 + \omega^2}{\sigma_{\max}^2}.$$

*Then, for both non-Markov sampling mechanisms $\mathcal{D}_{\mathsf{indep}}$ and $\mathcal{D}_{\mathsf{fixed}}$, we have the lower bound*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta_s}\mathbb{P}_\theta\left[\|\widehat{\theta}-\theta\|_\infty \geq c\gamma_\ell(\mathcal{D})\frac{s}{p\sqrt{T}}\right] \geq \frac{1}{2}$$

*whereas for Markov sampling $\mathcal{D}_{\mathsf{Markov}}$ we have a weaker lower bound*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta_s}\mathbb{P}_\theta\left[\|\widehat{\theta}-\theta\|_\infty \geq c\gamma_\ell(\mathcal{D})\frac{s}{\sqrt{pq}\sqrt{T}}\right] \geq \frac{1}{2} \quad \textit{with} \quad q = \max\{1-b, 2p-(1-b)\}.$$

Note that our choice of norm is not only useful for the proof but coherent with our railway application: since the $\ell_\infty$ norm of a congestion vector $X_t$ represents a maximum amount of lost minutes on the network at time $t$, the induced operator norm $\|\theta\|_\infty$ controls the evolution of this maximum congestion through time.

The proof of this bound is based on *Fano's method*, which we sum up in Lemma 34. For a detailed presentation, we refer the reader to Tsybakov (2008, Chapter 2). Note that Wainwright (2019, Chapter 15) and Duchi (2019, Chapter 7) also offer good treatments of the subject.

Fano's method relies on choosing a set of parameters $\theta_0, \theta_1, ..., \theta_M$ satisfying two seemingly contradictory conditions: their induced distributions must be hard to distinguish, yet they must lie as fart apart from one another as possible. In particular, the crucial requirement of Fano's method is a tight upper bound on the KL divergence between two distributions generated by different parameters $\theta_i$ and $\theta_0$. More precisely, we want to bound

$$\begin{aligned}\frac{1}{M+1}\sum_{i=1}^{M}\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi, Y) \parallel \mathbb{P}_{\theta_0}(\Pi, Y)\right\} &\leq \max_i \mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi, Y) \parallel \mathbb{P}_{\theta_0}(\Pi, Y)\right\} \\ &\leq \max_i\left(\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi) \parallel \mathbb{P}_{\theta_0}(\Pi)\right\} + \mathbb{E}_\Pi\left[\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \parallel \mathbb{P}_{\theta_0}(Y|\Pi)\right\}\right]\right),\end{aligned}$$

where we used Lemma 35 for the last step. Since $\theta_i$ does not affect the distribution of the index set $\Pi$, the first term of the RHS is zero, and we will concentrate on the second term. First, we will upper-bound the random variable inside the expectation for a fixed value of $\Pi$, and then we will average said bound over all possible sampling matrices.

We now give the structure of the proof in a coherent order, along with the most important intermediate results:

1. Compute the conditional covariance $\mathrm{Cov}_\theta[Y|\Pi]$ and decompose it into a constant term $Q_\Pi$ (corresponding to the independent case $\theta = 0$) plus a residual $R_\Pi(\theta)$ (Lemma 12).

2. Upper-bound the conditional KL divergence $\mathrm{KL}\left\{\mathbb{P}_\theta[Y|\Pi] \parallel \mathbb{P}_0[Y|\Pi]\right\}$ using the "deviations from the identity" $\Delta_\Pi(\theta) = Q_\Pi^{-1/2}R_\Pi(\theta)Q_\Pi^{-1/2}$ (Lemma 13).

3. Control $\Delta_\Pi(\theta)$ using features of $R(\theta)$ scaled by factors that depend on the distribution of $\Pi$ (Lemmas 14, 15 and 16).

4. Deduce an upper bound on the KL divergence $\mathbb{E}_\Pi[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi) \parallel \mathbb{P}_0(Y|\Pi)\right\}]$ (Lemma 17).

5. Apply Fano's method to a set of parameters $\theta_i$ constructed from a pruned binary hypercube of well-chosen radius.

See Appendix B for the details of the argument.

### 3.2 Non-Asymptotic Error of the Sparse Estimator

Let us now move on to a positive result on the specific estimator we introduced earlier. We start by quantifying the non-asymptotic error of the covariance estimator.

**Theorem 2 (Convergence rate of the covariance estimator)** *Consider the covariance matrix estimator $\widehat{\Gamma}_h$ defined by Equation* (8) *and suppose that $T$ is "large enough", as specified by Equations* (16) *and* (18)*. Then for some well-chosen value of $\lambda_0$, the estimator $\widehat{\Gamma}_h$ satisfies*

$$\|\widehat{\Gamma}_h - \Gamma_h\|_{\max} \leq c\frac{\sigma_{\max}^2 + \omega^2}{(1-\vartheta)^2}\frac{\sqrt{\log(D/\delta)}}{p\sqrt{T}}$$

*with probability greater than $1-\delta$.*

From this, we deduce the convergence rate of the transition matrix estimator.

**Theorem 3 (Convergence rate of the transition matrix estimator)** *Consider the transition matrix estimator $\widehat{\theta}$ defined by the optimization problem* (LP) *with $h_0 = 0$. We impose the same conditions on $T$ as in Theorem 2, and we define*

$$\gamma_u(\theta) = \frac{\|\theta\|_\infty + 1}{(1-\|\theta\|_2)^2}\frac{\sigma_{\max}^2 + \omega^2}{\|\Gamma_0(\theta)^{-1}\|_1^{-1}}.$$

*Then for some well-chosen value of $\lambda_0$, the estimator $\widehat{\theta}$ satisfies*

$$\|\widehat{\theta} - \theta\|_\infty \leq c\gamma_u(\theta)\frac{s\sqrt{\log(D/\delta)}}{p\sqrt{T}}$$

*with probability greater than $1-\delta$.*

Our proof goes through the following steps:

1. Justify the formula for $\widehat{\Gamma}_h$ by computing $S(h)$ and $C(h)$ such that Equation (8) defines an unbiased estimator of $\Gamma_h$ (Lemma 18).

2. Fixing $d_1$ and $d_2$, reformulate $(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}$ using quadratic forms $g_i'\Psi_i'L_{ij}\Psi_j g_j$ of standard Gaussian vectors (Lemma 19).

3. Control the norms and traces of the $L_{ij}$ matrices using discrete concentration inequalities (Lemmas 20, 21, 22 and 23).

4. Apply a modified version of the Hanson-Wright inequality to bound the deviation of $g_i' \Psi_i' L_{ij} \Psi_j g_j$ using what we know about the $L_{ij}$ (Lemmas 24 and 25).
5. Conclude on $(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}$ and deduce a high-probability bound on $\|\widehat{\Gamma}_h - \Gamma_h\|_{\max}$.
6. Using an argument from Han et al. (2015) that exploits the optimization problem defining $\widehat{\theta}$, extract a high-probability bound on $\|\widehat{\theta} - \theta\|_\infty$ (Lemmas 26 and 27).

See Appendix C for the details.

### 3.3 Extension to VAR Processes of Higher Order

Although our theorems only apply to state-space models based on an underlying VAR process of order 1, we could try to extend them to the more general case of VAR($P$) processes. Just for this Section, suppose $X_t$ is no longer given by Equation (4), but instead satisfies:

$$X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + ... + \theta_P X_{t-P} + \varepsilon_t.$$

Then we can represent this as a VAR(1) process using augmented variables (Lütkepohl, 2005, Equation 2.1.8):

$$\widetilde{X}_t = \widetilde{\theta}\widetilde{X}_{t-1} + \widetilde{\varepsilon}_t \quad \text{with} \quad \widetilde{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-P+1} \end{pmatrix}, \widetilde{\varepsilon}_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and where the augmented transition matrix is given by a block companion matrix of dimension $\widetilde{D} = PD$:

$$\widetilde{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_{P-1} & \theta_P \\ I_D & 0 & \cdots & 0 & 0 \\ 0 & I_D & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_D & 0 \end{bmatrix}.$$

Unfortunately, by this reasoning, the distribution $\mathcal{D}$ chosen for $\Pi$ gives rise to a new distribution $\widetilde{\mathcal{D}}$ for $\widetilde{\Pi}$ which is not among $\{\mathcal{D}_{\mathsf{indep}}, \mathcal{D}_{\mathsf{fixed}}, \mathcal{D}_{\mathsf{Markov}}\}$. For instance, when $\mathcal{D} = \mathcal{D}_{\mathsf{indep}}$, the order-$P$ sampling matrix $\widetilde{\Pi}$ defines a stochastic process $\widetilde{\Pi}_t$ that is no longer independent but instead has a memory of size $P$. Therefore, the adaptation is not straightforward and would require a more careful proof.

### 3.4 Comments on Both Bounds

Let us now compare the convergence rate of Theorem 3, which is an upper bound on the error of the optimal algorithm, with the minimax lower bound of Theorem 1.

Our first and most important remark is that $s$ and $T$ play exactly the same roles in both bounds, which proves that *the dependency of the error in $s/\sqrt{T}$ is right* (up to a logarithmic factor in the upper bound).

### 3.4.1 Role of the Observation Probability

The fraction of observations $p$ appears as $1/p$ in the upper bound, regardless of the sampling mechanism. This dependency matches the minimax lower bound for $\mathcal{D}_{\mathsf{indep}}$ and $\mathcal{D}_{\mathsf{fixed}}$. However, for the Markov sampling mechanism $\mathcal{D}_{\mathsf{Markov}}$, the lower bound scales as $1/\sqrt{pq}$ instead. Subsequently, for this scenario, *we have not proven the optimality of either bound.* The numerical experiments of Section 4 will help shed light on this phenomenon.

However, it is reassuring to notice that when $a = 1 - b = p$, that is, when $\mathcal{D}_{\mathsf{Markov}}$ boils down to $\mathcal{D}_{\mathsf{indep}}$, the lower bound simplifies into the $1/p$ dependency we expect (since $q = \max\{1 - b, 2p - (1 - b)\} = p$).

### 3.4.2 Role of the Maximum Norm

Quite surprisingly, the maximum $\ell_2$ operator norm of the transition matrix, which we called $\vartheta$, plays opposite roles in both bounds. In the minimax lower bound, $(1 - \vartheta)$ appears in the numerator, whereas in the estimator's convergence rate it appears in the denominator. We do not know whether the respective exponents are optimal, but at least they are compatible with one another: as $\vartheta \to 1$, that is, as the VAR process becomes unstable, the lower bound tends to 0 and the upper bound to $+\infty$. This is simply a reflection of the fact that our proofs make heavy use of the distance between $\theta$ and the unit circle, which means they become invalid when $\theta$ gets too large.

### 3.4.3 Role of the Variances

The variances $\Sigma$ and $\omega^2$ appear in $\gamma_\ell(\mathcal{D})$ for the lower bound, and in $\gamma_u(\theta)$ for the upper bound. In both cases, the ratio $\gamma$ tells us if the underlying process is big enough to be detected among the noise. Roughly speaking, the magnitude of $X$ is related to the spectrum of $\Sigma$, while the magnitude of $Y$ is related to the spectrum of $\Sigma + \omega^2 I$. If the latter is significantly bigger than the former, recovering $X$ is a hopeless endeavor.

To simplify the comparison of both bounds, we assume that $\Sigma = \sigma^2 I$ is proportional to the identity matrix. We also assume that $\theta$ is normal, i.e. that it commutes with its transpose. This enables us to simplify the factor $\|\Gamma_0(\theta)^{-1}\|_1^{-1}$ appearing in $\gamma_u(\theta)$:

$$\|\Gamma_0^{-1}(\theta)\|_1^{-1} = \left\| \left( \sigma^2 (I - \theta\theta')^{-1} \right)^{-1} \right\|_1^{-1} = \sigma^2 \|I - \theta\theta'\|_1^{-1}.$$

We can then give a more precise expression of $\gamma_\ell$ and $\gamma_u$ (for all $\mathcal{D} \neq \mathcal{D}_{\mathsf{fixed}}$):

$$\gamma_\ell(\mathcal{D}) = (1 - \vartheta)^{3/2} \frac{\sigma_{\min}^2 + \omega^2}{\sigma_{\max}^2} = (1 - \vartheta)^{3/2} \frac{\sigma^2 + \omega^2}{\sigma^2}$$
$$\gamma_u(\theta) = \frac{\|\theta'\|_1 + 1}{(1 - \vartheta)^2} \frac{\sigma_{\max}^2 + \omega^2}{\|\Gamma_0(\theta)^{-1}\|_1^{-1}} = \frac{(\|\theta'\|_1 + 1)\|I - \theta\theta'\|_1}{(1 - \vartheta)^2} \frac{\sigma^2 + \omega^2}{\sigma^2}.$$

As we can see, in this simple case, we retrieve the same dependency in both bounds, namely

$$\gamma \propto 1 + \frac{\sigma^2}{\omega^2}.$$

We will now give a heuristic argument suggesting it may be optimal.

For that, let us consider the one-dimensional setting ($D = 1$) with full observations ($p = 1$), since we are mainly interested in the role of the parameters $\sigma^2$ and $\omega^2$. In this case, Theorem 1 argues that the error of any estimator should grow at least like $\gamma_\ell = 1 + \frac{\omega^2}{\sigma^2}$. We also note that in this simple case, Theorem 3 states that $\gamma_u \propto \gamma_\ell$.

We will compare this to the asymptotic error of the *Maximum Likelihood Estimator* (MLE) $\widehat{\theta}$, which (for well-behaved models) is given by the inverse of the *Fisher information matrix*. To make this statement more precise, we will invoke Douc et al. (2014, Proposition 2.14). Let us verify the conditions:

- The process is stable, i.e. $\rho(\theta) < 1$. We made sure of that by assuming $\|\theta\|_2 \leq \vartheta < 1$.
- The sampling matrix $\Pi_t$ is constant across time. Although this assumption is not essential, it is true here since $p = 1$ and $D = 1$ hence $\Pi_t = I_1$.
- The model has the smallest possible dimension.
- The true parameter $\theta$ is identifiable and does not lie on the boundary of $\Theta_s$. Identifiability is easily deduced from Lemma 11 by observing that $\theta = \Gamma_1(\theta)\Gamma_0(\theta)^{-1}$ can be entirely deduced from distribution moments.

Since all of these prerequisites hold in our simple setting, Douc et al. (2014, Proposition 2.14) gives us a Central Limit Theorem for the MLE of linear Gaussian models:

$$\sqrt{T}(\widehat{\theta} - \theta) \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathcal{I}_\infty(\theta)^{-1}) \quad \text{where} \quad \mathcal{I}_\infty(\theta) = \lim_{T\to\infty} \frac{\mathcal{I}_T(\theta)}{T}.$$

We only have to compute the Fisher information matrix $\mathcal{I}_T(\theta)$. The covariance matrix of $Y$ is given by Lemma 12, but in our case the sampling matrix is constant, and we obtain the simpler (unconditional) result

$$\mathrm{Cov}_\theta[Y] = (\sigma^2 + \omega^2) I_T + R(\theta),$$

where the residual $R(\theta)$ is of order 1 in $\theta$. Indeed, our simplifying assumptions imply $\Gamma_0(\theta) = \frac{\sigma^2}{1-\theta^2}$ and therefore

$$R(\theta) = \frac{\sigma^2}{1-\theta^2} \begin{pmatrix} \theta^2 & \theta^1 & \theta^2 & \cdots \\ \theta^1 & \theta^2 & \theta^1 & \\ \theta^2 & \theta^1 & \theta^2 & \\ \vdots & & & \ddots \end{pmatrix} \qquad \partial_\theta R(\theta) = \sigma^2 \begin{pmatrix} 0 & 1 & 0 & \cdots \\ 1 & 0 & 1 & \\ 0 & 1 & 0 & \\ \vdots & & & \ddots \end{pmatrix} + \mathcal{O}(\theta).$$

The Fisher information of $Y$ with respect to $\theta$ has an explicit formula (Malagò and Pistone, 2015, Section 3.5):

$$\begin{aligned} \mathcal{I}_T(\theta) &= \frac{1}{2} \operatorname{Tr}\left[\mathrm{Cov}_\theta[Y]^{-1} \partial_\theta \mathrm{Cov}_\theta[Y] \, \mathrm{Cov}_\theta[Y]^{-1} \partial_\theta \mathrm{Cov}_\theta[Y]\right] \\ &= \frac{1}{2} \operatorname{Tr}\left[\left(I + \frac{R(\theta)}{\sigma^2+\omega^2}\right)^{-1} \frac{\partial_\theta R(\theta)}{\sigma^2+\omega^2} \left(I + \frac{R(\theta)}{\sigma^2+\omega^2}\right)^{-1} \frac{\partial_\theta R(\theta)}{\sigma^2+\omega^2}\right]. \end{aligned}$$

If assume $\theta$ is small and perform a Taylor expansion, we get:

$$\mathcal{I}_T(\theta) \approx \frac{1}{2(\sigma^2+\omega^2)^2} \operatorname{Tr}\left[(\partial_\theta R(\theta))^2\right].$$

Incidentally, we also note that at the lowest order in $\theta$,

$$\operatorname{Tr}[(\partial_\theta R(\theta))^2] = \|\partial_\theta R(\theta)\|_F^2 \approx 2\sigma^4(T-1).$$

Which gives us an approximate information matrix for $T$ steps:

$$\mathcal{I}_T(\theta) \approx \frac{\operatorname{Tr}[(\partial_\theta R(\theta))^2]}{2(\sigma^2+\omega^2)^2} \approx \frac{T}{2}\left(\frac{\sigma^2}{\sigma^2+\omega^2}\right)^2.$$

Taking the temporal limit yields:

$$\mathcal{I}_\infty(\theta) = \lim_{T\to\infty} \frac{\mathcal{I}_T(\theta)}{T} \approx \frac{1}{2}\left(\frac{\sigma^2}{\sigma^2+\omega^2}\right)^2.$$

In conclusion, this informal analysis reveals an asymptotic error equivalent to

$$\frac{1}{\sqrt{T}}\sqrt{\mathcal{I}_\infty(\theta)^{-1}} \approx \frac{\sqrt{2}}{\sqrt{T}}\left(1+\frac{\omega^2}{\sigma^2}\right),$$

which is coherent with the dependency we identified in Theorem 1.

## 4. Numerical Illustrations

We start by illustrating our results on simulated data, and then move on to a real case study of railway event times.

All experiments were run on a Dell Precision 5530 mobile workstation with Intel Core i7-8850H CPU (2.60GHz × 12) and 31 GiB of RAM, running under Ubuntu 20.04. The code was written in Python (version 3.8) and it is available on GitHub[3] and Zenodo (Dalle and Castro, 2021), along with instructions to download the dataset we used.

Data preparation was performed using Pandas (Wes McKinney, 2010; Reback et al., 2021), graph stuctures were represented within NetworkX (Hagberg et al., 2008) while linear optimization problems were modeled using CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018) and solved with the ECOS solver (Domahidi et al., 2013).

### 4.1 Simulated Dataset

Simulating a partially-observed VAR process with known transition matrix $\theta$ allows us to compute the estimation error $\|\widehat{\theta}-\theta\|_\infty$ and study its dependence on parameters such as $T$, $D$, $p$, $\omega$, etc.

3. `https://github.com/gdalle/PartiallyObservedVectorAutoRegressions`

### 4.1.1 Data Generation

Each point on the result graphs below corresponds to one run of the algorithm, aimed at estimating a random value of $\theta$. These values for $\theta$ were drawn using independent standard Gaussian distributions for each matrix coefficient. They were subsequently normalized to satisfy $\|\theta\|_2 = \vartheta = 1/2$.

To simplify comparison with the theoretical bounds, we used a diagonal covariance matrix $\Sigma = \sigma^2 I$. The default sampling mechanism is $\mathcal{D}_{\mathsf{indep}}$. When not mentioned or plotted explicitly, all simulation parameters are equal to their default values given below:

$$T = 10000 \qquad D = 5 \qquad \sigma = 1.0 \qquad \omega = 0.1$$

Many of the following graphs are displayed with logarithmic scaling, in order to highlight the exponent of the dependencies. When a straight line is present on a plot, it is the result of a Siegel regression (a robust form of linear regression, see Siegel, 1982) applied to the points of the same color: its slope is denoted as $\alpha$ in the legend.

Most of the simulations are run in a dense estimation scenario. For those who require the sparse procedure, selecting a good regularization parameter $\lambda$ is paramount: indeed, Theorem 3 is only valid for a specific choice of $\lambda$ (which is not known in practice, but we can hope to approximate this near-optimal choice).

A standard way to tune $\lambda$ would be cross-validation. However, evaluating a choice of $\lambda$ (and the resulting estimate $\widehat{\theta}^{\lambda}$) requires inferring the hidden state sequence $X_t$ from the observations $Y$. If the sampling matrices $\Pi_t$ were known constants, the inference could be performed with Kalman filtering (Kalman, 1960), but since they are random, the distribution of $(X, Y)$ is no longer a Gaussian and the whole procedure breaks down. Finding an appropriate inference method in our setting will be the topic of future studies.

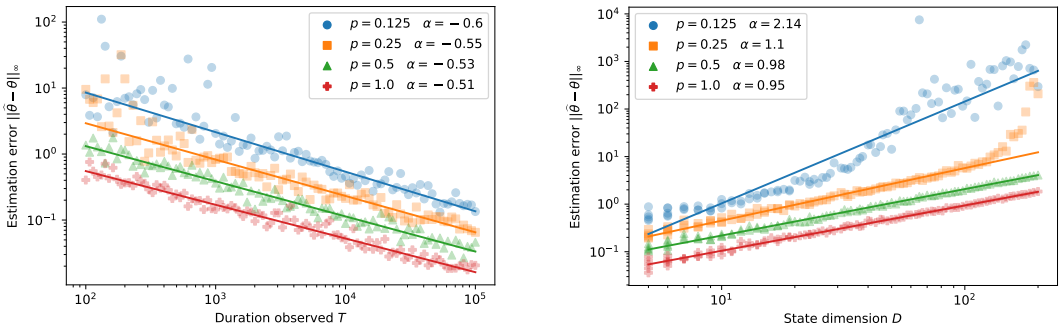
In the meantime, to tune $\lambda$, we resorted to a slightly unrealistic method which requires knowing (or guessing) the sparsity level of the real transition matrix $\theta$. Suppose we have an estimate $s_{\text{guess}}$ for the number of non-zero coefficients in each row of $\theta$: we can then use dichotomy on $\lambda$ to find a transition matrix estimate $\widehat{\theta}$ whose row sparsity level is also close to $s_{\text{guess}}$. Incidentally, this procedure allows us to study the effect of a wrong guess on the obtained sparsity level of $\widehat{\theta}$.

Another parameter we have to guess is the observation noise level $\omega$. Unless otherwise specified, we will assume that it is known to the estimator.

### 4.1.2 Results

We start by studying the impact of dimension parameters, as shown on Figure 1. As the number of time steps $T$ goes up, the estimation error goes down proportionally to $1/\sqrt{T}$. Note that this is only true because the sampling probability $p$ remains constant. If instead we had a limited observation budget but an increasing time precision, we would have $p \propto 1/T$, in which case the error would increase with $\sqrt{T}$ instead of decreasing.

The dimension $D$ of the transition matrix has the opposite effect and makes the error increase linearly (recall that here $s = D$). This is not surprising, since we measure the error with the $\ell_\infty$ operator norm, and this norm scales with the dimension of the matrix.




(a) Influence of $T$: error $\propto T^{-1/2}$

(b) Influence of $D$: error $\propto D^1$

Figure 1: Influence of dimension parameters on the estimation error (dense case)

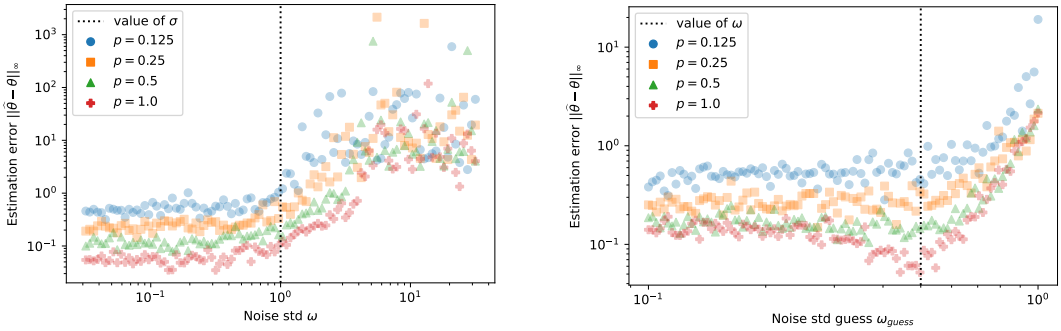
When $p$ is too small, the extreme parts of the curves (when $T$ is small or when $D$ is large) sometimes deviate from the linear dependency predicted by our analysis. We should expect this, since our theorems are only valid for a certain range of parameters where the error is small enough, which probably excludes these hard edge cases.

We briefly comment on the influence of the standard deviations $\sigma$ (for the innovations) and $\omega$ (for the observation noise). Since they only play a role through their ratio $\omega/\sigma$, we can keep $\sigma$ fixed and vary only $\omega$. As we can see on Figure 2, the higher the noise, the harder it becomes to extract a meaningful signal from the observations. More precisely, three phases are clearly identifiable. In the first one, corresponding to $\omega/\sigma \ll 1$, the error remains small and constant. Then, the error increases (presumably with slope 2) when $\omega/\sigma \sim 1$. In the third phase, corresponding to $\omega/\sigma \gg 1$, the error remains high and volatile, which is the only case not predicted by our theoretical analysis.

Other insights can be gained from looking at the role of our noise level guess $\omega_{\text{guess}}$. When much information is available, we see a clear optimum around the true value of $\omega$ (a region in which the error drops). However, in harder estimation settings (for instance with lower $p$), this effect is drowned by the noise, and we might as well pick $\omega_{\text{guess}} = 0$. This justifies our approach on the real-life data set.

Moving on to the influence of the sampling fraction $p$, we are pleased to notice that the slopes of Figure 3 reflect a linear dependency of the error in $1/p$ for both non Markovian sampling mechanisms. However, when it comes to Markov sampling, the convergence rate depends not only upon $p$, but also upon $b$, as can be deduced from the second row of plots.

This raises the question of the correct order of magnitude for the error in this scenario: is it really $1/p$, as our upper bound calculations suggest? Is it rather $1/\sqrt{pq}$, as we could wonder by looking at the minimax lower bound? Or is the true convergence rate somewhere in between? Figure 4 provides a partial answer by showing both the evolution of the error as a function of $a$ and $b$ (on the heatmap) and the level sets for $p$ and $\sqrt{pq}$ (with the white lines). The color map and the level values are represented with a logarithmic scale, as are the axes, to enable visual comparison. As we can see, neither of our candidate convergence rates perfectly fits the shape of the error, which suggests the real order of magnitude might be a more complex function of $(a, b)$.




(a) Influence of $\omega$: error $\propto 1 + \frac{\omega^2}{\sigma^2}$ (at first)

(b) Influence of $\omega_{\text{guess}}$

Figure 2: Influence of standard deviations on the estimation error (dense case)

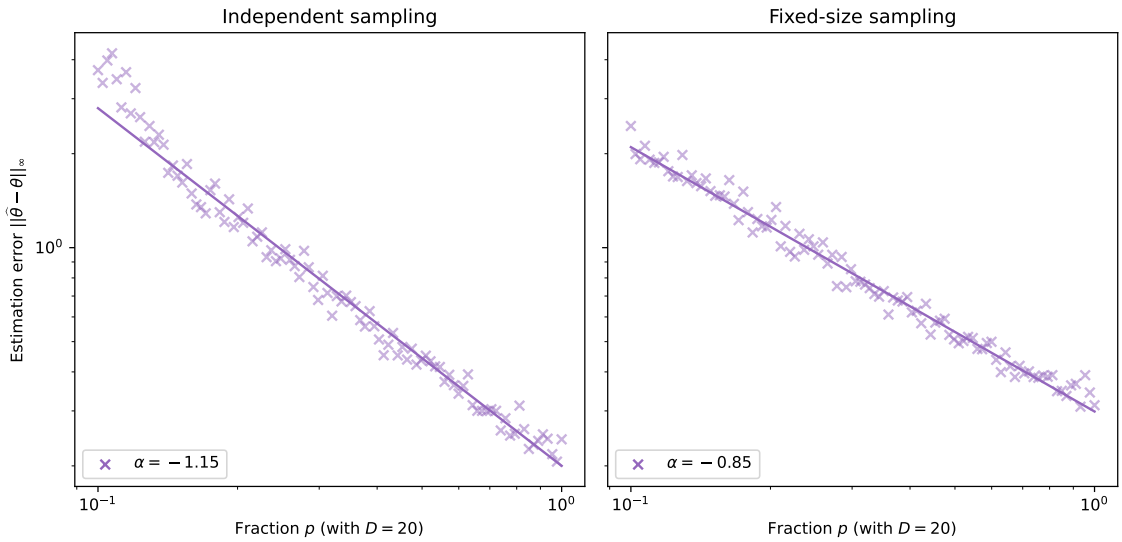
In particular, the beneficial influence of $1-b$ on Figure 3 suggests that the global number of observations (given by $p$) is not the only thing that matters. Indeed, if these observations are close to one another, they might bring more information regarding the transition structure than when they are equally numerous but more widely spaced. Since $1-b$ is the probability for a given dimension $d$ to remain active between $t$ and $t+1$, it measures the tendency of observations to cluster together, hence this interpretation is coherent with our numerical findings.

We continue with the influence of the sparsity level, both the real value $s$ and the guess $s_{\text{guess}}$ used to tune $\lambda$ in the estimation procedure. Figure 5 shows what happens when $D = 5$ is kept constant while the value of $s$ is increased (with $s_{\text{guess}} = s$). When we do that, the error rises as expected, but not with a slope of 1, which seemingly contradicts the linear relationship demonstrated by Theorem 3. Our interpretation of this behavior is that the function $\gamma_u(\theta)$ also depends upon the sparsity level in complicated ways, through the various norm terms (which are also affected by our renormalization to $\|\theta\|_2 = 1/2$). This may interfere with the linear dependency between $s$ and $\|\widehat{\theta} - \theta\|_\infty$ predicted by the theory.

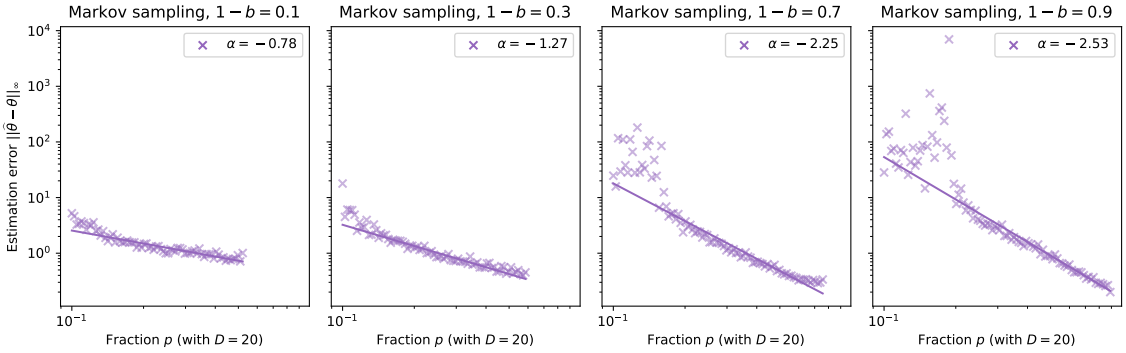
Meanwhile, Figure 6 displays another experiment where the true value of $s = 5$ is kept constant while $D$ is increased. We compare two scenarios: one in which we assume $\theta$ is dense ($s_{\text{guess}} = D$) and one in which we know $\theta$ is sparse ($s_{\text{guess}} = s$). As we can see, knowing the true sparsity level of $\theta$ enables the estimation error to rise much slower when compared to a naive dense estimation. However, Theorem 3 suggests that in the sparse case, the estimation error should actually be completely independent of $D$, which is not the case here. We think it may be due to our heuristic choice of $\lambda$ (selected by density matching) which does not guarantee an optimal speed of convergence.

We conclude with the influence of $h_0$, which is the smallest covariance lag used for estimating $\theta$. Our default procedure requires the estimators $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$, which means $h_0 = 0$, but nothing prevents us from using higher lags.

The upside of basing our estimation on (for instance) $\widehat{\Gamma}_1$ and $\widehat{\Gamma}_2$ is that we do not need to guess a value for $\omega$: indeed, the observation noise only appears in the formula for $\widehat{\Gamma}_0$, see Equation (8). Unfortunately, the higher the lag, the harder it becomes to correctly infer the




(a) Influence of $p$ for non-Markov sampling mechanisms




(b) Influence of $p$ for Markov sampling with different values of $b$

Figure 3: Influence of sampling parameters on the estimation error (dense case)

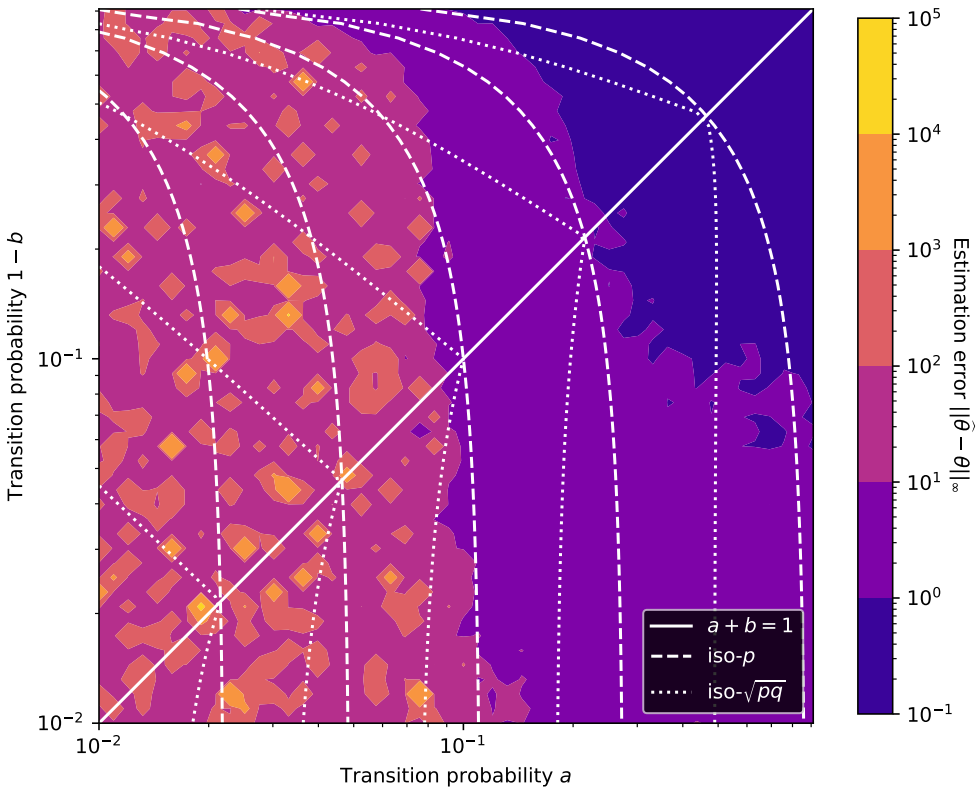



Figure 4: Joint influence of $a$ and $1-b$ for Markov sampling

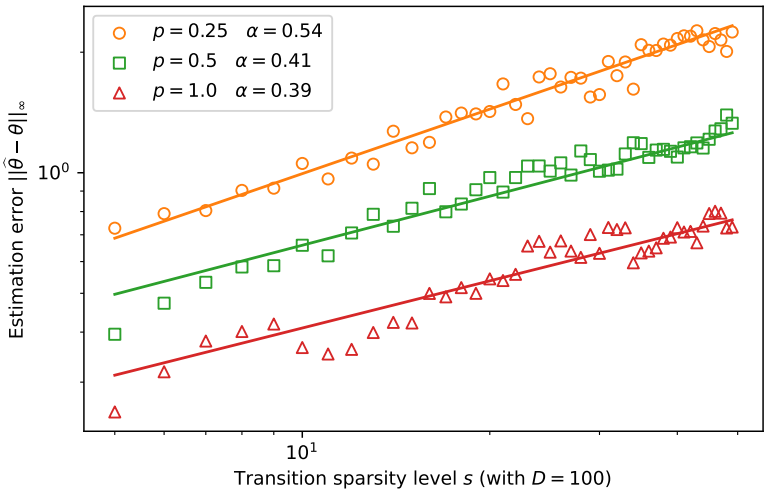



Figure 5: Influence of $s$ with fixed $D$

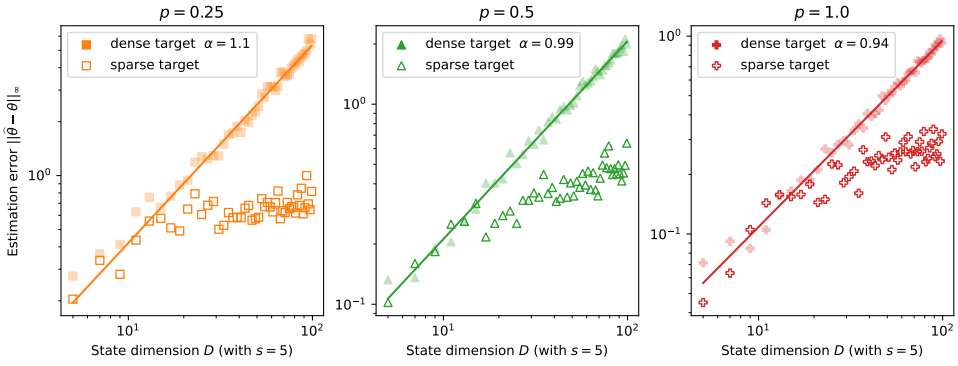



Figure 6: Influence of $D$ with fixed $s$

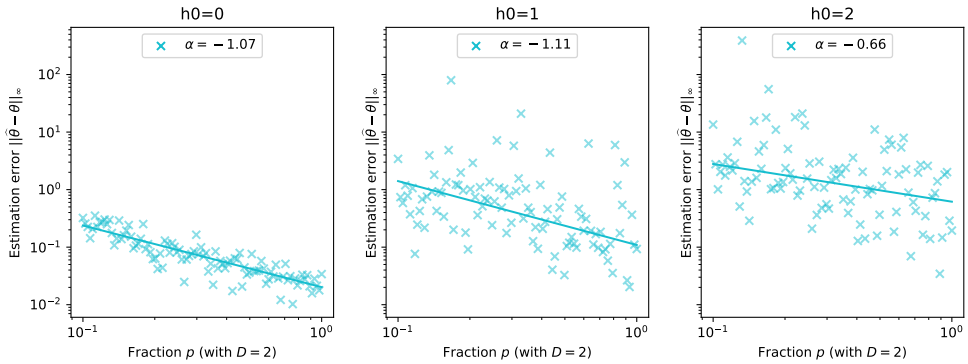



Figure 7: Influence of the covariance lag $h_0$

value of $\theta$, as can be seen on Figure 7: when $h_0 > 0$, both the error and its variance increase significantly. Therefore, when applying our method to the real data set, we will stick to the lowest lag $h_0 = 0$ and accept a small bias (due to the unknown value of $\omega$) in the hope of reducing the variance.

## 4.2 Real Dataset

We now turn to a real-life use case of partially-observed VAR processes: railway delay modeling. This application was described in Section 1.2, but now we go into more details regarding the data set and our analysis method.

### 4.2.1 Data Description and Preprocessing

Public transport agencies often make their theoretical transportation plan available (for instance using the General Transit Feed Specification format developed by Google), and many also provide an Application Programming Interface to query real-time traffic information. However, it is much harder to find large historical archives of *realized* event times. One such data set is available on the open data platform Mobility Switzerland[4].

Starting in January 2018, an increasing number of train arrival and departure times were pulled from the customer information systems of railway companies operating in Switzerland. These event times were then stored into daily CSV files, along with other useful information regarding each train[5]: company, line, trip and stop ID, possible perturbations (like cancelled trips or skipped stops).

Our intuition tells us that a congestion model such as ours best applies to a dense network with frequent trips, for instance that of a large urban or suburban area. As a consequence, we chose to focus on the tramway network of Zürich, operated by *Verkehrsbetriebe Zürich*[6]. We further restricted ourselves to the years 2018 and 2019, since data before 2018 is incomplete and data from 2020 onwards is likely to be affected by the ongoing Covid-19 crisis.

4. `https://opentransportdata.swiss/en/dataset`
5. `https://opentransportdata.swiss/en/dataset/istdaten`
6. `https://www.stadt-zuerich.ch/vbz/en/index.html`

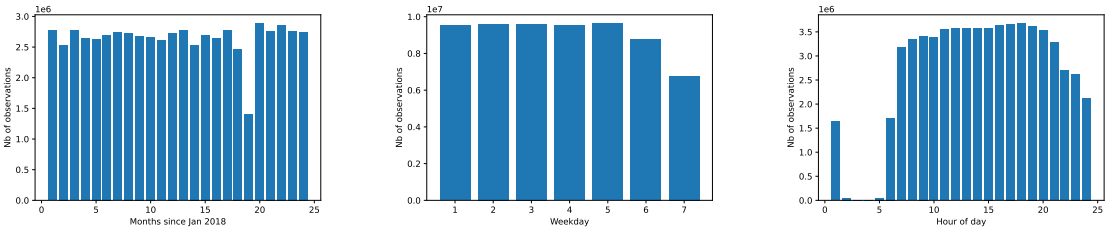
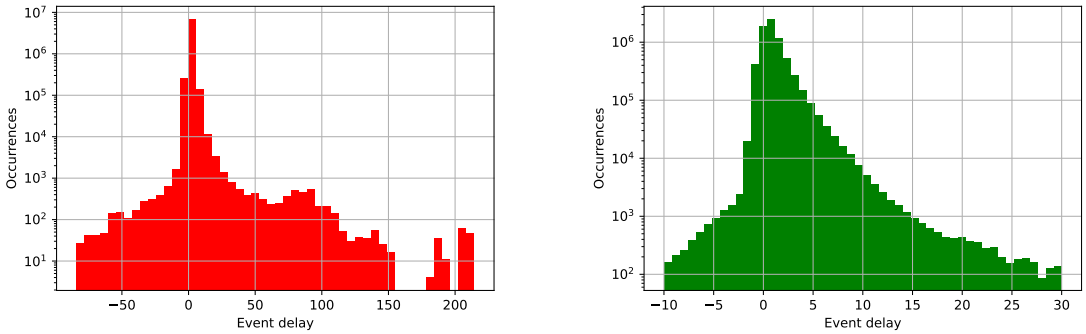



Figure 8: Number of observations for each month, weekday and hour on the Zürich tram data set




(a) Histogram of measured arrival delays before cleaning

(b) Histogram of measured arrival delays after cleaning

Figure 9: Effect of outlier filtering

Figure 8 gives an overview of the quantity of data available for these two years: these visualizations are important to prepare a homogeneous data set in terms of data quantity. Indeed, if trains are less frequent on a significant portion of the period we study, the congestion will propagate differently and our estimation procedure will be biased.

We notice that apart from July 2020, the months are relatively similar to one another. As for the weekdays, Saturdays and Sundays have fewer observations, which is why we exclude them from analysis. Finally, train frequency is zero at night but otherwise relatively constant through the day. Still, we choose to focus on what would be the evening "peak hour" in a typical urban network, that is from 5 pm to 8 pm.

Beyond this initial filtering, we applied a few more preprocessing steps. First, we removed skipped stops, unplanned and cancelled trips. We then removed departures to keep only arrival events. An important step consisted in detecting outliers by imposing limits on the minimum and maximum values for edge durations, arrival delays and additional edge delays. This is illustrated on Figure 9.

It may seem strange to keep slightly negative edge durations or delay values. We made this decision because the planned arrival times are rounded to the minute, while the actual event times are recorded with second-level precision. As a consequence, a train could appear to be a few tens of seconds late or early simply due to rounding phenomena.

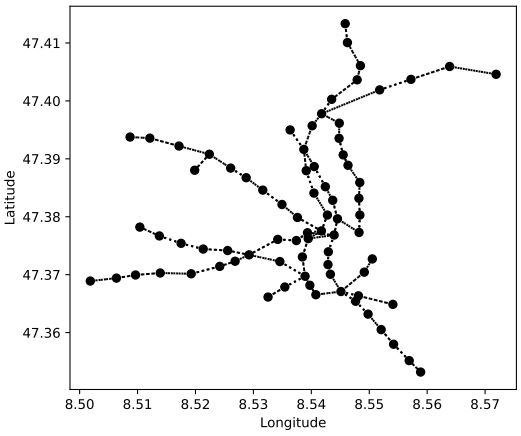



Figure 10: Map of the Zürich tram network generated from consecutive events

Then, we had to construct the graph representation $\mathcal{G}$ of our data set, with one vertex per stop and edges corresponding to railway tracks. We could have used a network map, but since we wanted the process to be automated, we chose to build the graph directly from our event data.

To each arrival event, we map the next event for the same train journey, which gives us a collection of station couples, also known as directed edges. Some of these edges are crossed frequently, whereas others are only used a few times over the two-year period. Since the precision of estimation relies on a good approximation of the congestion $X_{t,e}$ on each edge $e$, we must get rid of these infrequent edges. Indeed, keeping every single edge we obtained (there are over 2500) would result in a much higher dimension for the underlying process, but without sufficient data to exploit it.

Our pruning method consisted in selecting the 200 most frequently crossed edges in the complete network, and then retrieving the largest connected component of the resulting subgraph: it has 78 nodes and $D = 163$ edges. A graphical representation is given on Figure 10 (some edges are denser because of superposition).

The final preprocessing step required centering the data at expectation by removing additional delay averages for each edge. Indeed, for the real process, we suspected that the noise $\eta$ may not be zero-mean, so this was our method to standardize it.

### 4.2.2 Results

As we mentioned during the discussion on simulated data, cross-validation is difficult to achieve here due to the lack of a standard inference algorithm for hidden congestion values. Since in this case we don't have access to the sparsity level of the "true" $\theta$, we tried several values of the regularization parameter $\lambda$ and plotted the behavior of the resulting estimator $\widehat{\theta}^{\lambda}$. The main features of interest are presented on Figure 11. As expected, the

first graph shows the fraction of non-zero coefficients in $\widehat{\theta}^{\lambda}$ decreasing as the penalization $\lambda$ increases.

The second graph is more interesting: it depicts the evolution of a quantity characterizing the typical distance at which edges seem to interact, based on the estimated transition matrix. This quantity is computed as a weighted average of distances $d^{\mathcal{G}}_{e_1,e_2}$ between edge couples, each distance being weighed by the absolute value $|\widehat{\theta}^{\lambda}_{e_1,e_2}|$ of the relevant transition coefficient. In other words, this "average interaction distance" is given by the formula

$$\text{AID} = \frac{\sum_{e_1,e_2} |\widehat{\theta}^{\lambda}_{e_1,e_2}| d^{\mathcal{G}}_{e_1,e_2}}{\sum_{e_1,e_2} |\widehat{\theta}^{\lambda}_{e_1,e_2}|}.$$

Since graph distances are only defined between vertices, we need to specify what we mean with $d^{\mathcal{G}}_{e_1,e_2}$. If $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$, we define it as $d^{\mathcal{G}}_{e_1,e_2} = \min\{d^{\mathcal{G}}_{u_1,u_2}, d^{\mathcal{G}}_{u_2,u_1}\}$. It makes sense because we use the first vertex $u$ of an edge to determine the time step at which a train reaches it. Since $G$ was chosen to be connected, at least one of those two distances $d^{\mathcal{G}}_{u_1,u_2}$ and $d^{\mathcal{G}}_{u_2,u_1}$ will be finite. And since we use mean edge durations as weights, the resulting interaction distance will be expressed in minutes. One could say it measures the distance traveled by the congestion signal during a period of $\Delta t$, except that the distance is expressed in minutes (using the average train speed) instead of kilometers.

Our initial intuition for this propagation model is that the network congestion should propagate locally, from one edge to its neighbors, as time flows. And there is one clue on Figure 11 that supports this intuition: the fact that interaction distance decreases as the penalization becomes stronger.

For small values of $\lambda_0$, the average interaction distance stabilizes around a high value, which is close to the unweighted average of the distances between all pairs of edges $(e_1, e_2)$. In other words, no signal is captured. But as $\lambda_0$ rises, we see that the average interaction distance decreases, which suggests that *local effects start to prevail*. This behavior would not be observed if the transition matrix $\theta$ was completely independent of the graph structure $\mathcal{G}$.

In addition, when we pick a small interval $\Delta t$, the average interaction distance seems to stabilize at the end of the curve, whereas it quickly drops to zero for larger intervals. This suggests that picking $\Delta t$ close to the typical duration of an edge (1.5 min in our data set) may be a good idea.

Of course, there are still things we do not understand, such as the increasing behavior of the curve for $\Delta t = 5$ min or the sudden jump at the end of the one for $\Delta t = 7$ min. We assume these must be due to outliers in the data that only appear at a specific sampling frequency, not unlike a resonance phenomenon. We are also unsure how to compare these curves with one another quantitatively, since each of them captures interactions at a different timescale: if the "true model" was the one with $\Delta t = 1$, then each of these curves would roughly correspond to an estimation of $\theta^{\Delta t}$.

At any rate, we should keep in mind that our procedure is still just a linear Gaussian model, with very little fine-tuning for this specific use case. The fact that we recover a real-world intuition from railway experts is very encouraging, and suggests that we may be on the right path. However, building a more sophisticated predictor that takes into account more network and timetable features would undoubtedly lead to a better understanding of this phenomenon.

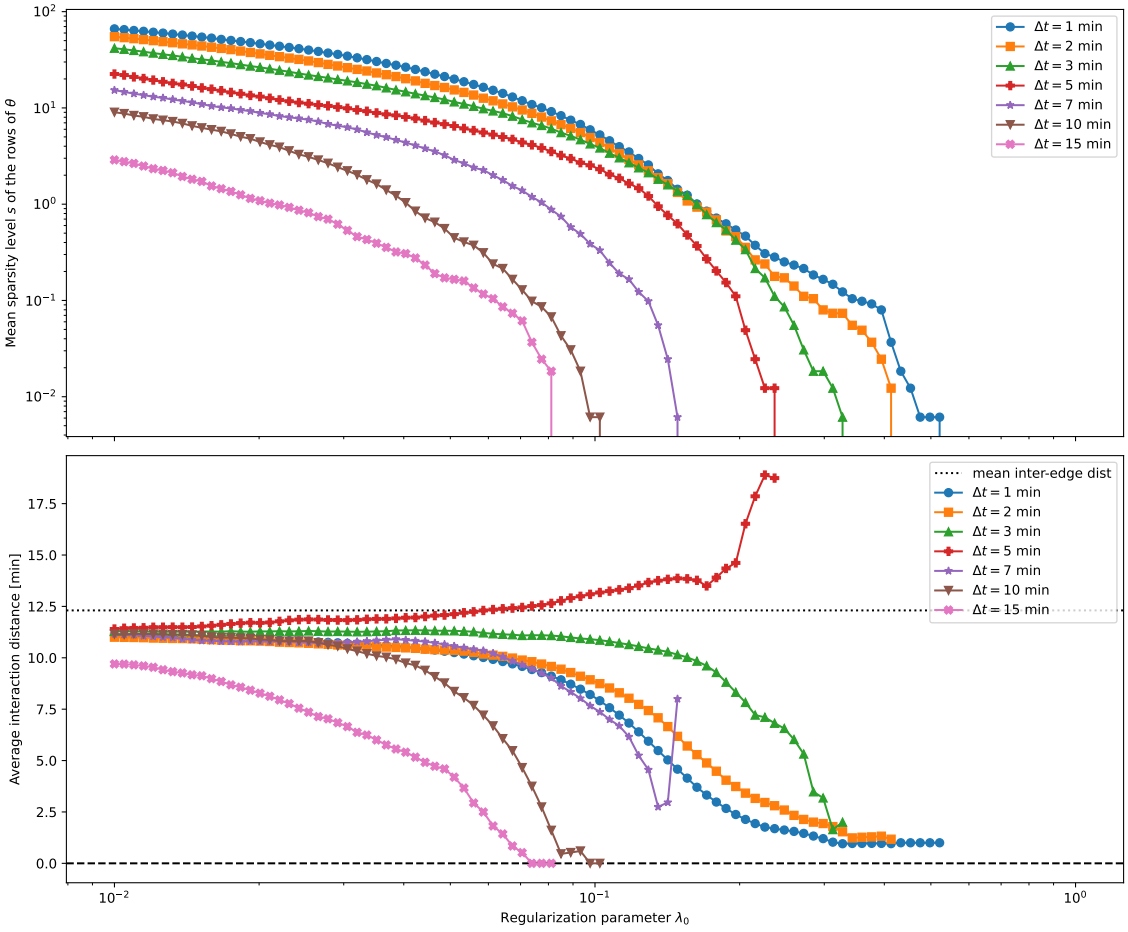



Figure 11: Effect of regularization and time discretization interval on some features of the estimate $\widehat{\theta}$

## 5. Conclusion and Perspectives

In this paper, we studied a partially-observed VAR process, whose coordinates are randomly sampled and corrupted with noise. The spatial and temporal correlations within the sampling matrices are a novel feature of our work, and combining both sources of randomness (discrete and continuous) required the use of tailored probabilistic methods. We provided upper and lower bounds for the optimal estimation error on the transition matrix, and found that these bounds roughly match. This analysis, supported by empirical results, sheds light on the intrinsic difficulty of such statistical problems, which arise naturally when analyzing several types of dynamic network processes.

However, our study leaves many questions open for future work:

- In public transport applications such as the one mentioned in Section 1.2, the sampling process may even be dependent from the underlying state. In this case, new methods must be introduced to handle such dependencies in error estimates.
- The simulations involving Markov sampling suggest that the true convergence rate of the transition estimator lies somewhere in the gap between our lower and upper bound. Finding more advanced methods to capture the effect of temporal dependency in sampling could provide more insight into the behavior of such processes.
- For practical purposes, finding a good compromise between model simplicity and statistical power is paramount to foster adoption in the field. As a consequence, more research is needed to transfer these theoretical ideas into fully usable real-time delay predictors that make use of all the available information.

More generally, we believe that the study of partially-observed evolving networks has a steadily growing number of applications, and the development of adequate mathematical tools will be a fruitful area of research for years to come.

## Acknowledgments

The authors would like to thank their colleague Axel Parmentier for his collaboration and careful proofreading. Clément Mantoux, Éloïse Berthier, Maxime Godin and Pierre Marion (by alphabetical order of first names) provided more support and advice than can be described in such a constrained space. We also thank Emeline Luirard for her help with a critical Lemma, and Mathieu Besançon for his last-minute look at the draft.

We are very grateful to the SNCF, especially its departments DGEX Solutions (SNCF Réseau) and Transilien (SNCF Voyageurs) for providing us with the inspiration behind this work. In particular, the opportunity to spend two days in a traffic regulation center was a great opportunity to gain a better understanding of how delays evolve. More generally, the collaboration with Bertrand Houzel and his team is always interesting and challenging. In particular, the models in this paper were very much improved thanks to fruitful discussions with Clément Raoux and Florian Chassagne.

## Appendix A. Proof Preliminaries

Here we present a few useful building blocks of the main proofs.

### A.1 Moments of the Sampling Distributions

For the rest of the appendix, we denote by $\kappa_{t,d}$ the integer random variable equal to the number of observations (that is, rows $o$ of $\Pi_t$) where column $d$ is activated (that is, in which coefficient $(\Pi_t)_{o,d}$ is equal to 1). We also define $\pi_{t,d}$ as the binary random variable equal to 1 if the column $d$ is activated in at least one observation row $o$ of $\Pi_t$. Since $\Pi_t$ is binary, these definitions amount to

$$\kappa_{t,d} = \sum_o (\Pi_t)_{o,d} \quad \text{and} \quad \pi_{t,d} = \mathbf{1}\{\kappa_{t,d} > 0\}. \tag{9}$$

We recall that the number of rows in $\Pi$ is itself random and corresponds to the number of observations. Moreover, each row of $\Pi_t$ has exactly one non-zero entry.

**Lemma 4** *The first- and second-order moments of $\kappa$ and $\pi$ are given by Tables 1 and 2 for all sampling mechanisms. Regardless of the sampling mechanism, every coefficient from the scaling matrix $S(h) = \mathbb{E}[\pi_{t+h}\pi_t']$ satisfies $S(h)_{d_1,d_2} \geq cp^2 =: S(h)_{\min}$.*

| $\mathbb{E}[\pi_{t,d}]$ | | | $\mathbb{E}[\kappa_{t,d}]$ |
|---|---|---|---|
| $\mathcal{D}_{\mathsf{indep}}$ | $\mathcal{D}_{\mathsf{fixed}}$ | $\mathcal{D}_{\mathsf{Markov}}$ | $\mathcal{D}_{\mathsf{fixed}}$ |
| $p$ | $1 - \left(1 - \frac{1}{D}\right)^{pD}$ | $p$ | $p$ |

Table 1: First-order moments of $\pi$ and $\kappa$ for all sampling mechanisms

| | $S(h)_{d_1,d_2} = \mathbb{E}\left[\pi_{t+h,d_1}\pi_{t,d_2}\right]$ | | | $\mathbb{E}\left[\kappa_{t+h,d_1}\kappa_{t,d_2}\right]$ |
|---|---|---|---|---|
| **Conditions** | $\mathcal{D}_{\mathsf{indep}}$ | $\mathcal{D}_{\mathsf{fixed}}$ | $\mathcal{D}_{\mathsf{Markov}}$ | $\mathcal{D}_{\mathsf{fixed}}$ |
| $d_1 = d_2$ and $h = 0$ | $p$ | $1 - \left(1 - \frac{1}{D}\right)^{pD}$ | $p$ | $p\left(1 - \frac{1}{D}\right) + p^2$ |
| $d_1 \neq d_2$ and $h = 0$ | $p^2$ | $1 - \left(1 - \frac{2}{D}\right)^{pD}$ | $p^2$ | $p^2 - \frac{p}{D}$ |
| $d_1 = d_2$ and $h \geq 1$ | $p^2$ | $\left(1 - \left(1 - \frac{1}{D}\right)^{pD}\right)^2$ | $p^2 + p(1-p)\times$ $(1-a-b)^h$ | $p^2$ |
| $d_1 \neq d_2$ and $h \geq 1$ | $p^2$ | $\left(1 - \left(1 - \frac{1}{D}\right)^{pD}\right)^2$ | $p^2$ | $p^2$ |

Table 2: Second-order moments of $\pi$ and $\kappa$ for all sampling mechanisms

**Proof** We must distinguish between each sampling mechanism. Let $i = (t+h, d_1)$ and $j = (t, d_2)$ be two indices in $[T] \times [D]$.

### A.1.1 Independent Sampling

For the independent sampling mechanism, each component of $X$ can be sampled at most once, hence $\kappa = \pi$ takes values in $\{0, 1\}$. Since $\pi_i$ has a Bernoulli distribution, we obviously have

$$\mathbb{E}[\pi_i] = \mathbb{E}[\pi_i^2] = p.$$

And if $i \neq j$, independence yields

$$\mathbb{E}[\pi_i \pi_j] = \mathbb{E}[\pi_i]\mathbb{E}[\pi_j] = p^2.$$

### A.1.2 Fixed-size Sampling

For the fixed-size sampling mechanism with replacement, we explore a case where $\kappa \neq \pi$, since the same component of $X$ can be sampled multiple times. We start with $\kappa_i$, which follows a binomial distribution $\mathcal{B}(pD, 1/D)$. In particular,

$$\mathbb{E}[\kappa_i] = \frac{pD}{D} = p.$$

As for the second-order moments, we can deduce $\mathbb{E}[\kappa_i^2]$ from the decomposition of the variance:

$$\mathbb{E}[\kappa_i^2] = \mathrm{Var}[\kappa_i] + \mathbb{E}[\kappa_i]^2 = pD\frac{1}{D}\left(1 - \frac{1}{D}\right) + p^2 = p\left(1 - \frac{1}{D}\right) + p^2.$$

For $i \neq j$, we have to consider the value of $h$. If $h \geq 1$, then $\kappa_{t+h}$ and $\kappa_t$ are independent, hence

$$\mathbb{E}[\kappa_i \kappa_j] = \mathbb{E}[\kappa_i]\mathbb{E}[\kappa_j] = p^2.$$

If $h = 0$, we remark that the whole vector $\kappa_t$ has a multinomial distribution with $pD$ trials and individual success probabilities $1/D$ for each dimension:

$$\mathbb{E}[\kappa_i \kappa_j] = \mathrm{Cov}[\kappa_i, \kappa_j] + \mathbb{E}[\kappa_i]\mathbb{E}[\kappa_j] = -pD\frac{1}{D}\frac{1}{D} + p^2 = -\frac{p}{D} + p^2.$$

We now turn to the variable $\pi$, which is only zero if every one of the $pD$ draws at time $t$ fails to select dimension $d_1$. As a consequence,

$$\mathbb{E}[\pi_i] = \mathbb{E}[\pi_i^2] = 1 - \left(\frac{D-1}{D}\right)^{pD}.$$

For $i \neq j$, we have a similar disjunction as before. If $h \geq 1$ then independence yields

$$\mathbb{E}[\pi_i \pi_j] = \mathbb{E}[\pi_i]\mathbb{E}[\pi_i] = \left(1 - \left(1 - \frac{1}{D}\right)^{pD}\right)^2,$$

whereas if $h = 0$ we have the slightly different expression

$$\mathbb{E}[\pi_i \pi_j] = 1 - \left(\frac{D-2}{D}\right)^{pD} = 1 - \left(1 - \frac{2}{D}\right)^{pD}.$$

We finally note that all the values in this column of Table 2 are greater than a constant times $p^2$. Indeed,

$$\left(1 - \left(1 - \frac{1}{D}\right)^{pD}\right)^2 \geq (1 - e^{-p})^2 \geq \left(-p + \frac{p^2}{2}\right)^2 = p^2(1 - p/2)^2 \geq cp^2.$$

### A.1.3 Markov Sampling

For the Markov sampling mechanism, each component of $X$ can be sampled at most once, hence $\kappa = \pi$. We once again have $\mathbb{E}[\pi_i] = \mathbb{E}[\pi_i^2] = p$. If $d_1 \neq d_2$, then the variables $\pi_i$ and $\pi_j$ belong to independent Markov chains, and thus $\mathbb{E}[\pi_i \pi_j] = p^2$. Otherwise, we have $i = (t+h, d)$ and $j = (t, d)$, which means these two variables are part of the same Markov chain. Stationarity yields

$$\mathbb{E}[\pi_i \pi_j] = \mathbb{P}[\pi_{t,d} = 1] \times \mathbb{P}[\pi_{t+h,d} = 1 | \pi_{t,d} = 1] = p(\mathcal{Q}^h)_{11}.$$

When diagonalizing $\mathcal{Q}$, we see that the bottom-right coefficient of $\mathcal{Q}^h$ is given by

$$(\mathcal{Q}^h)_{11} = \frac{a + b(1-a-b)^h}{a+b} = p + (1-p)(1-a-b)^h.$$

Plugging this in, we get

$$\mathbb{E}[\pi_i \pi_j] = p^2 + p(1-p)(1-a-b)^h.$$

■

## A.2 Concentration of $\frac{1}{T-h}\sum_t \pi_{t+h,d_1}\pi_{t,d_2}$

Here we study the product variables $\pi_{t+h,d_1}\pi_{t,d_2}$, seen as a stochastic process indexed by the time $t$. We aim to prove the following concentration result:

**Lemma 5 (Concentration of the sampling Bernoullis)** *There exist constants $c_1$ and $c_2$ such that for any $d_1, d_2$, for any bounded $h$, for any sampling mechanism, for all $u \in [0,1]$ (this restriction is important),*

$$\mathbb{P}\left(\left|\frac{1}{T-h}\sum_{t=1}^{T-h} \pi_{t+h,d_1}\pi_{t,d_2} - S(h)_{d_1,d_2}\right| \geq uS(h)_{d_1,d_2}\right) \leq c_1 \exp(-c_2 u^2 T S(h)_{d_1,d_2}).$$

**Proof** Once again, we must distinguish between sampling mechanisms. Before delving into the details, note that we only consider bounded values of $h$, so that $T - h \geq cT$. This is why the $T-h$ factor in the exponential can be simplified to $T$, as long as we are ready to accept a slightly smaller constant in front of it.

### A.2.1 Independent Sampling

We start by assuming $\Pi \sim \mathcal{D}_{\mathsf{indep}}$. When $d_1 = d_2$ and $h = 0$, or when $d_1 \neq d_2$, the process $\pi_{t+h,d_1}\pi_{t,d_2}$ is composed of independent Bernoulli variables. If $d_1 = d_2$ and $h \geq 1$ however, the Bernoulli variables $\pi_{t+h,d_1}\pi_{t,d_2}$ are no longer mutually independent.

To tackle this difficulty, we restrict ourselves to the subprocesses with indices that have the same remainder modulo $h+1$. Let us define $[T]_r^{h+1} = \{t \in [T] : t = r \mod h+1\}$ We easily remark that mutual independence holds again for the subprocesses $(\pi_{t+h,d_1}\pi_{t,d_2})_{t\in[T-h]_r^{h+1}}$ for each $r \in \{0, \dots, h\}$. This is illustrated on Figure 12: two pairs of variables linked with

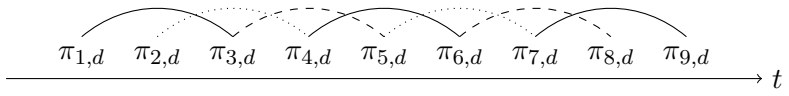



Figure 12: Illustration of alternate independence for $\mathcal{D}_{\mathsf{indep}}$ sampling with $h = 2$

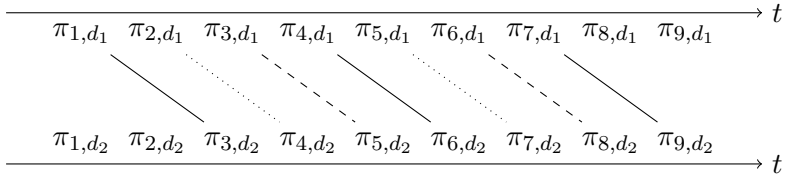



Figure 13: Illustration of alternate independence for $\mathcal{D}_{\mathsf{fixed}}$ sampling with $h = 2$

the same line style have empty intersection. Each of these $h+1$ separate subprocesses satisfies the Chernoff bound of Lemma 38: for all $u \in [0,1]$ and all $r \in [h+1]$,

$$\mathbb{P}\left(\left|\frac{h+1}{T-h}\sum_{t\in[T-h]_r^{h+1}} \pi_{t+h,d_1}\pi_{t,d_2} - S(h)_{d_1,d_2}\right| \geq uS(h)_{d_1,d_2}\right) \leq c_1 \exp\left(-c_2 u^2 \frac{T-h}{h+1} S(h)_{d_1,d_2}\right).$$

By the union bound,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{T-h}\sum_{t=1}^{T-h} \pi_{t+h,d_1}\pi_{t,d_2} - S(h)_{d_1,d_2}\right| \geq uS(h)_{d_1,d_2}\right) \\
&\quad\leq \sum_{r=0}^{h} \mathbb{P}\left(\left|\frac{h+1}{T-h}\sum_{t\in[T-h]_r^{h+1}} \pi_{t+h,d_1}\pi_{t,d_2} - S(h)_{d_1,d_2}\right| \geq \frac{u}{h+1}S(h)_{d_1,d_2}\right) \\
&\quad\leq c_1(h+1)\exp\left(-2\left(\frac{u}{h+1}\right)^2 \frac{T-h}{h+1}S(h)_{d_1,d_2}\right) \\
&\quad\leq c_1 \exp\left(-c_2 u^2 T S(h)_{d_1,d_2}\right)
\end{aligned}$$

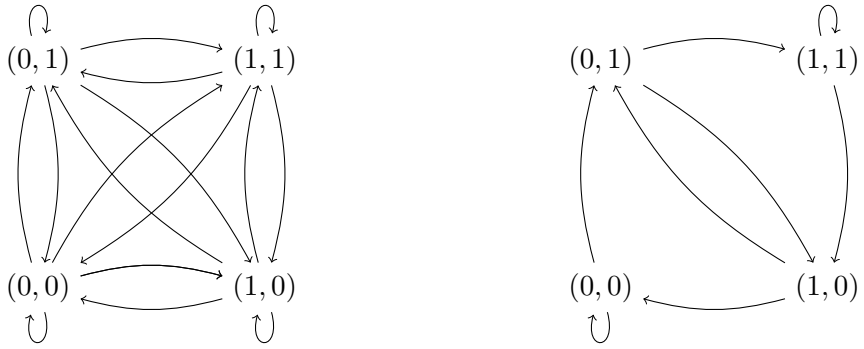
since we only consider bounded values of $h$.

### A.2.2 Fixed-size Sampling

Let us now assume that $\Pi \sim \mathcal{D}_{\mathsf{fixed}}$. When $d_1 = d_2$, or when $d_1 \neq d_2$ and $h = 0$, the situation is identical to the previous one, since we also have to deal with a sequence of independent Bernoulli variables. When $d_1 \neq d_2$ and $h \geq 1$ however, the variables $\pi_{t+h,d_1}\pi_{t,d_2}$ are no longer independent across time. Once again, considering $h+1$ subprocesses separately solves the issue. This is illustrated on Figure 13: although the variables in each column are (negatively) correlated, skipping enough steps restores independence.

### A.2.3 Markov Sampling

Finally, we delve into the scenario $\Pi \sim \mathcal{D}_{\mathsf{Markov}}$.




(a) When $d_1 \neq d_2$: transition matrix $\mathcal{Q} \otimes \mathcal{Q}$ (b) When $d_1 = d_2$: transition matrix $\mathcal{R}(1)$

Figure 14: State space and transitions for the Markov chain $(\pi_{t,d_2}, \pi_{t+1,d_1})_t$

- When $d_1 = d_2 = d$ and $h = 0$, the product $\pi_{t+h,d_1}\pi_{t,d_2}$ boils down to $\pi_{t,d}$, which is a 2-state Markov chain with transition matrix $\mathcal{Q}$.
- When $d_1 \neq d_2$, the couple $(\pi_{t,d_2}, \pi_{t+h,d_1})$ is a 4-state Markov chain with transition matrix $\mathcal{Q} \otimes \mathcal{Q}$ since the chains $\pi_{t+h,d_1}$ and $\pi_{t,d_2}$ along different dimensions are independent. Its state space diagram for $h = 1$ is given on Figure 14a.
- When $d_1 = d_2$ and $h \geq 1$, we must study the $(h+1)$-tuple $(\pi_{t,d_1}, \pi_{t+1,d_1}, ..., \pi_{t+h,d_1})$. It is a $2^{h+1}$-state Markov chain with transition matrix $\mathcal{R}(h)$. Its state space diagram for $h = 1$ is given on Figure 14b.

In all of these cases, our variable of interest $\pi_{t+h,d_1}\pi_{t,d_2}$ is a function of the underlying Markov chain. The relevant functions are:

$$f_1 : x \mapsto x \qquad f_2 : (x, y) \mapsto yx \qquad f_3 : (x_0, ..., x_h) \mapsto x_h x_0.$$

We note that since all the coefficients of $\mathcal{Q}$ are greater than $\chi$, all the coefficients of $\mathcal{Q} \otimes \mathcal{Q}$ are greater than $\chi^2$. We can even go further and state that the coefficients of $\mathcal{R}(h)^{h+1}$ are greater than $\chi^{h+1}$, because all pairs of states are connected after $h+1$ steps. Let us illustrate this phenomenon with $h = 1$:

$$\mathcal{R}(1) = \begin{pmatrix} 1-a & a & 0 & 0 \\ 0 & 0 & b & 1-b \\ 1-a & a & 0 & 0 \\ 0 & 0 & b & 1-b \end{pmatrix} \qquad \mathcal{R}(1)^2 = \begin{pmatrix} (1-a)^2 & a(1-a) & ab & a(1-b) \\ (1-a)b & ab & (1-b)b & (1-b)^2 \\ (1-a)^2 & a(1-a) & ab & a(1-b) \\ (1-a)b & ab & (1-b)b & (1-b)^2 \end{pmatrix}.$$

Subsequently, all the transition matrices $\mathcal{T}$ we are interested in, namely $\mathcal{T} \in \{\mathcal{Q}, \mathcal{Q} \otimes \mathcal{Q}, \mathcal{R}(h)^{h+1}\}$, satisfy the Doeblin condition with $r = h + 1$ and $\delta = \chi^{h+1}$:

$$\mathcal{T}^{h+1} \geq \chi^{h+1} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

Since we will only consider bounded values of $h$, and since $\chi$ is fixed for our purposes, these chains fulfill the conditions of Lemma 41. We thus conclude:

$$\mathbb{P}\left(\left|\frac{1}{T-h}\sum_{t=1}^{T-h}\pi_{t+h,d_1}\pi_{t,d_2} - S(h)_{d_1,d_2}\right| \geq uS(h)_{d_1,d_2}\right) \leq c_1\exp\left(-c_2u^2TS(h)_{d_1,d_2}\right).$$

■

### A.3 Conditional Gaussian Concentration

Here, we present a conditional version of a very useful Gaussian concentration inequality. We start with the unconditional case.

**Lemma 6 (Hanson-Wright inequality: Gaussian case)** *Let $A$ be a square matrix. If $X$ and $Y$ are two independent standard Gaussian vectors, we have:*

$$\mathbb{P}\left(|X'AX - \mathbb{E}[X'AX]| \geq u\right) \leq 2\exp\left(-c\min\left\{\frac{u^2}{\|A\|_F^2}, \frac{u}{\|A\|_2}\right\}\right)$$
$$\mathbb{P}\left(|X'AY - \mathbb{E}[X'AY]| \geq u\right) \leq 2\exp\left(-c\min\left\{\frac{u^2}{\|A\|_F^2}, \frac{u}{\|A\|_2}\right\}\right).$$

**Proof** See Vershynin (2018, Theorem 6.2.1) for the first inequality. We will see that it implies the second one. Let us define

$$\widetilde{A} = \begin{bmatrix} 0 & A \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \widetilde{X} = \begin{bmatrix} X \\ Y \end{bmatrix}.$$

We note that $\|\widetilde{A}\|_F = \|A\|_F$ and $\|\widetilde{A}\|_2 = \|A\|_2$. Applying the first inequality to $\widetilde{X}'\widetilde{A}\widetilde{X} = X'AY$ yields the expected result. ■

Now we move on to our custom conditional version.

**Lemma 7 (Conditional Hanson-Wright inequality)** *Let $A$ be a random square matrix such that with probability $1-\delta$,*

$$\|A\|_2 \leq M_2 \qquad \textit{and} \qquad \|A\|_F^2 \leq M_F^2.$$

*If $X$ and $Y$ are two independent standard Gaussian vectors independent of $A$, we have:*

$$\mathbb{P}\left(|X'AX - \mathbb{E}[X'AX]| \geq u\right) \leq \delta + 2\exp\left(-c\min\left\{\frac{u^2}{M_F^2}, \frac{u}{M_2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(A - \mathbb{E}[A])| \geq u/2\right)$$
$$\mathbb{P}\left(|X'AY - \mathbb{E}[X'AY]| \geq u\right) \leq \delta + 2\exp\left(-c\min\left\{\frac{u^2}{M_F^2}, \frac{u}{M_2}\right\}\right).$$

The additional trace term that appears inside Lemma 7 (as opposed to the non-conditional version of Lemma 6) is absent from the papers by Rao et al. (2017a,b), which is why we think their upper bound proofs are incomplete.

**Proof** We start with the first case. Since $A$ is a discrete random matrix with a finite set $\mathcal{A}$ of possible values,

$$\begin{aligned}\mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u) &= \sum_{a\in\mathcal{A}} \mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u \cap A = a) \\ &= \sum_{a\in\mathcal{A}} \mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u \cap A = a).\end{aligned}$$

Using independence between $X$ and $A$ gives us

$$\mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u) = \sum_{a\in\mathcal{A}} \mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u)\mathbb{P}(A = a).$$

We now split the set of feasible values $\mathcal{A}$ into

$$\mathcal{A}_{\leq} = \{a \in \mathcal{A} : \|a\|_F^2 \leq M_F^2\} \quad \text{and} \quad \mathcal{A}_{>} = \{a \in \mathcal{A} : \|a\|_F^2 > M_F^2\}.$$

Since we assumed $\mathbb{P}(A \in \mathcal{A}_{>}) = \sum_{a\in A_{>}} \mathbb{P}(A = a) \leq \delta$, we get:

$$\mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u) \leq \delta + \sum_{a\in\mathcal{A}_{\leq}} \mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u)\mathbb{P}(A = a).$$

Unfortunately, Lemma 6 only lets us bound

$$\mathbb{P}(|X'aX - \mathbb{E}[X'aX]| \geq u) \quad \text{and not} \quad \mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u)$$

(notice the change inside the expectation), which means we need an additional step. For a fixed $a \in \mathcal{A}_{\leq}$, we use independence and normality to obtain

$$\begin{aligned}\mathbb{E}[X'aX] - \mathbb{E}[X'AX] &= \mathbb{E}[\mathrm{Tr}(X'(a - A)X)] = \mathrm{Tr}(\mathbb{E}[XX'(a - A)]) \\ &= \mathrm{Tr}(\mathbb{E}[XX']\mathbb{E}[a - A]) = \mathrm{Tr}(a - \mathbb{E}[A]).\end{aligned}$$

We are now ready to decompose, with the help of the union bound:

$$\begin{aligned}\mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u) &= \mathbb{P}\left(|X'aX - \mathbb{E}[X'aX] + \mathbb{E}[X'aX] - \mathbb{E}[X'AX]| \geq u\right) \\ &\leq \mathbb{P}\left(|X'aX - \mathbb{E}[X'aX]| \geq u/2\right) + \mathbb{P}\left(|\mathbb{E}[X'aX] - \mathbb{E}[X'AX]| \geq u/2\right) \\ &\leq 2\exp\left(-c\min\left\{\frac{u^2}{\|a\|_F^2}, \frac{u}{\|a\|_2}\right\}\right) + \mathbf{1}\left\{|\,\mathrm{Tr}(a - \mathbb{E}[A])| \geq u/2\right\}.\end{aligned}$$

This implies:

$$\begin{aligned}\mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u) &\leq \delta + \sum_{a\in\mathcal{A}_{\leq}} \mathbb{P}(A = a)\mathbb{P}(|X'aX - \mathbb{E}[X'AX]| \geq u) \\ &\leq \delta + \sum_{a\in\mathcal{A}_{\leq}} \mathbb{P}(A = a) \times 2\exp\left[-c\min\left\{\frac{u^2}{\|a\|_F^2}, \frac{u}{\|a\|_2}\right\}\right] \\ &\quad + \sum_{a\in\mathcal{A}_{\leq}} \mathbb{P}(A = a) \times \mathbf{1}\left\{|\,\mathrm{Tr}(a - \mathbb{E}[A])| \geq u/2\right\}.\end{aligned}$$

By definition of $\mathcal{A}_{\leq}$,

$$\begin{aligned}\mathbb{P}(|X'AX - \mathbb{E}[X'AX]| \geq u) &\leq \delta + \sum_{a \in \mathcal{A}_{\leq}} \mathbb{P}(A = a) \times 2\exp\left(-c\min\left\{\frac{u^2}{M_F^2}, \frac{u}{M_2}\right\}\right) \\ &\quad + \mathbb{P}\left(|\operatorname{Tr}(A - \mathbb{E}[A])| \geq u/2\right) \\ &\leq \delta + 2\exp\left(-c\min\left\{\frac{u^2}{M_F^2}, \frac{u}{M_2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(A - \mathbb{E}[A])| \geq u/2\right).\end{aligned}$$

The proof for $X'AY$ follows the same lines, except that we replace $\mathbb{E}[XX'] = I$ by $\mathbb{E}[XY'] = 0$, which removes the trace term in the final expression. ■

### A.4 A Useful Pseudo-Inverse

Here we explore some properties of our sampling matrix $\Pi$ which do not depend upon the sampling mechanism.

**Lemma 8** *Let $A$ be a binary matrix with at most one 1 per row. Then we have:*

$$A'A = \operatorname{diag}\left\{\sum_k A_{k,j} : j \in [n]\right\}.$$

**Proof** We first recall that

$$(A'A)_{i,j} = \sum_l A_{l,i}A_{l,j}.$$

If $i \neq j$, since $A$ has at most one 1 per row, either $A_{l,i} = 0$ or $A_{l,j} = 0$. We thus have $A_{l,i}A_{l,j} = 0$ for all values of $l$, which implies $(A'A)_{i,j} = 0$. Otherwise, $A_{l,i}A_{l,j} = A_{l,j}$ which yields the expected result. ■

**Lemma 9** *Let $A$ be a binary matrix with at most one 1 per row. Then its pseudo-inverse is given by*

$$A^+_{i,j} = \frac{A_{j,i}}{\sum_k A_{k,i}},$$

*where we decide that $0/0 = 0$. It satisfies*

$$A^+A = \operatorname{diag}\left\{\mathbf{1}\Big[\sum_k A_{k,j} > 0\Big] : j \in [n]\right\}.$$

**Proof** Let $B_{i,j} = A_{j,i}/\sum_k A_{k,i}$ be our pseudo-inverse candidate. We compute

$$\begin{aligned}(BA)_{i,j} &= \sum_l B_{i,l}A_{l,j} = \frac{\sum_l A_{l,i}A_{l,j}}{\sum_k A_{k,i}} \\ (AB)_{i,j} &= \sum_l A_{i,l}B_{l,j} = \sum_l \frac{A_{i,l}A_{j,l}}{\sum_k A_{k,l}}.\end{aligned}$$

It is clear that both $(AB)_{i,j}$ and $(BA)_{i,j}$ are symmetric expressions in $(i,j)$. We will use the first expression to compute the products $ABA$ and $BAB$. We have three cases to consider:

- If $i \neq j$, since $A$ has at most one 1 per row, either $A_{l,i} = 0$ or $A_{l,j} = 0$. This means $A_{l,i}A_{l,j} = 0$ for all values of $l$, which implies $(BA)_{i,j} = 0$.
- If $i = j$ and the column $A_{\cdot,i}$ contains only zeroes, then $A_{l,i} = A_{l,j} = 0$ for all values of $l$, which also implies $(BA)_{i,j} = 0$.
- If $i = j$ and the column $A_{\cdot,i}$ contains at least one 1, then

$$(BA)_{i,j} = \frac{\sum_l A_{l,i}^2}{\sum_k A_{k,i}} = \frac{\sum_l A_{l,i}}{\sum_k A_{k,i}} = 1,$$

  since both numerator and denominator contain the sum of the same non-empty column.

In conclusion,

$$(BA)_{i,j} = \begin{cases} 1 & \text{if } i = j \text{ and } \sum_k A_{k,i} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

From this, we deduce

$$\begin{aligned}(ABA)_{i,j} &= \sum_l A_{i,l}(BA)_{l,j} = \sum_l A_{i,l}\mathbf{1}\left[i = j \text{ and } \sum_k A_{k,l} > 0\right] \\ &= A_{i,j}\mathbf{1}[\Sigma A_{\cdot,j} > 0] = A_{i,j} \\ (BAB)_{i,j} &= \sum_l (BA)_{i,l}A_{l,j} = \sum_l \frac{\mathbf{1}[i = j \text{ and } \sum_k A_{k,l} > 0]A_{j,l}}{\sum_k A_{k,l}} \\ &= \frac{\mathbf{1}[\sum_k A_{k,i} > 0]A_{j,i}}{\sum_k A_{k,i}} = \frac{A_{j,i}}{\sum_k A_{k,i}} = B_{i,j}.\end{aligned}$$

We thus verified the following conditions:

1. $ABA = A$;
2. $BAB = B$;
3. $AB$ symmetric;
4. $BA$ symmetric.

This characterizes $B$ as the pseudo-inverse $A^+$ of $A$. The expression of $A^+A$ is a by-product of the proof. ■

**Lemma 10** *The matrix* $\Pi$ *and its transpose* $\Pi'$ *satisfy:*

$$\Pi'\Pi = \text{diag}(\kappa).$$

*The matrix* $\Pi$ *and its pseudo-inverse* $\Pi^+$ *satisfy:*

$$\Pi^+ = \text{diag}(\pi/\kappa)\Pi' \qquad \textit{and} \qquad \Pi^+\Pi = \text{diag}(\pi).$$

*With the appropriate indices, these equalities also hold for* $\Pi_t$, $\Pi_t'$ *and* $\Pi_t^+$.

**Proof** These equalities are straightforward consequences of Lemmas 8 and 9. ■

## Appendix B. Proof of the Minimax Lower Bound

We now present the detailed proof of Theorem 1.

### B.1 Covariance Matrices

As we saw in the proof sketch, the KL divergence will be a crucial ingredient of our information-theoretical argument. To compute it, we need to know the covariance matrix of the observations, but first we turn to the underlying VAR process.

**Lemma 11 (VAR covariance matrix)** *The blocks of the covariance matrix for the stationary VAR process defined by Equation* (4) *have the following expression:*

$$\Gamma_0(\theta) = \mathrm{Cov}_\theta[X_t] = \sum_{k=0}^{\infty} \theta^k \Sigma \theta'^k$$
$$\Gamma_h(\theta) = \mathrm{Cov}_\theta[X_{t+h}, X_t] = \theta^h \Gamma_0(\theta).$$

In our opinion, the derivation of this covariance matrix in the proofs for Rao et al. (2017b) was incorrect, which invalidates the rest of their argument. Note however that in their setting, $X_0 = 0$ while we assume stationarity of the process.

**Proof** We start by noting that according to Equation (4), the stacked vector $X = (X_{t,d})_{(t,d)\in[T]\times[D]}$ follows a $TD$-dimensional centered multivariate Gaussian distribution. The covariance matrix of $X_t$ can be deduced from the recursion:

$$\Gamma_0(\theta) = \mathrm{Cov}_\theta[X_t] = \mathrm{Cov}_\theta[\theta X_{t-1} + \varepsilon_t] = \theta \,\mathrm{Cov}_\theta[X_{t-1}]\theta' + \Sigma = \theta\Gamma_0(\theta)\theta' + \Sigma.$$

There is a unique stationary solution:

$$\Gamma_0(\theta) = \sum_{k=0}^{\infty} \theta^k \Sigma \theta'^k.$$

The covariance matrix between $X_{t+h}$ and $X_t$ is obtained similarly:

$$\begin{aligned}\Gamma_h(\theta) = \mathrm{Cov}_\theta[X_{t+h}, X_t] &= \mathbb{E}[X_{t+h}X_t'] = \mathbb{E}[(\theta X_{t+h-1} + \varepsilon_{t+h})X_t'] \\ &= \theta \,\mathrm{Cov}_\theta[X_{t+h-1}, X_t] = \theta^h \,\mathrm{Cov}_\theta[X_t, X_t] = \theta^h \Gamma_0(\theta).\end{aligned}$$

And $\mathrm{Cov}_\theta[X_t, X_{t+h}] = \mathrm{Cov}_\theta[X_{t+h}, X_t]'$. In other words, we just proved that

$$\mathrm{Cov}_\theta[X] = \begin{bmatrix} \Gamma_0(\theta) & \Gamma_0(\theta)\theta'^1 & \Gamma_0(\theta)\theta'^2 & \cdots & \Gamma_0(\theta)\theta'^{T-1} \\ \theta^1\Gamma_0(\theta) & \Gamma_0(\theta) & \Gamma_0(\theta)\theta'^1 & & \\ \theta^2\Gamma_0(\theta) & \theta^1\Gamma_0(\theta) & \Gamma_0(\theta) & & \\ \vdots & & & \ddots & \\ \theta^{T-1}\Gamma_0(\theta) & & & & \Gamma_0(\theta) \end{bmatrix}.$$

■

As we announced in the proof sketch, our reference parameter will be $\theta = 0$, which is why it makes sense to express the conditional covariance of $Y$ as a deviation from the case without interactions. This is the aim of the following result.

**Lemma 12 (Conditional covariance decomposition)** *The covariance matrix of $Y$ given $\Pi$ decomposes as*

$$\mathrm{Cov}_\theta[Y|\Pi] = Q_\Pi + R_\Pi(\theta),$$

*where $Q_\Pi$ is a constant term and $R_\Pi(\theta)$ is a residual which vanishes as $\theta \to 0$. They are defined as follows: the constant term is*

$$Q_\Pi = \Pi(\mathrm{bdiag}_T\,\Sigma)\Pi' + \omega^2 I$$

*whereas the residual equals*

$$R_\Pi(\theta) = \Pi R(\theta)\Pi' \qquad \textit{with} \qquad R(\theta) = \begin{bmatrix} \theta\Gamma_0(\theta)\theta' & \Gamma_0(\theta)\theta'^1 & \Gamma_0(\theta)\theta'^2 & \cdots \\ \theta^1\Gamma_0(\theta) & \theta\Gamma_0(\theta)\theta' & \Gamma_0(\theta)\theta'^1 & \\ \theta^2\Gamma_0(\theta) & \theta^1\Gamma_0(\theta) & \theta\Gamma_0(\theta)\theta' & \\ \vdots & & & \ddots \end{bmatrix}.$$

**Proof** We use Equation (6) to see that the conditional distribution $\mathbb{P}_\theta[Y|\Pi]$ is a centered multivariate Gaussian with covariance $\mathrm{Cov}_\theta[Y|\Pi] = \Pi\,\mathrm{Cov}_\theta[X]\Pi' + \omega^2 I$. We then use Lemma 11 to get an expression of $\mathrm{Cov}_\theta[X]$ and conclude that its zero-order term (in $\theta$) is $\Pi\,\mathrm{bdiag}_T(\Sigma)\Pi' + \omega^2 I =: Q_\Pi$.

Finally, we define $R(\theta) = \mathrm{Cov}_\theta[X] - \mathrm{bdiag}_T\,\Sigma$ and $R_\Pi(\theta) = \Pi R(\theta)\Pi'$ to obtain $\mathrm{Cov}_\theta[Y|\Pi] = Q_\Pi + R_\Pi(\theta)$. The diagonal blocks of $R(\theta)$ are easily computed by noticing that $\Gamma_0(\theta) - \Sigma = \theta\Gamma_0(\theta)\theta'$. ■

### B.2 From the KL Divergence to $\Delta_\Pi(\theta)$

Judging by Lemma 12, choosing a parameter $\theta$ close to 0 yields a conditional distribution for $Y$ whose covariance is close to $Q_\Pi$. In the next result, we translate this into a bound on the KL divergence between $\mathbb{P}_\theta(Y|\Pi)$ and $\mathbb{P}_0(Y|\Pi)$.

**Lemma 13** *Recall that $Q_\Pi$ and $R_\Pi(\theta)$ are defined in the covariance decomposition of Lemma 12. Let us define the* deviation from the identity*:*

$$\Delta_\Pi(\theta) := Q_\Pi^{-1/2} R_\Pi(\theta) Q_\Pi^{-1/2}.$$

*Then the conditional KL divergence is upper-bounded by:*

$$\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi) \parallel \mathbb{P}_0(Y|\Pi)\right\} \le \frac{\|\Delta_\Pi(\theta)\|_F^2}{2(1 + \lambda_{\min}(\Delta_\Pi(\theta)))}.$$

**Proof** The conditional KL divergence $\mathrm{KL}\left\{\mathbb{P}_\theta[Y|\Pi] \parallel \mathbb{P}_0[Y|\Pi]\right\}$ can be bounded using Lemma 37. Indeed, both conditional distributions are Gaussian and have the same expectation, and covariance matrices that are "close" in the following sense: by Lemma 12,

$$\begin{aligned} \mathrm{Cov}_0(Y|\Pi) &= Q_\Pi = Q_\Pi^{1/2}(Q_\Pi^{1/2})' \\ \mathrm{Cov}_\theta(Y|\Pi) &= Q_\Pi + R_\Pi(\theta) = Q_\Pi^{1/2}\Big(I + \underbrace{Q_\Pi^{-1/2} R_\Pi(\theta) Q_\Pi^{-1/2}}_{\Delta_\Pi(\theta)}\Big)(Q_\Pi^{1/2})'. \end{aligned}$$

Remember that $\mathrm{Cov}_\theta(Y|\Pi) \succeq \omega^2 I \succ 0$ and the same goes for $Q_\Pi$. By Lemma 29,

$$\exists r \in [\varsigma_{\min}(Q_\Pi^{1/2})^2, \varsigma_{\max}(Q_\Pi^{1/2})^2], \quad \lambda_{\min}(\mathrm{Cov}_\theta(Y|\Pi)) = r\lambda_{\min}(I + \Delta_\Pi(\theta)).$$

Therefore,

$$\lambda_{\min}(I + \Delta_\Pi(\theta)) = 1 + \lambda_{\min}(\Delta_\Pi(\theta)) > 0$$

and we can apply Lemma 37 with $\mathbb{P}_1 = \mathbb{P}_\theta[Y|\Pi]$ and $\mathbb{P}_0 = \mathbb{P}_0[Y|\Pi]$. ■

### B.3 From $\Delta_\Pi(\theta)$ to $R_\Pi(\theta)$

Lemma 13 strongly suggests studying a certain fraction involving $\Delta_\Pi(\theta)$. In the following result, we boil it down to a function of the residual term $R_\Pi(\theta)$.

**Lemma 14** *Assume $\|R(\theta)\|_2 \leq \frac{\sigma_{\min}^2}{2}$. We have the following upper bound, which depends on our choice of sampling distribution $\mathcal{D}$:*

$$\frac{\|\Delta_\Pi(\theta)\|_F^2}{2(1 + \lambda_{\min}(\Delta_\Pi(\theta)))} \leq \frac{\|R_\Pi(\theta)\|_F^2}{\left[\mathbf{1}_{\mathcal{D}\neq\mathcal{D}_{\mathsf{fixed}}}\sigma_{\min}^2 + \omega^2\right]^2}.$$

**Proof** Since the quantity $\lambda_{\min}(\Delta_\Pi(\theta))$ in the denominator is hard to control, we will work with the spectral norm instead, since whenever $\|\Delta_\Pi(\theta)\|_2 < 1$ we have

$$\frac{1}{1 - \lambda_{\min}(\Delta_\Pi(\theta))} \leq \frac{1}{1 - \|\Delta_\Pi(\theta)\|_2}.$$

Let us start by noticing that, thanks to Lemma 31,

$$\|\Delta_\Pi(\theta)\|_F^2 = \|Q_\Pi^{-1/2} R_\Pi(\theta) Q_\Pi^{-1/2}\|_F^2 \leq \|Q_\Pi^{-1/2}\|_2^4 \|R_\Pi(\theta)\|_F^2 = \|Q_\Pi^{-1}\|_2^2 \|R_\Pi(\theta)\|_F^2$$
$$\|\Delta_\Pi(\theta)\|_2 = \|Q_\Pi^{-1/2}\Pi R(\theta)\Pi' Q_\Pi^{-1/2}\|_2 \leq \|Q_\Pi^{-1/2}\Pi\|_2^2 \|R(\theta)\|_2.$$

We will later see how the spectral and Frobenius norms of the full residual can be controlled as a function of $\theta$. For now, we must work to upper bound $\|Q_\Pi^{-1}\|_2$ and $\|Q_\Pi^{-1}\Pi\|_2^2$. To simplify the following proof, we write $\Sigma_{\mathrm{d}} := \mathrm{bdiag}_T\, \Sigma$. Since $\Sigma_{\mathrm{d}}$ is block-diagonal, its spectrum is the same as the spectrum of $\Sigma$ repeated $T$ times, hence $\lambda_{\min}(\Sigma_{\mathrm{d}}) = \sigma_{\min}^2$.

We start with $\|Q_\Pi^{-1}\|_2$. Since $Q_\Pi \succeq \omega^2 I \succ 0$ is non-singular and symmetric,

$$\|Q_\Pi^{-1}\|_2 = \lambda_{\max}(Q_\Pi^{-1}) = \frac{1}{\lambda_{\min}(Q_\Pi)} = \frac{1}{\lambda_{\min}(\Pi\Sigma_{\mathrm{d}}\Pi' + \omega^2 I)} = \frac{1}{\lambda_{\min}(\Pi\Sigma_{\mathrm{d}}\Pi') + \omega^2}.$$

Since $\Sigma_{\mathrm{d}} \succeq \sigma_{\min}^2 I$, we have $\Pi\Sigma_{\mathrm{d}}\Pi' \succeq \sigma_{\min}^2 \Pi\Pi'$ and thus

$$\|Q_\Pi^{-1}\|_2 \leq \frac{1}{\lambda_{\min}(\Pi\Pi')\sigma_{\min}^2 + \omega^2}.$$

We now continue with $\|Q_\Pi^{-1/2}\Pi\|_2^2$. By definition of the spectral norm,

$$\|Q_\Pi^{-1/2}\Pi\|_2^2 = \lambda_{\max}\left[\Pi' Q_\Pi^{-1}\Pi\right] = \lambda_{\max}\left[\Pi'(\Pi\Sigma_{\mathrm{d}}\Pi' + \omega^2 I)^{-1}\Pi\right].$$

We use the observation

$$\Pi\Sigma_{\mathrm{d}}\Pi' + \omega^2 I \succeq \sigma_{\min}^2 \Pi\Pi' + \omega^2 I$$

to deduce that, since matrix inversion is decreasing w.r.t. the Loewner order on positive semi-definite matrices,

$$(\Pi\Sigma_{\mathrm{d}}\Pi' + \omega^2 I)^{-1} \preceq (\sigma_{\min}^2 \Pi\Pi' + \omega^2 I)^{-1}$$
$$\Pi'(\Pi\Sigma_{\mathrm{d}}\Pi' + \omega^2 I)^{-1}\Pi \preceq \Pi'(\sigma_{\min}^2 \Pi\Pi' + \omega^2 I)^{-1}\Pi.$$

It follows that

$$\begin{aligned}\|Q_\Pi^{-1/2}\Pi\|_2^2 &\leq \lambda_{\max}\left[\Pi'(\sigma_{\min}^2\Pi\Pi' + \omega^2 I)^{-1}\Pi\right] \\ &= \frac{1}{\sigma_{\min}^2}\lambda_{\max}\left[\Pi'\left(\Pi\Pi' + \frac{\omega^2}{\sigma_{\min}^2} I\right)^{-1}\Pi\right].\end{aligned}$$

By Lemma 33,

$$\|Q_\Pi^{-1/2}\Pi\|_2^2 \leq \frac{1}{\sigma_{\min}^2}\frac{\lambda_{\max}(\Pi\Pi')}{\frac{\omega^2}{\sigma_{\min}^2} + \lambda_{\max}(\Pi\Pi')} = \frac{\lambda_{\max}(\Pi\Pi')}{\lambda_{\max}(\Pi\Pi')\sigma_{\min}^2 + \omega^2} \leq \frac{1}{\sigma_{\min}^2}.$$

Another partial conclusion is within reach:

$$\begin{aligned}\frac{\|\Delta_\Pi(\theta)\|_F^2}{1 + \lambda_{\min}(\Delta_\Pi(\theta))} &\leq \frac{\|\Delta_\Pi(\theta)\|_F^2}{1 - \|\Delta_\Pi(\theta)\|_2} \leq \frac{\|Q_\Pi^{-1}\|_2^2\|R_\Pi(\theta)\|_F^2}{1 - \|Q_\Pi^{-1/2}\Pi\|_2^2\|R(\theta)\|_2} \\ &\leq \frac{\|Q_\Pi^{-1}\|_2^2\|R_\Pi(\theta)\|_F^2}{1 - \frac{1}{\sigma_{\min}^2}\|R(\theta)\|_2} \leq \frac{\|Q_\Pi^{-1}\|_2^2\|R_\Pi(\theta)\|_F^2}{1 - \frac{1}{2}} \\ &= \frac{2\|R_\Pi(\theta)\|_F^2}{\left[\lambda_{\min}(\Pi\Pi')\sigma_{\min}^2 + \omega^2\right]^2}.\end{aligned}$$

To make sure that this holds, we only need to assume that $\|\Delta_\Pi(\theta)\|_2 < 1$, which is implied by $\|R(\theta)\|_2 \leq \frac{1}{2}\sigma_{\min}^2 \leq \frac{1}{2}\|\Sigma\|_2$.

The final step requires lower-bounding the eigenvalue $\lambda_{\min}(\Pi\Pi')$. Let us first recall that $\Pi\Pi'$ and $\Pi'\Pi$ have the same non-zero eigenvalues, and that $\Pi'\Pi = \mathrm{diag}(\kappa)$. Thus, if $\Pi\Pi'$ is non-singular then

$$\lambda_{\min}(\Pi\Pi') = \min\{\kappa_{t,d} : \kappa_{t,d} > 0\} \geq 1.$$

For the sampling mechanisms $\mathcal{D}_{\mathsf{indep}}$ and $\mathcal{D}_{\mathsf{Markov}}$, each state component is sampled at most once, hence $\Pi\Pi'$ is almost surely non-singular. For $\mathcal{D}_{\mathsf{fixed}}$ on the other hand, the probability of singularity is non-zero. To simplify the rest of the argument, we will use the crude lower bound $\lambda_{\min}(\Pi\Pi') \geq 0$ for this sampling mechanism, even though it is clearly suboptimal. We obtain the final bound:

$$\frac{\|\Delta_\Pi(\theta)\|_F^2}{1 + \lambda_{\min}(\Delta_\Pi(\theta))} \leq \frac{2\|R_\Pi(\theta)\|_F^2}{\left[\mathbf{1}_{\mathcal{D}\neq\mathcal{D}_{\mathsf{fixed}}}\sigma_{\min}^2 + \omega^2\right]^2}.$$

■

### B.4 From $R_\Pi(\theta)$ to $R(\theta)$

As the previous lemma underlines, the last step we need to get rid of the dependency in $\Pi$ is to study the average norm of $R_\Pi(\theta)$.

**Lemma 15** *If $\Pi$ is distributed according to $\mathcal{D}_{\mathsf{indep}}$ or $\mathcal{D}_{\mathsf{fixed}}$ then*

$$\mathbb{E}\left[\|R_\Pi(\theta)\|_F^2\right] \leq p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + p^2\|R(\theta)\|_F^2,$$

*whereas if $\Pi$ follows $\mathcal{D}_{\mathsf{Markov}}$ then*

$$\mathbb{E}\left[\|R_\Pi(\theta)\|_F^2\right] \leq p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + pq\|R(\theta)\|_F^2.$$

Note that when $p = 1 = a = 1 - b$ (full observation), our bounds have the same order of magnitude as for the full residual.

**Proof** We first notice that for any matrix $A$,

$$\begin{aligned}\mathbb{E}[\|\Pi A\Pi'\|_F^2] &= \mathbb{E}\,\mathrm{Tr}\left[\Pi A\Pi'\Pi A'\Pi\right] = \mathbb{E}\,\mathrm{Tr}\left[\,\mathrm{diag}(\kappa)A\,\mathrm{diag}(\kappa)A'\right]\\ &= \mathbb{E}\sum_i \kappa_i (A\,\mathrm{diag}(\kappa)A')_{i,i} = \mathbb{E}\sum_i \kappa_i \sum_j A_{i,j}\kappa_j A'_{j,i} \sum_{i,j}\mathbb{E}[\kappa_i\kappa_j]A_{i,j}^2.\end{aligned}$$

We can apply this to $R_\Pi(\theta) = \Pi R(\theta)\Pi'$:

$$\mathbb{E}[\|R_\Pi(\theta)\|_F^2] = \sum_{i,j}\mathbb{E}[\kappa_i\kappa_j]R(\theta)_{i,j}^2.$$

The rest of the proof consists in instantiating this formula using the moments computed in Lemma 4. For independent sampling, we have

$$\mathbb{E}_{\mathcal{D}_{\mathsf{indep}}}[\|R_\Pi(\theta)\|_F^2] = \sum_{\substack{t_1,t_2,d_1,d_2\\(t_1,d_1)=(t_2,d_2)}} pR(\theta)^2_{(t_1,d_1),(t_2,d_2)} + \sum_{\substack{t_1,t_2,d_1,d_2\\(t_1,d_1)\neq(t_2,d_2)}} p^2R(\theta)^2_{(t_1,d_1),(t_2,d_2)},$$

which we can concisely upper bound as:

$$\mathbb{E}_{\mathcal{D}_{\mathsf{indep}}}[\|R_\Pi(\theta)\|_F^2] \leq p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + p^2\|R(\theta)\|_F^2.$$

We move on to fixed-size sampling, for which we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\mathsf{fixed}}}[\|R_\Pi(\theta)\|_F^2] = &\sum_{\substack{t_1,t_2,d_1,d_2\\(t_1,d_1)=(t_2,d_2)}} \left(p\left(1-\frac{1}{D}\right)+p^2\right)R(\theta)^2_{(t_1,d_1),(t_2,d_2)}\\ &+ \sum_{\substack{t_1,t_2,d_1,d_2\\t_1\neq t_2}} p^2R(\theta)^2_{(t_1,d_1),(t_2,d_2)}\\ &+ \sum_{\substack{t_1,t_2,d_1,d_2\\t_1=t_2,d_1\neq d_2}} \left(p^2-\frac{p}{D}\right)R(\theta)^2_{(t_1,d_1),(t_2,d_2)},\end{aligned}$$

which has the same concise upper bound:

$$\mathbb{E}_{\mathcal{D}_{\mathsf{fixed}}}\big[\|R_\Pi(\theta)\|_F^2\big] \le p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + p^2\|R(\theta)\|_F^2.$$

Finally, in the case of Markov sampling,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\mathsf{Markov}}}\big[\|R_\Pi(\theta)\|_F^2\big] = &\sum_{\substack{t_1,t_2,d_1,d_2\\(t_1,d_1)=(t_2,d_2)}} pR(\theta)^2_{(t_1,d_1),(t_2,d_2)}\\ &+\sum_{\substack{t_1,t_2,d_1,d_2\\d_1\neq d_2}} p^2R(\theta)^2_{(t_1,d_1),(t_2,d_2)}\\ &+\sum_{\substack{t_1,t_2,d_1,d_2\\d_1=d_2,t_1\neq t_2}} (p^2+p(1-p)(1-a-b)^{|t_1-t_2|})R(\theta)^2_{(t_1,d_1),(t_2,d_2)}.\end{aligned}$$

The sum in the last term can be crudely controlled as follows:

$$\begin{aligned}\sum_{\substack{t_1,t_2,d\\t_1\neq t_2}}(1-a-b)^{|t_1-t_2|}R(\theta)^2_{(t_1,d),(t_2,d)} &\le |1-a-b|\sum_{\substack{t_1,t_2\\t_1\neq t_2}}\sum_d (R(\theta)_{[t_1,t_2]})^2_{d,d}\\ &\le |1-a-b|\sum_{t_1\neq t_2}\|R(\theta)_{[t_1,t_2]}\|_F^2\\ &\le |1-a-b|\cdot\|R(\theta)\|_F^2\end{aligned}$$

This yields a shorter, but probably looser bound:

$$\mathbb{E}_{\mathcal{D}_{\mathsf{Markov}}}\big[\|R_\Pi(\theta)\|_F^2\big] \le p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + (p^2+p(1-p)|1-a-b|)\|R(\theta)\|_F^2.$$

Finally, we remember our hypothesis $a+b\le 1$, which will be useful to simplify the previous expression:

$$\begin{aligned}p+(1-p)(1-a-b) &= \frac{a}{a+b}+\frac{b}{a+b}(1-a-b)\\ &= \frac{a+b-ab-b^2}{a+b} = \frac{a(1-b)+b(1-b)}{a+b}\\ &= 1-b\\ p+(1-p)(a+b-1) &= \frac{a}{a+b}+\frac{b}{a+b}(a+b-1)\\ &= \frac{a+ba+b^2-b}{a+b} = \frac{a(1+b)-b(1-b)}{a+b}\\ &= p(1+b)-(1-p)(1-b) = 2p-(1-b).\end{aligned}$$

As a consequence,

$$p+(1-p)|1-a-b| = \max\{1-b, 2p-(1-b)\} =: q.$$

■

### B.5 Bounding $R(\theta)$

Lemma 15 relates the bounds involving $R_\Pi(\theta)$ to features of the full residual $R(\theta)$, which we now study.

**Lemma 16** *The residual $R(\theta)$ satisfies the following inequalities:*

$$\|R(\theta)\|_2 \leq \frac{2\sigma_{\min}^2}{(1-\vartheta)^2}\|\theta\|_2$$
$$\|R(\theta)\|_F^2 \leq \frac{2T\sigma_{\min}^4}{(1-\vartheta)^3}\|\theta\|_F^2$$
$$\mathrm{Tr}[R(\theta)\odot R(\theta)] \leq \frac{T\sigma_{\min}^4}{(1-\vartheta)^2}\|\theta\|_2^2\|\theta\|_F^2.$$

**Proof** In what follows, we will heavily use the following remark

$$\|\Gamma_0(\theta)\|_2 \leq \sum_{k=0}^{\infty}\|\theta^k\Sigma\theta'^k\|_2 \leq \|\Sigma\|_2\sum_{k=0}^{\infty}\|\theta\|_2^{2k} = \frac{\|\Sigma\|_2}{1-\|\theta\|_2^2} \leq \frac{\|\Sigma\|_2}{1-\vartheta^2}.$$

We start by giving a formula for the blocks of $R(\theta)$: by Lemma 12,

$$R(\theta)_{[t,s]} = \begin{cases}\theta^{t-s}\Gamma_0(\theta) & \text{if } s\in[1,t-1]\\ \theta\Gamma_0(\theta)\theta' & \text{if } s=t\\ \Gamma_0(\theta)\theta'^{t-s} & \text{if } s\in[t+1,T].\end{cases}$$

These individual blocks can be bounded using Lemma 31: if $r\geq 1$, then

$$\|\theta^r\Gamma_0(\theta)\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2\|\theta^{r-1}\|_2^2 \leq \frac{\|\Sigma\|_2^2}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^{2(r-1)}$$
$$\|\Gamma_0(\theta)\theta'^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2\|\theta^{r-1}\|_2^2 \leq \frac{\|\Sigma\|_2^2}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^{2(r-1)}$$
$$\|\theta\Gamma_0(\theta)\theta'\|_F^2 \leq \|\theta\|_2^2\|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2 \leq \frac{\|\Sigma\|_2^2}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^2.$$

Since we control the norm of each block of $R(\theta)$, we control the norm of the whole matrix:

$$\begin{aligned}\|R(\theta)\|_F^2 &= \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\|\theta^{t-s}\Gamma_0(\theta)\|_F^2 + \|\theta\Gamma_0(\theta)\theta\|_F^2 + \sum_{s=t+1}^{T}\|\Gamma_0(\theta)\theta^{s-t}\|_F^2\right)\\ &\leq \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2}\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\|\theta\|_2^{2(t-s-1)} + \|\theta\|_2^2 + \sum_{s=t+1}^{T}\|\theta\|_2^{2(s-t-1)}\right)\\ &\leq \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2}\sum_{t=1}^{T}\left(\sum_{s=-\infty}^{t-1}\|\theta\|_2^{2(t-1-s)} + \|\theta\|_2^2 + \sum_{s=t+1}^{+\infty}\|\theta\|_2^{2(s-1-t)}\right)\\ &= \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{1}{1-\|\theta\|_2^2} + \|\theta\|_2^2 + \frac{1}{1-\|\theta\|_2^2}\right)\end{aligned}$$

We now remember our hypothesis $\|\theta\|_2 \leq \vartheta < 1$:

$$\begin{aligned}\|R(\theta)\|_F^2 &\leq \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2} T\left(\frac{1}{1-\vartheta^2} + \vartheta^2 + \frac{1}{1-\vartheta^2}\right) \\ &= \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2} T\left(\frac{2+\vartheta^2(1-\vartheta^2)}{1-\vartheta^2}\right) \\ &\leq \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2} T\left(\frac{2+2\vartheta}{1-\vartheta^2}\right) = \frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta^2)^2} T\left(\frac{2}{1-\vartheta}\right) \\ &= 2T\frac{\|\Sigma\|_2^2\|\theta\|_F^2}{(1-\vartheta)^3}.\end{aligned}$$

Now that we have a handle on the Frobenius norm of $R(\theta)$, we move on to its spectral norm. Notice that $R(\theta)$ can be written as a sum of Kronecker products with the subdiagonal matrices $J_t$:

$$R(\theta) = I \otimes \theta\Gamma_0(\theta)\theta' + \sum_{t=1}^{T-1}\left[J_t \otimes \theta^t\Gamma_0(\theta) + J_t' \otimes \Gamma_0(\theta)\theta'^t\right].$$

We can use Lemma 30 and write:

$$\begin{aligned}\|R(\theta)\|_2 &\leq \|I\|_2 \times \|\theta\Gamma_0(\theta)\theta'\|_2 + \sum_{t=1}^{T-1}\left[\|J_t\|_2 \times \|\theta^t\Gamma_0(\theta)\|_2 + \|J_t'\|_2 \times \|\Gamma_0(\theta)\theta'^t\|_2\right] \\ &\leq \|\Gamma_0(\theta)\|_2\left(\|\theta\|_2^2 + 2\sum_{t=1}^{T-1}\|\theta\|_2^t\right) \leq \frac{\|\Sigma\|_2}{1-\vartheta^2}\left(\|\theta\|_2^2 + 2\frac{\|\theta\|_2}{1-\|\theta\|_2}\right) \\ &\leq \frac{\|\Sigma\|_2\|\theta\|_2}{1-\vartheta^2}\left(\vartheta + \frac{2}{1-\vartheta}\right) \leq \frac{\|\Sigma\|_2\|\theta\|_2}{1-\vartheta}\left(\frac{2+2\vartheta}{1-\vartheta^2}\right) \\ &= 2\frac{\|\Sigma\|_2\|\theta\|_2}{(1-\vartheta)^2}.\end{aligned}$$

We finish with the trace of the Hadamard product $R(\theta) \odot R(\theta)$.

$$\begin{aligned}\operatorname{Tr}[R(\theta) \odot R(\theta)] &= T\operatorname{Tr}[(\theta\Gamma_0(\theta)\theta') \odot (\theta\Gamma_0(\theta)\theta')] \\ &\leq T\|\theta\Gamma_0(\theta)\theta'\|_F^2 \leq T\|\Sigma\|_2^2\frac{\|\theta\|_2^2\|\theta\|_F^2}{(1-\vartheta)^2}.\end{aligned}$$

The last step is replacing $\|\Sigma\|_2 = \lambda_{\max}(\Sigma) = \sigma_{\max}^2$ in all the previous bounds. ■

### B.6 Upper Bound on the KL Divergence

We now have all the tools in hand to extract a KL divergence bound.

**Lemma 17 (Final KL bound)** *Assume $\theta$ satisfies*

$$\|\theta\|_2 \leq \min\left\{\vartheta, \frac{(1-\vartheta)^2}{4}\right\},$$

*then the expected conditional divergence is upper-bounded as follows:*

$$\mathbb{E}_\Pi\left[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right]\leq \mathrm{KL}_{\max}(\|\theta\|_2,\|\theta\|_F)$$

*where we defined*

$$\mathrm{KL}_{\max}(\|\theta\|_2,\|\theta\|_F)=\frac{2}{\gamma_\ell(\mathcal{D})}Tp(\|\theta\|_2^2+\mathbf{1}_{\mathcal{D}\neq\mathcal{D}_{\mathsf{Markov}}}p+\mathbf{1}_{\mathcal{D}=\mathcal{D}_{\mathsf{Markov}}}q)\|\theta\|_F^2$$

*and*

$$\gamma_\ell(\mathcal{D})=(1-\vartheta)^{3/2}\frac{\mathbf{1}_{\mathcal{D}\neq\mathcal{D}_{\mathsf{fixed}}}\sigma_{\min}^2+\omega^2}{\sigma_{\max}^2}.$$

**Proof** We only provide the proof for $\Pi\sim\mathcal{D}_{\mathsf{Markov}}$. Let us start with Lemma 13 on the conditional KL divergence between $\mathbb{P}_\theta[Y|\Pi]$ and $\mathbb{P}_0[Y|\Pi]$:

$$\mathbb{E}\left[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right]\leq\mathbb{E}\left[\frac{\|\Delta_\Pi(\theta)\|_F^2}{2(1+\lambda_{\min}(\Delta_\Pi(\theta)))}\right]$$

Our hypothesis on the spectral norm of $\theta$ is useful to bound the spectral norm of $R(\theta)$ using Lemma 16:

$$\|R(\theta)\|_2\leq\frac{2\sigma_{\min}^2}{(1-\vartheta)^2}\|\theta\|_2\leq\frac{2\sigma_{\min}^2}{(1-\vartheta)^2}\times\frac{(1-\vartheta)^2}{4}\leq\frac{1}{2}\sigma_{\min}^2$$

Since this condition is satisfied, we can move on with Lemma 14 linking $\Delta_\Pi(\theta)$ to $R_\Pi(\theta)$:

$$\mathbb{E}\left[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right]\leq\frac{\mathbb{E}\left[\|R_\Pi(\theta)\|_F^2\right]}{\left[\sigma_{\min}^2+\omega^2\right]^2}.$$

We follow up with Lemma 15 to get rid of the expectation:

$$\mathbb{E}\left[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right]\leq\frac{p\,\mathrm{Tr}[R(\theta)\odot R(\theta)]+pq\|R(\theta)\|_F^2}{\left[\sigma_{\min}^2+\omega^2\right]^2}.$$

By Lemma 16, both terms in the numerator can be controlled using the norms of $\theta$:

$$\|R(\theta)\|_F^2\leq\frac{2T\sigma_{\min}^4}{(1-\vartheta)^3}\|\theta\|_F^2$$
$$\mathrm{Tr}[R(\theta)\odot R(\theta)]\leq\frac{T\sigma_{\min}^4}{(1-\vartheta)^2}\|\theta\|_2^2\|\theta\|_F^2.$$

This leaves us with:

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right]&\leq\frac{p\times\frac{T\sigma_{\min}^4}{(1-\vartheta)^2}\|\theta\|_2^2\|\theta\|_F^2+pq\times\frac{2T\sigma_{\min}^4}{(1-\vartheta)^3}\|\theta\|_F^2}{\left[\sigma_{\min}^2+\omega^2\right]^2}\\&\leq\left(\frac{\sigma_{\min}^2}{\sigma_{\min}^2+\omega^2}\right)^2\frac{2Tp(\|\theta\|_2^2+q)}{(1-\vartheta)^3}\|\theta\|_F^2.\end{aligned}$$

We obtain the result we expected by noting the presence of $\gamma_\ell(\mathcal{D}_{\mathsf{Markov}})$ in the denominator of the previous expression. ■

### B.7 Application of Fano's Method

Given the KL bound we just obtained, we are finally able to prove Theorem 1.

**Proof** Again, we only provide the proof for Markov sampling. Fano's method requires finding $M+1$ parameters $\theta_i$ such that $\theta_0 = 0$ and $\|\theta_i - \theta_j\|_F \geq 2\tau$ for $i \neq j$ (with $\tau$ to be specified), while keeping control upon the average KL divergence between the probability distributions $\mathbb{P}_{\theta_i}$ and $\mathbb{P}_0$. Judging by Lemma 17, one way to achieve this control on the KL divergence is to bound the $\|\theta_i\|_F$ uniformly in $i$ (in other words, to choose them all inside a ball of fixed radius). We will then have to see how many $2\tau$-separated matrices we can fit in such a ball.

Let us consider the set $\mathcal{H}(r)$ of all block-diagonal $D \times D$ matrices with coefficients in $\{0, r\}$ such that each block has size $s \times s$ (we assume $s$ divides $D$). In particular, these matrices are all column-sparse, with no more than $s$ non-zero coefficients per column. In terms of dimensionality, we are dealing with the (scaled) matrix equivalent of a $Ds$-dimensional hypercube, hence the notation $\mathcal{H}(r)$. It has cardinality $2^{Ds}$ and for every $\theta \in \mathcal{H}$, we have the following norm bounds:

$$\|\theta\|_2 \leq rs \qquad \text{and} \qquad \|\theta\|_F \leq r\sqrt{Ds}.$$

The spectral norm bound on $\theta$ is obtained as the maximum spectral norm of each block, which we in turn control using the Frobenius norm of each block.

Unfortunately, in this hypercube, not all pairs of vertices are well-separated. That is why we need the Gilbert-Varshamov bound of Lemma 42: according to this result, there exists a pruned subset $\mathcal{K}(r) \subset \mathcal{H}(r)$ containing 0 and such that

$$|\mathcal{K}(r)| \geq |\mathcal{H}(r)|^{1/8} = 2^{Ds/8} \qquad \text{and} \qquad \|\mathrm{vec}(\theta_i) - \mathrm{vec}(\theta_j)\|_1 \geq \frac{rDs}{8}$$

for all pairs of distinct vertices $\theta_i$ and $\theta_j$ in $\mathcal{K}(r)$. We choose our set of parameters $\theta_0, \theta_1, ..., \theta_M$ to be exactly this pruned subset $\mathcal{K}(r)$, in particular $M+1 = |\mathcal{K}(r)|$.

The missing ingredient is an upper bound on the maximum average KL divergence between $\mathbb{P}_{\theta_i}$ and $\mathbb{P}_0$: we can obtain it using Lemma 17. We only need to assume

$$\|\theta_i\|_F \leq r\sqrt{Ds} \leq \min\left\{\vartheta, \frac{(1-\vartheta)^2}{4}\right\}$$

to get

$$\max_i \mathbb{E}_\Pi \left[\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \parallel \mathbb{P}_{\theta_0}(Y|\Pi)\right\}\right] \leq \max_i \mathrm{KL}_{\max}(\|\theta_i\|_2, \|\theta_i\|_F) \leq \mathrm{KL}_{\max}(rs, r\sqrt{Ds}).$$

Since we must satisfy the constraint from Equation (19) in Fano's method, we will choose $r$ so that:

$$\mathrm{KL}_{\max}(rs, r\sqrt{Ds}) \leq \alpha \log(M) = \alpha \log\left(2^{Ds/8} - 1\right)$$

with $\alpha = \frac{\log 3 - \log 2}{\log 2}$. We want to solve the previous inequality for $r$, and for that we start by replacing $\mathrm{KL}_{\max}(rs, r\sqrt{Ds})$ with its value from Lemma 17, replacing $\gamma_\ell(\mathcal{D})$ with $\gamma_\ell$ to

lighten notations:

$$\mathrm{KL}_{\max}(rs, r\sqrt{Ds}) \leq \alpha \log\left(2^{Ds/8} - 1\right) \iff \frac{2}{\gamma_\ell} Tp\left((rs)^2 + q\right)(r\sqrt{Ds})^2 \leq cDs$$
$$\iff Ds^3 r^4 + qDsr^2 - c\frac{\gamma_\ell^2 Ds}{Tq} \leq 0.$$

If we consider this as a degree two polynomial in the variable $r^2$, its determinant is

$$\Delta = q^2 D^2 s^2 + 4Ds^3 c\frac{\gamma_\ell^2 Ds}{Tp}.$$

For $\beta$ to be small enough, $r^2$ must remain below the only positive root of the polynomial, namely

$$r^2 \leq \frac{-qDs + \sqrt{q^2 D^2 s^2 + c\frac{\gamma_\ell^2 D^2 s^4}{Tp}}}{2Ds^3} = \frac{q}{2s^2}\left(\sqrt{1 + c\frac{\gamma_\ell^2 s^2}{Tpq^2}} - 1\right).$$

If we assume the quantity $c\frac{\gamma_\ell^2 s^2}{Tpq^2}$ inside the square root is smaller than 1, i.e.

$$\frac{\gamma_\ell s}{\sqrt{p}q\sqrt{T}} \leq c, \tag{10}$$

then we can lower-bound $\sqrt{1+x}$ by its chord $(\sqrt{2}-1)x$. In other words, a sufficient condition for $r^2$ to remain small enough is given by

$$r^2 \leq \frac{q}{2s^2} \times (\sqrt{2} - 1)c\frac{\gamma_\ell^2 s^2}{Tpq^2} = c\frac{\gamma_\ell^2}{Tpq}.$$

To sum up, we have three constraints on $r$:

$$\begin{cases} rs \leq \vartheta \\ rs \leq \dfrac{(1-\vartheta)^2}{4} \\ r \leq \sqrt{c\dfrac{\gamma_\ell^2}{Tpq}}. \end{cases}$$

We can therefore choose $r$ as the largest value satisfying all three of them:

$$r := \min\left\{\frac{\vartheta}{s}, \frac{(1-\vartheta)^2}{4s}, c\frac{\gamma_\ell}{\sqrt{pq}\sqrt{T}}\right\}$$

To reach our conclusion, we simply need to remark that the vectorized $\ell_1$ distance between any two matrices in $\mathcal{K}(r)$ gives us a lower bound on the operator $\ell_\infty$ distance that separates them:

$$\|\theta_i - \theta_j\|_\infty = \max_{k\in[D]} \sum_{l\in[D]} |(\theta_i - \theta_j)|_{k,l} \geq \frac{1}{D} \sum_{1\leq k,l\leq D} |(\theta_i - \theta_j)|_{k,l}$$
$$= \frac{1}{D}\|\mathrm{vec}(\theta_i) - \mathrm{vec}(\theta_j)\|_1 \geq \frac{rDs}{8D} = \frac{rs}{8}$$

Subsequently, our parameters $\theta_i$ are $2\tau$-separated (in $\ell_\infty$ operator distance) with

$$\tau = \frac{rs}{8} = \min\left\{\frac{\vartheta}{8}, \frac{(1-\vartheta)^2}{32}, c\frac{\gamma_\ell s}{\sqrt{pq}\sqrt{T}}\right\}.$$

As soon as

$$c\frac{\gamma_\ell s}{\sqrt{pq}\sqrt{T}} \leq c\min\{\vartheta, (1-\vartheta)^2\} \tag{11}$$

we can simplify this expression as

$$\tau = c\frac{\gamma_\ell s}{\sqrt{pq}\sqrt{T}}.$$

In this case, by Lemma 34, we can conclude:

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta_s} \mathbb{P}_\theta\left[\|\widehat{\theta} - \theta\|_\infty \geq c\frac{\gamma_\ell s}{\sqrt{pq}\sqrt{T}}\right] \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha \geq \frac{1}{2}.$$

■

## Appendix C. Proof of the Estimator's Convergence Rate

Here we present the detailed proofs of Theorems 2 and 3.

For this part, we slightly change the previous conventions: we now assume that $\Pi_t$ has a fixed size of $O$, with a variable number of non-trivial observation rows stacked at the top, followed if necessary by a block of rows that are full of zeroes. This allows us to fully decouple $\eta$ from $\Pi$, but it doesn't change the heart of our problem since we do not have more information available, just a number of rows containing only noise.

In this new setting, it is crucial to notice that Lemma 10 still applies, because the matrices $\Pi_t$ now have *at most* one 1 per row instead of exactly one.

### C.1 Building an Unbiased Estimator

The first step consists in justifying the construction of our estimator for $\Gamma_h$.

**Lemma 18 (Expectation of the covariance estimator)** *Let us define*

$$S(h) := \mathbb{E}[\pi_{t+h}\pi_t'] \qquad \text{and} \qquad C(h) := \mathbf{1}_{\{h=0\}}\omega^2 \operatorname{diag}\left(\frac{\mathbb{E}[\pi_t/\kappa_t]}{\mathbb{E}[\pi_t]}\right).$$

*Then the estimator $\widehat{\Gamma}_h$ given by Equation* (8) *for the covariance matrix $\Gamma_h$ of rank $h$ is unbiased.*

**Proof** By Equation (6),

$$\begin{aligned}(\Pi_{t+h}^+ Y_{t+h})(\Pi_t^+ Y_t)' &= \Pi_{t+h}^+(\Pi_{t+h}X_{t+h} + \eta_{t+h})(X_t'\Pi_t' + \eta_t')\Pi_t^{+\prime} \\ &= \Pi_{t+h}^+\Pi_{t+h}X_{t+h}X_t'\Pi_t'\Pi_t^{+\prime} + \Pi_{t+h}^+\Pi_{t+h}X_{t+h}\eta_t'\Pi_t^{+\prime} \\ &\quad + \Pi_{t+h}^+\eta_{t+h}X_t'\Pi_t'\Pi_t^{+\prime} + \Pi_{t+h}^+\eta_{t+h}\eta_t'\Pi_t^{+\prime},\end{aligned} \tag{12}$$

so that

$$\begin{aligned}\mathbb{E}[(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'|\Pi] &= \mathbb{E}\left[\Pi_{t+h}^{+}(\Pi_{t+h}X_{t+h}+\eta_{t+h})(X_t'\Pi_t'+\eta_t')\Pi_t^{+\prime}|\Pi\right]\\ &= \Pi_{t+h}^{+}\Pi_{t+h}\mathbb{E}\left[X_{t+h}X_t'\right]\Pi_t'\Pi_t^{+\prime}\\ &\quad+\Pi_{t+h}^{+}\Pi_{t+h}\mathbb{E}\left[X_{t+h}\eta_t'\right]\Pi_t^{+\prime}\\ &\quad+\Pi_{t+h}^{+}\mathbb{E}\left[\eta_{t+h}X_t'\right]\Pi_t'\Pi_t^{+\prime}\\ &\quad+\Pi_{t+h}^{+}\mathbb{E}\left[\eta_{t+h}\eta_t'\right]\Pi_t^{+\prime}.\end{aligned}$$

The two cross-product terms in the middle are zero because $X$ and $\eta$ are centered at expectation and independent (given $\Pi$). Since $\mathbb{E}\left[X_{t+h}X_t'\right]=\Gamma_h$ and $\mathbb{E}[\eta_{t+h},\eta_t]=\mathbf{1}_{\{h=0\}}\omega^2 I$, we are left with:

$$\mathbb{E}[(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'|\Pi] = (\Pi_{t+h}^{+}\Pi_{t+h})\Gamma_h(\Pi_t^{+}\Pi_t)' + \mathbf{1}_{\{h=0\}}\omega^2\Pi_t^{+}\Pi_t^{+\prime}.$$

Using Lemma 10, we find $\Pi_t^{+}\Pi_t=\operatorname{diag}(\pi_t)$ and

$$\Pi_t^{+}\Pi_t^{+\prime} = \operatorname{diag}\left(\frac{\pi_t}{\kappa_t}\right)\Pi_t'\Pi_t\operatorname{diag}\left(\frac{\pi_t}{\kappa_t}\right) = \operatorname{diag}\left(\frac{\pi_t}{\kappa_t}\right)\operatorname{diag}(\kappa_t)\operatorname{diag}\left(\frac{\pi_t}{\kappa_t}\right) = \operatorname{diag}\left(\frac{\pi_t}{\kappa_t}\right). \quad (13)$$

Plugging this in yields

$$\begin{aligned}\mathbb{E}[(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'|\Pi] &= \operatorname{diag}(\pi_{t+h})\Gamma_h\operatorname{diag}(\pi_t)+\mathbf{1}_{\{h=0\}}\omega^2\operatorname{diag}(\pi_t/\kappa_t)\\ &= (\pi_{t+h}\pi_t')\odot\Gamma_h+\mathbf{1}_{\{h=0\}}\omega^2\operatorname{diag}(\pi_t/\kappa_t).\end{aligned}$$

We now take the expectation w.r.t. $\Pi$:

$$\begin{aligned}\mathbb{E}[(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'] &= \mathbb{E}\left[\mathbb{E}[(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'|\Pi]\right]\\ &= \mathbb{E}[\pi_{t+h}\pi_t']\odot\Gamma_h+\mathbf{1}_{\{h=0\}}\omega^2\mathbb{E}[\operatorname{diag}(\pi_t/\kappa_t)],\end{aligned}$$

Dividing by $\mathbb{E}[\pi_{t+h}\pi_t']$, we get

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\mathbb{E}[\pi_{t+h}\pi_t']}\odot(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)'\right] &= \Gamma_h+\mathbf{1}_{\{h=0\}}\omega^2\mathbb{E}[\operatorname{diag}(\pi_t/\kappa_t)]\odot\frac{1}{\mathbb{E}[\pi_t\pi_t']},\\ &= \Gamma_h+\mathbf{1}_{\{h=0\}}\omega^2\operatorname{diag}\left(\frac{\mathbb{E}[\pi_t/\kappa_t]}{\mathbb{E}[\pi_t]}\right)\end{aligned}$$

which shows that our estimator

$$\widehat{\Gamma}_h := \frac{1}{T-h}\sum_{t=1}^{T-h}\frac{1}{\mathbb{E}[\pi_{t+h}\pi_t']}\odot(\Pi_{t+h}^{+}Y_{t+h})(\Pi_t^{+}Y_t)' - \mathbf{1}_{\{h=0\}}\omega^2\operatorname{diag}\left(\frac{\mathbb{E}[\pi_t/\kappa_t]}{\mathbb{E}[\pi_t]}\right)$$

is unbiased. Since the process $(\Pi_t)$ is stationary, the values of

$$S(h)=\mathbb{E}[\pi_{t+h}\pi_t'] \qquad \text{and} \qquad C(h)=\mathbf{1}_{\{h=0\}}\omega^2\operatorname{diag}\left(\frac{\mathbb{E}[\pi_t/\kappa_t]}{\mathbb{E}[\pi_t]}\right)$$

do not depend on $t$, and we get:

$$\widehat{\Gamma}_h = \frac{1}{S(h)} \odot \frac{1}{T-h} \sum_{t=1}^{T-h} (\Pi^+_{t+h} Y_{t+h})(\Pi^+_t Y_t)' - C(h)$$

The reasoning above immediately entails that $\widehat{\Gamma}_h$ is unbiased. ■

Note that for distributions $\mathcal{D}_{\mathsf{indep}}$ and $\mathcal{D}_{\mathsf{Markov}}$, since $\pi_t = \kappa_t$, the fraction in $C(h)$ is a simple Bernoulli variable $\frac{\pi_{t,d}}{\kappa_{t,d}} = \pi_{t,d} \sim \mathcal{B}(p)$ whose expectation is known. Meanwhile, for $\mathcal{D}_{\mathsf{fixed}}$, the fraction $\frac{\pi_{t,d}}{\kappa_{t,d}}$ (whenever it is nonzero) is the inverse of a binomial variable $\mathcal{B}(pD, 1/D)$, whose expectation has no closed form but can easily be approximated numerically.

### C.2 Gaussian Concentration, Episode 1

From now on, our goal will be to quantify the concentration of $\widehat{\Gamma}_h$ around its expectation. We will do this coefficient by coefficient: let us fix two indices $d_1$ and $d_2$. Our goal is to control the deviation of $(\widehat{\Gamma}_h)_{d_1,d_2}$ around its mean.

**Lemma 19 (Deviation of $(\widehat{\Gamma}_h)_{d_1,d_2}$)** *The deviation probability for $(\widehat{\Gamma}_h)_{d_1,d_2}$ can be decomposed as follows:*

$$\begin{aligned}
\mathbb{P}(|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \geq u) \leq \quad & \mathbb{P}\left(|g'_\varepsilon \Psi'_\varepsilon L_{\varepsilon\varepsilon} \Psi_\varepsilon g_\varepsilon - \mathbb{E}\left[g'_\varepsilon \Psi'_\varepsilon L_{\varepsilon\varepsilon} \Psi_\varepsilon g_\varepsilon\right]| \geq u/4\right) \\
& + \mathbb{P}\left(|g'_\eta \Psi'_\eta L_{\eta\varepsilon} \Psi_\varepsilon g_\varepsilon - \mathbb{E}\left[g'_\eta \Psi'_\eta L_{\eta\varepsilon} \Psi_\varepsilon g_\varepsilon\right]| \geq u/4\right) \\
& + \mathbb{P}\left(|g'_\varepsilon \Psi'_\varepsilon L_{\varepsilon\eta} \Psi_\eta g_\eta - \mathbb{E}\left[g'_\varepsilon \Psi'_\varepsilon L_{\varepsilon\eta} \Psi_\eta g_\eta\right]| \geq u/4\right) \\
& + \mathbb{P}\left(|g'_\eta \Psi'_\eta L_{\eta\eta} \Psi_\eta g_\eta - \mathbb{E}\left[g'_\eta \Psi'_\eta L_{\eta\eta} \Psi_\eta g_\eta\right]| \geq u/4\right)
\end{aligned}$$

*where the (random) $L$ matrices are defined in Equation* (14) *and the $\Psi$ matrices are defined in Equation* (15)*.*

**Proof** By Equation (8),

$$\begin{aligned}
(\widehat{\Gamma}_h + C(h))_{d_1,d_2} &= \frac{1}{T-h} \sum_{t=1}^{T-h} \left(\frac{1}{S(h)} \odot (\Pi^+_{t+h} Y_{t+h})(\Pi^+_t Y_t)'\right)_{d_1,d_2} \\
&= \frac{1}{T-h} \sum_{t=1}^{T-h} \frac{1}{S(h)_{d_1,d_2}} \mathbf{1}'_{d_1} (\Pi^+_{t+h} Y_{t+h})(\Pi^+_t Y_t)' \mathbf{1}_{d_2} \\
&= \frac{1}{T-h} \sum_{t=1}^{T-h} \mathrm{Tr}\left[\frac{\mathbf{1}_{d_2} \mathbf{1}'_{d_1}}{S(h)_{d_1,d_2}} (\Pi^+_{t+h} Y_{t+h})(\Pi^+_t Y_t)'\right]
\end{aligned}$$

Equation (12) allows us to rewrite $(\Pi^+_{t+h}Y_{t+h})(\Pi^+_t Y_t)'$:

$$\begin{aligned}(\widehat{\Gamma}_h + C(h))_{d_1,d_2} = &\frac{1}{T-h}\sum_{t=1}^{T-h} X_t'\Pi_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}\Pi_{t+h}X_{t+h}\\ &+\frac{1}{T-h}\sum_{t=1}^{T-h}\eta_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}\Pi_{t+h}X_{t+h}\\ &+\frac{1}{T-h}\sum_{t=1}^{T-h} X_t'\Pi_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}\eta_{t+h}\\ &+\frac{1}{T-h}\sum_{t=1}^{T-h}\eta_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}\eta_{t+h}\end{aligned}$$

Let us denote by $P_t$ the projection keeping only the components associated with time $t$, i.e. such that $X_t = P_t X$ and $\eta_t = P_t\eta$. We then have

$$\begin{aligned}(\widehat{\Gamma}_h + C(h))_{d_1,d_2} = &X'\underbrace{\Big(\frac{1}{T-h}\sum_{t=1}^{T-h}P_t'(\Pi_t^+\Pi_t)'\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}(\Pi^+_{t+h}\Pi_{t+h})P_{t+h}\Big)}_{L_{\varepsilon\varepsilon}}X\\ &+\eta'\underbrace{\Big(\frac{1}{T-h}\sum_{t=1}^{T-h}P_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}(\Pi^+_{t+h}\Pi_{t+h})P_{t+h}\Big)}_{L_{\eta\varepsilon}}X\\ &+X'\underbrace{\Big(\frac{1}{T-h}\sum_{t=1}^{T-h}P_t'(\Pi_t^+\Pi_t)'\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}P_{t+h}\Big)}_{L_{\varepsilon\eta}}\eta\\ &+\eta'\underbrace{\Big(\frac{1}{T-h}\sum_{t=1}^{T-h}P_t'\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi^+_{t+h}P_{t+h}\Big)}_{L_{\eta\eta}}\eta\end{aligned} \tag{14}$$

Since $X$) and $\eta$ both follow centered multivariate Gaussian distributions, we can write them as linear combinations of standard Gaussian vectors $g_\varepsilon$ and $g_\eta$ (indexed by the source of randomness):

$$\begin{aligned}X = \Psi_\varepsilon g_\varepsilon \quad &\text{with} \quad \begin{cases} g_\varepsilon \sim \mathcal{N}(0, I_{TD}) \\ \Psi_\varepsilon := \mathrm{Cov}[X]^{1/2} \text{ (see Lemma 11)}\end{cases}\\ \eta = \Psi_\eta g_\eta \quad &\text{with} \quad \begin{cases} g_\eta \sim \mathcal{N}(0, I_{TO}) \\ \Psi_\eta := \mathrm{Cov}[\eta]^{1/2} = \omega I.\end{cases}\end{aligned} \tag{15}$$

We replace $X$ and $\eta$ to get:

$$(\widehat{\Gamma}_h + C(h))_{d_2,d_1} = g_\varepsilon'\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon g_\varepsilon + g_\eta'\Psi_\eta'L_{\eta\varepsilon}\Psi_\varepsilon g_\varepsilon + g_\varepsilon'\Psi_\varepsilon'L_{\varepsilon\eta}\Psi_\eta g_\eta + g_\eta'\Psi_\eta'L_{\eta\eta}\Psi_\eta g_\eta,$$

which implies

$$\begin{aligned}(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2} = {} & g_\varepsilon' \Psi_\varepsilon' L_{\varepsilon\varepsilon} \Psi_\varepsilon g_\varepsilon - \mathbb{E}[g_\varepsilon' \Psi_\varepsilon' L_{\varepsilon\varepsilon} \Psi_\varepsilon g_\varepsilon] \\ & + g_\eta' \Psi_\eta' L_{\eta\varepsilon} \Psi_\varepsilon g_\varepsilon - \mathbb{E}[g_\eta' \Psi_\eta' L_{\eta\varepsilon} \Psi_\varepsilon g_\varepsilon] \\ & + g_\varepsilon' \Psi_\varepsilon' L_{\varepsilon\eta} \Psi_\eta g_\eta - \mathbb{E}[g_\varepsilon' \Psi_\varepsilon' L_{\varepsilon\eta} \Psi_\eta g_\eta] \\ & + g_\eta' \Psi_\eta' L_{\eta\eta} \Psi_\eta g_\eta - \mathbb{E}[g_\eta' \Psi_\eta' L_{\eta\eta} \Psi_\eta g_\eta].\end{aligned}$$

The union bound gives us the expected result. ■

### C.3 Interlude: Discrete Concentration

Now, our goal is to control the $L$ matrices, in order to apply a conditional version of the Hanson-Wright inequality (Lemma 7) to these deviation probabilities. Since these $L$ are random and built from the binary sampling matrices $\Pi_t$, we need to take a little detour through discrete concentration inequalities.

We first notice that each one of these $L$ matrices can be written as

$$L = \frac{1}{T-h} \sum_{t=1}^{T-h} P_t' L_{[t,t+h]} P_{t+h}$$

which means they are block-superdiagonal of rank $h$ and satisfy

$$\|L\|_2 = \frac{1}{T-h} \max_{t \in [T-h]} \|L_{[t,t+h]}\|_2 \qquad \text{and} \qquad \|L\|_F^2 = \frac{1}{(T-h)^2} \sum_{t=1}^{T-h} \|L_{[t,t+h]}\|_F^2$$

**Lemma 20 (Spectral norm bound for the $L$ matrices)** *The spectral norms of all $L$ matrices are bounded by the same deterministic quantity:*

$$\max\{\|L_{\varepsilon\varepsilon}\|_2, \|L_{\eta\varepsilon}\|_2, \|L_{\varepsilon\eta}\|_2, \|L_{\eta\eta}\|_2\} \leq \frac{c}{p^2 T}.$$

**Proof** We must bound the spectral norm of $L_{[t,t+h]}$. By Lemma 10, we know that $\|\Pi_t^+ \Pi_t\|_2 = \|\mathrm{diag}(\pi_t)\|_2 \leq 1$. Meanwhile, Equation (13) gives us

$$\|\Pi_t^+\|_2 = \sqrt{\lambda_{\max}\left(\mathrm{diag}(\pi_t/\kappa_t)\right)} \leq 1$$

And of course $\|\mathbf{1}_{d_2} \mathbf{1}_{d_1}'\|_2 = 1$. Judging by Equation (14), this means that $L_{[t,t+h]}$ can be written as $1/S(h)_{d_1,d_2}$ times the product of three matrices with spectral norm smaller than 1. As a consequence,

$$\|L_{[t,t+h]}\|_2 \leq \frac{1}{S(h)_{d_1,d_2}} \leq \frac{1}{S(h)_{\min}} \leq \frac{c}{p^2}$$

So we can conclude

$$\max\{\|L_{\varepsilon\varepsilon}\|_2, \|L_{\eta\varepsilon}\|_2, \|L_{\varepsilon\eta}\|_2, \|L_{\eta\eta}\|_2\} \leq \frac{1}{T-h} \max_{t \in [T-h]} \frac{c}{p^2} \leq \frac{c}{Tp^2}.$$

■

**Lemma 21 (Frobenius norm comparison for the $L$ matrices)** *The Frobenius norm of all $L$ matrices is controlled by that of $L_{\varepsilon\varepsilon}$:*

$$\max\{\|L_{\varepsilon\varepsilon}\|_F^2, \|L_{\eta\varepsilon}\|_F^2, \|L_{\varepsilon\eta}\|_F^2, \|L_{\eta\eta}\|_F^2\} = \|L_{\varepsilon\varepsilon}\|_F^2.$$

**Proof** In this whole proof, we will be using Lemma 10 to simplify expressions involving $\Pi_t^+$. Let us start with the simplest one, namely $L_{\varepsilon\varepsilon}$:

$$\begin{aligned}
\|L_{\varepsilon\varepsilon}\|_F^2 &= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|(\Pi_t^+\Pi_t)'\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}(\Pi_{t+h}^+\Pi_{t+h})\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\operatorname{diag}(\pi_t)\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}}\mathbf{1}_{d_2}\mathbf{1}_{d_1}'\right\|_F^2 \\
&= \frac{1}{(T-h)^2 S(h)_{d_1,d_2}^2}\sum_{t=1}^{T-h}\pi_{t+h,d_1}\pi_{t,d_2}.
\end{aligned}$$

Now we move on to $L_{\eta\varepsilon}$ (the reasoning for $L_{\varepsilon,\eta}$ is the same):

$$\begin{aligned}
\|L_{\eta\varepsilon}\|_F^2 &= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}(\Pi_{t+h}^+\Pi_{t+h})\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\Pi_t\operatorname{diag}(\pi_t/\kappa_t)\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\frac{\pi_{t+h,d_1}(\pi_{t,d_2}/\kappa_{t,d_2})}{S(h)_{d_1,d_2}}\Pi_t\mathbf{1}_{d_2}\mathbf{1}_{d_1}'\right\|_F^2 \\
&= \frac{1}{(T-h)^2 S(h)_{d_1,d_2}^2}\sum_{t=1}^{T-h}\pi_{t+h,d_1}\pi_{t,d_2}\frac{\left\|\Pi_t\mathbf{1}_{d_2}\mathbf{1}_{d_1}'\right\|_F^2}{\kappa_{t,d_2}^2}.
\end{aligned}$$

Let us now finish with $L_{\eta\eta}$:

$$\begin{aligned}
\|L_{\eta\eta}\|_F^2 &= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\Pi_t^{+\prime}\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\Pi_{t+h}^+\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\Pi_t\operatorname{diag}(\pi_t/\kappa_t)\frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h}/\kappa_{t+h})\Pi_{t+h}'\right\|_F^2 \\
&= \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\left\|\frac{(\pi_{t+h,d_1}/\kappa_{t+h,d_1})(\pi_{t,d_2}/\kappa_{t,d_2})}{S(h)_{d_1,d_2}}\Pi_t\mathbf{1}_{d_2}\mathbf{1}_{d_1}'\Pi_{t+h}'\right\|_F^2 \\
&= \frac{1}{(T-h)^2 S(h)_{d_1,d_2}^2}\sum_{t=1}^{T-h}\pi_{t+h,d_1}\pi_{t,d_2}\frac{\left\|\Pi_t\mathbf{1}_{d_2}\mathbf{1}_{d_1}'\Pi_{t+h}'\right\|_F^2}{\kappa_{t+h,d_1}^2\kappa_{t,d_2}^2}.
\end{aligned}$$

To prove that $L_{\varepsilon,\varepsilon}$ has the largest Frobenius norm of them all, we only need the following insight:

$$\frac{\left\|\Pi_t \mathbf{1}_{d_2} \mathbf{1}'_{d_1}\right\|_F^2}{\kappa_{t,d_2}^2} \leq 1 \qquad \text{and} \qquad \frac{\left\|\Pi_t \mathbf{1}_{d_2} \mathbf{1}'_{d_1} \Pi'_{t+h}\right\|_F^2}{\kappa_{t,d_2}^2 \kappa_{t+h,d_1}^2} \leq 1$$

Indeed, $\Pi_t \mathbf{1}_{d_2}$ is the $d_2$-th column from $\Pi_t$, and the number of non-zero coefficients in this column is given by $\kappa_{t,d_2}$. As a consequence, its Frobenius (Euclidean) norm is $\kappa_{t,d_2}$. The same goes for $\Pi_{t+h} \mathbf{1}_{d_1}$. ■

**Lemma 22 (Frobenius norm bound for the $L$ matrices)** *For any $\delta$ such that condition* (16) *holds, the Frobenius norm of $L_{\varepsilon\varepsilon}$ is bounded by:*

$$\|L_{\varepsilon\varepsilon}\|_F^2 \leq \frac{c}{p^2 T}$$

*with probability at least $1 - \delta$.*

**Proof** We will exploit discrete concentration inequalities to get a high-probability bound on

$$\|L_{\varepsilon\varepsilon}\|_F^2 = \frac{1}{(T-h)^2 S(h)_{d_1,d_2}^2} \sum_{t=1}^{T-h} \pi_{t+h,d_1} \pi_{t,d_2}.$$

By Lemma 5, for all $u \in [0,1]$,

$$\mathbb{P}\left( \frac{1}{T-h} \sum_{t=1}^{T-h} \pi_{t+h,d_1} \pi_{t,d_2} \geq (1+u) S(h)_{d_1,d_2} \right) \leq c_1 \exp(-c_2 u^2 T S(h)_{d_1,d_2}).$$

Which implies, by rescaling,

$$\mathbb{P}\left( \frac{1}{T-h} \sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1} \pi_{t,d_2}}{S(h)_{d_1,d_2}} \geq 1+u \right) \leq c_1 \exp(-c_2 u^2 T S(h)_{d_1,d_2}).$$

The conclusion about $\|L_{\varepsilon\varepsilon}\|_F^2$ is now within reach:

$$\begin{aligned}
&\mathbb{P}\left( \|L_{\varepsilon\varepsilon}\|_F^2 \geq \frac{1+u}{(T-h) S(h)_{d_1,d_2}} \right) \\
&\quad = \mathbb{P}\left( \frac{1}{(T-h) S(h)_{d_1,d_2}} \left( \frac{1}{T-h} \sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1} \pi_{t,d_2}}{S(h)_{d_1,d_2}} \right) \geq \frac{1+u}{(T-h) S(h)_{d_1,d_2}} \right) \\
&\quad \leq c_1 \exp(-c_2 u^2 T S(h)_{d_1,d_2}).
\end{aligned}$$

We finally remember that $S(h)_{d_1,d_2} \geq S(h)_{\min} \geq c p^2$, so that

$$\mathbb{P}\left( \|L_{\varepsilon\varepsilon}\|_F^2 \geq \frac{c_3(1+u)}{T p^2} \right) \leq c_1 \exp\left( -c_2 u^2 p^2 T \right).$$

All we need to make sure that $\mathbb{P}\left(\|L_{\varepsilon\varepsilon}\|_F^2 \geq \frac{c_3(1+u)}{Tp^2}\right) \leq \delta$ is to choose $u$ such that

$$\begin{aligned} & c_1 \exp\left(-c_2 u^2 p^2 T\right) \leq \delta \\ \iff \quad & -c_2 u^2 p^2 T \leq \log(\delta/c_1) \\ \iff \quad & u \geq \left(\frac{1}{c_2 p^2 T}\log(c_1/\delta)\right)^{1/2} = \frac{c_4 \log(1/\delta)}{p\sqrt{T}}. \end{aligned}$$

However, for this to be valid, we must assume that our choice of $u$ is smaller than 1, i.e.

$$\frac{c\log(1/\delta)}{p\sqrt{T}} \leq 1. \tag{16}$$

If this holds, we might as well replace $u$ with 1 directly in the concentration result, which yields the simpler result announced above. ■

**Lemma 23 (Trace bound for the $L$ matrices)** *The trace of $\Psi_\eta' L_{\eta\eta}\Psi_\eta$ is always zero. The trace of $\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon$ concentrates around its mean as follows: for every $u \in [0,1]$,*

$$\mathbb{P}(|\operatorname{Tr}(\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon - \mathbb{E}[\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon])| \geq u) \leq c_1 \exp\left(-\frac{c_2 u^2 p^2 T}{\|\Gamma_h\|_2^2}\right).$$

**Proof** By definition, $\Psi_\eta = \omega I$ is diagonal. For $h \geq 1$, $L_{\eta\eta}$ is superdiagonal of rank $h$ by blocks, so $\Psi_\eta' L_{\eta\eta}\Psi_\eta$ is too. This means that the trace of $\Psi_\eta' L_{\eta\eta}\Psi_\eta$ is zero almost surely. The case of $\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon$ is harder since $\Psi_\varepsilon = \operatorname{Cov}[X]^{1/2}$ is not block-diagonal.

Luckily, we can compute an explicit formula for the trace, again thanks to Lemma 10:

$$\begin{aligned} \operatorname{Tr}(\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon) &= \operatorname{Tr}\left(\frac{1}{T-h}\sum_{t=1}^{T-h} \Psi_\varepsilon' P_t' \operatorname{diag}(\pi_t) \frac{\mathbf{1}_{d_2}\mathbf{1}_{d_1}'}{S(h)_{d_1,d_2}} \operatorname{diag}(\pi_{t+h}) P_{t+h}\Psi_\varepsilon\right) \\ &= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \operatorname{Tr}\left(\Psi_\varepsilon' P_t' \mathbf{1}_{d_2}\mathbf{1}_{d_1}' P_{t+h}\Psi_\varepsilon\right). \end{aligned}$$

Remembering the definition of $\Psi_\varepsilon$ in Equation (15) gives

$$\begin{aligned} \operatorname{Tr}(\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon) &= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \left(\mathbf{1}_{d_1}' P_{t+h} \operatorname{Cov}[X] P_t' \mathbf{1}_{d_2}\right) \\ &= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} (\Gamma_h)_{d_1,d_2}. \end{aligned}$$

And therefore,

$$\operatorname{Tr}(\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon - \mathbb{E}[\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon]) = (\Gamma_h)_{d_1,d_2}\left(\frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} - 1\right)$$

Like before, we apply Lemma 5: for all $u \in [0,1]$,

$$\mathbb{P}\left(\left|\frac{1}{T-h}\sum_{t=1}^{T-h}\frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}}-1\right|\geq u\right)\leq c_1\exp(-c_2u^2TS(h)_{d_1,d_2}).$$

Since $|(\Gamma_h)_{d_1,d_2}|\leq\|\Gamma_h\|_2$, we can deduce

$$\mathbb{P}\left(|\operatorname{Tr}(\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon-\mathbb{E}[\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon])|\geq u\|\Gamma_h\|_2\right)\leq c_1\exp(-c_2u^2TS(h)_{d_1,d_2})$$

which is equivalent to:

$$\mathbb{P}\left(|\operatorname{Tr}(\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon-\mathbb{E}[\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon])|\geq v\right)\leq c_1\exp\left(-\frac{c_2v^2p^2T}{\|\Gamma_h\|_2^2}\right).$$

■

### C.4 Gaussian Concentration, Episode 2

We will now apply a Gaussian concentration inequality that exploits our knowledge of the $L$ matrices.

**Lemma 24 (Applying Hanson-Wright)** *The deviation probability for $(\widehat{\Gamma}_h)_{d_1,d_2}$ satisfies*

$$\mathbb{P}(|(\widehat{\Gamma}_h-\Gamma_h)_{d_1,d_2}|\geq u)\leq 4\delta+c_1\exp\left(-\frac{c_2p^2Tu^2}{(\|\Psi_\varepsilon\|_2^2+\|\Psi_\eta\|_2^2)^2}\right)+c_1\exp\left(-\frac{c_2p^2Tu^2}{\|\Gamma_h\|_2^2}\right).$$

**Proof** The bound we had reached before our discrete concentration detour is given by Lemma 19, and we can rewrite it as

$$\mathbb{P}(|(\widehat{\Gamma}_h-\Gamma_h)_{d_1,d_2}|\geq u)\leq p_{\varepsilon\varepsilon}+p_{\eta\varepsilon}+p_{\varepsilon\eta}+p_{\eta\eta},$$

where each $p_{ij}$ represents a deviation probability for a specific quadratic form $g_i'\Psi_i'L_{ij}\Psi_jg_j$. We control the norms of these quadratic forms as follows: by Lemmas 20 and 22, with probability at least $1-\delta$, the following eight inequalities occur at the same time if condition (16) holds:

$$\begin{aligned}
\|\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon\|_2&\leq\frac{\|\Psi_\varepsilon\|_2^2}{p^2T} & \|\Psi_\varepsilon'L_{\varepsilon\varepsilon}\Psi_\varepsilon\|_F^2&\leq\frac{\|\Psi_\varepsilon\|_2^4}{p^2T}\\
\|\Psi_\eta'L_{\eta\varepsilon}\Psi_\varepsilon\|_2&\leq\frac{\|\Psi_\eta\|_2\|\Psi_\varepsilon\|_2}{p^2T} & \|\Psi_\eta'L_{\eta\varepsilon}\Psi_\varepsilon\|_F^2&\leq\frac{\|\Psi_\eta\|_2^2\|\Psi_\varepsilon\|_2^2}{p^2T}\\
\|\Psi_\varepsilon'L_{\varepsilon\eta}\Psi_\eta\|_2&\leq\frac{\|\Psi_\varepsilon\|_2\|\Psi_\eta\|_2}{p^2T} & \|\Psi_\varepsilon'L_{\varepsilon\eta}\Psi_\eta\|_F^2&\leq\frac{\|\Psi_\varepsilon\|_2^2\|\Psi_\eta\|_2^2}{p^2T}\\
\|\Psi_\eta'L_{\eta\eta}\Psi_\eta\|_2&\leq\frac{\|\Psi_\eta\|_2^2}{p^2T} & \|\Psi_\eta'L_{\eta\eta}\Psi_\eta\|_F^2&\leq\frac{\|\Psi_\eta\|_2^4}{p^2T}.
\end{aligned}$$

Lemma 7 (applied with $X = g_a$, $Y = g_b$ and $A = \Psi_a' L \Psi_b$) now provides the concentration result we need:

$$p_{\varepsilon\varepsilon} \leq \delta + 2\exp\left(-c_1 p^2 T \min\left\{\frac{(u/4)^2}{\|\Psi_\varepsilon\|_2^4}, \frac{(u/4)}{\|\Psi_\varepsilon\|_2^2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\varepsilon\varepsilon}\Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon])| \geq u/2\right)$$
$$p_{\eta\varepsilon} \leq \delta + 2\exp\left(-c_1 p^2 T \min\left\{\frac{(u/4)^2}{\|\Psi_\eta\|_2^2\|\Psi_\varepsilon\|_2^2}, \frac{(u/4)}{\|\Psi_\eta\|_2\|\Psi_\varepsilon\|_2}\right\}\right)$$
$$p_{\varepsilon\eta} \leq \delta + 2\exp\left(-c_1 p^2 T \min\left\{\frac{(u/4)^2}{\|\Psi_\varepsilon\|_2^2\|\Psi_\eta\|_2^2}, \frac{(u/4)}{\|\Psi_\varepsilon\|_2\|\Psi_\eta\|_2}\right\}\right)$$
$$p_{\eta\eta} \leq \delta + 2\exp\left(-c_1 p^2 T \min\left\{\frac{(u/4)^2}{\|\Psi_\eta\|_2^4}, \frac{(u/4)}{\|\Psi_\eta\|_2^2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\eta\eta}\Psi_\eta) - \mathbb{E}[\Psi_\eta' L_{\eta\eta}\Psi_\eta])| \geq u/2\right).$$

The denominators inside the minimums can be controlled as follows:

$$\max\left\{\|\Psi_\varepsilon\|_2^4, \|\Psi_\varepsilon\|_2^2\|\Psi_\eta\|_2^2, \|\Psi_\eta\|_2^4\right\} \leq \left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right)^2$$
$$\max\left\{\|\Psi_\varepsilon\|_2^2, \|\Psi_\varepsilon\|_2\|\Psi_\eta\|_2, \|\Psi_\eta\|_2^2\right\} \leq \left(\|\Psi_\varepsilon\|_2 + \|\Psi_\eta\|_2\right)^2 \leq 2\left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right).$$

This means we can upper bound each of the 4 minimums by

$$\min\left\{\left(\frac{u/4}{\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2}\right)^2, \frac{u/4}{\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2}\right\}.$$

From now on, we additionally suppose that

$$\frac{u/4}{\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2} \leq 1 \tag{17}$$

This enables us to get rid of the $\min\{\cdot,\cdot\}$ by reducing it to the (smaller) quadratic term only. We end up with

$$p_{\varepsilon\varepsilon} \leq \delta + 2\exp\left(-c_1 p^2 T \frac{u^2}{\left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right)^2}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\varepsilon\varepsilon}\Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon])| \geq u/2\right)$$
$$p_{\eta\varepsilon} \leq \delta + 2\exp\left(-c_1 p^2 T \frac{u^2}{\left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right)^2}\right)$$
$$p_{\varepsilon\eta} \leq \delta + 2\exp\left(-c_1 p^2 T \frac{u^2}{\left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right)^2}\right)$$
$$p_{\eta\eta} \leq \delta + 2\exp\left(-c_1 p^2 T \frac{u^2}{\left(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2\right)^2}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\eta\eta}\Psi_\eta) - \mathbb{E}[\Psi_\eta' L_{\eta\eta}\Psi_\eta])| \geq u/2\right).$$

As for the trace terms, they are taken care of by Lemma 23:

$$\mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\eta\eta}\Psi_\eta) - \mathbb{E}[\Psi_\eta' L_{\eta\eta}\Psi_\eta])| \geq u/2\right) = 0$$
$$\mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L_{\varepsilon\varepsilon}\Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L_{\varepsilon\varepsilon}\Psi_\varepsilon])| \geq u/2\right) \leq c_3\exp\left(-c_4\frac{(u/2)^2 p^2 T}{\|\Gamma_h\|_2^2}\right).$$

We plug this in and merge some constants to obtain the expected result:

$$p_{\varepsilon\varepsilon} + p_{\eta\varepsilon} + p_{\varepsilon\eta} + p_{\eta\eta} \leq 4\delta + 8\exp\left(-\frac{c_1 p^2 T u^2}{(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2)^2}\right) + c_3 \exp\left(-\frac{c_4 p^2 T u^2}{\|\Gamma_h\|_2^2}\right).$$

■

**Lemma 25 (Spectral norms of $\Psi_\varepsilon$, $\Psi_\eta$ and $\Gamma_h$)** *The matrices $\Psi_\varepsilon$ and $\Psi_\eta$ satisfy:*

$$\|\Psi_\varepsilon\|_2^2 \leq \frac{\sigma_{\min}^2}{(1-\vartheta)^2} \qquad \|\Psi_\eta\|_2^2 = \omega^2 \qquad \|\Gamma_h\|_2 \leq \frac{\vartheta^h}{1-\vartheta}\sigma_{\min}^2.$$

**Proof** We can write $\Psi_\varepsilon$ as a sum of Kronecker products:

$$\Psi_\varepsilon^2 = \mathrm{Cov}[X] = I \otimes \Gamma_0(\theta) + \sum_{t=1}^{T-1}\left[J_t \otimes \theta^t \Gamma_0(\theta) + J_t' \otimes \Gamma_0(\theta)\theta'^t\right]$$

As a consequence, we have control over its spectral norm thanks to Lemma 30:

$$\begin{aligned}
\|\Psi_\varepsilon\|_2^2 = \|\Psi_\varepsilon^2\|_2 &\leq \|I\|_2 \times \|\Gamma_0(\theta)\|_2 + \sum_{t=1}^{T-1}\left[\|J_t\|_2 \times \|\theta^t\Gamma_0(\theta)\|_2 + \|J_t'\|_2 \times \|\Gamma_0(\theta)\theta'^t\|_2\right] \\
&\leq \|\Gamma_0(\theta)\|_2\left(1 + 2\sum_{t=1}^{T-1}\|\theta\|_2^t\right) \leq \frac{\|\Sigma\|_2}{1-\vartheta^2}\left(1 + 2\frac{\|\theta\|_2}{1-\|\theta\|_2}\right) \\
&\leq \frac{\sigma_{\min}^2}{1-\vartheta^2}\frac{1+\vartheta}{1-\vartheta} = \frac{\sigma_{\min}^2}{(1-\vartheta)^2}.
\end{aligned}$$

The spectral norm of $\Psi_\eta$ is easily seen to equal $\|\Psi_\eta\|_2^2 = \|\omega^2 I\|_2 = \omega^2$. Finally, we turn to $\Gamma_h$ using Lemma 11:

$$\|\Gamma_h\|_2 = \|\theta^h\Gamma_0(\theta)\|_2 \leq \frac{\vartheta^h}{1-\vartheta^2}\|\Sigma\|_2 \leq \frac{\vartheta^h}{1-\vartheta}\sigma_{\min}^2.$$

■

We can now finish the proof of Theorem 2.

**Proof** Let us plug Lemma 25 into Lemma 24

$$\begin{aligned}
\mathbb{P}(|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \geq u) &\leq 4\delta + c_1\exp\left(-\frac{c_2 p^2 T u^2}{\left(\frac{\sigma_{\min}^2}{(1-\vartheta)^2} + \omega^2\right)^2}\right) + c_1\exp\left(-\frac{c_2 p^2 T u^2}{\left(\frac{\vartheta^h\sigma_{\min}^2}{1-\vartheta}\right)^2}\right) \\
&\leq 4\delta + c_1\exp\left(-\frac{c_2(1-\vartheta)^4 p^2 T u^2}{(\sigma_{\min}^2 + \omega^2)^2}\right) + c_1\exp\left(-\frac{c_2(1-\vartheta)^2 p^2 T u^2}{\vartheta^{2h}\sigma_{\min}^4}\right).
\end{aligned}$$

We merge both exponential terms by keeping the least negative exponent:

$$\mathbb{P}(|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \geq u) \leq 4\delta + c_1\exp\left(-\frac{c_2(1-\vartheta)^4 p^2 T}{(\sigma_{\min}^2 + \omega^2)^2}u^2\right).$$

All that is left to do is choose $u$ such that

$$\mathbb{P}(|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \geq u) \leq 8\delta,$$

which will be true if

$$c_1 \exp\left(-\frac{c_2(1-\vartheta)^4 p^2 T}{(\sigma_{\min}^2 + \omega^2)^2} u^2\right) \leq 4\delta \quad \iff \quad u \geq c\frac{\sqrt{\log(D/\delta)}(\sigma_{\min}^2 + \omega^2)}{(1-\vartheta)^2 p\sqrt{T}}.$$

With this choice of $u$, the assumption we made in Equation (17) translates into

$$\frac{\sqrt{\log(1/\delta)}(\sigma_{\min}^2 + \omega^2)}{(1-\vartheta)^2 p\sqrt{T}(\|\Psi_\varepsilon\|_2^2 + \|\Psi_\eta\|_2^2)} \leq c$$

Using Lemma 25 again yields the condition

$$\frac{\sqrt{\log(1/\delta)}}{(1-\vartheta)^2 p\sqrt{T}}\left(1 + \frac{\sigma_{\min}^2}{\omega^2}\right) \leq c. \tag{18}$$

Summing up, we just proved that with probability at least $1 - 8\delta$,

$$|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \leq c\frac{\sigma_{\min}^2 + \omega^2}{(1-\vartheta)^2}\frac{\sqrt{\log(1/\delta)}}{p\sqrt{T}}.$$

We finish with a union bound, applying the previous result to all pairs $(d_1, d_2) \in [D]^2$. With probability greater than $1 - 2D^2\delta$, we have:

$$\max_{d_1,d_2} |(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| = \|\widehat{\Gamma}_h - \Gamma_h\|_{\max} \leq c\frac{\sigma_{\min}^2 + \omega^2}{(1-\vartheta)^2}\frac{\sqrt{\log(1/\delta)}}{p\sqrt{T}}.$$

Replacing $\delta$ with $D^2\delta$ gives us the result we wanted: with probability greater than $1 - \delta$,

$$\|\widehat{\Gamma}_h - \Gamma_h\|_{\max} \leq c\frac{\sigma_{\min}^2 + \omega^2}{(1-\vartheta)^2}\frac{\sqrt{\log(D/\delta)}}{p\sqrt{T}}.$$

■

### C.5 Behavior of the Dantzig selector

We now walk the final steps from the error on $\widehat{\Gamma}_h$ to the error on $\widehat{\theta}$. In order to obtain Theorem 2, we adapt the convergence proof from Han et al. (2015, Appendix A.1). However, we use our own notations and our custom concentration results for $\widehat{\Gamma}_h$. To make comparison between both papers easier, we provide a dictionary of the main notations in Table 3.

**Lemma 26 (Feasibility of the real $\theta$)** *If we define*

$$\mathrm{err}(\delta) := c\frac{\sigma_{\min}^2 + \omega^2}{(1-\vartheta)^2}\frac{\sqrt{\log(D/\delta)}}{p\sqrt{T}}$$

*and select the penalization level*

$$\lambda_0 := (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta),$$

*then the real $\theta$ is a feasible solution to the optimization problem* (LP) *with probability* $1 - \delta$.

| | **This paper** | **Han et al. (2015)** |
|---|---|---|
| VAR def | $X_t = \theta X_{t-1} + \varepsilon_t$ | $X_t = A_1' X_{t-1} + Z_t$ |
| Covariance | $\Gamma_h = \mathrm{Cov}(X_h, X_0)$ | $\Sigma_i = \mathrm{Cov}(X_0, X_i)$ |
| Yule-Walker | $\Gamma_h = \theta^h \Gamma_0$ | $\Sigma_i = \Sigma_0 A_1^i$ |
| Covariance estimate | $\widehat{\Gamma}_h$ | $S_i$ |
| Covariance error | $\mathrm{err}(\delta)$ | $\zeta_i$ |
| Optimization constraint | $\|M\widehat{\Gamma}_0 - \widehat{\Gamma}_1\|_{\max} \le \lambda_0$ | $\|S_0 M - S_1\|_{\max} \le \lambda_0$ |
| Optimization objective | $\|\mathrm{vec}(M)\|_1$ | $\|\mathrm{vec}(M)\|_1$ |
| Threshold in proof | $\nu$ | $\lambda_1$ |

Table 3: Notation correspondence between this paper and Han et al. (2015)

**Proof** Our sparse transition estimator is defined as a solution to (LP). The end goal is to control the error $\|\widehat{\theta} - \theta\|_1$, where $\theta = \Gamma_1 \Gamma_0^{-1}$ is the true transition matrix. We start by choosing a specific $\lambda_0$ such that $\theta$ is feasible with high probability:

$$\begin{aligned}
\|\theta\widehat{\Gamma}_0 - \widehat{\Gamma}_1\|_{\max} &= \|\Gamma_1\Gamma_0^{-1}\,\widehat{\Gamma}_0 - \widehat{\Gamma}_1\|_{\max} \\
&= \|\Gamma_1\Gamma_0^{-1}\,\widehat{\Gamma}_0 - \Gamma_1 + \Gamma_1 - \widehat{\Gamma}_1\|_{\max} \\
&\le \|\Gamma_1\Gamma_0^{-1}\,\widehat{\Gamma}_0 - \Gamma_1\Gamma_0^{-1}\Gamma_0\|_{\max} + \|\Gamma_1 - \widehat{\Gamma}_1\|_{\max} \\
&= \|\theta(\widehat{\Gamma}_0 - \Gamma_0)\|_{\max} + \|\Gamma_1 - \widehat{\Gamma}_1\|_{\max}
\end{aligned}$$

By Lemma 32,

$$\|\theta(\widehat{\Gamma}_0 - \Gamma_0)\|_{\max} \le \|\theta\|_\infty \|\widehat{\Gamma}_0 - \Gamma_0\|_{\max}$$

By Theorem 2, with probability greater than $1-\delta$ (more precisely greater than $1-2\delta$, but we ignore constants here),

$$\|\widehat{\Gamma}_0 - \Gamma_0\|_{\max} \le \mathrm{err}(\delta) \quad \text{and} \quad \|\widehat{\Gamma}_1 - \Gamma_1\|_{\max} \le \mathrm{err}(\delta)$$

Thus, with probability greater than $1-\delta$,

$$\|\theta\widehat{\Gamma}_0 - \widehat{\Gamma}_1\|_{\max} \le (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta)$$

which is exactly the feasibility criterion for (LP) if $\lambda_0 = (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta)$. ■

**Lemma 27 (Error in max norm)** *If we choose $\lambda_0 = (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta)$, then with probability at least $1-\delta$,*

$$\|\widehat{\theta} - \theta\|_{\max} \le 2\lambda_0 \|\Gamma_0^{-1}\|_1$$

**Proof**

$$\begin{aligned}
\|\widehat{\theta} - \theta\|_{\max} &= \|\widehat{\theta} - \Gamma_1\Gamma_0^{-1}\|_{\max} \\
&= \|(\widehat{\theta}\Gamma_0 - \Gamma_1)\Gamma_0^{-1}\|_{\max} \\
&= \|(\widehat{\theta}\Gamma_0 - \widehat{\theta}\widehat{\Gamma}_0 + \widehat{\theta}\widehat{\Gamma}_0 - \widehat{\Gamma}_1 + \widehat{\Gamma}_1 - \Gamma_1)\Gamma_0^{-1}\|_{\max} \\
&\le \|(\widehat{\theta}\Gamma_0 - \widehat{\theta}\widehat{\Gamma}_0)\Gamma_0^{-1}\|_{\max} + \|(\widehat{\theta}\widehat{\Gamma}_0 - \widehat{\Gamma}_1)\Gamma_0^{-1}\|_{\max} + \|(\widehat{\Gamma}_1 - \Gamma_1)\Gamma_0^{-1}\|_{\max}
\end{aligned}$$

By Lemma 32,

$$\begin{aligned}\|\widehat{\theta}-\theta\|_{\max} &\leq \left(\|\widehat{\theta}(\Gamma_0-\widehat{\Gamma}_0)\|_{\max}+\|\widehat{\theta}\widehat{\Gamma}_0-\widehat{\Gamma}_1\|_{\max}+\|\widehat{\Gamma}_1-\Gamma_1\|_{\max}\right)\|\Gamma_0^{-1}\|_1\\ &\leq \left(\|\widehat{\theta}\|_\infty\|\Gamma_0-\widehat{\Gamma}_0\|_{\max}+\|\widehat{\theta}\widehat{\Gamma}_0-\widehat{\Gamma}_1\|_{\max}+\|\widehat{\Gamma}_1-\Gamma_1\|_{\max}\right)\|\Gamma_0^{-1}\|_1\end{aligned}$$

We want to control $\|\widehat{\theta}\|_\infty$ using $\|\theta\|_\infty$. Let us recall that the operator $\ell_\infty$ norm is equal to the maximum $\ell_1$ norm of the rows of a matrix. To control the rows of $\widehat{\theta}$, we notice that the optimization problem defining $\widehat{\theta}$, namely

$$\min_{M\in\mathbb{R}^{D\times D}}\|\mathrm{vec}(M)\|_1 \quad \text{s.t.} \quad \|M\widehat{\Gamma}_0-\widehat{\Gamma}_1\|_{\max}\leq\lambda_0$$

is equivalent to the row-wise minimization

$$\forall i, \quad \min_{M_{i,\cdot}\in\mathbb{R}^{1\times D}}\|M_{i,\cdot}\|_1 \quad \text{s.t.} \quad \|M_{i,\cdot}\widehat{\Gamma}_0-(\widehat{\Gamma}_1)_{i,\cdot}\|_{\max}\leq\lambda_0$$

From this, we deduce that each row of the optimum $\widehat{\theta}$ satisfies $\|\widehat{\theta}_{i,\cdot}\|_1\leq\|\theta_{i,\cdot}\|_1$, which implies $\|\widehat{\theta}\|_\infty\leq\|\theta\|_\infty$. Going back to our error estimate, we get:

$$\|\widehat{\theta}-\theta\|_{\max}\leq\left(\|\theta\|_\infty\|\Gamma_0-\widehat{\Gamma}_0\|_{\max}+\|\widehat{\theta}\widehat{\Gamma}_0-\widehat{\Gamma}_1\|_{\max}+\|\widehat{\Gamma}_1-\Gamma_1\|_{\max}\right)\|\Gamma_0^{-1}\|_1$$

Note that the middle term is smaller than $\lambda_0$ because the optimum $\widehat{\theta}$ is a feasible solution. Meanwhile, the first and third term are smaller than $\mathrm{err}(\delta)$ with probability $1-\delta$:

$$\|\widehat{\theta}-\theta\|_{\max}\leq\left(\|\theta\|_\infty\,\mathrm{err}(\delta)+\lambda_0+\mathrm{err}(\delta)\right)\|\Gamma_0^{-1}\|_1=2\lambda_0\|\Gamma_0^{-1}\|_1$$

■

To complete the proof of Theorem 3, we simply need to go from the max norm to the $\ell_\infty$ operator norm.

**Proof** Let $\nu>0$ be a threshold (to be chosen later). We define

$$s_1=\max_i\sum_j\min\left\{\frac{|\theta_{i,j}|}{\nu},1\right\} \quad \text{and} \quad \mathcal{T}_i=\{j:|\theta_{i,j}|\geq\nu\}$$

With high probability, the following holds for any row $i$:

$$\begin{aligned}\|\widehat{\theta}_{i,\cdot}-\theta_{i,\cdot}\|_1 &\leq \|\widehat{\theta}_{i,\mathcal{T}_i^c}-\theta_{i,\mathcal{T}_i^c}\|_1+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &\leq \|\widehat{\theta}_{i,\mathcal{T}_i^c}\|_1+\|\theta_{i,\mathcal{T}_i^c}\|_1+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &= (\|\widehat{\theta}_{i,\cdot}\|_1-\|\widehat{\theta}_{i,\mathcal{T}_i}\|_1)+\|\theta_{i,\mathcal{T}_i^c}\|_1+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &\leq \|\theta_{i,\cdot}\|_1-\|\widehat{\theta}_{i,\mathcal{T}_i}\|_1+\|\theta_{i,\mathcal{T}_i^c}\|_1+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &= (\|\theta_{i,\mathcal{T}_i}\|_1+\|\theta_{i,\mathcal{T}_i^c}\|_1)-\|\widehat{\theta}_{i,\mathcal{T}_i}\|_1+\|\theta_{i,\mathcal{T}_i^c}\|_1+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &= 2\|\theta_{i,\mathcal{T}_i^c}\|_1+(\|\theta_{i,\mathcal{T}_i}\|_1-\|\widehat{\theta}_{i,\mathcal{T}_i}\|_1)+\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\\ &\leq 2\|\theta_{i,\mathcal{T}_i^c}\|_1+2\|\widehat{\theta}_{i,\mathcal{T}_i}-\theta_{i,\mathcal{T}_i}\|_1\end{aligned}$$

By definition of $\mathcal{T}_i$, for all $j \in \mathcal{T}_i^c$, $|\theta_{i,j}| \leq \nu$, hence

$$\|\theta_{i,\mathcal{T}_i^c}\|_1 = \sum_{j \in \mathcal{T}_i^c} |\theta_{i,j}| = \sum_{j \in \mathcal{T}_i^c} \min\{|\theta_{i,j}|, \nu\} \leq \sum_j \min\{|\theta_{i,j}|, \nu\} \leq \nu s_1$$

Meanwhile, the second term satisfies

$$\|\widehat{\theta}_{i,\mathcal{T}_i} - \theta_{i,\mathcal{T}_i}\|_1 \leq |\mathcal{T}_i| \times \|\widehat{\theta} - \theta\|_{\max}$$

And by definition of $\mathcal{T}_i$, for all $j \in \mathcal{T}_i$, $|\theta_{i,j}| \geq \nu$, hence

$$|\mathcal{T}_i| = \sum_{j \in \mathcal{T}_i} 1 = \sum_{j \in \mathcal{T}_i} \min\left\{\frac{|\theta_{i,j}|}{\nu}, 1\right\} \leq \sum_j \min\left\{\frac{|\theta_{i,j}|}{\nu}, 1\right\} \leq s_1$$

Combining all of this, we get that with high probability,

$$\|\widehat{\theta}_{i,\cdot} - \theta_{i,\cdot}\|_1 \leq 2(\nu + 2\lambda_0 \|\Gamma_0^{-1}\|_1) s_1$$

Judging by the last Equation, it makes sense to choose $\nu = 2\lambda_0 \|\Gamma_0^{-1}\|_1$. Furthermore, our sparsity hypothesis on $\theta$ implies that for all but $s$ of the coefficients of any row $i$, $\min\{|\theta_{i,j}|, \nu\} = |\theta_{i,j}| = 0$. We deduce that for every $i$,

$$\sum_j \min\{|\theta_{i,j}|, \nu\} \leq s \max_j \min\{|\theta_{i,j}|, \nu\} \leq \nu s$$

which directly implies

$$\nu s_1 = \max_i \sum_j \min\{|\theta_{i,j}|, \nu\} \leq \nu s$$

We finally find that with high probability,

$$\|\widehat{\theta}_{i,\cdot} - \theta_{i,\cdot}\|_1 \leq 4\nu s_1 \leq 4\nu s = 8\lambda_0 \|\Gamma_0^{-1}\|_1 s$$

With the help of a union bound, again with high probability,

$$\|\widehat{\theta} - \theta\|_\infty = \max_i \|\widehat{\theta}_{i,\cdot} - \theta_{i,\cdot}\|_1 \leq 8\lambda_0 \|\Gamma_0^{-1}\|_1 s$$

We replace the value of $\lambda_0$ and obtain

$$\|\widehat{\theta} - \theta\|_\infty \leq 8(\|\theta\|_\infty + 1)\,\mathrm{err}(\delta) \|\Gamma_0^{-1}\|_1 s$$

Once we plug in the value of $\mathrm{err}(\delta)$, the resulting high-probability error bound reads

$$\|\widehat{\theta} - \theta\|_\infty \leq c \frac{\|\theta\|_\infty + 1}{(1-\vartheta)^2} \frac{\sigma_{\min}^2 + \omega^2}{\|\Gamma_0^{-1}\|_1^{-1}} \frac{s\sqrt{\log(D/\delta)}}{p\sqrt{T}}$$

Since $\vartheta$ only acted as an upper bound on $\|\theta\|_2$ in this proof, we can define

$$\gamma_u(\theta) = \frac{\|\theta\|_\infty + 1}{(1 - \|\theta\|_2)^2} \frac{\sigma_{\min}^2 + \omega^2}{\|\Gamma_0^{-1}\|_1^{-1}}$$

to obtain the compressed expression

$$\|\widehat{\theta} - \theta\|_\infty \leq c\gamma_u(\theta) \frac{s\sqrt{\log(D/\delta)}}{p\sqrt{T}}.$$

■

## Appendix D. Independent Lemmas

### D.1 Linear Algebra

The following set of results will sometimes be used in matrix calculations without explicit justifications.

**Lemma 28 (Weyl's inequality)** *Let $A$ and $B$ be two $n \times n$ symmetric matrices. Then for all $i$ we have:*

$$\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_1(B).$$

*In particular,*

$$\lambda_{\min}(A) + \lambda_{\min}(B) \leq \lambda_{\min}(A+B).$$

**Proof** See Horn and Johnson (2012, Theorem 4.3.1). ■

**Lemma 29 (Ostrowski)** *Let $S$ and $A$ be two $n \times n$ matrices with $S$ symmetric. For all $i$, there is a real number $r_i \in [\varsigma_{\min}(A)^2, \varsigma_{\max}(A)^2]$ such that $\lambda_i(ASA') = r_i\lambda_i(S)$.*

**Proof** See Horn and Johnson (2012, Theorem 4.5.9 and Corollary 4.5.11) ■

**Lemma 30 (Singular values of the Kronecker product)** *Let $A$ and $B$ be two matrices. Then*

$$\|A \otimes B\|_2 \leq \|A\|_2 \|B\|_2.$$

**Proof** See Horn and Johnson (1994, Theorem 4.2.15). ■

**Lemma 31** *For any two matrices $A$ and $B$, we have:*

$$\|AB\|_F \leq \min\left\{\|A\|_2\|B\|_F, \|A\|_F\|B\|_2\right\}$$

**Proof** The Loewner order on symmetric matrices satisfies the following properties:

$$\begin{aligned} \forall (A,B) \in \mathcal{S}_n(\mathbb{R}), \forall C, \quad A \preceq B &\implies C'AC \preceq C'BC \\ \forall (A,B) \in \mathcal{S}_n(\mathbb{R}), \quad A \preceq B &\implies \operatorname{Tr}(A) \leq \operatorname{Tr}(B). \end{aligned}$$

The first inequality is true because if $x$ is a vector, $x'C'(B-A)Cx = (Cx)'(B-A)(Cx) \geq 0$ due to the Loewner positivity of $B - A$. The second inequality can be directly deduced from the relation between the spectra of $A$ and $B$. Therefore, since $A'A$ is symmetric,

$$B'A'AB \leq \lambda_{\max}(A'A)B'B$$

which implies

$$\|AB\|_F^2 = \operatorname{Tr}(B'A'AB) \leq \lambda_{\max}(AA')\operatorname{Tr}(B'B) = \|A\|_2^2\|B\|_F^2.$$

The proof for the other inequality is identical. ■

**Lemma 32** *Let $A$ and $B$ be two matrices with compatible sizes: then*

$$\|AB\|_{\max} \leq \min\{\|A\|_\infty \|B\|_{\max}, \|A\|_{\max}\|B\|_1\}.$$

**Proof**

$$\|AB\|_{\max} = \max_{i,j} |(AB)_{i,j}| = \max_{i,j} \left| \sum_k A_{i,k} B_{k,j} \right|$$

We easily deduce:

$$\|AB\|_{\max} \leq \max_i \left| \sum_k A_{i,k} \right| \times \|B\|_{\max} = \|A\|_\infty \|B\|_{\max}$$

$$\|AB\|_{\max} \leq \|A\|_{\max} \times \max_j \left| \sum_k B_{k,j} \right| = \|A\|_{\max}\|B\|_1$$

■

**Lemma 33** *Let $A$ be an $m \times n$ rectangular matrix and $b > 0$ be a positive real number. Then the eigenvalues of $M = A'(bI + AA')^{-1}A$ are given by*

$$\lambda_i(M) = \frac{\varsigma_i(A)^2}{b + \varsigma_i(A)^2}.$$

*In particular, its spectral norm is*

$$\|A'(bI + AA')^{-1}A\|_2 = \frac{\|A\|_2^2}{b + \|A\|_2^2}.$$

**Proof** Let $A = USV'$ be the singular value decomposition of $A$, in which $S = \mathrm{diag}(\varsigma_i)$ is rectangular of size $m \times n$ while $U$ and $V$ are both square orthogonal matrices. We have

$$\begin{aligned}
A'(bI + AA')^{-1}A &= VSU' \left(bI + USS'U'\right)^{-1} USV' \\
&= VSU' \left(U(bI + S^2)U'\right)^{-1} USV' \\
&= VS(U'U)(bI + S^2)^{-1}(U'U)SV' \\
&= VS(bI + S^2)^{-1}SV' \\
&= V \,\mathrm{diag}\left(\frac{\varsigma_i(A)^2}{b + \varsigma_i(A)^2}\right) V'.
\end{aligned}$$

■

### D.2 Probability

**Lemma 34 (Fano's method)** *Let $\theta_0, ..., \theta_M$ be $M+1$ parameters that are $2\tau$-separated w.r.t. a distance $d$*

$$\forall i \neq j, \quad d(\theta_i, \theta_j) \geq 2\tau$$

*and such that the average KL divergence between $\mathbb{P}_{\theta_i}$ and $\mathbb{P}_{\theta_0}$ is small enough*

$$\frac{1}{M+1}\sum_{i=1}^{M} \mathrm{KL}\left\{\mathbb{P}_{\theta_i} \parallel \mathbb{P}_{\theta_0}\right\} \leq \alpha \log M \quad \textit{with} \quad 0 < \alpha < 1 \tag{19}$$

*Then the minimax probability of an error at threshold $\tau$ satisfies:*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_s} \mathbb{P}_\theta \left[ d\left(\widehat{\theta}, \theta\right) \geq \tau \right] \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

**Proof** See Tsybakov (2008, Section 2.2 + Corollary 2.6). In particular, since $M \mapsto \frac{\log(M+1)-\log 2}{\log M}$ is increasing, setting $\alpha = \frac{\log(3)-\log(2)}{2\log(2)}$ is enough to obtain a minimax risk $\geq \alpha \geq 1/2$. ■

**Lemma 35 (Chain rule for KL divergence)** *If $\mathbb{P}_0$ and $\mathbb{P}_1$ are probability densities on a product space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$ discrete, then:*

$$\mathrm{KL}\left\{\mathbb{P}_0[X,Y] \parallel \mathbb{P}_1[X,Y]\right\} = \mathrm{KL}\left\{\mathbb{P}_0[X] \parallel \mathbb{P}_1[X]\right\} + \mathbb{E}_X\left[\mathrm{KL}\left\{\mathbb{P}_0[Y|X] \parallel \mathbb{P}_1[Y|X]\right\}\right].$$

**Proof** See Cover and Thomas (2012, Theorem 2.5.3). ■

**Lemma 36 (KL divergence between Gaussians)** *The KL divergence between two multivariate Gaussian distributions $\mathbb{P}_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ of dimension $n$ is*

$$\mathrm{KL}\left\{\mathbb{P}_0 \parallel \mathbb{P}_1\right\} = \frac{1}{2}\left(\mathrm{Tr}(\Sigma_0\Sigma_1^{-1}) + (\mu_1-\mu_0)'\Sigma_1^{-1}(\mu_1-\mu_0) - n + \mathrm{logdet}(\Sigma_1\Sigma_0^{-1})\right).$$

**Proof** See Duchi (2007, page 13). ■

**Lemma 37 (KL divergence between close Gaussians)** *Let $\Delta$ be a symmetric matrix of size $n$ such that $\lambda_{\min}(\Delta) > -1$, and let $M$ be a rectangular matrix such that $MM' \succ 0$. Then the KL divergence between*

$$\mathbb{P}_1 = \mathcal{N}(\mu, M(I+\Delta)M') \quad \textit{and} \quad \mathbb{P}_0 = \mathcal{N}(\mu, MM')$$

*satisfies*

$$\mathrm{KL}\left\{\mathbb{P}_1 \parallel \mathbb{P}_0\right\} \leq \frac{\|\Delta\|_F^2}{2(1+\lambda_{\min}(\Delta))}.$$

**Proof** From Lemma 36 (beware of the switch between $\mathbb{P}_0$ and $\mathbb{P}_1$) we get:

$$\begin{aligned}\mathrm{KL}\left\{\mathbb{P}_1 \parallel \mathbb{P}_0\right\} &= \frac{1}{2}\left(\mathrm{Tr}(\Sigma_1\Sigma_0^{-1}) + (\mu_0-\mu_1)'\Sigma_0^{-1}(\mu_0-\mu_1) - n + \mathrm{logdet}(\Sigma_0\Sigma_1^{-1})\right)\\ &= \frac{1}{2}\left(\mathrm{Tr}(M(I+\Delta)M^{-1}) - n - \mathrm{logdet}(M(I+\Delta)M^{-1})\right)\\ &= \frac{1}{2}\left(\mathrm{Tr}(\Delta) - \mathrm{logdet}(I+\Delta)\right).\end{aligned}$$

As it happens, for small deviations from the identity, the log-determinant is almost equal to the trace. Indeed, since

$$\forall x > -1, \quad \log(1+x) \geq \frac{x}{1+x},$$

we have

$$\begin{aligned}\mathrm{Tr}(\Delta) - \mathrm{logdet}(I+\Delta) &= \sum_{k=1}^n \lambda_k(\Delta) - \sum_{k=1}^n \log(1+\lambda_k(\Delta))\\ &\leq \sum_{k=1}^n \lambda_k(\Delta) - \sum_{k=1}^n \frac{\lambda_k(\Delta)}{1+\lambda_k(\Delta)}\\ &= \sum_{k=1}^n \frac{\lambda_k(\Delta)^2}{1+\lambda_k(\Delta)} \leq \frac{1}{\min_k(1+\lambda_k(\Delta))}\sum_{k=1}^n \lambda_k(\Delta)^2\\ &= \frac{\|\Delta\|_F^2}{1+\lambda_{\min}(\Delta)}.\end{aligned}$$

■

**Lemma 38 (Chernoff inequality for Bernoulli variables)** *Let $(X_t)$ be sequence of independent $\mathcal{B}(p)$ variables. Their average satisfies*

$$\forall u \in [0,1], \quad \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T X_t - p\right| \geq up\right) \leq c_1 \exp\left(-c_2 u^2 T p\right).$$

**Proof** See Dubhashi and Panconesi (2009, Theorem 1.1). ■

**Lemma 39 (Doeblin condition and mixing time)** *Let $(X_t)$ be an irreducible aperiodic Markov chain with state space $\mathcal{X}$, transition matrix $P$ and stationary distribution $\mu$. Suppose that $(X_t)$ satisfies the Doeblin condition:*

$$\exists r \in \mathbb{N}, \exists \delta > 0, \forall (x,y) \in \mathcal{X}^2, \quad P^r(x,y) \geq \delta\mu(y).$$

*Then the mixing time of $X_t$, defined as*

$$t_{\mathrm{mix}}(\epsilon) = \min\left\{t \in \mathbb{N} : \max_{x\in\mathcal{X}} \left\|P^t(x,\cdot) - \mu\right\|_{\mathrm{TV}} \leq \epsilon\right\},$$

*satisfies:*

$$t_{\mathrm{mix}}(\epsilon) \geq r\left(1 + \frac{\log\frac{1}{\epsilon}}{\log\frac{1}{1-\delta}}\right).$$

**Proof** The proof of Levin and Peres (2017, Theorem 5.4) shows that with our assumptions,

$$\forall x \in \mathcal{X}, \quad \left\| P^t(x, \cdot) - \mu \right\|_{\mathrm{TV}} \leq (1-\delta)^{\lfloor t/r \rfloor}.$$

From which we can deduce a sufficient condition for $\epsilon$-mixing:

$$(1-\delta)^{\lfloor t/r \rfloor} \leq \epsilon \quad \iff \quad \left\lfloor \frac{t}{r} \right\rfloor \geq \frac{\log(\epsilon)}{\log(1-\delta)} \quad \Longleftarrow \quad \frac{t}{r} - 1 \geq \frac{\log \frac{1}{\epsilon}}{\log \frac{1}{1-\delta}}.$$

The result follows easily. ■

**Lemma 40 (Chernoff inequality for Markov chains)** *Let $(X_t)$ be an ergodic stationary Markov chain with finite state space $\mathcal{X}$. We consider a function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}[f(X_t)] = \mu$. Then:*

$$\forall u \in [0,1], \quad \mathbb{P}\left( \left| \frac{1}{T} \sum_{t=1}^{T} X_t - \mu \right| \geq u\mu \right) \leq c_1 \exp\left( -c_2 \frac{u^2 T \mu}{t_{\mathrm{mix}}(1/8)} \right)$$

**Proof** See Chung et al. (2012, Theorem 3) ■

**Lemma 41 (Chernoff inequality for Markov chains under Doeblin condition)** *Under the hypotheses of the previous two Lemmas (39 and 40), if the parameters $r$ and $\delta$ in the Doeblin condition are* constants*, then we have:*

$$\forall u \in [0,1], \quad \mathbb{P}\left( \left| \frac{1}{T} \sum_{t=1}^{T} X_t - \mu \right| \geq u\mu \right) \leq c_1 \exp\left( -c_2 u^2 T \mu \right)$$

**Proof** By Lemma 39, since $r$ and $\delta$ are constants, the $\frac{1}{8}$-mixing time of $(X_t)$ can be bounded by a constant

$$t_{\mathrm{mix}}(1/8) \leq r\left( 1 + \frac{\log(8)}{\log \frac{1}{1-\delta}} \right) \leq c_3,$$

which we merge with the $c_2$ inside the exponential of Lemma 40. ■

**Lemma 42 (Gilbert-Varshamov)** *Let $\mathcal{H} = \{0,1\}^d$ be the $d$-dimensional binary hypercube. If $d \geq 8$, there exists a pruned subset $\mathcal{K} \subset \mathcal{H}$ such that*

$$\forall (x,y) \in \mathcal{K}, \ \|x - y\|_1 \geq \frac{d}{8} \quad \textit{and} \quad |\mathcal{K}| \geq 2^{d/8}.$$

**Proof** See Tsybakov (2008, Lemma 2.9) ■

## Appendix E. Glossary

Here is a list of the most frequent symbols and their meaning.

Dimensions:

- $t \in [T]$: time step
- $d \in [D]$ or $e \in \mathcal{E}$: dimension
- $n \in [N]$: number of days

State process:

- $X_t$: state process / network congestion
- $\theta$: transition matrix
- $\varepsilon_t$: innovations
- $\Sigma$: covariance matrix of $\varepsilon_t$
- $\sigma^2_{\min}, \sigma^2_{\max}$: extremal eigenvalues of $\Sigma$
- $s$: sparsity level of $\theta$ (number of non-zero coefficients in each row)
- $\vartheta$: maximum $\ell_2$ norm for $\theta$
- $\Theta_s$: set of feasible values for $\theta$
- $\Gamma_h(\theta)$: covariance between $X_{t+h}$ and $X_t$

Observations:

- $\Pi_t$: random sampling matrix
- $\mathcal{D}$: distribution of the sampling matrix
- $p$: fraction of state components activated by observations
- $\mathcal{Q}$: transition matrix for Markov sampling
- $a, b$: transition probabilities for Markov sampling
- $\chi$: minimum distance between $a$ or $b$ and $\{0, 1\}$ (considered constant)
- $\pi_{t,d}$: indicator of whether $(t, d)$ is activated
- $\kappa_{t,d}$: counter of how many times $(t, d)$ was activated
- $Y_t$: observations / train arrival times
- $\eta_t$: noise
- $\omega^2$: variance of $\eta_t$

Estimation:

- $h$: covariance time lag
- $S(h)$: multiplicative scaling for covariance estimation
- $S(h)_{\min}$: smallest coefficient of the scaling matrix
- $C(h)$: additive noise correction for covariance estimation
- $h_0$: minimum covariance time lag for transition estimation

Other:

- $\mathcal{G} = (\mathcal{V}, \mathcal{E})$: network graph
- $\Delta t$: time interval for discretization
- $g$: standard Gaussian vector
- $\mathcal{H}$ (resp. $\mathcal{K}$): binary hypercube (resp. pruned hypercube)
- $\mathcal{I}$: Fisher information matrix
- $L$: bilinear form
- $r$: small radius
- $R(\theta)$: non-constant term in the covariance of $Y$
- $u$: threshold in concentration inequalities
- $\delta$: small probability
- $Q_\Pi$: constant term in the conditional variance of $Y$
- $R_\Pi(\theta)$: varying term in the conditional variance of $Y$
- $\Delta_\Pi(\theta)$: deviation from the identity
- $\gamma_\ell(\mathcal{D})$ (resp. $\gamma_u(\theta)$): signal-to-noise ratio in the lower bound (resp. the upper bound)
- $\tau$: separation of the parameters in Fano's method
- $\Psi_\varepsilon$ (resp. $\Psi_\eta$): link between $X$ (resp. $\eta$) and a standard Gaussian vector

## References

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018. doi: 10/ggkj28.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567, August 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1315. URL https://projecteuclid.org/euclid.aos/1434546214.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, August 2009. ISSN 0090-5364, 2168-8966. doi: 10/fmxd7h. URL https://projecteuclid.org/journals/annals-of-statistics/volume-37/issue-4/Simultaneous-analysis-of-Lasso-and-Dantzig-selector/10.1214/08-AOS620.full.

John P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. CRC Press, March 2010. ISBN 978-1-4200-6658-6.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007. ISSN 0090-5364, 2168-8966. doi: 10/b4rfq6. URL https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-6/The-Dantzig-selector--Statistical-estimation-when-p-is-much/10.1214/009053606000001523.full.

Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models*. Springer Science & Business Media, April 2006. ISBN 978-0-387-28982-3.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-Hoeffding Bounds for Markov Chains: Generalized and Simplified. In Christoph Dürr and Thomas Wilke, editors, *29th International Symposium on Theoretical Aspects of Computer Science (STACS 2012)*, volume 14 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 124–135, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-35-4. doi: 10.4230/LIPIcs.STACS.2012.124. URL http://drops.dagstuhl.de/opus/volltexte/2012/3437.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, November 2012. ISBN 978-1-118-58577-1.

Guillaume Dalle and Yohann De Castro. gdalle/PartiallyObservedVectorAutoRegressions: Preprint. Zenodo, June 2021. URL https://zenodo.org/record/4969054.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.

Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press, January 2014. ISBN 978-1-4665-0225-3.

Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, July 2000. ISSN 1573-1375. doi: 10.1023/A:1008935410038. URL https://doi.org/10.1023/A:1008935410038.

Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, June 2009. ISBN 978-0-521-88427-3.

John Duchi. Derivations for linear algebra and optimization. 2007. URL https://web.stanford.edu/~jduchi/projects/general_notes.pdf.

John Duchi. Statistics 311/Electrical Engineering 377: Information Theory and Statistics. 2019. URL https://stanford.edu/class/stats311/lecture-notes.pdf.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.

Fang Han, Huanran Lu, and Han Liu. A Direct Estimation of High Dimensional Stationary Vector Autoregressions. *Journal of Machine Learning Research*, 16(97):3115–3150, 2015. ISSN 1533-7928. URL http://jmlr.org/papers/v16/han15a.html.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 978-1-4987-1216-3.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, June 1994. ISBN 978-0-521-46713-1.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, October 2012. ISBN 978-1-139-78888-5.

Amin Jalali and Rebecca Willett. Missing Data in Sparse Transition Matrix Estimation for Sub-Gaussian Vector Autoregressive Processes. *arXiv:1802.09511 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1802.09511.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, March 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://asmedigitalcollection.asme.org/fluidsengineering/article/82/1/35/397706/A-New-Approach-to-Linear-Filtering-and-Prediction.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, June 2015. ISSN 0304-4076. doi: 10.1016/j.jeconom.2015.02.013. URL http://www.sciencedirect.com/science/article/pii/S0304407615000378.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Soc., October 2017. ISBN 978-1-4704-2962-1.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40 (3):1637–1664, June 2012. ISSN 0090-5364, 2168-8966. doi: 10/ggx5bj. URL https://projecteuclid.org/journals/annals-of-statistics/volume-40/issue-3/High-dimensional-regression-with-noisy-and-missing-data--Provable/10.1214/12-AOS1018.full.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, December 2005. ISBN 978-3-540-27752-1.

Luigi Malagò and Giovanni Pistone. Information Geometry of the Gaussian Distribution in View of Stochastic Optimization. In *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*, FOGA '15, pages 150–162, Aberystwyth, United Kingdom, January 2015. Association for Computing Machinery. ISBN 978-1-4503-3434-1. doi: 10.1145/2725494.2725510. URL https://doi.org/10.1145/2725494.2725510.

Igor Melnyk and Arindam Banerjee. Estimating Structured Vector Autoregressive Models. In *International Conference on Machine Learning*, pages 830–839, June 2016. URL http://proceedings.mlr.press/v48/melnyk16.html.

K. B. Petersen and M. S. Pedersen. The matrix cookbook. Technical report, 2012. URL https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

Milind Rao, Tara Javidi, Yonina C. Eldar, and Andrea Goldsmith. Estimation in autoregressive processes with partial observations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4212–4216, March 2017a. doi: 10.1109/ICASSP.2017.7952950.

Milind Rao, Tara Javidi, Yonina C. Eldar, and Andrea Goldsmith. Fundamental estimation limits in autoregressive processes with compressive measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2895–2899, June 2017b. doi: 10.1109/ISIT.2017.8007059.

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, gfyoung, Sinhrks, Matthew Roeschke, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Shahar Naveh, patrick, Marc Garcia, Jeremy Schendel, Andy Hayden, Daniel Saxton, Vytautas Jancauskas, Marco Gorelli, Richard Shadrach, Ali McMaster, Pietro Battiston, Skipper Seabold, Kaiqi Dong, chris-b1, and h-vetinari. Pandas-dev/pandas: Pandas 1.2.4. Zenodo, April 2021. URL https://zenodo.org/record/4681666.

R. H. Shumway and D. S. Stoffer. An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1982.tb00349.x. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1982.tb00349.x.

Andrew F. Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, April 1982. ISSN 0006-3444. doi: 10/c942t3. URL https://doi.org/10.1093/biomet/69.1.242.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://www.jstor.org/stable/2346178.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, October 2008. ISBN 978-0-387-79052-7.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, September 2018. ISBN 978-1-108-24454-1.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, February 2019. ISBN 978-1-108-49802-9.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. doi: 10/ggr6q3.