Minimax Estimation of Partially-Observed Vector AutoRegressions

Guillaume Dalle, Yohann de Castro

MINIMAX ESTIMATION OF PARTIALLY-OBSERVED VECTOR AUTOREGRESSIONS

Guillaume Dalle
CERMICS, École des Ponts
Marne-la-Vallée, France
guillaume.dalle@enpc.fr

Yohann De Castro
Institut Camille Jordan, École Centrale Lyon
Écully, France
yohann.de-castro@ec-lyon.fr

April 2022

Abstract

High-dimensional time series are a core ingredient of the statistical modeling toolkit, for which numerous estimation methods are known. But when observations are scarce or corrupted, the learning task becomes much harder. The question is: how much harder?

In this paper, we study the properties of a partially-observed Vector AutoRegressive process, which is a state-space model endowed with a stochastic observation mechanism. Our goal is to estimate its sparse transition matrix, but we only have access to a small and noisy subsample of the state components. Interestingly, the sampling process itself is random and can exhibit temporal correlations, a feature shared by many realistic data acquisition scenarios.

We start by describing an estimator based on the Yule-Walker equation and the Dantzig selector, and we give an upper bound on its non-asymptotic error. Then, we provide a matching minimax lower bound, thus proving near-optimality of our estimator. The convergence rate we obtain sheds light on the role of several key parameters such as the sampling ratio, the amount of noise and the number of non-zero coefficients in the transition matrix. These theoretical findings are commented and illustrated by numerical experiments on simulated data.

1 INTRODUCTION

Time series provide a natural representation for periodic measurements of a stochastic process. In particular, those defined by linear Gaussian recursions may be the most widely used and the easiest to study. Well-known examples include the AutoRegressive (AR) process and its multivariate counterpart, the Vector AutoRegressive (VAR) process.

Industrial applications of these models encounter two main challenges. First, they often involve signals in high dimension, which means sparsity assumptions play an important role. Second, the variables of interest are rarely measured exactly or entirely. Indeed, physical constraints such as the cost of sensors can make it impossible to capture every component of the system's state at all times. It is therefore natural to ask: *how much harder does high-dimensional learning become when one only observes a fraction of the relevant values?*

1.1 CONTEXT OF THE STUDY

To answer this question, we study a state-space model where the state $X_t \in \mathbb{R}^D$ follows a VAR process of order 1 over a period of length T. Since the dimension D of X_t is high, we assume that its transition matrix $\theta \in \mathbb{R}^{D \times D}$ is s-sparse (there are no more than s non-zero coefficients in each row). However, we do not observe the state itself: our observations Y_t only involve the subset of components $X_{t,d}$ for which $\pi_{t,d} = 1$, where π_t is a vector of Bernoulli variables. To make matters worse, this subset is corrupted with noise, which leads to the following generative procedure:

$$X_t = \theta X_{t-1} + \mathcal{N}(0, \sigma^2 I) \qquad \pi_{t,d} \sim \mathcal{B}(p) \qquad Y_t = \operatorname{diag}(\pi_t) X_t + \mathcal{N}(0, \omega^2 I). \tag{1}$$

When we write $\pi_{t,d} \sim \mathcal{B}(p)$, we mean that the marginals of the sampling variables are identical, which requires that every state component be sampled with equal probability p. However, we reject the standard independence assumption in favor of temporal dependencies between the Bernoulli variables $\pi_{t,d}$ (see Section 1.2 for a practical justification).

To shed light on the properties of our model, we start by constructing a sparse estimator for θ, whose non-asymptotic error we upper bound. We complement this finding with a lower bound on the minimax error that does not depend on the choice of estimator. Upper and lower bound match in most regards, which proves their optimality. A rough summary of our analysis is that the best possible estimator $\widehat{\theta}$ satisfies

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

with high probability. We observe that the error does not depend on the state dimension D, but only on the sparsity s of the transition matrix. As expected, it decreases linearly as p grows, since more information becomes available. Lastly, it is a function of ω^2/σ^2, which means that precise recovery of θ is only possible when the noise is not too much larger than the signal.

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 temporal correlations within the sampling process, the combination of discrete and continuous concentration inequalities to obtain error estimates, as well as detailed numerical experiments on simulated data.

1.2 Example of Application

Our study was inspired by concrete questions related to delay propagation on railway networks, which came up during a collaboration with a leading railway company. When external factors (weather, passenger behavior, mechanical failures) trigger a primary delay, resource conflicts between trains can amplify the initial incident and send ripple effects through the whole network. Understanding and predicting this propagation phenomenon is a crucial task for traffic management and robust scheduling.

To model it, we construct a network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ linking the railway stations, and we assume the existence of a hidden congestion variable $X_{t,d}$ that lives on the edges $d \in \mathcal{E}$. This congestion evolves according to a VAR process, whose transition matrix θ represents pairwise interactions between edges. The sparsity structure of θ expresses the local nature of delay propagation, which is why it is closely related to the adjacency structure of $\mathcal{G}$. Indeed, between times t and $t+1$, edges are expected to transmit congestion to their close neighbors, and not to regions of the network that are very far away.

Unfortunately for us, X_t is never observed directly. The only information we have is collected by the trains whenever they cross an edge of the network. The crossing time of a train is influenced by the congestion, but also by other individual factors: in this sense, our observations Y_t are a noisy version of the underlying process X_t. Furthermore, the observations are limited in size: the dimension of X_t is the number of edges $D = |\mathcal{E}|$, while the dimension of Y_t is linked to the number of trains on the timetable and the length of their respective journeys. We can thus define a random variable $\pi_{t,d}$ equal to 1 if a train crosses edge d between t and $t+1$, and 0 otherwise. A more realistic model would account for the possibility of multiple trains crossing an edge in the same time step, especially if the discretization interval is large. However, our binary assumption greatly simplifies exposition without betraying the qualitative behavior of the system. Crucially, this sampling mechanism exhibits temporal correlations: periods of dense traffic are likely to be followed by dense traffic, which means that the sequence of sampling variables $\pi_{t,d}$ is not independently distributed.

We recognize the framework of Equation (1), and can therefore apply the theoretical result of Equation (2). This error quantification provides useful insight on the estimation of θ, which is essential to help railway operatives dimension their data sets or evaluate prediction uncertainty.

1.3 Related Works

The theory of VAR processes has been known for a long time: the book of Lütkepohl [2005] provides a detailed account. If we have full and noiseless observations of the process X_t, we can use conditional Least Squares to

estimate θ by minimizing the quadratic error $\sum_t \|X_t - \theta X_{t-1}\|_2^2$. This is equivalent to solving the Yule-Walker equation $\Gamma_h = \theta \Gamma_{h-1}$, where we replace the autocovariance matrix $\Gamma_h = \mathrm{Cov}[X_{t+h}, X_t]$ with its empirical counterpart $\widehat{\Gamma}_h$. In the case of Gaussian innovations, both approaches coincide with the Maximum Likelihood Estimator (MLE).

Neither of these methods was initially designed for missing or noisy data. Luckily, statistical estimation with imprecise measurements has been thoroughly studied [Buonaccorsi, 2010]. The same goes for incomplete data sets ; an extensive survey was recently published by Little and Rubin [2019]. According to their terminology, our work deals with data that is missing completely at random (MCAR), which means that the projection π_t is independent from the underlying process X_t. We also assume to know the distribution of the missingness indicators $1 - \pi_{t,d}$, which is not necessarily true for other applications (e.g. clinical trials).

A principled approach to deal with missing data would require extending the MLE to partially-observed time series, also known as 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 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 that use covariance estimates within the Yule-Walker equation have been quite popular in the machine learning community.

In this line of work, the core challenge is the high dimension D of the VAR process X_t. To address it, many authors use sparsity-inducing penalties as a way to reduce data requirements and computational workload. In the last ten years, the LASSO [Tibshirani, 1996] has been increasingly applied to random designs exhibiting correlations or missing data. This trend started with the seminal work of Loh and Wainwright [2012], and numerous other papers followed [see for example Basu and Michailidis, 2015, Kock and Callot, 2015, Melnyk and Banerjee, 2016, Jalali and Willett, 2018].

As an alternative to the LASSO, the Dantzig selector [Candes and Tao, 2007] enforces sparsity in the objective and data fidelity in the constraints. While the LASSO requires solving a Quadratic Program (QP), for instance with proximal methods, the Dantzig selector gives rise to a Linear Program (LP) which can be parallelized across dimensions. Han et al. [2015] studied its application to VAR estimation, obtaining finite-sample error bounds with very natural hypotheses. A little later, Rao et al. [2017a] extended these results to the more general scenario in which a hidden VAR process is randomly sampled or projected, and then corrupted with noise. This last work is quite similar to ours, but we think that the proof they present to control the non-asymptotic error is incomplete at best[1].

Another salient feature of our paper is the search for a minimax lower bound, which allows us to prove the optimality of our convergence rates. To the best of our knowledge, this was only attempted once for partially-observed VAR processes. Rao et al. [2017b] presented a lower bound on the minimax error in a setting very similar to ours, but their result is less generic in several regards. Indeed, we account for the possibility of temporal correlations within sampling, as well as observation noise. Moreover, unlike the one proposed by Rao et al. [2017b], our proof focuses on geometric properties and doesn't make use of the admissible set of transition matrices until the very end. This makes it easy to handle many different types of structured transitions without additional work: sparse, Toeplitz, banded, etc.

Finally, the error bounds we obtain are backed up by detailed numerical experiments on simulated data, which allow us to visualize the influence of every parameter of interest.

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 of the transition matrix. We then state both of our theoretical results in Section 3: an upper-bound on the error of our specific estimator, complemented by a minimax lower bound on the error of any estimation algorithm. Section 4 contains numerical experiments demonstrating the impact of various parameters, which lead to the conclusion in Section 5.

[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, this introduces an additional difficulty and forces us to use a more complex Gaussian concentration result (Lemma 37). See https://web.stanford.edu/~milind/papers/system_id_icassp_proof.pdf for the supplementary material in question.

Appendix A is dedicated to proving the convergence rate of the sparse estimator, while Appendix B contains the derivation of the minimax lower bound. A number of useful results from linear algebra and probability are presented in Appendix C to make the paper as self-contained as possible. Most of them are well-known, some were obtained or adapted specifically for our proof. Appendix D contains a summary of the main notations and symbols.

2 The Partially-Observed VAR Process and its Sparse Estimator

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

2.1 Model Definition

The model we study was described approximately in the introduction. We now fill the gaps of the generative procedure it relies on.

The underlying state $X = (X_t)_{t\in[T]}$ follows a stationary VAR process of order 1. This process has dimension D and the following recursive definition:

$$X_t = \theta X_{t-1} + \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma). \tag{3}$$

Here $\theta \in \mathbb{R}^{D\times D}$ is the transition matrix and $\Sigma \in \mathbb{R}^{D\times D}$ is the covariance matrix of the innovations (in the introduction, we assumed $\Sigma = \sigma^2 I$).

To ensure stationarity of the VAR process, we must constrain the spectral radius of θ to satisfy $\rho(\theta) < 1$. Throughout the paper, we actually make the following (slightly stronger) assumption on the spectral norm of θ: there exists $\vartheta \in (0, 1)$ such that for all the values of θ we consider, $\|\theta\|_2 \leq \vartheta < 1$. Furthermore, we only study row-sparse transition matrices, having at most s nonzero coefficients in each row. In other words, we restrict our choice of parameters to

$$\theta \in \Theta_s \quad \text{where} \quad \Theta_s = \{\theta \in \mathbb{R}^{D\times D} : \|\theta\|_2 \leq \vartheta < 1 \quad \text{and} \quad \forall i, \|\theta_{i,\cdot}\|_0 \leq s\}. \tag{4}$$

We denote by $\sigma^2_{\min} = \lambda_{\min}(\Sigma)$ and $\sigma^2_{\max} = \lambda_{\max}(\Sigma)$ the minimum and maximum eigenvalues of the covariance matrix Σ.

The observation mechanism we chose implies that we do not have direct access to the latent process X_t. To construct the observations Y_t, we sample a subset of state components according to the binary vectors π_t. Then, independent Gaussian noise with variance ω^2 is added to these selected components, and we observe the result. If we denote by $\Pi_t = \operatorname{diag}(\pi_t)$ the diagonal projection matrix, we have

$$Y_t = \Pi_t X_t + \eta_t \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \omega^2 I). \tag{5}$$

An essential hypothesis we make is the mutual independence between our three sources of randomness: the innovations ε_t, the projections π_t and the observation noise η_t.

A major feature of the present work is the non-deterministic selection of observed state components, that is, the fact that π_t is a random sequence of Bernoulli vectors following a known distribution. In order to sum up the amount of information available using one parameter $p \in (0, 1)$, we want this distribution to satisfy the following condition: each component $X_{t,d}$ of the latent state must be sampled with the same marginal probability $p = \mathbb{P}(\pi_{t,d} = 1)$.

On the other hand, we also want to introduce temporal dependencies between the projections. The simplest way to achieve that is through a Markovian hypothesis: independently along each dimension d, time indices t are selected for observation according to a binary-valued Markov chain with transition matrix $\mathcal{T} = \left(\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right)$. Its coefficients are chosen to make the chain stationary with invariant measure $(\frac{b}{a+b}, \frac{a}{a+b}) = (1-p, p)$. Note that when $a = 1 - b = p$, this reduces to independent sampling of each component with probability p. We also assume there exists a universal constant χ such that $0 < \chi \leq a, b \leq 1 - \chi < 1$: this means that the chain does not transition too fast nor too slowly.

Our data set is built from N independent realizations of this process. For the sake of simplicity however, we will prove all convergence theorems in the case $N = 1$: extending those results to the general case simply amounts to replacing T with NT in the resulting error bounds.

2.2 Sparse Estimator for the Transition Matrix

We now introduce the estimation method chosen for this problem.

THE TRANSITION ESTIMATOR presented here is a straightforward generalization of the one used by Rao et al. [2017a]. The lag-h covariance matrix of the VAR process X_t is given by the Yule-Walker recursion (see Lemma 1):

$$\Gamma_h(\theta) = \mathrm{Cov}_\theta[X_{t+h}, X_t] = \theta\Gamma_{h-1}(\theta) = \theta^h\Gamma_0(\theta) \tag{6}$$

We can use it to define a simple two-step procedure:

1. For a given h_0, build estimators $\widehat{\Gamma}_{h_0}$ and $\widehat{\Gamma}_{h_0+1}$ of the covariances Γ_{h_0} and Γ_{h_0+1}.
2. Use them to approximate the transition matrix by inverting Equation (6).

A simple inversion technique uses the Moore-Penrose pseudoinverse (just in case $\widehat{\Gamma}_{h_0}$ is singular):

$$\widehat{\theta}^{\text{dense}} = \widehat{\Gamma}_{h_0+1}\widehat{\Gamma}^\dagger_{h_0}. \tag{7}$$

The problem with this procedure is that is does not guarantee sparsity of $\widehat{\theta}$. To obtain a sparse result, we follow Han et al. [2015] and cast Equation (6) as a soft constraint enforcing proximity between $\widehat{\Gamma}_{h_0+1}$ and $\widehat{\theta}\widehat{\Gamma}_{h_0}$. This amounts to solving the following constrained optimization problem:

$$\widehat{\theta} \in \underset{M\in\mathbb{R}^{D\times D}}{\operatorname{argmin}} \|\mathrm{vec}(M)\|_1 \quad \text{subject to} \quad \|M\widehat{\Gamma}_{h_0} - \widehat{\Gamma}_{h_0+1}\|_{\max} \leq \lambda. \tag{8}$$

Here $\|\mathrm{vec}(\cdot)\|_1$ denotes the sum of the absolute values of all the coefficients of a matrix, while $\|\cdot\|_{\max}$ is the maximum of these absolute values. Given that both of these norms are piecewise linear, the problem of Equation (8) can be reformulated as an LP. It can even be 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 Γ_h.

THE COVARIANCE ESTIMATOR we use is a variant of the empirical covariance. Since $Y_t = \Pi_t X_t + \eta_t$ where η_t is zero-mean, a natural proxy for X_t is obtained by inverting the sampling operator: $\widehat{X}_t = \Pi_t^\dagger Y_t$. It would therefore seem logical to build an estimator of Γ_h by plugging this proxy into the empirical covariance between X_{t+h} and X_t. However, in order for this idea to work, we must make two small adjustments.

To account for the random sampling, the plug-in empirical covariance must be scaled elementwise by a matrix $S(h) = \mathbb{E}[\pi_{t+h}\pi_t']$. 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. Furthermore, to account for the observation noise, we must incorporate an additive correction $-\omega^2 I$. This correction becomes unnecessary for $h \geq 1$ since the observation noise η_t is independent across time.

In conclusion, we obtain the following covariance estimator:

$$\widehat{\Gamma}_h = \frac{1}{S(h)} \odot \frac{1}{T-h}\sum_{t=1}^{T-h}\left(\Pi_{t+h}^\dagger Y_{t+h}\right)\left(\Pi_t^\dagger Y_t\right)' - \mathbf{1}_{\{h=0\}}\omega^2 I. \tag{9}$$

The coefficients of the scaling matrix $S(h)$ are computed in Lemma 4.

3 Lower and Upper Bound on the Estimation Error

We now have the necessary background to formulate our theoretical results. In all the following statements (and their proofs), the letter c denotes 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"...

3.1 Main Theorems

We start by bounding the non-asymptotic error of the estimator we just introduced.

Theorem 1 (Error upper bound)**.** *Consider the partially-observed VAR model defined in Section 2.1. We use the estimator $\widehat{\theta}$ of Section 2.2 with $h_0 = 0$, and we suppose that T is "large enough", as specified by Equations (20) and (23). Let us define*

$$\gamma_u(\theta) = \frac{\|\theta\|_\infty + 1}{(1 - \|\theta\|_2)^2} \frac{\sigma_{\max}^2 + \omega^2}{\|\Gamma_0(\theta)^{-1}\|_1^{-1}} \qquad \text{and} \qquad q_u = \min\{p, 1 - b\} \leq p. \tag{10}$$

Then there is a value of λ such that the following upper bound holds with probability at least $1 - \delta$:

$$\|\widehat{\theta} - \theta\|_\infty \leq c \frac{\gamma_u(\theta) s}{\sqrt{T p q_u}} \sqrt{\log(D/\delta)}. \tag{11}$$

Proof. The argument combines discrete and continuous concentration inequalities, to account for both the Bernoulli sampling and the Gaussian noise. More precisely, we exploit a recent Chernoff bound that applies to non-reversible Markov chains, and we plug it into a conditional version of the Hanson-Wright inequality that we derived specifically for our purposes. See Appendix A for more details. □

We now move on to a minimax lower bound which is estimator-independent, and quantifies the intrinsic difficulty of our statistical problem. The term minimax means that we study the probability of making an error of magnitude ζ, when we pick the best possible estimator $\widehat{\theta}$ and nature replies by choosing the worst possible parameter θ:

$$\mathfrak{P}(\zeta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_s} \mathbb{P}_\theta \left[\|\widehat{\theta} - \theta\|_\infty \geq \zeta \right]. \tag{12}$$

More precisely, we want to find a threshold ζ such that the probability of exceeding it is non-negligible, for instance $\mathfrak{P}(\zeta) \geq \frac{1}{2}$. The evolution of this threshold will tell us how the error behaves with respect to the various problem parameters.

Theorem 2 (Error lower bound)**.** *Consider the partially-observed VAR model defined in Section 2.1. We suppose that T is "large enough", as specified by Equations (25) and (27). Let us define*

$$\gamma_\ell = (1 - \vartheta)^{3/2} \frac{\sigma_{\min}^2 + \omega^2}{\sigma_{\max}^2} \qquad \text{and} \qquad q_\ell = \max\{1 - b, 2p - (1 - b)\} \geq p. \tag{13}$$

Then the following minimax lower bound holds:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_s} \mathbb{P}_\theta \left[\|\widehat{\theta} - \theta\|_\infty \geq c \frac{\gamma_\ell s}{\sqrt{T p q_\ell}} \right] \geq \frac{1}{2}. \tag{14}$$

Proof. The argument is based on an information-theoretical result known as Fano's inequality. To apply it, we need to upper bound the Kullback-Leibler (KL) divergence between the distributions $\mathbb{P}_{\theta_0}(\Pi, Y)$ and $\mathbb{P}_{\theta_1}(\Pi, Y)$, where θ_0 and θ_1 are sufficiently far apart. See Appendix B for more details. □

3.2 Influence of the Problem Parameters

Let us now compare the error bounds of Theorems 1 and 2. Our first remark is that s and T play exactly the same roles in both bounds (up to a logarithmic factor), which shows that the dependency of the error in $s/\sqrt{T}$ is optimal.

The sampling parameters appear as $1/\sqrt{p q_u}$ in the upper bound, whereas the lower bound scales as $1/\sqrt{p q_\ell}$ instead. This means that we have not proven the optimality of either bound with respect to p or b. However, it is reassuring to note that there is no conflict between them since $q_\ell \geq p \geq q_u$. Furthermore, when $a = 1 - b = p$ (that is, when Markov sampling boils down to independent sampling), both bounds simplify into the $1/p$ dependency we would expect (since $q_u = q_\ell = p$). So in the case of independent sampling, $1/p$ is indeed the optimal rate.

THE ℓ_2 NORM OF THE TRANSITION MATRIX plays opposite roles on each side. In the lower bound, $1 - \vartheta = 1 - \max_{\theta \in \Theta_s} \|\theta\|_2$ appears in the numerator, whereas in the upper bound, $1 - \|\theta\|_2$ appears in the denominator. It is likely that these dependencies are suboptimal, but at least they are compatible with one another: as $\|\theta\|_2 \to 1$, that is, as the VAR process becomes unstable, the lower bound tends to 0 and the upper bound to $+\infty$. This is a reflection of the fact that our proofs make heavy use of the distance between θ and the unit sphere, which means they become meaningless when θ gets too large.

THE VARIANCES Σ AND ω^2 are involved in γ_ℓ for the lower bound, and in $\gamma_u(\theta)$ for the upper bound. In both cases, the ratio γ tells us whether the underlying process is large enough to be detected among the noise. Roughly speaking, the magnitude of X_t is related to the spectrum of Σ, while the magnitude of Y_t is related to the spectrum of $\Sigma + \omega^2 I$. If the latter is significantly larger than the former, recovering X_t (and thus θ) is a hopeless endeavor.

To simplify the comparison, let us assume in this discussion that $\Sigma = \sigma^2 I$, and that θ commutes with its transpose. Then we have $\|\Gamma_0^{-1}(\theta)\|_1^{-1} = \left\|\left(\sigma^2(I - \theta\theta')^{-1}\right)^{-1}\right\|_1^{-1} = \sigma^2 \|I - \theta\theta'\|_1^{-1}$, and we can give a simpler expression of γ_ℓ and $\gamma_u(\theta)$:

$$\gamma_u(\theta) = \frac{(\|\theta'\|_1 + 1)\|I - \theta\theta'\|_1}{(1 - \vartheta)^2} \frac{\sigma^2 + \omega^2}{\sigma^2} \qquad\qquad \gamma_\ell = (1 - \vartheta)^{3/2} \frac{\sigma^2 + \omega^2}{\sigma^2}.$$

We recognize the same dependency in both bounds, namely $\gamma \propto 1 + \frac{\sigma^2}{\omega^2}$. Lemma 38 gives a heuristic argument linking this functional form to the asymptotic behavior of the MLE.

3.3 EXTENSION TO VAR PROCESSES OF HIGHER ORDER

Although our results 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(K) processes. Just for this Section, suppose X_t is no longer given by Equation (3), but instead satisfies:

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

Then we can represent this as a VAR(1) process using augmented variables [Lütkepohl, 2005]. Indeed, observe that defining $\widetilde{X}_t = \begin{pmatrix} X_t & X_{t-1} & \cdots & X_{t-K+1} \end{pmatrix}'$ and $\widetilde{\varepsilon}_t = \begin{pmatrix} \varepsilon_t & 0 & \cdots & 0 \end{pmatrix}'$ yields

$$\widetilde{X}_t = \widetilde{\theta}\widetilde{X}_{t-1} + \widetilde{\varepsilon}_t \qquad \text{with} \qquad \widetilde{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_{K-1} & \theta_K \\ I_D & 0 & \cdots & 0 & 0 \\ 0 & I_D & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_D & 0 \end{bmatrix}.$$

Unfortunately, by this reasoning, the Markov sampling mechanism that generates Π_t gives rise to a new distribution for $\widetilde{\Pi}_t$ which is no longer part of the same family. Indeed, the augmented sampling process $\widetilde{\Pi}_t$ is still Markovian but with a memory of size K instead of 1. Therefore, the adaptation would require new arguments and we leave it for future work.

4 NUMERICAL ILLUSTRATIONS

We now illustrate our results on simulated data. All experiments were performed on a Dell Precision 5530 mobile workstation with Intel Core i7-8850H CPU (2.60GHz $\times$ 12) and 31 GiB of RAM, running under Ubuntu 20.04. Our code was written in Julia [Bezanson et al., 2017], linear optimization problems were modeled using JuMP [Dunning et al., 2017] and solved with the COIN-OR Clp solver [Forrest et al., 2022]. The reproducible Pluto notebook used to generate all the plots will be made available on GitHub as soon as the review procedure is complete and anonymity is no longer required.

4.1 Data Generation

Simulating a partially-observed VAR process with known transition matrix θ allows us to compute the estimation error $\|\widehat{\theta} - \theta\|_\infty$ and study the influence of parameters such as T, D, s, p, ω, etc. Real values for θ were drawn using independent standard Gaussian distributions for each coefficient, and then normalized to satisfy $\|\theta\|_2 = \vartheta = \frac{1}{2}$. To simplify comparison with the theoretical bounds, we used a diagonal innovation covariance $\Sigma = \sigma^2 I$ and set the sampling parameters to $a = 1-b = p$, which amounts to independent sampling (except for the experiment that focuses specifically on the influence of b). When not mentioned explicitly, all other parameters are equal to their default values given below (we assume ω is known):

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

Most of the simulations are run in a dense estimation scenario. For those that require the sparse procedure, selecting a good regularization parameter λ is paramount: indeed, Theorem 1 is only valid for a specific value of λ (which is not known in practice, but we can hope to approximate this near-optimal choice).

A standard way to tune λ would be cross-validation. However, evaluating a choice of λ (and the resulting estimate $\widehat{\theta}$) requires inferring the hidden state sequence X_t from the observations Y_t. If the projection matrices Π_t were deterministic, the inference could be performed with Kalman filtering [Kalman, 1960], but since they are stochastic, the distribution of (X, Y) is no longer jointly Gaussian and the justification behind the Kalman filter breaks down. Finding an appropriate inference method in our setting will be the topic of future studies.

In the meantime, to tune λ, we suppose that the sparsity level of the real transition matrix θ is known. We then use this target sparsity $\widehat{s}$ to guide a dichotomy search on λ, until we find a transition matrix estimate $\widehat{\theta}$ whose row sparsity level is sufficiently close to $\widehat{s}$.

4.2 Results

The main results are presented on Figure 1. With the exception of 1f, all plots have the estimation error $\|\widehat{\theta} - \theta\|_\infty$ on their y-axis, and some parameter of interest on their x-axis. The axes are displayed with logarithmic scaling, in order to highlight the exponent of the dependencies. Each point corresponds to one run of the algorithm, aimed at estimating a single random value of θ. When a straight line is added to a scatter plot, it is the result of a Theil-Sen regression [Sen, 1968] applied to the points of the same color: its slope is denoted by α in the legend.

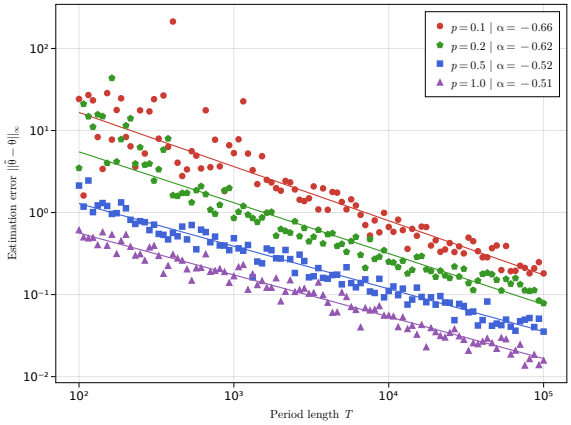
Figure 1a confirms that the error decreases as $1/\sqrt{T}$. This is only true because the sampling probability p remains constant. If instead we had a limited observation budget but an increasing temporal precision, we would have $p \propto 1/T$, in which case the error would increase as $\sqrt{T}$ instead of decreasing.

Figure 1b exhibits three clearly identifiable regimes with respect to the noise variance. In the first one, corresponding to $\omega/\sigma \ll 1$, the error remains small and constant. Then, the error increases when $\omega/\sigma \simeq 1$. In the third phase, corresponding to $\omega/\sigma \gg 1$, the error remains high and volatile. This is consistent with the theoretical dependency in $1 + \omega^2/\sigma^2$.

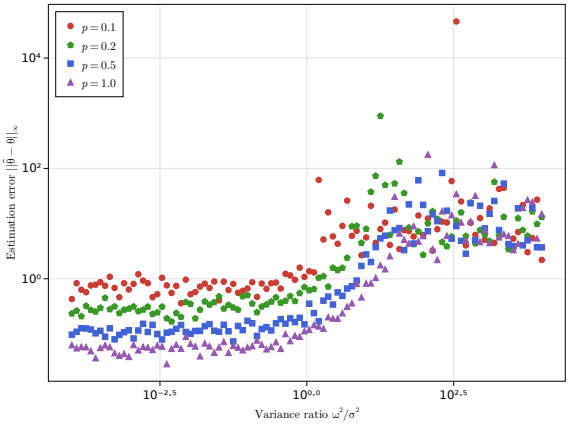
Figure 1c compares the respective benefits of sparse and dense estimation by increasing the ambient dimension D while keeping the true sparsity level s constant. The error for $\widehat{\theta}^{\text{dense}}$ scales linearly with D, while its sparse counterpart $\widehat{\theta}$ achieves a much slower error growth. As a side note, the fact that the error grows with D is not surprising. Indeed, we measure it with the ℓ_∞ operator norm, which scales with the dimension of the matrix.

Figure 1d takes the opposite perspective by increasing the number of nonzero coefficients in a space of fixed dimension. In this case, the theory predicts that the error should scale linearly with s, but the slope we observe is below 1. Our interpretation is that the function $\gamma_u(\theta)$ also depends on the sparsity level in complicated ways through θ, especially since the real values are renormalized to satisfy $\|\theta\|_2 = \frac{1}{2}$.

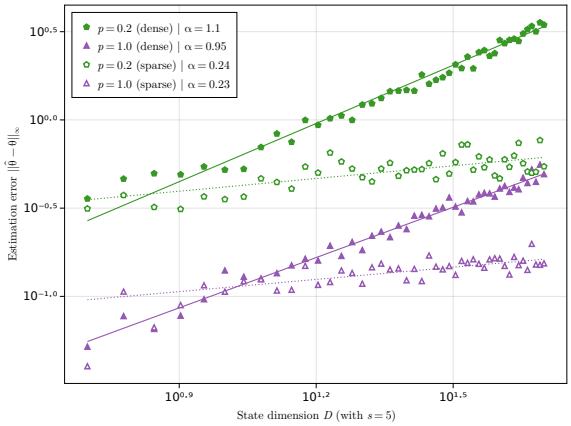
Figure 1e shows that the error evolves as $1/p$, which is consistent with our upper bound. It is also informative w.r.t. the choice of h_0. Choosing $h_0 = 0$ means we need to know ω to perform estimation. If this parameter is unknown, we can choose $h_0 \geq 1$, which leads to a much higher variance of the estimator (this is not visible in our results since we wrote the proof in the case where $h_0 = 0$). An alternate solution would be to keep $h_0 = 0$ and plug in a guess such as $\omega = 0$, effectively trading lower variance for a higher bias.



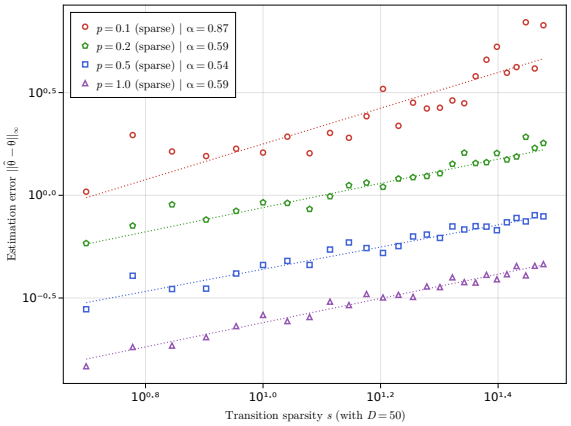
(a) Influence of T



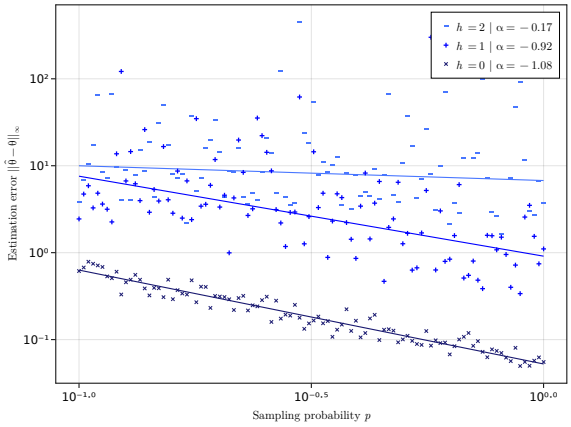
(b) Influence of ω



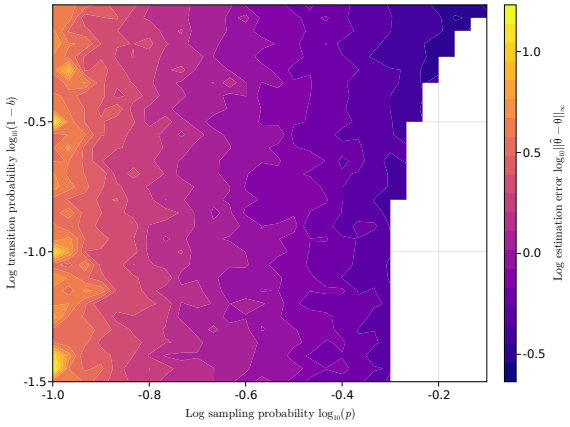
(c) Influence of D with fixed s



(d) Influence of s with fixed D



(e) Influence of p and h_0



(f) Influence of (p, b)

Figure 1: Impact of model parameters on the estimation error

Figure 1f takes a closer look at the role of the Markov sampling parameter b. The white region corresponds to values of b for which there is no a such that $p = a/(a+b)$. On this logarithmic heatmap, we see regularly-spaced and nearly vertical contour lines, which is consistent with a convergence rate of $1/p$ that does not depend on b. We conjecture that $1/p$ is the true order of magnitude for the optimal error, and that the dependencies $1/\sqrt{pq_u}$ and $1/\sqrt{pq_\ell}$ from our Theorems could be refined and brought together with a more careful theoretical analysis.

5 Conclusion and Perspectives

In this paper, we studied a partially-observed VAR process, whose latent state components are randomly projected and corrupted with noise before being observed. The temporal correlations within the sampling process are a novel feature, 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. Our analysis, supported by empirical results, sheds light on the intrinsic difficulty of such statistical problems, which arise naturally when analyzing several types of network processes.

However, our study leaves many questions open for future work. On the theoretical side, bridging the gap between our bounds will probably require more sophisticated tools to capture the precise behavior of Markov sampling. Going from the uniform case, where the sampling probability equals p everywhere, to more realistic heterogeneous settings, is also a worthy avenue to explore. On the practical side, this linear Gaussian model may not perform well when applied to real prediction problems. Finding ways to enhance it will be necessary if we want to gather insights on complex high-dimensional dynamics, especially for graph-structured data.

Acknowledgements

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.

References

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

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98, Jan. 2017. ISSN 0036-1445. doi: 10.1137/141000671. URL https://epubs.siam.org/doi/10.1137/141000671.

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

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, Dec. 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.

O. Cappé, É. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer Science & Business Media, Apr. 2006. ISBN 978-0-387-28982-3.

K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher. Chernoff-Hoeffding Bounds for Markov Chains: Generalized and Simplified. In C. Dürr and T. 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.

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

R. Douc, É. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press, Jan. 2014. ISBN 978-1-4665-0225-3.

A. Doucet, S. Godsill, and C. 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.

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

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

J. Duchi. Information Theory and Statistics. 2019. URL https://stanford.edu/class/stats311/lecture-notes.pdf.

I. Dunning, J. Huchette, and M. Lubin. JuMP: A Modeling Language for Mathematical Optimization. *SIAM Review*, 59(2):295–320, Jan. 2017. ISSN 0036-1445. doi: 10/gftshn. URL https://epubs.siam.org/doi/abs/10.1137/15M1020575.

J. Forrest, S. Vigerske, T. Ralphs, L. Hafer, J. Forrest, jpfasano, H. G. Santos, M. Saltzman, Jan-Willem, B. Kristjansson, h-i-gassmann, A. King, pobonomo, S. Brito, and to-st. Coin-or/Clp: Release releases/1.17.7. Zenodo, Jan. 2022. URL https://zenodo.org/record/5839302.

F. Han, H. Lu, and H. 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.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, June 1994. ISBN 978-0-521-46713-1.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Oct. 2012. ISBN 978-1-139-78888-5.

A. Jalali and R. Willett. Missing Data in Sparse Transition Matrix Estimation for Sub-Gaussian Vector Autoregressive Processes. *arXiv:1802.09511 [cs, stat]*, Feb. 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, Mar. 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.

A. B. Kock and L. 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.

D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. American Mathematical Soc., Oct. 2017. ISBN 978-1-4704-2962-1.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Apr. 2019. ISBN 978-0-470-52679-8.

P.-L. Loh and M. 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.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, Dec. 2005. ISBN 978-3-540-27752-1.

L. Malagò and G. 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, Jan. 2015. Association for Computing Machinery. ISBN 978-1-4503-3434-1. doi: 10.1145/2725494.2725510. URL https://doi.org/10.1145/2725494.2725510.

I. Melnyk and A. 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.

M. Rao, T. Javidi, Y. C. Eldar, and A. Goldsmith. Estimation in autoregressive processes with partial observations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4212–4216, Mar. 2017a. doi: 10.1109/ICASSP.2017.7952950.

M. Rao, T. Javidi, Y. C. Eldar, and A. 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.

P. K. Sen. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324):1379–1389, Dec. 1968. ISSN 0162-1459. doi: 10/gfxz87. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480934.

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.

R. 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.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, Oct. 2008. ISBN 978-0-387-79052-7.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Sept. 2018. ISBN 978-1-108-24454-1.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Feb. 2019. ISBN 978-1-108-49802-9.

A Proof of the Estimator's Convergence Rate

Here we present the detailed proof of Theorem 1.

A.1 Overview

The main steps of the argument are the following:

1. Prove the Yule-Walker Equation (6) and deduce an expression for the covariance matrix of X (Lemmas 1 and 2).
2. Justify the formula of Equation (9) for $\widehat{\Gamma}_h$ by showing that it defines an unbiased estimator of Γ_h (Lemmas 3 and 4).
3. Fixing two indices d_1 and d_2, rewrite $(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}$ using quadratic forms $g_a' \Psi_a' L \Psi_b g_b$ of standard Gaussian vectors (Lemma 5).
4. Control the deviation of the matrix L using discrete concentration inequalities (Lemmas 6, 7, 8, 9 and 10).
5. Apply a conditional version of the Hanson-Wright inequality (Lemma 37) to the quadratic forms $g_a' \Psi_a' L \Psi_b g_b$ (Lemma 11).
6. Obtain a high-probability control on $\|\widehat{\Gamma}_h - \Gamma_h\|_{\max}$ with a union bound (Lemma 13).
7. Deduce the error of $\widehat{\theta}$ from the error of $\widehat{\Gamma}_{h_0}$ and $\widehat{\Gamma}_{h_0+1}$ by drawing inspiration from Han et al. [2015] (Lemmas 14 and 15).

A.2 Covariance Matrices

The Yule-Walker equation is a direct consequence of the VAR recursion, as can be seen from this Lemma.

Lemma 1 (VAR covariance matrices)**.** *The autocovariance matrices of the stationary VAR process defined by Equation (3) have the following expressions:*

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

Proof. We start by noting that according to Equation (3), the stacked vector $X = (X_t)_{t\in[T]}$ 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

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

□

The following result will come in handy later.

Lemma 2 (Norm of $\Gamma_0(\theta)$)**.** *The covariance matrix $\Gamma_0(\theta)$ satisfies*

$$\|\Gamma_0(\theta)\|_2 \leq \frac{\sigma^2_{\max}}{1-\vartheta^2}$$

Proof. By Lemma 1,

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

□

A.3 Construction of the Covariance Estimator

Now we justify the construction of our covariance estimator. Let $h_0 = 0$: for most of the proof, we fix a lag value $h \in \{h_0, h_0+1\} = \{0,1\}$.

Lemma 3 (Bias of the covariance estimator)**.** *The estimator $\widehat{\Gamma}_h$ given by Equation (9) for the covariance matrix Γ_h is unbiased.*

Proof. First, let us remember that since $\Pi_t = \operatorname{diag}(\pi_t)$ is diagonal and binary, we also have $\Pi_t^\dagger = \Pi_t' = \Pi_t$. By Equation (5),

$$\begin{aligned}(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)' &= \Pi_{t+h}^\dagger (\Pi_{t+h} X_{t+h} + \eta_{t+h})(X_t' \Pi_t' + \eta_t') \Pi_t^{\dagger\prime} \\ &= \operatorname{diag}(\pi_{t+h}) \left(X_{t+h} X_t' + X_{t+h}\eta_t' + \eta_{t+h} X_t' + \eta_{t+h}\eta_t'\right) \operatorname{diag}(\pi_t).\end{aligned} \tag{15}$$

Taking the conditional expectation and removing the cross-product terms (by independence of X and Π), we get:

$$\mathbb{E}[(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'|\Pi] = \operatorname{diag}(\pi_{t+h}) \left(\mathbb{E}[X_{t+h}X_t'] + \mathbb{E}[\eta_{t+h}\eta_t']\right) \operatorname{diag}(\pi_t).$$

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:

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

where $\odot$ is the elementwise Hadamard product. We now take the expectation w.r.t. Π:

$$\mathbb{E}[(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'] = \mathbb{E}[\pi_{t+h}\pi_t'] \odot \Gamma_h + \mathbf{1}_{\{h=0\}}\omega^2 \mathbb{E}[\operatorname{diag}(\pi_t)].$$

Dividing elementwise by the scaling matrix $S(h) = \mathbb{E}[\pi_{t+h}\pi_t']$, we get

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

which shows that our estimator

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

is unbiased. □

Note that since the process (Π_t) is stationary, the coefficients of $S(h)$ do not depend on t. They are computed in the next Lemma.

Lemma 4. *The second-order moments of π are given by*

$$S(h)_{d_1,d_2} = \mathbb{E}\left[\pi_{t+h,d_1}\pi_{t,d_2}\right] = \begin{cases} p^2 & \text{if } d_1 \neq d_2 \\ p & \text{if } d_1 = d_2 \text{ and } h = 0 \\ p^2 + p(1-p)(1-a-b)^h & \text{if } d_1 = d_2 \text{ and } h \geq 1 \end{cases}$$

In particular, every coefficient of the scaling matrix $S(h)$ is lower-bounded by

$$\min_{d_1,d_2,h} S(h)_{d_1,d_2} = \min\{p^2, p(1-b)\} = pq_u \quad \text{where} \quad q_u = \min\{p, 1-b\}.$$

Proof. Let $i = (t+h, d_1)$ and $j = (t, d_2)$ be two indices in $[T]\times[D]$. We have $\mathbb{E}[\pi_i] = \mathbb{E}[\pi_i^2] = p$. If $d_1 \neq d_2$, then the variables π_i and π_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

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

When diagonalizing the transition matrix $\mathcal{T}$, we see that the bottom-right coefficient of $\mathcal{T}^h$ is

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

Plugging this in, we get

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

Among all the possible values of $S(h)_{d_1,d_2}$, the smallest one is p^2 if $1-a-b \geq 0$, and $p^2 + p(1-p)(1-a-b)$ otherwise. But since

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

we conclude

$$\min_{d_1,d_2,h} S(h)_{d_1,d_2} = \min\{p^2, p^2 + p(1-p)(1-a-b)\} = \min\{p^2, p(1-b)\}.$$

□

A.4 Gaussian Concentration, Episode 1

From now on, we will study the concentration of $\widehat{\Gamma}_h$, 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 5 (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:*

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

where the random matrix L is defined in Equation (16), Ψ_ε and Ψ_η are defined in Equation (18), and g_ε and g_η are standard Gaussian vectors.

Proof. We denote by e_d the basis vector filled with zeros except for a 1 in position d. By Equation (9),

$$\begin{aligned}(\widehat{\Gamma}_h + \mathbf{1}_{\{h=0\}}\omega^2 I)_{d_1,d_2} &= \frac{1}{T-h}\sum_{t=1}^{T-h}\left(\frac{1}{S(h)} \odot (\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'\right)_{d_1,d_2} \\ &= \frac{1}{T-h}\sum_{t=1}^{T-h}\frac{1}{S(h)_{d_1,d_2}}\mathrm{e}_{d_1}'(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'\mathrm{e}_{d_2} \\ &= \frac{1}{T-h}\sum_{t=1}^{T-h}\mathrm{Tr}\left[\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'\right]\end{aligned}$$

Equation (15) allows us to rewrite $(\Pi_{t+h}^\dagger Y_{t+h})(\Pi_t^\dagger Y_t)'$:

$$\begin{aligned}(\widehat{\Gamma}_h + \mathbf{1}_{\{h=0\}}\omega^2 I)_{d_1,d_2} = &\frac{1}{T-h}\sum_{t=1}^{T-h} X_t' \operatorname{diag}(\pi_t)\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})X_{t+h} \\ &+ \frac{1}{T-h}\sum_{t=1}^{T-h} \eta_t' \operatorname{diag}(\pi_t)\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})X_{t+h} \\ &+ \frac{1}{T-h}\sum_{t=1}^{T-h} X_t' \operatorname{diag}(\pi_t)\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})\eta_{t+h} \\ &+ \frac{1}{T-h}\sum_{t=1}^{T-h} \eta_t' \operatorname{diag}(\pi_t)\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})\eta_{t+h}\end{aligned}$$

Let us denote by P_t the projection of $\mathbb{R}^{TD}$ keeping only the components associated with time t, i.e. such that $X_t = P_t X$ and $\eta_t = P_t \eta$. We recognize the following matrix L in all four lines of the expression above:

$$\begin{aligned}L &= \frac{1}{T-h}\sum_{t=1}^{T-h} P_t' \operatorname{diag}(\pi_t)\frac{\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}\operatorname{diag}(\pi_{t+h})P_{t+h} \\ &= \frac{1}{T-h}\sum_{t=1}^{T-h} P_t'\frac{\pi_{t+h,d_1}\pi_{t,d_2}\mathrm{e}_{d_2}\mathrm{e}_{d_1}'}{S(h)_{d_1,d_2}}P_{t+h}\end{aligned} \tag{16}$$

This leads to:

$$(\widehat{\Gamma}_h + \mathbf{1}_{\{h=0\}}\omega^2 I)_{d_1,d_2} = X'LX + \eta'LX + X'L\eta + \eta'L\eta$$

Since X and η both follow centered multivariate Gaussian distributions, we can express them as linear combinations of standard Gaussian vectors g_ε and g_η of dimension TD (indexed by the source of randomness):

$$X = \Psi_\varepsilon g_\varepsilon \qquad \text{and} \qquad \eta = \Psi_\eta g_\eta \tag{17}$$

where Ψ_ε and Ψ_η are the square roots of the respective covariance matrices

$$\Psi_\varepsilon = \mathrm{Cov}[X]^{1/2} \qquad \text{and} \qquad \Psi_\eta = \mathrm{Cov}[\eta]^{1/2} = \omega I. \tag{18}$$

We substitute X and η to get:

$$(\widehat{\Gamma}_h + \mathbf{1}_{\{h=0\}}\omega^2 I)_{d_2,d_1} = g_\varepsilon'\Psi_\varepsilon' L\Psi_\varepsilon g_\varepsilon + g_\eta'\Psi_\eta' L\Psi_\varepsilon g_\varepsilon + g_\varepsilon'\Psi_\varepsilon' L\Psi_\eta g_\eta + g_\eta'\Psi_\eta' L\Psi_\eta g_\eta,$$

which implies

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

The union bound gives us the expected result. □

Now, our goal is to apply a Gaussian concentration inequality to these deviation probabilities. However, since L is generated by the discrete sampling process π, it is random, and so are the products $\Psi_a' L \Psi_b$ (where $a, b \in \{\varepsilon, \eta\}$). We thus need a conditional version of the Hanson-Wright inequality (Lemma 37), in which the following random variables will come into play:

- The spectral norm $\|\Psi_a' L \Psi_b\|_2$
- The Frobenius norm $\|\Psi_a' L \Psi_b\|_F^2$
- The shifted trace $\mathrm{Tr}(\Psi_a' L \Psi_b - \mathbb{E}[\Psi_a' L \Psi_b])$

A.5 Interlude: Discrete Concentration

We exploit discrete concentration results to bound the deviations of the three quantities we just mentioned, starting with the norms.

Lemma 6 (Norm reformulation for L)**.** *The spectral and Frobenius norms of L are given by*

$$\|L\|_2 = \frac{\max_{t\in[T-h]} \pi_{t+h,d_1}\pi_{t,d_2}}{(T-h)S(h)_{d_1,d_2}} \quad \textit{and} \quad \|L\|_F^2 = \frac{1}{(T-h)^2 S(h)_{d_1,d_2}} \sum_{t=1}^{T-h} \pi_{t+h,d_1}\pi_{t,d_2}.$$

Proof. We first notice that L has a block-superdiagonal structure of rank h:

$$L = \frac{1}{T-h}\sum_{t=1}^{T-h} P_t' L_{[t,t+h]} P_{t+h} \qquad \text{with} \qquad L_{[t,t+h]} = \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \mathrm{e}_{d_2}\mathrm{e}_{d_1}'. \tag{19}$$

The spectral and Frobenius norms of such a matrix can easily be deduced from those of its blocks. Since $\|\mathrm{e}_{d_2}\mathrm{e}_{d_1}'\|_2 = \|\mathrm{e}_{d_2}\mathrm{e}_{d_1}'\|_F = 1$ and the π_t are binary-valued, this leads to the following formulas:

$$\|L\|_2 = \frac{1}{T-h}\max_{t\in[T-h]}\|L_{[t,t+h]}\|_2 = \frac{1}{(T-h)S(h)_{d_1,d_2}}\max_{t\in[T-h]}\pi_{t+h,d_1}\pi_{t,d_2}$$
$$\|L\|_F^2 = \frac{1}{(T-h)^2}\sum_{t=1}^{T-h}\|L_{[t,t+h]}\|_F^2 = \frac{1}{(T-h)^2 S(h)_{d_1,d_2}^2}\sum_{t=1}^{T-h}\pi_{t+h,d_1}\pi_{t,d_2}.$$

□

We can bound the spectral norm for free.

Lemma 7 (Spectral norm bound for L)**.** *With probability* 1*, the spectral norm $\|L\|_2$ satisfies*

$$\|L\|_2 \le \frac{c}{Tpq_u}$$

Proof. Note that $S(h)_{d_1,d_2} \ge pq_u$, and since $h \in \{0,1\}$, we can state that $T - h \ge cT$. By Lemma 6, we deduce

$$\|L\|_2 = \frac{\max_{t\in[T-h]}\pi_{t+h,d_1}\pi_{t,d_2}}{(T-h)S(h)_{d_1,d_2}} \le \frac{1}{(T-h)S(h)_{d_1,d_2}} \le \frac{1}{(T-h)pq_u} \le \frac{1}{cTpq_u}.$$

□

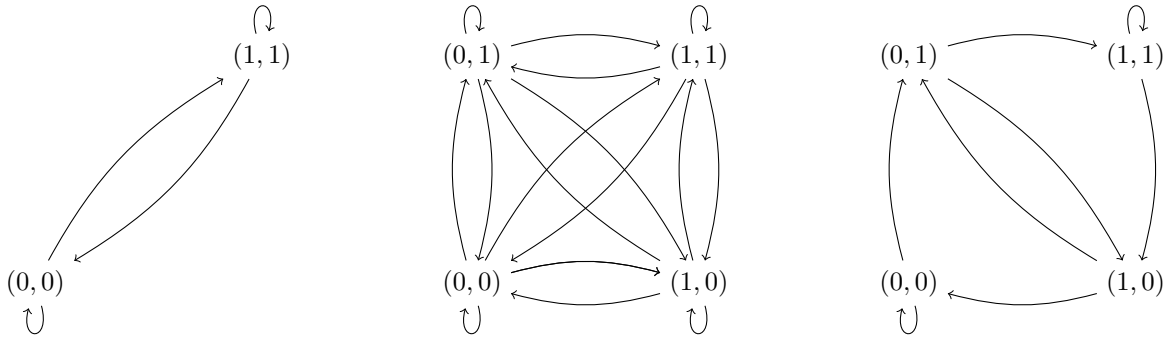
The Frobenius norm requires a little more work because of the sum it contains.

Lemma 8 (Concentration of the sampling Bernoullis)**.** *For all $u \in [0,1]$,*

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

Proof. We distinguish three cases:

- When $d_1 = d_2$ and $h = 0$, we have $\pi_{t+h,d_1} = \pi_{t,d_2}$, which is a 2-state Markov chain with transition matrix $\mathcal{T} \otimes I$, depicted on Figure 2a.
- 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 $\mathcal{T} \otimes \mathcal{T}$ since the chains π_{t+h,d_1} and π_{t,d_2} evolve along independent dimensions. It is shown on Figure 2b.
- 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 $\mathcal{S}(h)$, whose non-reversible transition diagram can be seen on Figure 2c.



(a) $d_1 = d_2$ & $h = 0$: transition $\mathcal{T}$ (b) $d_1 \neq d_2$: transition $\mathcal{T} \otimes \mathcal{T}$ (c) $d_1 = d_2$ & $h = 1$: transition $\mathcal{S}(1)$.

Figure 2: State space and transitions for the Markov chains used in the discrete concentration result

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 $\chi \leq a, b \leq 1 - \chi$, all the coefficients of $\mathcal{T}$ are greater than χ. Furthermore, all the coefficients of $\mathcal{T} \otimes \mathcal{T}$ are greater than χ^2. Finally, all the coefficients of $\mathcal{S}(h)^{h+1}$ are greater than χ^{h+1}, because all pairs of states are connected after $h + 1$ steps. Let us illustrate this with $h = 1$:

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

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

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

Since we only consider $h \in \{0, 1\}$ and since χ is fixed for our purposes, these chains fulfill the assumptions of Lemma 34. We thus conclude:

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

We finally replace $T - h$ with cT in the exponential, leading to the result we announced. □

Based on this concentration property, we can now bound the norms of the random matrix L with high probability.

Lemma 9 (Frobenious norm bound for L)**.** *For any δ such that Equation (20) holds, with probability at least $1-\delta$, the Frobenius norm $\|L\|_F^2$ satisfies*

$$\|L\|_F^2 \leq \frac{c}{Tpq_u}.$$

Proof. By Lemma 8: for all $u \in [0,1]$,

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

We remember the expression of Lemma 6 for $\|L\|_F^2$ and notice that:

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

We finally recall that $S(h)_{d_1,d_2} \geq pq_u$ and $T-h \geq cT$, so that

$$\begin{aligned}
\mathbb{P}\left(\|L\|_F^2 \geq \frac{1+u}{cTpq_u}\right) &\leq \mathbb{P}\left(\|L\|_F^2 \geq \frac{1+u}{(T-h)pq_u}\right) \\
&\leq \mathbb{P}\left(\|L\|_F^2 \geq \frac{1+u}{(T-h)S(h)_{d_1,d_2}}\right) \\
&\leq c_1\exp\left(-c_2u^2TS(h)_{d_1,d_2}\right) \\
&\leq c_1\exp\left(-c_2u^2Tpq_u\right).
\end{aligned}$$

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

$$c_1\exp\left(-c_2u^2Tpq_u\right) \leq \delta \quad\iff\quad u \geq \sqrt{\frac{\log(c_1/\delta)}{c_2Tpq_u}}$$

Note that we can replace $\log(c_1/\delta)$ by a constant times $\log(1/\delta)$ to simplify expressions: this is possible as long as δ is chosen "small enough" (i.e. smaller than some universal constant). We will assume this fairly often in the rest of the proof.

For Lemma 8 to apply, we must ensure that our choice of u is smaller than 1. With the previous discussion in mind, $u \leq 1$ is implied by

$$\sqrt{\frac{\log(1/\delta)}{Tpq_u}} \leq c. \tag{20}$$

If this holds, then we have

$$\mathbb{P}\left(\|L\|_F^2 \geq \frac{2}{cTpq_u}\right) \leq \mathbb{P}\left(\|L\|_F^2 \geq \frac{1+u}{cTpq_u}\right) \leq \delta.$$

This yields the result we wanted. □

We now move on to studying the shifted trace of $\Psi_a' L \Psi_b$, which is the last ingredient we need for our application of Lemma 37.

Lemma 10 (Trace bound for the L matrices). *For all $u \in [0,1]$,*

$$\mathbb{P}(|\operatorname{Tr}(\Psi_\varepsilon' L \Psi_\varepsilon - \mathbb{E}[\Psi_\varepsilon' L \Psi_\varepsilon])| \geq u) \leq c_1 \exp\left(-\frac{c_2 u^2 T p q_u}{\|\Gamma_h\|_2^2}\right)$$
$$\mathbb{P}(|\operatorname{Tr}(\Psi_\eta' L \Psi_\eta - \mathbb{E}[\Psi_\eta' L \Psi_\eta])| \geq u) \leq c_1 \exp\left(-\frac{c_2 u^2 T p q_u}{\omega^4}\right).$$

Proof. We can compute an explicit formula thanks to Equation (19): if $a \in \{\varepsilon, \eta\}$ then

$$\begin{aligned}
\operatorname{Tr}(\Psi_a' L \Psi_a) &= \operatorname{Tr}\left(\frac{1}{T-h}\sum_{t=1}^{T-h} \Psi_a' P_t' \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \mathrm{e}_{d_2}\mathrm{e}_{d_1}' P_{t+h} \Psi_a\right) \\
&= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \operatorname{Tr}\left(\Psi_a' P_t' \mathrm{e}_{d_2}\mathrm{e}_{d_1}' P_{t+h} \Psi_a\right) \\
&= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \left(\mathrm{e}_{d_1}' P_{t+h} \Psi_a \Psi_a' P_t' \mathrm{e}_{d_2}\right) \\
&= \frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} \left((\Psi_a \Psi_a')_{[t+h,t]}\right)_{d_1,d_2}
\end{aligned}$$

where $\left((\Psi_a \Psi_a')_{[t+h,t]}\right)_{d_1,d_2}$ denotes the (d_1,d_2) coefficient of the $(t+h,t)$ block of $\Psi_a \Psi_a'$. Now is the time to look back on Equation (18), which tells us that both $\Psi_\varepsilon \Psi_\varepsilon'$ and $\Psi_\eta \Psi_\eta'$ are constant along their superdiagonal of rank h. We thus find that

$$\operatorname{Tr}(\Psi_\varepsilon' L \Psi_\varepsilon - \mathbb{E}[\Psi_\varepsilon' L \Psi_\varepsilon]) = (\Gamma_h)_{d_1,d_2}\left(\frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} - 1\right)$$
$$\operatorname{Tr}(\Psi_\eta' L \Psi_\eta - \mathbb{E}[\Psi_\eta' L \Psi_\eta]) = (\mathbf{1}_{\{h=0\}}\omega^2 I)_{d_1,d_2}\left(\frac{1}{T-h}\sum_{t=1}^{T-h} \frac{\pi_{t+h,d_1}\pi_{t,d_2}}{S(h)_{d_1,d_2}} - 1\right)$$

Like before, we can apply Lemma 8: for all $u \in [0,1]$,

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

Since $|(\Gamma_h)_{d_1,d_2}| \leq \|\Gamma_h\|_2$ and $(\mathbf{1}_{\{h=0\}}\omega^2 I)_{d_1,d_2} \leq \omega^2$, we can deduce

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

which, after rescaling, yields the result we announced. □

A.6 Gaussian Concentration, Episode 2

We are now ready to apply our conditional concentration result.

Lemma 11 (Applying Hanson-Wright). *Let $\delta > 0$ and $u \in [0,1]$. Assume that Equations (20) and (21) hold. Then the deviation probability for $(\widehat{\Gamma}_h)_{d_1,d_2}$ satisfies*

$$\mathbb{P}(|(\widehat{\Gamma}_h - \Gamma_h)_{d_1,d_2}| \geq u) \leq 4\delta + c_1 \exp\left(-\frac{c_2 u^2 T p q_u}{\max\left\{(\|\Psi_\varepsilon\|_2^2 + \omega^2)^2, \|\Gamma_h\|_2^2, \omega^4\right\}}\right).$$

Proof. The conclusion we had reached before our discrete interlude is given by Lemma 5, 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_{ab} represents a deviation probability for a specific quadratic form $g_a' \Psi_a' L \Psi_b g_b$. Let us choose δ such that Equation (20) holds. By Lemmas 7 and 9, with probability at least $1 - \delta$, the following eight inequalities occur at the same time (we use Lemma 25 to split the products):

$$\begin{aligned}
\|\Psi_\varepsilon' L \Psi_\varepsilon\|_F^2 &\leq \frac{c\|\Psi_\varepsilon\|_2^4}{Tpq_u} & \|\Psi_\varepsilon' L \Psi_\varepsilon\|_2 &\leq \frac{c\|\Psi_\varepsilon\|_2^2}{Tpq_u} \\
\|\Psi_\eta' L \Psi_\varepsilon\|_F^2 &\leq \frac{c\|\Psi_\eta\|_2^2\|\Psi_\varepsilon\|_2^2}{Tpq_u} & \|\Psi_\eta' L \Psi_\varepsilon\|_2 &\leq \frac{c\|\Psi_\eta\|_2\|\Psi_\varepsilon\|_2}{Tpq_u} \\
\|\Psi_\varepsilon' L \Psi_\eta\|_F^2 &\leq \frac{c\|\Psi_\varepsilon\|_2^2\|\Psi_\eta\|_2^2}{Tpq_u} & \|\Psi_\varepsilon' L \Psi_\eta\|_2 &\leq \frac{c\|\Psi_\varepsilon\|_2\|\Psi_\eta\|_2}{Tpq_u} \\
\|\Psi_\eta' L \Psi_\eta\|_F^2 &\leq \frac{c\|\Psi_\eta\|_2^4}{Tpq_u} & \|\Psi_\eta' L \Psi_\eta\|_2 &\leq \frac{c\|\Psi_\eta\|_2^2}{Tpq_u}.
\end{aligned}$$

The spectral norm of Ψ_η is easily seen to equal $\|\Psi_\eta\|_2 = \|\omega^2 I\|_2^{1/2} = \omega$, which allows us to lighten these expressions. From there, Lemma 37 (applied with $X = g_a$, $Y = g_b$ and $A = \Psi_a' L \Psi_b$) provides the concentration bounds we need[2]:

$$\begin{aligned}
p_{\varepsilon\varepsilon} &\leq \delta + 2\exp\left(-cTpq_u \min\left\{\frac{(u/4)^2}{\|\Psi_\varepsilon\|_2^4}, \frac{(u/4)}{\|\Psi_\varepsilon\|_2^2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L \Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L \Psi_\varepsilon])| \geq u/8\right) \\
p_{\eta\varepsilon} &\leq \delta + 2\exp\left(-cTpq_u \min\left\{\frac{(u/4)^2}{\omega^2\|\Psi_\varepsilon\|_2^2}, \frac{(u/4)}{\omega\|\Psi_\varepsilon\|_2}\right\}\right) \\
p_{\varepsilon\eta} &\leq \delta + 2\exp\left(-cTpq_u \min\left\{\frac{(u/4)^2}{\|\Psi_\varepsilon\|_2^2\omega^2}, \frac{(u/4)}{\|\Psi_\varepsilon\|_2\omega}\right\}\right) \\
p_{\eta\eta} &\leq \delta + 2\exp\left(-cTpq_u \min\left\{\frac{(u/4)^2}{\omega^4}, \frac{(u/4)}{\omega^2}\right\}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L \Psi_\eta) - \mathbb{E}[\Psi_\eta' L \Psi_\eta])| \geq u/8\right).
\end{aligned}$$

The denominators inside the minima can be unified: for the left column,

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

and for the right column,

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

This means we can upper bound each of the four minima by

$$\min\left\{\left(\frac{u/4}{\|\Psi_\varepsilon\|_2^2 + \omega^2}\right)^2, \frac{u/8}{\|\Psi_\varepsilon\|_2^2 + \omega^2}\right\}.$$

From now on, we additionally suppose that

$$\frac{u/4}{\|\Psi_\varepsilon\|_2^2 + \omega^2} \leq \frac{1}{2} \tag{21}$$

[2]The additional trace terms that appear when applying Lemma 37 (as opposed to the non-conditional version of Lemma 36) are absent from the papers by Rao et al. [2017a,b], which is why we think their upper bound proofs are incomplete.

This enables us to get rid of these minima by reducing them to the (smaller) quadratic term on the left. We end up with

$$
\begin{aligned}
p_{\varepsilon\varepsilon} &\leq \delta + 2\exp\left(-\frac{cu^2Tpq_u}{\left(\|\Psi_\varepsilon\|_2^2+\omega^2\right)^2}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\varepsilon' L\Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L\Psi_\varepsilon])| \geq u/8\right)\\
p_{\eta\varepsilon} &\leq \delta + 2\exp\left(-\frac{cu^2Tpq_u}{\left(\|\Psi_\varepsilon\|_2^2+\omega^2\right)^2}\right)\\
p_{\varepsilon\eta} &\leq \delta + 2\exp\left(-\frac{cu^2Tpq_u}{\left(\|\Psi_\varepsilon\|_2^2+\omega^2\right)^2}\right)\\
p_{\eta\eta} &\leq \delta + 2\exp\left(-\frac{cu^2Tpq_u}{\left(\|\Psi_\varepsilon\|_2^2+\omega^2\right)^2}\right) + \mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L\Psi_\eta) - \mathbb{E}[\Psi_\eta' L\Psi_\eta])| \geq u/8\right).
\end{aligned}
$$

As for the trace terms, they are taken care of by Lemma 10:

$$
\begin{aligned}
\mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L\Psi_\varepsilon) - \mathbb{E}[\Psi_\varepsilon' L\Psi_\varepsilon])| \geq u/8\right) &\leq c_3\exp\left(-c_4\frac{(u/8)^2Tpq_u}{\|\Gamma_h\|_2^2}\right)\\
\mathbb{P}\left(|\operatorname{Tr}(\Psi_\eta' L\Psi_\eta) - \mathbb{E}[\Psi_\eta' L\Psi_\eta])| \geq u/8\right) &\leq c_3\exp\left(-c_4\frac{(u/8)^2Tpq_u}{\omega^4}\right)
\end{aligned}
$$

We plug this in and rearrange to get:

$$
p_{\varepsilon\varepsilon} + p_{\eta\varepsilon} + p_{\varepsilon\eta} + p_{\eta\eta} \leq 4\delta + c_1\exp\left(-\frac{c_2u^2Tpq_u}{\max\left\{(\|\Psi_\varepsilon\|_2^2+\omega^2)^2, \|\Gamma_h\|_2^2, \omega^4\right\}}\right).
$$

□

The following result will simplify the denominator inside the exponential.

Lemma 12 (Spectral norms of Ψ_ε and Γ_h)**.** *The matrices Ψ_ε and Γ_h satisfy:*

$$
\|\Psi_\varepsilon\|_2^2 \leq \frac{\sigma_{\max}^2}{(1-\vartheta)^2} \qquad \textit{and} \qquad \|\Gamma_h\|_2 \leq \frac{\vartheta^h\sigma_{\max}^2}{1-\vartheta}.
$$

As a consequence,

$$
\max\left\{(\|\Psi_\varepsilon\|_2^2+\omega^2)^2, \|\Gamma_h\|_2^2, \omega^4\right\} \leq \frac{(\sigma_{\max}^2+\omega^2)^2}{(1-\vartheta)^4}
$$

Proof. By Lemma 1, we can write Ψ_ε^2 as a sum of Kronecker products (one for each block). Let J_t he a matrix full of zeros, except for the subdiagonal of rank t, which is full of ones. Then we have:

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

This gives us control over its spectral norm thanks to Lemma 24:

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

We now use Lemma 2:

$$\|\Psi_\varepsilon\|_2^2 \leq \frac{\sigma_{\max}^2}{1-\vartheta^2}\left(1+2\frac{\vartheta}{1-\vartheta}\right) = \frac{\sigma_{\max}^2}{(1-\vartheta)^2}.$$

We now turn to Γ_h with Lemmas 1 and 2:

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

In particular, we have

$$\begin{aligned}\max\left\{(\|\Psi_\varepsilon\|_2^2+\omega^2)^2, \|\Gamma_h\|_2^2, \omega^4\right\} &\leq \max\left\{\left(\frac{\sigma_{\max}^2}{(1-\vartheta)^2}+\omega^2\right)^2, \left(\frac{\vartheta^h\sigma_{\max}^2}{1-\vartheta^2}\right)^2, \omega^4\right\} \\ &\leq \frac{(\sigma_{\max}^2+\omega^2)^2}{(1-\vartheta)^4}\end{aligned}$$

□

We can now control the error of the covariance estimator:

Lemma 13 (Max norm convergence rate of the covariance estimator)**.** *Let $\delta > 0$ be small enough. Assume that Equations (20) and (23) hold. Then the covariance estimator $\widehat{\Gamma}_h$ from Equation (9) satisfies*

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

with probability greater than $1-\delta$.

Proof. Let us plug Lemma 12 into Lemma 11

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

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

$$c_1 \exp\left(-\frac{c_2(1-\vartheta)^4 Tpq_u}{(\sigma_{\max}^2+\omega^2)^2}u^2\right) \leq 4\delta \quad \Longleftrightarrow \quad u \geq \sqrt{\frac{\log(c_1/4\delta)(\sigma_{\max}^2+\omega^2)^2}{c_2(1-\vartheta)^4 Tpq_u}}.$$

As long as δ is small enough, we can take

$$u = c\frac{\sqrt{\log(1/\delta)}(\sigma_{\max}^2+\omega^2)}{(1-\vartheta)^2\sqrt{Tpq_u}}. \tag{22}$$

For Lemma 11 to apply, we must verify that $u \in [0,1]$ and that Equation (21) is satisfied. In other words, we have to ensure that

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

Using Lemma 12, this is implied by the condition

$$\frac{\sqrt{\log(1/\delta)}\max\{1,(\sigma_{\max}^2+\omega^2)^{-1}\}}{(1-\vartheta)^2\sqrt{Tpq_u}} \leq c \tag{23}$$

	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	$\mathrm{err}(\delta)$	ζ_i
Optimization constraint	$\|M\widehat{\Gamma}_0 - \widehat{\Gamma}_1\|_{\max} \leq \lambda$	$\|S_0 M - S_1\|_{\max} \leq \lambda$
Optimization objective	$\|\mathrm{vec}(M)\|_1$	$\|\mathrm{vec}(M)\|_1$
Threshold in proof	ν	λ_1

Table 1: Notation correspondence between this paper and Han et al. [2015]

Under these hypotheses, we just proved that with probability at least $1 - 8\delta$,

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

We finish with a union bound, applying the previous result to all pairs $(d_1, d_2) \in [D]^2$. With probability greater than $1 - 8D^2\delta$, we have:

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

Replacing δ with $8D^2\delta$ gives us the result we wanted: with probability greater than $1 - \delta$,

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

□

A.7 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 recover Theorem 1, 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 1.

Our sparse transition estimator is defined as a solution to (8). 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 λ such that θ is feasible with high probability.

Lemma 14 (Feasibility of the real θ)**.** *If we select the penalization level*

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

then with probability at least $1 - \delta$, the real θ is a feasible solution to the optimization problem (8)*.*

Proof.

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

By Lemma 26,

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

By Lemma 13, with probability greater than $1 - 2\delta$,

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

which implies

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

This is exactly the feasibility criterion for (8) if $\lambda = (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta)$. □

Lemma 15 (Error on $\widehat{\theta}$ in max norm)**.** *If we select $\lambda = (\|\theta\|_\infty + 1)\,\mathrm{err}(\delta)$, then with probability at least $1 - \delta$, the max norm error of $\widehat{\theta}$ satisfies*

$$\|\widehat{\theta} - \theta\|_{\max} \leq 2\lambda \|\Gamma_0^{-1}\|_1.$$

Proof.

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

By Lemma 26,

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

We want to control $\|\widehat{\theta}\|_\infty$ using $\|\theta\|_\infty$. Let us recall that the operator ℓ_∞ norm is equal to the maximum ℓ_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

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

is equivalent to the row-wise minimization

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

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

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

□

To complete the proof of Theorem 1, we simply need to go from the max norm to the ℓ_∞ operator norm.

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

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

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

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

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

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

Meanwhile, the second term satisfies

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

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

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

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

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

Judging by the last Equation, it makes sense to choose $\nu = 2\lambda\|\Gamma_0^{-1}\|_1$. Furthermore, our sparsity hypothesis on θ 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,

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

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

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

We substitute the value of λ 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

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

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

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

to obtain the compressed expression

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

□

B Proof of the Minimax Lower Bound

We now present the detailed proof of Theorem 2.

B.1 Overview

Our argument is based on Fano's method, which we sum up in Lemma 27. 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 θ_i and θ_0. Taking the latter to be 0, we actually want to bound

$$\frac{1}{M+1} \sum_{i=1}^{M} \mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi, Y) \,\|\, \mathbb{P}_0(\Pi, Y)\right\} \leq \max_i \mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi, Y) \,\|\, \mathbb{P}_0(\Pi, Y)\right\}$$

By Lemma 28,

$$\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi, Y) \,\|\, \mathbb{P}_0(\Pi, Y)\right\} = \mathrm{KL}\left\{\mathbb{P}_{\theta_i}(\Pi) \,\|\, \mathbb{P}_0(\Pi)\right\} + \mathbb{E}_\Pi\left[\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \,\|\, \mathbb{P}_0(Y|\Pi)\right\}\right]$$

Since θ_i does not affect the distribution of the sampling process Π, the first term of the right-hand side is zero, and we will concentrate on the second term. First, we will upper-bound the random variable inside the expectation for a fixed realization of Π, and then we will average said bound over all possible projections.

We now give the structure of the argument 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_Π (corresponding to the independent case $\theta = 0$) plus a residual $R_\Pi(\theta)$ (Lemma 16).
2. Upper-bound the conditional KL divergence $\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi) \,\|\, \mathbb{P}_0(Y|\Pi)\right\}$ using the "deviations from the identity" $\Delta_\Pi(\theta) = Q_\Pi^{-1/2} R_\Pi(\theta) Q_\Pi^{-1/2}$ (Lemma 17).
3. Control $\Delta_\Pi(\theta)$ using features of $R(\theta)$ scaled by sampling-related factors (Lemmas 18, 19 and 20).
4. Deduce an upper bound on the KL divergence $\mathbb{E}_\Pi[\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi) \,\|\, \mathbb{P}_0(Y|\Pi)\right\}]$ (Lemma 21).
5. Apply Fano's method to a set of parameters θ_i constructed from a pruned binary hypercube of well-chosen radius.

B.2 Change of Notations

For this part, we slightly modify the previous conventions: we now assume that all the rows of Π_t that contain only zeros are removed. In other words, Π_t is no longer the diagonal matrix $\mathrm{diag}(\pi_t)$ but instead becomes a wide rectangular matrix with exactly one 1 per row and at most one 1 per column. We thus have $\Pi\Pi' = I$ unless all of the $\pi_{t,d}$ are zero, in which case the matrix Π is empty, and so are the observations Y. Let us denote this very unlikely event by E, and its complement by E^c. If Π is such that E happens, we obviously have $\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \parallel \mathbb{P}_0(Y|\Pi)\right\} = 0$, which means that

$$\mathbb{E}_\Pi\left[\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \parallel \mathbb{P}_0(Y|\Pi)\right\}\right] = \mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\,\mathrm{KL}\left\{\mathbb{P}_{\theta_i}(Y|\Pi) \parallel \mathbb{P}_0(Y|\Pi)\right\}\right] \tag{24}$$

For the beginning of the proof, we consider a fixed, non-empty realization of Π.

B.3 Covariance Decomposition

As we announced in the proof sketch, our reference parameter will be $\theta_0 = 0$, 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 16 (Conditional covariance decomposition)**.** *The covariance matrix of Y given Π decomposes as*

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

where Q_Π 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 (5) to see that the conditional distribution $\mathbb{P}_\theta(Y|\Pi)$ is a centered multivariate Gaussian with covariance

$$\mathrm{Cov}_\theta[Y|\Pi] = \omega^2 I + \Pi\,\mathrm{Cov}_\theta[X]\Pi'.$$

We then use Lemma 1 to get an expression of $\mathrm{Cov}_\theta[X]$ and deduce that its constant term (w.r.t to θ) is a block-diagonal matrix filled with copies of Σ:

$$\mathrm{Cov}_\theta[Y|\Pi] = \omega^2 I + \Pi\,\mathrm{bdiag}_T(\Sigma)\Pi' + \Pi\left(\mathrm{Cov}_\theta[X] - \mathrm{bdiag}_T(\Sigma)\right)\Pi'.$$

Finally, we define $Q_\Pi = \omega^2 I + \Pi\,\mathrm{bdiag}_T(\Sigma)\Pi'$, $R(\theta) = \mathrm{Cov}_\theta[X] - \mathrm{bdiag}_T(\Sigma)$ and $R_\Pi(\theta) = \Pi R(\theta)\Pi'$ to obtain the decomposition we announced. The diagonal blocks of $R(\theta)$ are easily computed by noticing that $\Gamma_0(\theta) - \Sigma = \theta\Gamma_0(\theta)\theta'$. □

B.4 From the KL Divergence to $\Delta_\Pi(\theta)$

Judging by Lemma 16, choosing a parameter θ close to 0 yields a conditional distribution for Y whose covariance is close to Q_Π. 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 17. *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:

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

Proof. The conditional KL divergence $\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi) \,\|\, \mathbb{P}_0(Y|\Pi)\right\}$ can be bounded using Lemma 30. Indeed, both conditional distributions are Gaussian and have the same expectation, and covariance matrices that are "close" in the following sense: by Lemma 16,

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

By Lemma 23, there exists a real number $r_{\min} \geq s_{\min}\left(Q_\Pi^{1/2}\right)^2 = s_{\min}(Q_\Pi)$ such that

$$\lambda_{\min}(\mathrm{Cov}_\theta(Y|\Pi)) = r_{\min}\lambda_{\min}(I + \Delta_\Pi(\theta)).$$

Since $Q_\Pi \succeq \omega^2 I \succ 0$, its minimum singular value satisfies $s_{\min}(Q_\Pi) > 0$, so that $r_{\min} > 0$. In addition, $\mathrm{Cov}_\theta(Y|\Pi) \succeq \omega^2 I \succ 0$, so that $\lambda_{\min}(\mathrm{Cov}_\theta(Y|\Pi)) > 0$. Therefore,

$$\lambda_{\min}(I + \Delta_\Pi(\theta)) = \frac{\lambda_{\min}(\mathrm{Cov}_\theta(Y|\Pi))}{r_{\min}} > 0 \quad \text{and} \quad \lambda_{\min}(\Delta_\Pi(\theta)) > -1,$$

which means we can apply Lemma 30 with $\mathbb{P}_1 = \mathbb{P}_\theta(Y|\Pi)$ and $\mathbb{P}_0 = \mathbb{P}_0(Y|\Pi)$. □

B.5 From $\Delta_\Pi(\theta)$ to $R_\Pi(\theta)$

Lemma 17 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 18. *Assume $\|R(\theta)\|_2 \leq (\sigma_{\min}^2 + \omega^2)/2$. We have the following upper bound:*

$$\frac{\|\Delta_\Pi(\theta)\|_F^2}{2(1 + \lambda_{\min}(\Delta_\Pi(\theta)))} \leq \frac{\|R_\Pi(\theta)\|_F^2}{(\sigma_{\min}^2 + \omega^2)^2}.$$

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

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

Let us start by noticing that, thanks to Lemma 25,

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

We will later see how the spectral and Frobenius norms of the residual $R(\theta)$ can be controlled as a function of θ. 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 Σ_{d} is block-diagonal, its spectrum is the same as the spectrum of Σ repeated T times, hence $\lambda_{\min}(\Sigma_{\mathrm{d}}) = \sigma_{\min}^2$. And since we assumed E^c happens (non empty projection), we have $\Pi\Pi' = I$ and $\Pi'\Pi = \mathrm{diag}(\pi)$, which has at least one entry equal to 1.

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

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

Remembering that $\Sigma_{\mathrm{d}} \succeq \sigma_{\min}^2 I$, we get

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

and thus

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

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

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

Because matrix inversion is decreasing w.r.t. the Loewner order on positive semi-definite matrices,

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

It follows that

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

The conclusion is within reach:

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

The last inequality is justified by our assumption $\|R(\theta)\|_2 \leq (\sigma_{\min}^2 + \omega^2)/2$. Another consequence of this assumption is that

$$\|\Delta_\Pi(\theta)\|_2 \leq \|Q_\Pi^{-1/2}\Pi\|_2^2\|R(\theta)\|_2 \leq \frac{1}{\sigma_{\min}^2 + \omega^2}\frac{\sigma_{\min}^2 + \omega^2}{2} = \frac{1}{2} < 1$$

which is sufficient for the first inequality to hold. □

B.6 From $R_\Pi(\theta)$ to $R(\theta)$

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

Lemma 19. *Let* $q_\ell = \max\{1 - b, 2p - (1 - b)\}$. *Then*

$$\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\|R_\Pi(\theta)\|_F^2\right] \leq p\,\mathrm{Tr}[R(\theta) \odot R(\theta)] + pq_\ell\|R(\theta)\|_F^2.$$

Proof. We first notice that for any matrix A,

$$\begin{aligned}\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\|\Pi A\Pi'\|_F^2\right] &= \mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\,\mathrm{Tr}\left[\Pi A\Pi'\Pi A'\Pi'\right]\right] \\ &= \mathbb{E}_\Pi\left[\mathrm{Tr}\left[\,\mathrm{diag}(\pi)A\,\mathrm{diag}(\pi)A'\right]\right] \\ &= \sum_{i,j}\mathbb{E}_\Pi[\pi_i\pi_j]A_{i,j}^2.\end{aligned}$$

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

$$\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\|R_\Pi(\theta)\|_F^2\right] = \sum_{i,j}\mathbb{E}_\Pi[\pi_i\pi_j]R(\theta)_{i,j}^2.$$

The rest of the proof consists in plugging in the moments $\mathbb{E}_\Pi[\pi_i\pi_j]$ from Lemma 4:

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

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

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

This yields a short, but probably suboptimal bound:

$$\mathbb{E}_\Pi\big[\mathbf{1}_{E^c}\|R_\Pi(\theta)\|_F^2\big] \leq p\,\mathrm{Tr}[R(\theta)\odot R(\theta)] + (p^2+p(1-p)|1-a-b|)\|R(\theta)\|_F^2.$$

In the previous part, we already saw that

$$p+(1-p)(1-a-b) = 1-b.$$

Similarly, we obtain

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

As a consequence,

$$p+(1-p)|1-a-b| = \max\{1-b, 2p-(1-b)\} = q_\ell,$$

which yields the expected result. □

B.7 Bounding $R(\theta)$

Lemma 19 relates the bounds involving $R_\Pi(\theta)$ to features of the full residual $R(\theta)$, which we now study.

Lemma 20. *The residual $R(\theta)$ satisfies the following inequalities:*

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

Proof. We start by giving a formula for the blocks of $R(\theta)$: by Lemma 16,

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

These individual blocks can be bounded using Lemmas 25 and 2: if $r \geq 1$, then

$$\|\theta^r\Gamma_0(\theta)\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2\|\theta^{r-1}\|_2^2 \leq \frac{\sigma_{\max}^4}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^{2(r-1)}$$

$$\|\Gamma_0(\theta)\theta'^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta^r\|_F^2 \leq \|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2\|\theta^{r-1}\|_2^2 \leq \frac{\sigma_{\max}^4}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^{2(r-1)}$$

$$\|\theta\Gamma_0(\theta)\theta'\|_F^2 \leq \|\theta\|_2^2\|\Gamma_0(\theta)\|_2^2\|\theta\|_F^2 \leq \frac{\sigma_{\max}^4}{(1-\vartheta)^2}\|\theta\|_F^2\|\theta\|_2^2.$$

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

$$\begin{aligned}
\|R(\theta)\|_F^2 &= \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\|\theta^{t-s}\Gamma_0(\theta)\|_F^2 + \|\theta\Gamma_0(\theta)\theta\|_F^2 + \sum_{s=t+1}^{T}\|\Gamma_0(\theta)\theta^{s-t}\|_F^2\right) \\
&\leq \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\|\theta\|_2^{2(t-s-1)} + \|\theta\|_2^2 + \sum_{s=t+1}^{T}\|\theta\|_2^{2(s-t-1)}\right) \\
&\leq \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}\sum_{t=1}^{T}\left(\sum_{s=-\infty}^{t-1}\|\theta\|_2^{2(t-1-s)} + \|\theta\|_2^2 + \sum_{s=t+1}^{+\infty}\|\theta\|_2^{2(s-1-t)}\right) \\
&= \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{1}{1-\|\theta\|_2^2} + \|\theta\|_2^2 + \frac{1}{1-\|\theta\|_2^2}\right)
\end{aligned}$$

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

$$\begin{aligned}
\|R(\theta)\|_F^2 &\leq \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{1}{1-\vartheta^2} + \vartheta^2 + \frac{1}{1-\vartheta^2}\right) \\
&= \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{2+\vartheta^2(1-\vartheta^2)}{1-\vartheta^2}\right) \\
&\leq \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{2+2\vartheta}{1-\vartheta^2}\right) = \frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta^2)^2}T\left(\frac{2}{1-\vartheta}\right) \\
&= 2T\frac{\sigma_{\max}^4\|\theta\|_F^2}{(1-\vartheta)^3}.
\end{aligned}$$

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 24 and write:

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

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

$$\begin{aligned}\operatorname{Tr}[R(\theta)\odot R(\theta)] &= T\operatorname{Tr}[(\theta\Gamma_0(\theta)\theta')\odot(\theta\Gamma_0(\theta)\theta')] \\ &\le T\|\theta\Gamma_0(\theta)\theta'\|_F^2 \le T\sigma_{\max}^4\frac{\|\theta\|_2^2\|\theta\|_F^2}{(1-\vartheta)^2}.\end{aligned}$$

□

B.8 Upper Bound on the KL Divergence

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

Lemma 21 (Final KL bound)**.** *Assume $\theta \in \Theta_s$ satisfies*

$$\|\theta\|_2 \le \frac{(1-\vartheta)^2(\sigma_{\min}^2+\omega^2)}{4\sigma_{\max}^2}$$

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

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

where we defined

$$\gamma_\ell = (1-\vartheta)^{3/2}\frac{\sigma_{\min}^2+\omega^2}{\sigma_{\max}^2} \qquad \textit{and} \qquad \mathrm{KL}_{\mathrm{avg}}(\|\theta\|_2,\|\theta\|_F) = \frac{2Tp(\|\theta\|_2^2+q_\ell)\|\theta\|_F^2}{\gamma_\ell}.$$

Proof. Let us start with Lemma 17 on the conditional KL divergence between $\mathbb{P}_\theta(Y|\Pi)$ and $\mathbb{P}_0(Y|\Pi)$: for any non-empty Π,

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

We continue with Lemma 18 linking $\Delta_\Pi(\theta)$ to $R_\Pi(\theta)$. As long as $\|R(\theta)\|_2 \le (\sigma_{\min}^2+\omega^2)/2$ (we will see to that at the end), we have

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

Taking the expectation on the event E^c yields:

$$\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\,\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right] \le \frac{\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\|R_\Pi(\theta)\|_F^2\right]}{(\sigma_{\min}^2+\omega^2)^2}$$

We can now apply Lemma 19:

$$\mathbb{E}_\Pi\left[\mathbf{1}_{E^c}\,\mathrm{KL}\left\{\mathbb{P}_\theta(Y|\Pi)\parallel\mathbb{P}_0(Y|\Pi)\right\}\right] \le \frac{p\operatorname{Tr}[R(\theta)\odot R(\theta)] + pq_\ell\|R(\theta)\|_F^2}{(\sigma_{\min}^2+\omega^2)^2}$$

We substitute the residual bounds from Lemma 20:

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

By Equation (24), this is equivalent to bounding the expected KL divergence regardless of the event E, hence the result. Note that our assumption on θ, combined with Lemma 20, implies

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

□

B.9 Application of Fano's Method

Given the KL bound we just obtained, we are finally able to prove Theorem 2.

Proof. Fano's method requires finding $M+1$ parameters θ_i such that $\theta_0=0$ and $\|\theta_i-\theta_j\|_F\geq 2\tau$ for $i\neq j$ (with τ 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 21, 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 2τ-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 row- and column-sparse, with no more than s non-zero coefficients per row or 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 θ 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 35: 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 θ_i and θ_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 21. We only need to assume

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

to get the upper bound

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

Since we must satisfy the constraint from Equation (28) in Fano's method, we will choose r so that:

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

with $\alpha = \frac{\log 3 - \log 2}{2 \log 2}$. We want to solve the previous inequality for r, and for that we start by replacing $\mathrm{KL}_{\mathrm{avg}}(rs, r\sqrt{Ds})$ with its value from Lemma 21, replacing γ_ℓ with γ_ℓ to lighten notations:

$$\begin{aligned}\mathrm{KL}_{\mathrm{avg}}(rs, r\sqrt{Ds}) \leq \alpha \log\left(2^{Ds/8} - 1\right) &\iff \frac{2}{\gamma_\ell} Tp\left((rs)^2 + q_\ell\right)(r\sqrt{Ds})^2 \leq cDs \\ &\iff Ds^3 r^4 + q_\ell Dsr^2 - c\frac{\gamma_\ell^2 Ds}{Tq_\ell} \leq 0.\end{aligned}$$

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

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

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

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

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

$$\frac{\gamma_\ell s}{\sqrt{p} q_\ell \sqrt{T}} \leq c, \tag{25}$$

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

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

To sum up, we have three constraints on r:

$$rs \leq \vartheta \qquad\qquad rs \leq \frac{(1-\vartheta)^2(\sigma_{\min}^2 + \omega^2)}{4\sigma_{\max}^2} = \frac{\sqrt{1-\vartheta}}{4}\gamma_\ell \qquad\qquad r \leq \sqrt{c\frac{\gamma_\ell^2}{Tpq_\ell}}.$$

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

$$r = \frac{1}{s}\min\left\{\vartheta, \frac{\gamma_\ell\sqrt{1-\vartheta}}{4}, c\frac{\gamma_\ell s}{\sqrt{Tpq_\ell}}\right\} \tag{26}$$

To reach our conclusion, we simply need to remark that the vectorized ℓ_1 distance between any two matrices in $\mathcal{K}(r)$ gives us a lower bound on the operator ℓ_∞ distance that separates them:

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

Subsequently, our parameters θ_i are 2τ-separated (in ℓ_∞ operator distance) with $\tau = rs/8$. As soon as the minimum in Equation (26) is reached by the third value, i.e. whenever

$$\frac{\gamma_\ell s}{\sqrt{Tpq_\ell}} \leq c\min\{\vartheta, \gamma_\ell\sqrt{1-\vartheta}\} \tag{27}$$

we can simplify the expression of τ:

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

In this case, by Lemma 27, we can conclude:

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

□

C Useful Lemmas

C.1 Linear Algebra

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

Lemma 22 (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 23 (Ostrowski)**.** *Let S and A be two $n\times n$ matrices with S symmetric. For all i, there is a real number $r_i \in [s_{\min}(A)^2, s_{\max}(A)^2]$ such that $\lambda_i(ASA') = r_i\lambda_i(S)$, where $s_{\min}$ (resp. $s_{\max}$) denotes the minimum (resp. maximum) singular value.*

Proof. See Horn and Johnson [2012, Theorem 4.5.9 and Corollary 4.5.11] □

Lemma 24 (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 25. *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:

$$\begin{aligned}\forall (P,Q)\in\mathcal{S}_n(\mathbb{R}), \forall R, \quad P\preceq Q &\implies R'PR \preceq R'QR\\ \forall (P,Q)\in\mathcal{S}_n(\mathbb{R}), \quad P\preceq Q &\implies \mathrm{Tr}(P)\leq \mathrm{Tr}(Q).\end{aligned}$$

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

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

which implies

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

The proof for the other inequality is identical. □

Lemma 26. *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:

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

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

□

C.2 PROBABILITY

Lemma 27 (Fano's method)**.** *Let $\theta_0, ..., \theta_M$ be $M+1$ parameters that are 2τ-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

$$\frac{1}{M+1} \sum_{i=1}^{M} \mathrm{KL}\left\{\mathbb{P}_{\theta_i} \parallel \mathbb{P}_{\theta_0}\right\} \leq \alpha \log M \quad \textit{with} \quad 0 < \alpha < 1 \tag{28}$$

Then the minimax probability of an error at threshold τ 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 $\alpha = \frac{\log(3)-\log(2)}{2\log(2)} \geq 1/2$ is enough to obtain a minimax risk greater than α, as soon as $M \geq 3$. □

Lemma 28 (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 29 (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 30 (KL divergence between close Gaussians)**.** *Let Δ 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 29 (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 31 (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 32 (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 μ. 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 ϵ-mixing:

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

The result follows easily. □

Lemma 33 (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 34 (Chernoff inequality for Markov chains under Doeblin condition)**.** *Under the hypotheses of the previous two Lemmas (32 and 33), if the parameters r and δ 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 32, since r and δ 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 33. □

Lemma 35 (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] □

Lemma 36 (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. □

Lemma 37 (Conditional Hanson-Wright inequality)**.** *Let A be a random square matrix such that with probability $1-\delta$,*

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

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

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

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 $\mathbb{P}(A \in \mathcal{A}_{>}) = \sum_{a\in\mathcal{A}_{>}} \mathbb{P}(A = a) \leq \delta$, 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 36 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

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

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

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

This implies:

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

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. □

Lemma 38 (Heuristic optimality of the signal-to-noise ratio)**.** *In the one-dimensional setting with full observations, the dependency of the error in* $1 + \frac{\sigma^2}{\omega^2}$ *is "coherent" with the asymptotic behavior of the MLE.*

Proof. Let us consider the case where $D = 1$ and $p = 1$, since we are mainly interested in the role of the parameters σ^2 and ω^2. In this case, Theorem 2 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 scenario, Theorem 1 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 Π_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 θ is identifiable and does not lie on the boundary of Θ_s. Identifiability is easily deduced from Lemma 1 by observing that $\theta = \Gamma_1(\theta)\Gamma_0(\theta)^{-1}$ can be entirely deduced from distribution moments.

Since all of these prerequisites hold here, 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 16, but in our case the sampling matrix is constant, and we obtain the simpler (unconditional) result

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

where the residual $R(\theta)$ is of order 1 in θ. 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 θ has an explicit formula [Malagò and Pistone, 2015, Section 3.5]:

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

If assume θ 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 θ,

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

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 2.

□

D GLOSSARY

D.1 NOTATIONS

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, $\operatorname{Var}[X]$ for a variance (scalar of vector) and $\operatorname{Cov}[X, Y]$ for a covariance (scalar or matrix). The symbols $\mathcal{B}(p)$ and $\mathcal{N}(\mu, \Sigma)$ denote a Bernoulli distribution and a (possibly multivariate) Gaussian distribution. 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. ℓ_1 norm, ℓ_∞ norm, number of nonzero entries). The notation e_i stands for a vector with a single non-zero coordinate at position 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 zeros, except for the subdiagonal of rank r which is filled with ones. The notation $\operatorname{diag}(\lambda)$ stands for the diagonal matrix with coefficients $\lambda_1, ..., \lambda_n$, while $\operatorname{bdiag}_T(M)$ stands for a block-diagonal matrix with T copies of M on the diagonal and zeros elsewhere. We write $\operatorname{vec}(M)$ 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, $M^\dagger$ its Moore-Penrose pseudo-inverse and M^{-1} its inverse;
- $\operatorname{Tr}(M)$ 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

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

- $s_{\max}(M)$ (resp. $s_{\min}(M)$, $s_i(M)$) 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 ℓ_1 norm, which is the maximum ℓ_1 norm of a column of M;
- $\|M\|_2 = \sup \frac{\|Mx\|_2}{\|x\|_2} = |s_{\max}(M)| = \sqrt{\lambda_{\max}(M'M)}$ its operator ℓ_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 ℓ_∞ norm, which is the maximum ℓ_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 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.

D.2 Frequent symbols

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

Dimensions:

- $t \in [T]$: time step
- $d \in [D]$: dimension

State process:

- X_t: state process
- θ: transition matrix
- ε_t: innovations
- Σ: covariance matrix of ε_t
- $\sigma^2_{\min}, \sigma^2_{\max}$: extremal eigenvalues of Σ
- s: sparsity level of θ (number of non-zero coefficients in each row)
- ϑ: maximum ℓ_2 norm for θ
- Θ_s: set of feasible values for θ
- $\Gamma_h(\theta)$: covariance between X_{t+h} and X_t

Observations:

- π_t: random sampling vector
- Π_t: diagonal random sampling matrix
- p: fraction of state components activated by observations

- $\mathcal{T}$: transition matrix for Markov sampling
- a, b: transition probabilities for Markov sampling
- χ: minimum distance between a or b and $\{0, 1\}$ (considered constant)
- Y_t: observations
- η_t: noise
- ω^2: variance of η_t

Estimation:

- h: covariance time lag
- h_0: minimum covariance time lag for transition estimation
- $S(h)$: scaling matrix for covariance estimation
- pq_u: smallest coefficient of the scaling matrix

Other:

- g: standard Gaussian vector
- Ψ_ε (resp. Ψ_η): link between X (resp. η) and a standard Gaussian vector
- L: random bilinear form
- u: threshold in concentration inequalities
- δ: small probability
- Q_Π: 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
- γ_ℓ (resp. $\gamma_u(\theta)$): signal-to-noise ratio in the lower bound (resp. the upper bound)