Matching correlated VAR time series

Anonymous authors
Paper under double-blind review

Abstract

We study the problem of matching correlated VAR time series databases, where a multivariate time series is observed along with a perturbed and permuted version, and the goal is to recover the unknown matching between them. To model this, we introduce a probabilistic framework in which two time series $(x_t)_{t\in[T]}, (x_t^{\#})_{t\in[T]}$ are jointly generated, such that $x_t^{\#} = x_{\pi^*(t)} + \sigma\tilde{x}_{\pi^*(t)}$, where $(x_t)_{t\in[T]}, (\tilde{x}_t)_{t\in[T]}$ are independent and identically distributed vector autoregressive (VAR) time series of order 1 with Gaussian increments, for a hidden π^*. The objective is to recover π^*, from the observation of $(x_t)_{t\in[T]}, (x_t^{\#})_{t\in[T]}$. This generalizes the classical problem of matching independent point clouds to the time series setting.

We derive the maximum likelihood estimator (MLE), leading to a quadratic optimization over permutations, and theoretically analyze an estimator based on linear assignment. For the latter approach, we establish recovery guarantees, identifying thresholds for σ that allow for perfect or partial recovery. Additionally, we propose solving the MLE by considering convex relaxations of the set of permutation matrices (e.g., over the Birkhoff polytope). This allows for efficient estimation of π^* and the VAR parameters via alternating minimization. Empirically, we find that linear assignment often matches or outperforms MLE relaxation based approaches.

1 Introduction

We consider the problem of matching two point clouds in $\mathbb{R}^d$. Let $X, X^{\#} \in \mathbb{R}^{d\times T}$ denote the matrices corresponding to the two point clouds, each containing T points in $\mathbb{R}^d$. We say that the point clouds are correlated if there exists a permutation[1] map $\pi^* : [T] \to [T]$ such that every column $x^{\#}_{\pi^*(t)}$ of $X^{\#}$ is correlated with the column x_t of X. Given X and $X^{\#}$, the goal then is to recover the unknown permutation π^*. This problem has a rich history with applications in computational geometry, computer vision, multi-object tracking, etc.

On the theoretical front, this has received attention recently when the columns of $X, X^{\#}$ are assumed to be random i.i.d vectors (Kunisky & Niles-Weed, 2022; da Rocha Schwengber & Oliveira, 2024). Specifically, Kunisky & Niles-Weed (2022) considered the setup where we first (a) draw $x_1, \ldots, x_T \sim \mathcal{N}(0, I_d)$ independently, then (b) draw the noise vectors $\tilde{x}_1, \ldots, \tilde{x}_T \sim \mathcal{N}(0, I_d)$ independently, and finally (c) obtain $x_t^{\#}$ as

$$x_t^{\#} = x_{\pi^*(t)} + \sigma\tilde{x}_{\pi^*(t)} \quad \text{for } t = 1, \ldots, T.$$

It was shown that the maximum likelihood estimator (MLE), which is essentially a linear assignment problem, provably recovers π^* provided the noise level σ is less than a threshold. This was shown for different recovery criteria such as exact recovery and partial recovery (with sublinear number of errors). The setting in da Rocha Schwengber & Oliveira (2024) extended these results to more general distributions, along with information-theoretic lower bounds; see Section 1.3 for a more detailed overview of related work.

[1] Extensions for which this holds only for a subset of $[m]$, considering sub-permutations, can be formulated analogously.

Motivation for matching time series data. The assumption that the points within a point cloud are drawn independently was motivated in Kunisky & Niles-Weed (2022) by a stylized model for multi-target tracking involving T independent standard Brownian motions. Here, $x_1, \ldots, x_T$ correspond to the position of the (unlabeled) particles at a given time instant, and $x_1^\#, \ldots, x_T^\#$ corresponds to their positions at the next time instant. In this paper, we consider a generalization of this setting to one where $(x_1, \ldots, x_T)$ and $(\tilde{x}_1, \ldots, \tilde{x}_T)$ are the realization of a stochastic process, hence the individual x_t's and $\tilde{x}_t$'s will be respectively dependent. The motivation for studying this *time-series* setting comes from the following applications.

- **Time stamp shuffling as a privacy mechanism.** One natural way to obfuscate temporal data is to release values while discarding or shuffling their time stamps. This was recently considered in the context of differential privacy for sensitive time series data such as health care records, financial transactions etc. (Ye et al., 2021; 2023). The intuition is that the resulting data remains useful for certain aggregate statistics (e.g., mean) while concealing the temporal structure. Consider wage data: observing a person's monthly income over several years without the ordering allows one to compute their average salary, but conceals whether their earnings followed a steady upward trend or fluctuated seasonally. From the shuffled sequence alone, these scenarios may look indistinguishable. This raises the question of whether time-stamp shuffling offers meaningful privacy protection: if an adversary has access to an auxiliary, correlated time series, they may be able to partially reconstruct the original ordering and recover sensitive temporal information.

- **Sensor fusion and lost timestamps.** In distributed sensing networks (seismology, wireless sensors, Internet of Things), different sensors may not be synchronized. One sensor provides a clean signal with timestamps, while another provides a noisy but related signal without ordering (e.g. due to packet loss, buffering, or asynchronous sampling). For e.g., in the Internet of Things context, one may have many cheap sensors scattered around, each with limited processing power and no global clock synchronization. The data streams can arrive out of order, delayed, or with missing timestamps, as reported in Keller et al. (2011) when analyzing data from a real time system (Beutel et al., 2009). A natural goal then would be to align unordered sensor readings with the reference signal by exploiting temporal correlations, thus improving reconstruction accuracy.

- **Time series alignment.** A fundamental problem in time series analysis is to align different, potentially misaligned sequences that reflect the same underlying phenomenon. Misalignment may be caused by temporal stretching, delays, or nonlinear warping. In the classical formulation, each series preserves its internal temporal structure, and the task is to find an appropriate monotone correspondence between time indices. In contrast, a shuffled time-stamp setting can model more severe distortions such as measurement defects, packet loss, or corrupted logging systems, where the ordering of observations is partially lost. This makes the alignment problem more challenging and closer in spirit to matching under unknown permutations. Applications of time series alignment are widespread, e.g., in neuroscience (Kawano et al., 2020), speech and gesture recognition (Sakoe & Chiba, 1978), and bioinformatics (Aach & Church, 2001) to name a few.

1.1 Correlated VAR model for planted matching

Given a matrix $A^* \in \mathbb{R}^{d\times d}$, where $d \in \mathbb{N}$, suppose $(x_t)_{t\in[T]}$ is generated as follows.

$$x_{t+1} = A^* x_t + \xi_{t+1}, \quad \text{for } t = 1, \ldots, T-1, \tag{1.1}$$

$$x_1 = \xi_1, \tag{1.2}$$

where $(\xi_t)_{t\in[T]}$ is a sequence of i.i.d. standard Gaussian vectors in $\mathbb{R}^d$. The above model is a *Vector Autoregressive* model of order 1, hereby referred to as $\mathbf{VAR}(1, d, T)$. The matrix A^* contains the coefficients that determine this temporal relationship and is referred to as the *system matrix*.

CVAR: A model for correlated time series. Let $(\tilde{x}_t)_{t\in[T]}$ be an independent copy of the "base" time series $(x_t)_{t\in[T]}$, drawn from $\mathbf{VAR}(1, d, T)$, but with an independent sequence of i.i.d. standard Gaussian

vectors $(\tilde{\xi}_t)_{t\in[T]}$. Given a noise parameter $\sigma \geq 0$, we first construct the perturbed time series

$$x'_t = x_t + \sigma\tilde{x}_t, \quad \text{for all } t \in [T]. \tag{1.3}$$

Here, $(x'_t)_{t\in[T]}$ is a noisy version of $(x_t)_{t\in[T]}$, with σ controlling the noise magnitude. Finally, given a permutation π^*, we define

$$x^{\#}_t = x'_{\pi^*(t)}, \quad \text{for all } t \in [T].$$

The above model is referred to as the correlated VAR model with parameters A^*, π^* and σ, or **CVAR**(1, d, T; A^*, π^*, σ) in short. The pair $\big((x_t)_{t\in[T]}, (x^{\#}_t)_{t\in[T]}\big)$ is then a realization from this model. Our goal is to recover the planted matching π^* given the observations $((x_t)_{t\in[T]}, (x^{\#}_t)_{t\in[T]})$, where the matrix A^* is unknown. Before delving into strategies for estimating π^*, the following remarks are worth noting.

1. The above setup is a generalization of the point clouds matching problem, presented in Kunisky & Niles-Weed (2022). Indeed, if $A^* = 0$, we have, for all $t \in [T]$,

   $$x_t = \xi_t \quad \text{and} \quad \tilde{x}_t = \tilde{\xi}_t.$$

   In this case, $x'_t = \xi_t + \sigma\tilde{\xi}_t$, which is identical to the setting in Kunisky & Niles-Weed (2022).

2. In the noiseless case ($\sigma = 0$) the permutation can be perfectly recovered with a simple algorithm. It suffices to match x_t with $x^{\#}_{t'}$ such that $x_t = x^{\#}_{t'}$. Given that, for any $s, t \in [T]$ with $t \neq s$, $\mathbb{P}(x_t = x_s) = 0$, the algorithm returns the correct permutation almost surely.

3. Notice that we assumed that the base time series is presented with its temporal ordering known. Hence, one may interpret the matching problem as that of recovering the temporal ordering of the unordered series $(x^{\#}_t)_{t\in[T]}$ using information available from $(x_t)_{t\in[T]}$. We touch upon this point briefly below, see Remarks 3 and 4 for a more detailed explanation.

Remark 1. *Interpreting the time-stamps as labels for each data point in the time series, the assumption that the correct temporal order is known for the base time series corresponds to a standard assumption in data privacy applications. Specifically, in data de-anonymization settings, one database is assumed to be public with known labels, while the other must remain private. The goal in this context is to recover the private labels using the public database as a reference (see, e.g., Narayanan & Shmatikov (2008) for a seminal work in this area).*

1.2 Contributions

Our contributions are outlined below.

1. We propose, to our knowledge, a novel statistical model for planted matching in the context of time series data. For this model, we derive the MLE estimator for estimating A^*, π^* which amounts to minimizing a biconvex objective subject to nonconvex constraints (due to the set of permutations). We formulate several methods to solve it based on the alternating minimization framework, by considering different convex relaxations of the set of permutation matrices. The empirical performance of these relaxations are compared through several experiments on synthetic data. Interestingly, the linear assignment (LA) estimator performs comparably to the best MLE relaxations even under high noise levels, when $\|A^*\|_2 \leq 1$, raising the question of whether it achieves optimal, or near-optimal, performance in this regime. Experimentally, we found that for some values $\|A^*\|_2 > 1$, MLE relaxations slightly outperform the LA estimator.

2. On the theoretical front, we analyze the statistical performance of the LA estimator for recovering π^*, which is well studied for geometric matching problems in the i.i.d setting (Kunisky & Niles-Weed, 2022; da Rocha Schwengber & Oliveira, 2024). This estimator is model agnostic and also does not need the base time series to be temporally ordered (see Remark 4). Assuming $\|A^*\|_2 < 1$, we derive bounds on the number of mismatched points by following the technique of counting augmenting

paths, considered recently in Kunisky & Niles-Weed (2022) (and later in da Rocha Schwengber & Oliveira (2024)); see Theorem 1. Theorem 1 shows various thresholds on the noise level σ which imply different levels of recovery of π^* (e.g., exact recovery, partial recovery). It's statement is analogous to that obtained in Kunisky & Niles-Weed (2022), up to an additional factor proportional to $(1 - \|A^*\|_2)^5$ appearing in the bounds; see Remark 5. The proof, while along the lines of that in Kunisky & Niles-Weed (2022) is considerably more challenging – not simply in the sense of more tedious calculations, but also in terms of technical difficulties imposed by the **CVAR** model; see Remark 7 for details.

1.3 Related work

Matching point clouds. As discussed earlier, our statistical model generalizes the i.i.d setting (within a point cloud) in Kunisky & Niles-Weed (2022) to the dependent setup, where each point cloud is a VAR time series. The work of da Rocha Schwengber & Oliveira (2024) extended the setup in Kunisky & Niles-Weed (2022) to handle more general classes of distributions, and also provides information-theoretic lower bounds on the expected error for any estimator. The latter was achieved by making a connection with matchings in random geometric graphs. Some of the results in Kunisky & Niles-Weed (2022) were shown to be information theoretically optimal in Wang et al. (2022). While we do not study lower bounds for the **CVAR** model, it is an interesting and non-trivial direction to pursue for future work.

Feature matching. A closely related problem referred to as the feature matching problem was studied in Collier & Dalalyan (2016). Here the goal is to match two sets of points in $\mathbb{R}^d$ (of potentially unequal size), and the proposed statistical model assumes all the points to be independently generated Gaussian's. The means of the Gaussian distributions for one point cloud are denoted $(\theta_i^*)_i$ while that of the other point cloud are $(\theta^*_{\pi^*(i)})_i$, with π^* the latent permutation. The performance of different estimators, including the LA estimator (referred therein as the least sum of squares estimator) was studied theoretically with aim of establishing the minimax rate of separation between the θ_i's for consistent recovery of π^*. This result was extended in Galstyan et al. (2022) to the setting where the second point set can contain outliers.

Covariance alignment. In Han et al. (2025), the authors, motivated by the feature matching problem, study the task of covariance matrix alignment. Specifically, two independent samples of i.i.d points are observed: one distributed as $\mathcal{N}(0, \Sigma)$, and the other as $\mathcal{N}(0, \Pi^* \Sigma {\Pi^*}^\top)$, where Π^* is a hidden permutation matrix. In this setting, Σ can be viewed as a nuisance parameter, and the goal is to align the sample covariance matrices of the two datasets. To recover Π^*, the authors derive a quasi–maximum likelihood estimator, which reduces to solving a quadratic optimization problem over the permutation set—an instance of the Quadratic Assignment Problem (QAP) (known to be NP-hard). They propose a relaxation over the Birkhoff polytope, referred to as the Gromov–Wasserstein estimator, due to its connection with optimal transport, and show that this estimator is minimax optimal.

Graph matching. In the graph matching problem, the goal is to find an assignment between the vertices of two graphs such that their edges are maximally aligned. This problem has found many applications in computer vision (Emmert-Streib et al., 2016), data de-anonymization (Narayanan & Shmatikov, 2008) and protein-protein interactions (Zaslavskiy et al., 2009), to name a few. In the statistical version of the problem, the pair of graphs are realizations of a probabilistic model for correlated random graphs. The most popular models are the *correlated Erdős-Rényi* model (Pedarsani & Grossglauser, 2011) and the *correlated Gaussian Wigner* model (Ding et al., 2021a). For both models, the MLE corresponds solving a QAP problem. Many algorithms rely on convex relaxations (Aflalo et al., 2015; Lyzinski et al., 2016; Fan et al., 2022; Araya & Tyagi, 2025), and some of our proposed relaxations for MLE in our setting draw inspiration from those of graph matching. A related line of work exists for recovering planted matchings in weighted bipartite graphs, without a latent geometric structure (Ding et al., 2021b; Semerjian et al., 2020; Moharrami et al., 2021).

Learning dynamical systems from non-sequenced data. The problem of learning dynamical systems from *non-sequenced* data has received significantly less attention than the setting of sequenced observations. This was first proposed in Huang & Schneider (2009) for linear VAR models where it was assumed that

multiple i.i.d realizations of the model are first generated and then a single state is sampled at random from each trajectory. An Expectation-Maximization (EM) algorithm was proposed and tested on synthetic data. This was extended to non-linear VAR models in Huang et al. (2010) where the single-trajectory setting was also considered. For this setup, Huang et al. (2010) proposed a convex program over the Birkhoff polytope for estimating the latent ordering of the points. The work of Huang & Schneider (2011) considered linear VAR models where a small amount of non-sequenced data drawn from the stationary distribution of the model is also available. A penalized least-squares method was proposed where the penalization is based on the Lyapunov equation concerning the covariance matrix of the stationary distribution. Finally, Huang & Schneider (2013) considered learning (first order or Hidden) Markov models from non-sequenced data, and proposed methods based on tensor decomposition along with theoretical guarantees.

1.4 Notation

We reserve lowercase letters for vectors and uppercase letters for matrices. For $x \in \mathbb{R}^d$, we write $\|x\|_2$ for the standard Euclidean (ℓ_2) norm. Similarly, for a matrix $X \in \mathbb{R}^{k\times k'}$, $\|X\|_F$ denotes its Frobenius norm while $\|X\|_2$ denotes its spectral norm. Given matrices $M \in \mathbb{R}^{k\times k'}$ and $N \in \mathbb{R}^{l\times l'}$, the standard Kronecker product between M and N is denoted by $M \otimes N \in \mathbb{R}^{kl\times k'l'}$. For a symmetric matrix $M \in \mathbb{R}^{k\times k}$, its eigenvalues are denoted by $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_k(M)$.

Given $T \in \mathbb{N}$, we let $\mathcal{S}_T$ denote the set of permutation maps on $\{1, \ldots, T\}$ and $\mathcal{P}_T$ the set of corresponding permutation matrices. Sets will be usually denoted by calligraphic letters.

The notation $x \sim \mathcal{N}(\mu, \Sigma)$ specifies that x is a Gaussian random vector with mean μ and covariance matrix Σ.

For negative functions f, g, we will often write $f(x) = O(g(x))$ if there exists $c > 0$ and x_0 such that $f(x) \leq cg(x)$ for all $x \geq x_0$. Moreover, we write $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$ holds, and write $f(x) = \Theta(g(x))$ if both $f(x) = O(g(x))$ and $g(x) = O(f(x))$ hold. Finally, we write $f(x) = o(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$ and $f(x) = \omega(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \infty$.

2 MLE and the linear assignment estimator

2.1 MLE for recovering π^* and A^* when σ is known

We first derive the MLE for π^* and A^*, given the observations $\big((x_t)_{t\in[T]}, (x_t^{\#})_{t\in[T]}\big)$ generated from **CVAR**(1, d, T; A^*, π^*, σ). Notice that although our main goal is to recover π^*, in principle, the parameters A^* and σ are also unobserved. For convenience, we will assume that σ is known. As we will see shortly, this assumption only enters the picture for obtaining the MLE of A^*. If A^* is known, then the MLE of π^* does not require σ to be known. In the work Kunisky & Niles-Weed (2022) where $A^* = 0$, the MLE for π^* (which is a linear assignment problem) was obtained for σ unknown.

Lemma 1 (MLE for CVAR given σ)**.** *Given σ, the MLE for (π^*, A^*) is found by solving*

$$\min_{\pi\in\mathcal{S}_T, A\in\mathbb{R}^{d\times d}} \left\{ \sum_{t=1}^{T} \left(\|x_t - Ax_{t-1}\|_2^2 + \frac{1}{\sigma^2}\|(x^{\#}_{\pi^{-1}(t)} - Ax^{\#}_{\pi^{-1}(t-1)}) - (x_t - Ax_{t-1})\|_2^2 \right) \right\}, \tag{2.1}$$

where we set $\pi^{-1}(0), x_0^{\#}, x_0 \equiv 0$ for notational convenience.

Proof. Let $(\xi_t)_{t\in[T]}, (\tilde{\xi}_t)_{t\in[T]}$ be two sequences of iid standard Gaussian vectors in dimension d. If $\big((x_t)_{t\in[T]}, (x_t^{\#})_{t\in[T]}\big)$ were generated according to the **CVAR**$(1, d, T; A, \pi, \sigma)$ model, we would have

$$x_t = Ax_{t-1} + \xi_t \quad \text{and} \quad \tilde{x}_t = A\tilde{x}_{t-1} + \tilde{\xi}_t \qquad \text{for } t = 2, \ldots, T,$$

with $x_1 = \xi_1$ and $\tilde{x}_1 = \tilde{\xi}_1$. On the other hand, $x'_t = x_t + \sigma\tilde{x}_t$ and $x_t^{\#} = x'_{\pi(t)}$. Define the negative log-likelihood function, including σ as unobserved, by

$$\mathcal{L}(A, \pi, \sigma) := -\log f_{A,\pi,\sigma}(x_1, x_1^{\#}, x_2, x_2^{\#}, \ldots, x_T, x_T^{\#}),$$

where $f_{A,\pi,\sigma}$ denotes the joint density of $x_1, x_1^\#, x_2, x_2^\#, \ldots, x_T, x_T^\#$, under the $\mathbf{CVAR}(1, d, T; A, \pi, \sigma)$ model. We use a similar notation to denote the density of marginals, e.g., $f_{A,\pi,\sigma}(x_t)$ denotes the (marginal) density of x_t. First, note that

$$\begin{aligned} x'_t = x_t + \sigma\tilde{x}_t &= Ax_{t-1} + \xi_t + \sigma A\tilde{x}_{t-1} + \sigma\tilde{\xi}_t \\ &= Ax'_{t-1} + \xi_t + \sigma\tilde{\xi}_t \\ &= Ax'_{t-1} + x_t - Ax_{t-1} + \sigma\tilde{\xi}_t. \end{aligned}$$

Denoting[2] $f_{A,\pi,\sigma}(x_t, x'_t | x_{t-1}, x'_{t-1})$ the density of (x_t, x'_t) given x_{t-1}, x'_{t-1}, we have for $t = 2, \ldots, T$,

$$\begin{aligned} f_{A,\pi,\sigma}(x_t, x'_t | x_{t-1}, x'_{t-1}) &= \underbrace{f_{A,\pi,\sigma}(x_t | x_{t-1}, x'_{t-1})}_{\sim\mathcal{N}(Ax_{t-1}, I_d)} \underbrace{f_{A,\pi,\sigma}(x'_t | x_t, x_{t-1}, x'_{t-1})}_{\sim\mathcal{N}(Ax'_{t-1} + x_t - Ax_{t-1}, \sigma^2 I_d)} \\ &= \frac{C}{\sigma^d} \exp\left(-\frac{\|x_t - Ax_{t-1}\|_2^2}{2}\right) \exp\left(-\frac{\|x'_t - Ax'_{t-1} - (x_t - Ax_{t-1})\|_2^2}{2\sigma^2}\right), \end{aligned}$$

where $C > 0$ is a constant. The above is also valid for $t = 1$, with the notation $x_0 = x'_0 = 0$. With that in mind, we have $f_{A,\pi,\sigma}(x_1, x'_1, \ldots, x_T, x'_T) = \prod_{t=1}^T f_{A,\pi,\sigma}(x_t, x'_t | x_{t-1}, x'_{t-1})$ which implies

$$\begin{aligned} -\log f_{A,\pi,\sigma}(x_1, x'_1, \ldots, x_T, x'_T) =& dT \log\sigma + \frac{1}{2}\sum_{t=1}^T \|x_t - Ax_{t-1}\|_2^2 + TC \\ &+ \frac{1}{2\sigma^2}\sum_{t=1}^T \|(x'_t - Ax'_{t-1}) - (x_t - Ax_{t-1})\|_2^2. \end{aligned}$$

Since $x_t^\# = x'_{\pi(t)}$, for $t = 1, \ldots, T$, we obtain

$$\begin{aligned} \mathcal{L}(A, \pi, \sigma) =& dT \log\sigma + \frac{1}{2}\sum_{t=1}^T \|x_t - Ax_{t-1}\|_2^2 + TC \\ &+ \frac{1}{2\sigma^2}\sum_{t=1}^T \|(x^\#_{\pi^{-1}(t)} - Ax^\#_{\pi^{-1}(t-1)}) - (x_t - Ax_{t-1})\|_2^2. \end{aligned}$$

For a given σ, we arrive at the formulation in (2.1) for finding the optimal A, π. □

Remark 2 (MLE when $\sigma = 0$)**.** *In the case $\sigma = 0$, the negative log-likelihood simplifies to*

$$\mathcal{L}(A, \pi, 0) = \sum_{t=1}^T \|x_t - Ax_{t-1}\|_2^2 - \sum_{t=1}^T \log \mathbb{1}_{\{x_t = x^\#_{\pi^{-1}(t)}\}}.$$

As a result, the optimization problem

$$\min_{\pi, A} \mathcal{L}(A, \pi, 0)$$

is separable. Consequently, the maximum likelihood estimate (MLE) for π^* can be determined independently of A^* using the simple algorithmic approach described in Section 1.1. On the other hand, the MLE for A^* is obtained by solving

$$\min_{A \in \mathbb{R}^{d\times d}} \sum_{t=1}^T \|x_t - Ax_{t-1}\|_2^2,$$

which corresponds to the problem of estimating the system matrix from a single observed realization of a time series. This is a well-known system identification *problem, which has been extensively studied in the literature (see e.g., Shirani Faradonbeh et al. (2018); Simchowitz et al. (2018); Sarkar & Rakhlin (2019); Jedra & Proutiere (2020a)). To obtain recovery guarantees for that problem, A^* is commonly assumed to be stable, i.e., its spectral radius lies within the unit circle. This is ensured by the stricter condition $\|A^*\|_2 < 1$. Such a condition will also be needed in our analysis later on, for the estimation of π^*.*

[2] We keep the notation $f_{A,\pi,\sigma}$ for the density of $x_1, x'_1, \ldots, x_T, x'_T$.

Notation. In order to rewrite (2.1) in a more convenient form, the following notation will be useful.

- The shift operator $S \in \{0,1\}^{T\times T}$,
$$S := \begin{bmatrix} 0 & I_{T-1} \\ 0 & 0 \end{bmatrix}.$$
- The data matrices $X, X^{\#} \in \mathbb{R}^{d\times T}$, where
$$X = [x_1\ x_2\ \dots\ x_T] \ \text{ and } \ X^{\#} = [x_1^{\#}\ x_2^{\#}\ \dots\ x_T^{\#}].$$
- The permutation matrix $\Pi \in \mathcal{P}_T$ corresponding to the map π,
$$\Pi = \begin{bmatrix} e_{\pi(1)}^{\top} \\ e_{\pi(2)}^{\top} \\ \vdots \\ e_{\pi(T)}^{\top} \end{bmatrix},$$
where we recall that e_t corresponds to the t-th canonical (column) vector in $\mathbb{R}^T$.

With this notation, (2.1) can be rewritten as
$$\min_{\Pi\in\mathcal{P}_T, A\in\mathbb{R}^{d\times d}} \left\{ \|X - AXS\|_F^2 + \frac{1}{\sigma^2}\|X^{\#}\Pi - AX^{\#}\Pi S - (X - AXS)\|_F^2 \right\}. \tag{2.2}$$
The following points are useful to note.

1. For a given A, observe from (2.2) that the optimal Π is given by
$$\widehat{\Pi}_{\text{MLE}}(A) \in \operatorname*{argmin}_{\Pi\in\mathcal{P}_T} \|X^{\#}\Pi - AX^{\#}\Pi S - (X - AXS)\|_F^2. \tag{2.3}$$
This formulation defines an optimization problem with a convex quadratic objective function in Π (as it corresponds to the squared norm of a linear function), subject to permutation constraints. Note that it does not require the knowledge of σ. While problem (2.3) is combinatorial in nature, it is unclear whether it is NP-hard. Later on in Section 4.1, we will consider solving its convex relaxations (see Algorithm 1) which can be solved efficiently and also perform well empirically. Note that if $A = 0$ then (2.3) reduces to the linear assignment problem (studied in Kunisky & Niles-Weed (2022)) which is efficiently solvable.
2. For a given Π, observe from (2.2) that the optimal A is given by
$$\widehat{A}_{\text{MLE}}(\Pi) \in \operatorname*{argmin}_{A\in\mathbb{R}^{d\times d}} \left\{ \|X - AXS\|_F^2 + \frac{1}{\sigma^2}\|X^{\#}\Pi - AX^{\#}\Pi S - (X - AXS)\|_F^2 \right\}, \tag{2.4}$$
which can be solved in closed form. This is proven in the following lemma, whose proof is deferred to Appendix A. Note that (2.4) requires knowledge of σ.

Lemma 2. *For a given $\Pi \in \mathcal{P}_T$, define $\widehat{A}_{\text{MLE}}(\Pi)$ as in (2.4). Then, it holds,*
$$\widehat{A}_{\text{MLE}}(\Pi) = \left[X(XS)^{\top} + \frac{1}{\sigma^2}\left(X^{\#}\Pi - X\right)\left(X^{\#}\Pi S - XS\right)^{\top} \right] \times \left[(XS)(XS)^{\top} + \frac{1}{\sigma^2}(X^{\#}\Pi S - XS)(X^{\#}\Pi S - XS)^{\top} \right]^{\dagger}.$$

As discussed in Section 4.2, one can formulate an efficient alternating minimization heuristic (see Algorithm 2) for solving (2.2), by iteratively solving (i) an efficiently solvable relaxation of (2.3) to first obtain $\widehat{\Pi}$, and (ii) then using $\widehat{\Pi}$ in (2.4) to obtain $\widehat{A}_{\text{MLE}}(\widehat{\Pi})$.

Remark 3 (Unordered base time-series)**.** *Suppose that the temporal ordering of the base time series* $(x_t)_{t=1}^T$ *was unknown, which simply means that we are presented with a sequence* $(y_t)_{t=1}^T$ *where* $y_t = x_{\pi_1^*(t)}$ *for an unknown permutation* π_1^**. Then, the unknown parameters are* $\pi_1^*, \pi^* \in \mathcal{S}_T$ *and* A^**. Denote* $Y = [y_1 \cdots y_T]$*, so that* $Y = X(\Pi_1^*)^\top$*, with* $\Pi_1^* \in \mathcal{P}_T$ *the permutation matrix corresponding to the map* π_1^**. Then it is easy to verify that the MLE (2.2) changes to*

$$\min_{\substack{\Pi_1, \Pi \in \mathcal{P}_T \\ A \in \mathbb{R}^{d\times d}}} \left\{ \|Y\Pi_1 - AY\Pi_1 S\|_F^2 + \frac{1}{\sigma^2}\|X^\#\Pi - AX^\#\Pi S - (Y\Pi_1 - AY\Pi_1 S)\|_F^2 \right\}. \tag{2.5}$$

As before, we can attempt solving (2.5) by alternating between updates to A *and* (Π, Π_1)*; note that for a given* A *the objective in (2.5) is convex in* (Π, Π_1)*. Clearly, the estimated maps* $\widehat{\pi}_1$*,* $\widehat{\pi}$ *can then be used to recover the correspondence between* $(y_t)_{t=1}^T$ *and* $(x_t^\#)_{t=1}^T$*.*

2.2 Linear assignment approach for estimating π^*

Given the difficulties of provably solving (and analyzing) the MLE given in (2.2), we consider estimating Π^* via linear assignment (LA). This consists in solving the (linear) optimization problem

$$\widehat{\Pi}_{\mathrm{LA}} \in \operatorname*{argmax}_{\Pi \in \mathcal{P}_T} \langle X^\#\Pi, X\rangle_F = \langle \Pi, \underbrace{(X^\#)^\top X}_{=:W}\rangle_F, \tag{2.6}$$

which does not require A^* or σ to be known. Recall that even if A^* was known, finding $\widehat{\Pi}_{\mathrm{MLE}}(A^*)$ would involve solving the quadratic problem (2.3), which in general appears to be a hard problem, as discussed earlier. In contrast, the problem (2.6) can be efficiently solved, using the Hungarian method Kuhn (1955), for instance. The matrix W can be viewed as a similarity matrix, whose entry (t, t'), defined as $\langle x_t^\#, x_{t'}\rangle$, represents the similarity between $x_t^\#$ and $x_{t'}$.

As remarked earlier, (2.6) was recently used in the problem of matching point clouds (see for instance Kunisky & Niles-Weed (2022); da Rocha Schwengber & Oliveira (2024)) which corresponds to our setup with $A^* = 0$. In particular, (2.6) is the MLE for π^* in that case for a fixed (not necessarily known) σ. As we will see below, this method is also meaningful in our more general setup where $A^* \neq 0$ necessarily. This is not obvious as the temporal correlation in our problem distinguishes it from the uncorrelated (i.e., $A^* = 0$) case. Also note that (2.6) does not require knowledge of A^*.

In that sense, our main guiding question is

> *"Under what conditions on* A^* *and* σ *can guarantees be established for the linear assignment estimator* $\widehat{\Pi}$*, defined in (2.6), to solve the VAR permutation regression problem?"*

We now give our main result regarding the recovery guarantees of the LA estimator. Following Kunisky & Niles-Weed (2022), we distinguish three regimes of recovery, ordered from stronger to weaker: perfect recovery, constant error, and sublinear error. For the upper bounds presented here, the applicable regime depends on the assumptions imposed on the noise level σ.

Remark 4. *Continuing Remark 3, note that LA is agnostic to the temporal nature of the data in terms of its formulation – it simply finds a correspondence between the columns of* X *and* $X^\#$*. This means that LA run on the matrices* Y *and* $X^\#$ *(where* Y *is an unknown column-shuffling of* X*) would find a matching between the columns of* Y *and* $X^\#$*. It will not recover the underlying temporal ordering for the respective time-series (unless of course* π_1^* *is known to be identity).*

Theorem 1. *Let* $s_0 := 2^{1/d}$*, and let* $A^* \in \mathbb{R}^{d\times d}$ *be such that* $\|A^*\|_2 < 1$*. Let* $X, X^\# \in \mathbb{R}^{d\times T}$ *be observed data from the* ***CVAR***$(1, d, T; A^*, \pi^*, \sigma)$ *model, and denote* $\widehat{\pi}_{\mathrm{LA}}$*, the permutation map corresponding to the linear assignment estimator defined in (2.6). Define,*

$$\mathcal{E} := \{t \in [T] : \widehat{\pi}_{\mathrm{LA}}(t) \neq \pi^*(t)\}$$

the set of mismatched indices by $\widehat{\pi}_{\mathrm{LA}}$*. Then the following three statements hold.*

(i) If

$$\sigma^2 \leq \frac{(1-\|A^*\|_2)^5}{4(s_0^{\omega(1)}T^{4/d}-1)}.$$

Then we have $\mathbb{E}[|\mathcal{E}|] \to 0$*, when* $T \to \infty$*. In particular,* $|\mathcal{E}| = 0$ *with high probability.*

(ii) If

$$\sigma^2 \leq \frac{(1-\|A^*\|_2)^5}{4(s_0^{O(1)}T^{4/d}-1)}.$$

Then, $\mathbb{E}[|\mathcal{E}|] = O(1)$*. In particular, for any function* $f(T) = w(1)$*, we have* $|\mathcal{E}| \leq f(T)$ *with high probability.*

(iii) If

$$\sigma^2 \leq \frac{(1-\|A^*\|_2)^5}{4(s_0^{\omega(1)}T^{2/d}-1)}$$

Then,

$$\mathbb{E}[|\mathcal{E}|] = \mathcal{O}\left(\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)^{-\frac{d}{2}} T^2\right).$$

Interestingly, the theorem recovers the results of Kunisky & Niles-Weed (2022), up to a multiplicative factor proportional to $(1-\left\|A^*\right\|_2)^5$. In particular, when $\|A^*\|_2 = 0$, parts (i) and (ii) yield the same noise condition as in their work, except for a factor of 4 in the denominator. In the case of part (iii), our condition is also of the same order as theirs.

We present the proof of Theorem 1 in the next section. As we will see, the time-series case introduces additional technical challenges, some of which lead to results that may be of independent interest.

Remark 5 (On the factor $(1-\|A^*\|_2)^5$)**.** *Theorem 1 requires* $\|A^*\|_2 < 1$*, which is stronger than requiring* $\rho(A^*) < 1$*. Such stability assumptions on* A^* *are common for parameter estimation problems for VAR models. While recent results have tackled this under weaker conditions – such as that of marginal stability wherein* $\rho(A^*) \leq 1$ *(see Simchowitz et al. (2018); Sarkar & Rakhlin (2019)), or even unstable systems (e.g., Shirani Faradonbeh et al. (2018); Sarkar & Rakhlin (2019)) – a number of prior results consist of estimation error bounds which tend to worsen as* $\|A^*\|_2$ *approaches one (e.g., Han et al. (2015); Loh & Wainwright (2012)). In that sense, we believe that the condition* $\|A^*\|_2 < 1$ *needed in our results is an artefact of the proof, and can be weakened as well. Establishing this is an interesting direction for future work.*

Remark 6 (Gaussian assumption on $(\xi_t)_{t=1}^T$)**.** *The assumption that* $\xi_t \overset{i.i.d}{\sim} \mathcal{N}(0, I_d)$ *is mainly for convenience for the theoretical analysis in Section 3, and also to be able to formulate the MLE as described already. This assumption was also made in Kunisky & Niles-Weed (2022) (for the i.i.d setting where* $A^* = 0$*), but was relaxed in da Rocha Schwengber & Oliveira (2024) to the more general class of sub-Gaussian distributions. We believe it should be similarly possible to extend our setup to the sub-Gaussian setting.*

3 Analysis for the linear assignment estimator

In this section, we present the key techniques used to prove Theorem 1. Section 3.1 introduces the first moment method and the counting of augmenting paths, techniques commonly used in matching problems (Kunisky & Niles-Weed, 2022; da Rocha Schwengber & Oliveira, 2024).

3.1 Augmenting cycles in the analysis of (2.6)

The technique used in Kunisky & Niles-Weed (2022); da Rocha Schwengber & Oliveira (2024) to obtain upper bounds on the error of the LA estimator $\widehat{\Pi}$ consists of counting the number of augmenting cycles. A

cycle C_t of length $t \leq T$ is a sequence $C_t = (i_1, \ldots, i_t)$ consisting of distinct indices $i_1, \ldots, i_t \in [T]$. We say that C_t is an augmenting t-cycle, if and only if (recall $W = (X^{\#})^\top X$)

$$\sum_{k=1}^{t} W_{i_k i_{k+1}} \geq \sum_{k=1}^{t} W_{i_k i_k} \,, \tag{3.1}$$

where $i_{t+1} := i_1$. To see the importance of the augmenting cycles for upper bounding[3] $|\mathcal{E}|$, assume without loss of generality[4] $\pi^* = \mathrm{id}$. It is easy to see that the elements of $\mathcal{E}$ belong to a union of disjoint augmenting t-cycles of $\widehat{\pi}$, for different $t \in \{2, \ldots, T\}$. This then implies

$$|\mathcal{E}| \leq \sum_{t=2}^{T} t N_t \quad \text{where} \quad N_t := \sum_{(i_1,\ldots,i_t)\text{is } t\text{-cycle}} \mathbb{1}_{\{(i_1,\ldots,i_t) \text{ is augmenting}\}}$$

represents the number of augmenting t-cycles. To guarantee perfect recovery, we will rely on the *first moment method*, which involves bounding the expected value of the error. For that, we have

$$\mathbb{E}[|\mathcal{E}|] \leq \sum_{t=2}^{T} t\mathbb{E}[N_t] = \sum_{t=2}^{T} t \sum_{(i_1,\ldots,i_t)\text{is } t\text{-cycle}} \mathbb{P}\big((i_1, \ldots, i_t) \text{ is augmenting}\big). \tag{3.2}$$

Therefore, a fundamental step is to bound the probability that a cycle C_t is augmenting. Section 3.2 analyzes this step for the case $t = 2$ to provide intuition for the general case in Section 3.3.

3.2 Warm-up: augmenting 2-cycles

We begin by tackling the case of augmenting 2-cycles, since it already contains the core of the argument. According to (3.1), a 2-cycle $C = (a\,, b)$, for a given pair $a, b \in [T]$, is augmenting if

$$W_{ab} + W_{ba} \geq W_{aa} + W_{bb}, \tag{3.3}$$

or equivalently

$$(x_a^{\#})^\top x_b + (x_b^{\#})^\top x_a \geq (x_a^{\#})^\top + (x_b^{\#})^\top x_b. \tag{3.4}$$

Recall that

$$x_a^{\#} = A^* x_{a-1}^{\#} + \xi_a + \sigma\tilde{\xi}_a \quad \text{and} \quad x_a = A^* x_{a-1} + \xi_a,$$

where we use the convention $x_0^{\#} = x_0 = 0$. Hence,

$$\begin{aligned}
(x_a^{\#})^\top x_b &= \left(A^* x_{a-1}^{\#} + \xi_a + \sigma\tilde{\xi}_a\right)^\top (A^* x_{b-1} + \xi_b) \\
(x_a^{\#})^\top x_a &= \left(A^* x_{a-1}^{\#} + \xi_a + \sigma\tilde{\xi}_a\right)^\top (A^* x_{a-1} + \xi_a).
\end{aligned}$$

It is easy to see that (3.4) it is equivalent to

$$\begin{aligned}
\|\xi_a - \xi_b\|_2^2 \leq\ & \sigma\langle\tilde{\xi}_a - \tilde{\xi}_b, \xi_b - \xi_a\rangle + \langle A^* x_{a-1}^{\#} - A^* x_{b-1}^{\#}, \xi_b - \xi_a\rangle \\
& + \langle A^* x_{a-1}^{\#} - A^* x_{b-1}^{\#}, A^* x_{b-1} - A_{a-1}^*\rangle \\
& + \langle \xi_b + \sigma\tilde{\xi}_b, A^* x_{a-1} - A^* x_{b-1}\rangle + \langle \xi_a + \sigma\tilde{\xi}_a, A^* x_{b-1} - A^* x_{a-1}\rangle.
\end{aligned} \tag{3.5}$$

Notice that when $A^* = 0$, we obtain $\|\xi_a - \xi_b\|_2^2 \leq \sigma\langle\tilde{\xi}_a - \tilde{\xi}_b, \xi_b - \xi_a\rangle$, which appears in the upper bound argument in Kunisky & Niles-Weed (2022); da Rocha Schwengber & Oliveira (2024). In comparison, here

[3] Recall the definition of $\mathcal{E}$ as the set of mismatched indices in Theorem 1, and note that $|\mathcal{E}|$ corresponds to the Hamming distance between $\widehat{\pi}$ and π^*.

[4] This is common in the analysis of matching problems.

we have to deal with more complicated expressions. Since $x_a^{\#} = x_a + \sigma \tilde{x}_a$, we have (after some algebra) that (3.4) is equivalent to

$$\sigma \langle \underbrace{A^*(\tilde{x}_{a-1} - \tilde{x}_{b-1}) + (\tilde{\xi}_a - \tilde{\xi}_b)}_{=:\tilde{y}_{ab}}, \underbrace{A^*(x_{b-1} - x_{a-1}) + \xi_b - \xi_a}_{=:y_{ab}} \rangle \geq \|A^*(x_{a-1} - x_{b-1}) - (\xi_b - \xi_a)\|_2^2. \tag{3.6}$$

Assuming w.l.o.g that $a > b$, we have that conditioned on $\xi_1, \ldots, \xi_a$ (so that y_{ab} is fixed) the left hand side of (3.6) is a Gaussian random variable. The following lemma more specifically characterizes this distribution. Its proof can be found in Appendix B.

Lemma 3. *Let $(x_t)_{t\in[a]}$,$(\tilde{x}_t)_{t\in[a]}$, $(\xi_t)_{t\in[a]}$ and $(\tilde{\xi}_t)_{t\in[a]}$ be as in **CVAR**$(1, d, T; A^*, \pi^*, \sigma)$. For $a > b$, define the variables*

$$y_{ab} := A^*(x_{b-1} - x_{a-1}) + (\xi_b - \xi_a),$$
$$\tilde{y}_{ab} := A^*(\tilde{x}_{a-1} - \tilde{x}_{b-1}) + (\tilde{\xi}_a - \tilde{\xi}_b).$$

Then, conditional on $(\xi_i)_{i=1}^a$, we have $\sigma\langle \tilde{y}_{ab}, y_{ab}\rangle \sim \mathcal{N}(0, \sigma^2\left(\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2\right))$, where

$$\tilde{\sigma}_1^2 = y_{ab}^\top \left(\sum_{i=0}^{b-1} (A^*)^i \left[(A^*)^{a-b} - I_d\right] \left[(A^*)^{a-b} - I_d\right]^\top (A^{*\top})^i \right) y_{ab}$$
$$\tilde{\sigma}_2^2 = y_{ab}^\top \left(\sum_{i=1}^{a-b} (A^*)^{a-b-i} (A^{*\top})^{a-b-i} \right) y_{ab}.$$

If $\|A^*\|_2 < 1$, we further have that

$$\sigma^2(\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2) \leq \frac{5\sigma^2 \|y_{ab}\|_2^2}{1 - \|A^*\|_2^2}. \tag{3.7}$$

From (3.6) and Lemma 3 we have that

$$\begin{aligned}
&\mathbb{P}\left(C = (a,b) \text{ is augmenting}\right) \\
&= \mathbb{E}_{\xi_1,\ldots,\xi_a}\left[\mathbb{P}\left(\sigma\langle \tilde{y}_{ab}, y_{ab}\rangle \geq \|y_{ab}\|_2^2 \,\middle|\, (\xi_i)_{i=1}^a\right)\right] \\
&= \mathbb{E}_{\xi_1,\ldots,\xi_a}\left[\mathbb{P}\left(\underbrace{\frac{\langle \tilde{y}_{ab}, y_{ab}\rangle}{\sqrt{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2}}}_{=:g} \geq \frac{\|y_{ab}\|_2^2}{\sigma\sqrt{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2}} \,\middle|\, (\xi_i)_{i=1}^a\right)\right] \\
&\leq \mathbb{E}_{\xi_1,\ldots,\xi_a}\left[\mathbb{P}\left(g \geq \sqrt{\frac{1 - \|A^*\|_2^2}{5\sigma^2}}\|y_{ab}\|_2^2 \,\middle|\, (\xi_i)_{i=1}^a\right)\right] && \text{(using (3.7))} \\
&\leq \mathbb{E}_{\xi_1,\ldots,\xi_a}\left[\exp\left(-\frac{(1 - \|A^*\|_2^2)}{10\sigma^2}\|y_{ab}\|_2^2\right)\right]. && \text{(conditional on } (\xi_i)_{i=1}^a,\ g \sim \mathcal{N}(0,1) \text{ by Lemma 3)}
\end{aligned}$$

Denoting $\xi_{1:a} := (\xi_1^\top, \xi_2^\top, \ldots, \xi_a^\top)^\top$, we have that $\|y_{ab}\|_2^2 = \xi_{1:a}^\top B B^\top \xi_{1:a}$ where

$$B := \begin{bmatrix} \left((A^*)^{a-b} - I_d\right)^\top (A^{*\top})^{b-1} \\ \vdots \\ \left((A^*)^{a-b} - I_d\right)^\top (A^{*\top}) \\ \left((A^*)^{a-b} - I_d\right)^\top \\ (A^*)^{a-b-1} \\ \vdots \\ A^{*\top} \\ I_d \end{bmatrix}.$$

Thus,

$$\begin{aligned}\mathbb{E}_{\xi_{1:a}}\left[\exp\left(-\frac{(1-\|A^*\|_2^2)}{10\sigma^2}\xi_{1:a}^\top BB^\top\xi_{1:a}\right)\right] &= \frac{1}{(2\pi)^{\frac{ad}{2}}}\int_{\mathbb{R}^{ad}}\exp\Big(-\frac{\xi^\top\xi}{2}\Big)\exp\left(-\underbrace{\frac{(1-\|A^*\|_2^2)}{10\sigma^2}}_{=:\alpha}\xi^\top BB^\top\xi\right)d\xi \\ &= \frac{1}{(2\pi)^{\frac{ad}{2}}}\int_{\mathbb{R}^{ad}}\exp\left(-\frac{\xi^\top}{2}(2\alpha BB^\top+I_{ad})\xi\right)d\xi \\ &= \det(2\alpha BB^\top+I_{ad})^{-\frac{1}{2}}\end{aligned}$$

and we obtain the bound

$$\mathbb{P}\left(C=(a,b)\text{ is augmenting}\right)\leq\det\left(\frac{(1-\|A^*\|_2^2)}{5\sigma^2}BB^\top+I_{ad}\right)^{-\frac{1}{2}}. \tag{3.8}$$

3.3 General augmenting t-cycles

The analysis of the general case $t\geq 2$ is similar to that of $t=2$, but involves cumbersome calculations. The following proposition, whose proof can be found in Section 3.4 summarizes our findings in this case. For $t=2$, note that the bound stated in Proposition 1 is not necessarily always worse than that in (3.8).

Proposition 1. *For $t\geq 2$, let $C_t=(i_1,\dots,i_t)$ be a t-cycle with i_1 the largest amongst the i_k's (w.lo.g). Then if $\|A^*\|_2<1$, it holds that*

$$\mathbb{P}(C_t\text{ is augmenting})\leq\det\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}L+I_{i_1d}\right)^{-\frac{1}{2}}, \tag{3.9}$$

where

$$L:=\sum_{k=1}^{t}B(\alpha_k,\beta_k)B^\top(\alpha_k,\beta_k),$$

with $B^\top(\alpha_k,\beta_k)\in\mathbb{R}^{d\times(i_1d)}$ defined as the matrix

$$\begin{bmatrix}(A^*)^{\alpha_k-1}-(A^*)^{\beta_k-1} & \dots & (A^*)^{\alpha_k-\beta_k}-I_d & (A^*)^{\alpha_k-\beta_k-1} & \dots & A^* & I_d & 0 & \dots & 0\end{bmatrix}.$$

Moreover, $\alpha_t:=i_1$, $\beta_t:=i_t$, and

$$\alpha_k:=\max\{i_k,i_{k+1}\},\ \beta_k:=\min\{i_k,i_{k+1}\},\ \text{for } 1\leq k\leq t-1.$$

Equipped with Proposition 1 and equation (3.2), it remains only to bound the right-hand side of (3.9) in order to obtain an upper bound on the expected error $\mathbb{E}[|\mathcal{E}|]$. The next proposition establishes the required estimate.

Proposition 2. *Under the assumptions and notation of Proposition 1, we have that*

$$\det\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}L+I_{i_1d}\right)^{-\frac{1}{2}}\leq\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)^{-(t-1)\frac{d}{2}}. \tag{3.10}$$

The proof of Proposition 2 is outlined in Section 3.5. Theorem 1 is then proved by combining Propositions 1 and 2 with (3.2). The reader is referred to Appendix D for a complete derivation.

Remark 7 (Comparison with the analysis in Kunisky & Niles-Weed (2022))**.** *Although our analysis for the LA estimator (in the VAR setting) is inspired by Kunisky & Niles-Weed (2022), it is useful to elaborate on the technical challenges arising here as compared to Kunisky & Niles-Weed (2022).*

1. *The proof of Proposition 1 follows, in spirit, along the same lines as that of (Kunisky & Niles-Weed, 2022, Prop. 3.1). However the underlying calculations in our case are considerably more involved and tedious, due to the VAR structure on the model.*

2. *Secondly, Proposition 2 has a considerably easy analogue in Kunisky & Niles-Weed (2022) – indeed, if $A^* = 0$, then the spectrum of L is known in closed form and bounding the* det *term is straightforward. In our setting, however, the spectrum of L is not known in closed form, and therefore, bounding the* det *term requires additional technical details. In particular, our argument relies only on information about the pseudo determinant of L.*

3. *Finally, once Propositions 2 and 1 are established, the remaining calculations needed to obtain Theorem 1 are essentially along the same lines as in Kunisky & Niles-Weed (2022).*

3.4 Proof of Proposition 1

To ease notation, we sometimes use A instead of A^* in this proof; especially during a long sequence of calculations. Suppose $t \geq 2$ and $C_t = (i_1, i_2, \ldots, i_t)$, where we assume w.l.o.g that i_1 is the largest among the indices in C_t. In this case, C_t is augmenting, if and only if

$$x_{i_t}^{\#\top} x_{i_1} + \sum_{k=1}^{t-1} x_{i_k}^{\#\top} x_{i_{k+1}} \geq \sum_{k=1}^{t} x_{i_k}^{\#\top} x_{i_k}. \tag{3.11}$$

Recalling that

$$\begin{cases} x_{i_k}^{\#} = A x_{i_k-1}^{\#} + \xi_{i_k} + \sigma \tilde{\xi}_{i_k} \\ x_{i_k} = A x_{i_k-1} + \xi_{i_k}, \end{cases}$$

we have

$$x_{i_k}^{\#\top} x_{i_{k+1}} = \langle A x_{i_{k-1}}^{\#}, A x_{i_{k+1}-1} \rangle + \langle A x_{i_{k-1}}^{\#}, \xi_{i_{k+1}} \rangle + \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, A x_{i_{k+1}-1} \rangle + \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, \xi_{i_{k+1}} \rangle.$$

Then (3.11) is equivalent to

$$\begin{aligned}
&\langle A x_{i_t-1}^{\#}, A x_{i_1-1} \rangle + \langle A x_{i_t-1}^{\#}, \xi_{i_1} \rangle + \langle \xi_{i_t} + \sigma \tilde{\xi}_{i_t}, A x_{i_1-1} \rangle + \langle \xi_{i_t} + \sigma \tilde{\xi}_{i_t}, \xi_{i_1} \rangle \\
&+ \sum_{k=1}^{t-1} \langle A x_{i_{k-1}}^{\#}, A x_{i_{k+1}-1} \rangle + \sum_{k=1}^{t-1} \langle A x_{i_k-1}^{\#}, \xi_{i_{k+1}} \rangle + \sum_{k=1}^{t-1} \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, A x_{i_{k+1}-1} \rangle + \sum_{k=1}^{t-1} \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, \xi_{i_{k+1}} \rangle \\
&\geq \langle A x_{i_t-1}^{\#}, A x_{i_t-1} \rangle + \langle A x_{i_t-1}^{\#}, \xi_{i_t} \rangle + \langle \xi_{i_t} + \sigma \tilde{\xi}_{i_t}, A x_{i_t-1} \rangle + \langle \xi_{i_t} + \sigma \tilde{\xi}_{i_t}, \xi_{i_t} \rangle \\
&+ \sum_{k=1}^{t-1} \langle A x_{i_k-1}^{\#}, A x_{i_k-1} \rangle + \sum_{k=1}^{t-1} \langle A x_{i_k-1}^{\#}, \xi_{i_k} \rangle + \sum_{k=1}^{t-1} \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, A x_{i_k-1} \rangle + \sum_{k=1}^{t-1} \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, \xi_{i_k} \rangle.
\end{aligned}$$

The previous inequality is equivalent to

$$\begin{aligned}
&\langle A x_{i_t-1}^{\#}, A(x_{i_1-1} - x_{i_t-1}) \rangle + \sum_{k=1}^{t-1} \langle A x_{i_k-1}^{\#}, A(x_{i_{k+1}-1} - x_{i_k-1}) \rangle + \langle A x_{i_t-1}, \xi_{i_1} - \xi_{i_t} \rangle + \sum_{k=1}^{t-1} \langle A x_{i_k-1}^{\#}, \xi_{i_{k+1}} - \xi_{i_k} \rangle \\
&+ \langle \xi_{i_t} + \sigma \tilde{\xi}_{i_t}, A(x_{i_1-1} - x_{i_t-1}) \rangle + \sigma \langle \tilde{\xi}_{i_t}, \xi_{i_1} - \xi_{i_t} \rangle + \sigma \sum_{k=1}^{t-1} \langle \tilde{\xi}_{i_k}, \xi_{i_{k+1}} - \xi_{i_k} \rangle + \sum_{k=1}^{t-1} \langle \xi_{i_k} + \sigma \tilde{\xi}_{i_k}, A^*(x_{i_{k+1}-1} - x_{i_k-1}) \rangle \\
&\geq \frac{1}{2} \left(\|\xi_{i_t} - \xi_{i_1}\|_2^2 + \sum_{k=1}^{t-1} \|\xi_{i_k} - \xi_{i_{k+1}}\|_2^2 \right).
\end{aligned}$$

Writing $x_k^{\#} = x_k + \sigma \tilde{x}_k$, for $k \in [t]$, we obtain after some re-shuffling and simple algebra, that (3.11) is equivalent to

$$\begin{aligned}
&\sigma \langle A \tilde{x}_{i_t-1} + \tilde{\xi}_{i_t}, A(x_{i_1-1} - x_{i_t-1}) + \xi_{i_1} - \xi_{i_t} \rangle + \sigma \sum_{k=1}^{t-1} \langle A \tilde{x}_{i_k-1} + \tilde{\xi}_{i_k}, A(x_{i_{k+1}-1} - x_{i_k-1}) + \xi_{i_{k+1}} - \xi_{i_k} \rangle \\
&\geq \frac{1}{2} \left(\|A(x_{i_t-1} - x_{i_1-1}) - (\xi_{i_1} - \xi_{i_t})\|_2^2 + \sum_{k=1}^{t-1} \|A(x_{i_k-1} - x_{i_{k+1}-1}) - (\xi_{i_{k+1}} - \xi_{i_k})\|_2^2 \right).
\end{aligned}$$

Now, denoting

$$y_t := A(x_{i_t-1} - x_{i_1-1}) - (\xi_{i_1} - \xi_{i_t}),$$
$$y_k := A(x_{i_{k+1}-1} - x_{i_k-1}) - (\xi_{i_{k+1}} - \xi_{i_k}), \text{ for } k = 1, \ldots, t-1,$$

we have that (3.11) is equivalent to

$$\sigma \sum_{k=1}^{t-1} \langle A\tilde{x}_{i_k-1} + \tilde{\xi}_{i_k}, y_k \rangle \geq \frac{1}{2} \left(\|y_t\|_2^2 + \sum_{k=1}^{t-1} \|y_k\|_2^2 \right). \tag{3.12}$$

Note that, conditioned on $\xi_1, \ldots, \xi_{i_1}$ the LHS above is a zero mean real Gaussian. To find its variance, we define $\pi : \{2, \ldots, t\} \to \{2, \ldots, t\}$ be the permutation that defines the ordering $i_1 > i_{\pi(2)} > i_{\pi(3)} > \ldots > i_{\pi(t)}$, i.e., $i_{\pi(2)}$ and $i_{\pi(t)}$ are, respectively, the largest and the smallest indices among $i_2, i_3, \ldots, i_t$. With this, we have

$$\begin{aligned}
g := \sigma \sum_{k=1}^{t} \langle A\tilde{x}_{i_k-1} + \tilde{\xi}_k, y_k \rangle &= \sigma \sum_{k=1}^{t} \sum_{j=1}^{i_k} y_k^\top A^{i_k - j} \tilde{\xi}_j \\
&= \sigma \sum_{k=2}^{t} \sum_{j=1}^{i_{\pi(k)}} y_{\pi(k)}^\top A^{i_{\pi(k)} - j} \tilde{\xi}_j + \sigma \sum_{j=1}^{i_1} y_1^\top A^{i_1 - j} \tilde{\xi}_j \\
&= \sigma \Bigg[\sum_{j=1}^{i_{\pi(t)}} \left(y_1^\top A^{i_1 - j} + \sum_{k=0}^{t-2} y_{\pi(t-k)}^\top A^{i_{\pi(t-k)} - j} \right) \tilde{\xi}_j \\
&+ \sum_{l=0}^{t-3} \left(\sum_{j=i_{\pi(t-l)}+1}^{i_{\pi(t-l-1)}} \left(y_1^\top A^{i_1 - j} + \sum_{k=l+1}^{t-2} y_{\pi(t-k)}^\top A^{i_{\pi(t-k)} - j} \right) \tilde{\xi}_j \right) \\
&+ \sum_{k=1}^{i_1 - i_{\pi(2)}} y_1^\top A^{i_1 - i_{\pi(2)} - k} \tilde{\xi}_{i_{\pi(2)+k}} \Bigg].
\end{aligned}$$

Hence, $g \sim \mathcal{N}(0, \sigma^2 \bar{\sigma}^2)$, where

$$\begin{aligned}
\bar{\sigma}^2 = &\underbrace{\sum_{j=1}^{i_{\pi(t)}} \left\| y_1^\top A^{i_1 - j} + \sum_{k=0}^{t-2} y_{\pi(t-k)}^\top A^{i_{\pi(t-k)} - j} \right\|_2^2}_{=:\bar{\sigma}_1^2} \\
&+ \underbrace{\sum_{l=0}^{t-3} \sum_{j=i_{\pi(t-l)}+1}^{i_{\pi(t-l-1)}} \left\| y_1^\top A^{i_1 - j} + \sum_{k=l+1}^{t-2} y_{\pi(t-k)}^\top A^{i_{\pi(t-k)} - j} \right\|_2^2}_{=:\bar{\sigma}_2^2} \\
&+ \underbrace{\sum_{k=1}^{i_1 - i_{\pi(2)}} \left\| y_1^\top A^{i_1 - i_{\pi(2)} - k} \right\|_2^2}_{=:\bar{\sigma}_3^2}.
\end{aligned}$$

We will now bound $\bar{\sigma}^2$, with the help of the following result. From now onwards, we will consider $\pi(1) = 1$.

Lemma 4. *For $\bar{\sigma}$ as defined above, it holds that*

$$\bar{\sigma}^2 \leq \left(\frac{1}{1 - \|A^*\|_2} \right)^3 \sum_{k=1}^{t} \|y_k\|_2^2.$$

Proof. The proof consists in bounding the terms $\bar{\sigma}_1^2, \bar{\sigma}_2^2, \bar{\sigma}_3^2$ separately, in three stages.

Bound on $\bar{\sigma}_1^2$. We define

$$M_0 := \begin{bmatrix} A^{i_{\pi(1)}-i_{\pi(t)}} & A^{i_{\pi(2)}-i_{\pi(t)}} & \dots & \underbrace{A^{i_{\pi(t)}-i_{\pi(t)}}}_{=I_d} \end{bmatrix}.$$

With this, we can rewrite and bound $\bar{\sigma}_1^2$ as follows.

$$\begin{aligned}\bar{\sigma}_1^2 = \sum_{j=1}^{i_{\pi(t)}} \left\| A^{j-1} M_0 \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t)} \end{bmatrix} \right\|_2^2 &\leq \left(\sum_{j=1}^{i_\pi(t)} \|A\|_2^{2(j-1)} \left\| M_0 \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t)} \end{bmatrix} \right\|_2^2 \right) \\ &\leq \frac{1}{1-\|A\|_2^2} \left\| M_0 \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t)} \end{bmatrix} \right\|_2^2.\end{aligned}$$

Bound on $\bar{\sigma}_2^2$. For $l = 1, \dots, t-2$, let

$$M_l := \begin{bmatrix} A^{i_{\pi(1)}-i_{\pi(t-l)}} & A^{i_{\pi(2)}-i_{\pi(t-l)}} & \dots & A^{i_{\pi(t-l)}-i_{\pi(t-l)}} \end{bmatrix}.$$

Then we have that

$$\bar{\sigma}_{2,l}^2 := \sum_{k=0}^{i_{\pi(t-l)}-i_{\pi(t-l+1)}-1} \left\| A^k M_l \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t-l)} \end{bmatrix} \right\|_2^2 \leq \frac{1}{1-\|A\|_2^2} \left\| M_l \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t-l)} \end{bmatrix} \right\|_2^2.$$

Notice that $\sum_{l=1}^{t-2} \bar{\sigma}_{2,l}^2 = \bar{\sigma}_2^2$. Hence,

$$\bar{\sigma}_2^2 \leq \frac{1}{1-\|A\|_2^2} \sum_{l=1}^{t-2} \left\| M_l \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t-l)} \end{bmatrix} \right\|_2^2.$$

Bounding $\bar{\sigma}_3^2$. We have

$$\begin{aligned}\bar{\sigma}_3^2 = \sum_{j=1}^{i_{\pi(1)}-i_{\pi(2)}} \left\| y_1^\top A^{i_{\pi(1)}-i_{\pi(2)}-j} \right\|_2^2 &\leq \|y_1\|_2^2 \sum_{j=1}^{i_{\pi(1)}-i_{\pi(2)}} \|A\|_2^{2(i_{\pi(1)}-i_{\pi(2)}-j)} \\ &\leq \frac{\|y_1\|_2^2}{1-\|A\|_2^2} = \frac{1}{1-\|A\|_2^2} \| \underbrace{M_{t-1}}_{=I_d} y_1 \|_2^2.\end{aligned}$$

Summarizing, we have

$$\bar{\sigma}^2 \leq \frac{1}{1-\|A\|_2^2} \sum_{k=0}^{t-1} \left\| M_k \begin{bmatrix} y_{\pi(1)} \\ y_{\pi(2)} \\ \vdots \\ y_{\pi(t-1)} \\ y_{\pi(t)} \end{bmatrix} \right\|_2^2.$$

Now,

$$\sum_{k=0}^{t-1}\left\|M_k\begin{bmatrix} y_{\pi(1)}\\ y_{\pi(2)}\\ \vdots\\ y_{\pi(t-1)}\\ y_{\pi(t)}\end{bmatrix}\right\|_2^2 = \left\|\underbrace{\begin{bmatrix} A^{i_{\pi(1)}-i_{\pi(t)}} & A^{i_{\pi(2)}-i_{\pi(t)}} & \cdots & \cdots & I_d\\ A^{i_{\pi(1)}-i_{\pi(t-1)}} & A^{i_{\pi(2)}-i_{\pi(t-1)}} & \cdots & I_d & 0\\ \vdots & & \iddots & & 0\\ \vdots & \iddots & & & \vdots\\ I_d & 0 & \cdots & & 0\end{bmatrix}}_{:=\Gamma_1}\begin{bmatrix} y_{\pi(1)}\\ y_{\pi(2)}\\ \vdots\\ \vdots\\ y_{\pi(t)}\end{bmatrix}\right\|_2^2.$$

Denote $j_1 = i_{\pi(1)}, j_2 = i_{\pi(2)}, \ldots, j_t = i_{\pi(t)}$. Then Γ_1 is a submatrix of Γ_2, where

$$\Gamma_2 = \begin{bmatrix} A^{j_1-j_t} & A^{j_1-j_t-1} & \cdots & \cdots & I_d\\ A^{j_1-j_t-1} & A^{j_1-j_t-2} & \cdots & I_d & 0\\ \vdots & & \iddots & & 0\\ \vdots & I_d & & & \vdots\\ I_d & 0 & \cdots & & 0\end{bmatrix}.$$

Hence, $\|\Gamma_1\|_2 \le \|\Gamma_2\|_2$ holds. In addition, $\Gamma_2 = \Pi\Gamma_3$, where

$$\Gamma_3 = \begin{bmatrix} I_d & & & \\ A & I_d & & \\ \vdots & & \ddots & \\ A^{j_1-j_t} & \cdots & \cdots & I_d\end{bmatrix}, \quad \Pi = \begin{bmatrix} & & & I_d\\ & & I_d & \\ & \iddots & & \\ I_d & & & \end{bmatrix}.$$

Thus, we have

$$\|\Gamma_2\|_2 = \|\Gamma_3\|_2 \le \sup_{x\in[0,1]}\left\|\sum_{s=0}^{j_1-j_t} A^s e^{\iota 2\pi s x}\right\|_2 \le \sum_{s=0}^{j_1-j_t}\|A^s\|_2 \le \frac{1}{1-\|A\|_2},$$

where the first inequality follows from a known result for banded Toeplitz matrices (see (Jedra & Proutiere, 2020b, Lemma 5) which in turn uses results from (Bottcher & Silbermann, 1999, Chapter 6)). This means we have shown that

$$\|\Gamma_1\|_2 \le \frac{1}{1-\|A\|_2}$$

which implies

$$\begin{aligned}\bar{\sigma}^2 \le \frac{1}{1-\|A\|_2^2}\|\Gamma_1\|_2^2\sum_{k=1}^{t}\|y_{\pi(k)}\|_2^2 &\le \frac{1}{1-\|A\|_2^2}\left(\frac{1}{1-\|A\|_2}\right)^2\sum_{k=1}^{t}\|y_{\pi(k)}\|_2^2\\ &\le \left(\frac{1}{1-\|A\|_2}\right)^3\sum_{k=1}^{t}\|y_k\|_2^2.\end{aligned}$$

□

Recall from (3.12) and the definition of g that (3.11) is equivalent to $g \ge \frac{1}{2}\left(\sum_{k=1}^{t}\|y_k\|_2^2\right)$. Hence, we have

$$\begin{aligned}\mathbb{P}\left(g \ge \frac{1}{2}\left(\sum_{k=1}^{t}\|y_k\|_2^2\right)\right) &= \mathbb{P}\left(\frac{g}{\sigma\bar{\sigma}} \ge \frac{1}{2\sigma\bar{\sigma}}\sum_{k=1}^{t}\|y_k\|_2^2\right)\\ &\le \mathbb{P}\left(g' \ge \frac{1}{2\sigma}\left(\sum_{k=1}^{t}\|y_k\|_2^2\right)^{\frac{1}{2}}(1-\|A\|_2)^{\frac{3}{2}}\right) \qquad \left(\text{since } \tfrac{g}{\sigma\bar{\sigma}} =: g' \sim \mathcal{N}(0,1)\right)\\ &\le \exp\left(-\frac{1}{8\sigma^2}\sum_{k=1}^{t}\|y_k\|_2^2(1-\|A\|_2)^3\right).\end{aligned}$$

Thus,

$$\mathbb{P}\left(C_t = (i_1, \ldots, i_t) \text{ is augmenting}\right) \leq \mathbb{E}_{\xi_1,\ldots,\xi_{i_1}}\left[\exp\left(-\frac{(1-\|A\|_2)^3}{8\sigma^2}\sum_{k=1}^{t}\|y_k\|_2^2\right)\right].$$

Now, note that

$$\begin{aligned}
y_t &= A(x_{i_1-1} - x_{i_t-1}) + \xi_{i_1} - \xi_{i_t} \\
&= \left(A^{i_1-1} - A^{i_t-1}\right)\xi_1 + \ldots + \left(A^{i_1-i_t} - I\right)\xi_{i_t} + (A^{i_1-i_t-1}\xi_{i_t+1} + \cdots + A\xi_{i_1-1}) + \xi_{i_1}, \\
y_1 &= A\left(x_{i_2-1} - x_{i_1-1}\right) + \xi_{i_2} - \xi_{i_1} \\
&= -\left(\left(A^{i_1-1} - A^{i_2-1}\right)\xi_1 + \ldots + \left(A^{i_1-i_2} - I\right)\xi_{i_2} + \left(A^{i_1-i_2-1}\xi_{i_2+1} + \ldots + A\xi_{i_1-1}\right) + \xi_{i_1}\right).
\end{aligned}$$

For $k = 1, \ldots, t-1$, we can write y_k as follows. Define

$$\alpha_k := \max\{i_k, i_{k+1}\},\ \beta_k := \min\{i_k, i_{k+1}\} \text{ and } s_k = \begin{cases} 1 \text{ if } i_{k+1} > i_k \\ -1 \text{ if } i_{k+1} < i_k \end{cases}.$$

Then, y_k can be written as

$$\begin{aligned}
y_k &= A(x_{i_{k+1}-1} - x_{i_k-1}) + \xi_{i_{k+1}} - \xi_{i_k} \\
&= s_k\left(\left(A^{\alpha_k-1} - A^{\beta_k-1}\right)\xi_1 + \ldots + \left(A^{\alpha_k-\beta_k} - I\right)\xi_{\beta_k} + \left(A^{\alpha_k-\beta_k-1}\xi_{\beta_{k+1}} + \ldots + A\xi_{\alpha_k-1}\right) + \xi_{\alpha_k}\right).
\end{aligned}$$

In fact, defining $\alpha_t = i_1$, $\beta_t = i_t$ and $s_t = 1$, the above expression holds for $k \in [t]$. Let

$$\begin{aligned}
&B^\top(\alpha_k, \beta_k) \\
&:= \begin{bmatrix} (A^*)^{\alpha_k-1} - (A^*)^{\beta_k-1} & \ldots & (A^*)^{\alpha_k-\beta_k} - I & (A^*)^{\alpha_k-\beta_k-1} & \ldots & A^* & I & \underbrace{0, \quad 0, \quad \cdots \quad 0,}_{i_1-\alpha_k \text{ times}} \end{bmatrix},
\end{aligned}$$

and $\xi := (\xi_1^\top, \xi_2^\top, \cdots, \xi_{i_1}^\top)^\top \in \mathbb{R}^{i_1 d}$. Then, we can write $y_k = s_k B^\top(\alpha_k, \beta_k)\xi$ and

$$\implies \sum_{k=1}^{t}\|y_k\|_2^2 = \xi^\top\left(\sum_{k=1}^{t} B(\alpha_k, \beta_k)B^\top(\alpha_k, \beta_k)\right)\xi = \xi^\top L \xi$$

for L as defined in Proposition 1. Hence we have (analogous to the case $t = 2$),

$$\mathbb{E}\left[\exp\left(-\frac{1}{8\sigma^2}(\xi^\top L\xi)(1-\|A^*\|_2)^3\right)\right] = \det\left(\frac{1}{4\sigma^2}(1-\|A^*\|_2)^3 L + I_{i_1 d}\right)^{-\frac{1}{2}},$$

since $\xi \sim \mathcal{N}(0, I_{i_1 d})$.

3.5 Proof of Proposition 2

To prove Proposition 2, we need to control the spectrum of the matrix L. To this end, we begin by expressing L in a more convenient form. Observe that

$$B(\alpha_k, \beta_k) = P_{\alpha_k} - P_{\beta_k},$$

where each $P_r \in \mathbb{R}^{i_1 d \times d}$ is a block matrix (with i_1 vertically stacked blocks of size $d \times d$) defined as

$$P_r = \begin{bmatrix} (A^*)^{r-1} \\ (A^*)^{r-2} \\ \vdots \\ A^* \\ I_d \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{(i_1 d)\times d},$$

with the final $i_1 - r$ blocks consisting of zero matrices. Using this representation, we can write $L = MM^\top$, where

$$M := \begin{bmatrix} P_{\alpha_1} - P_{\beta_1} & P_{\alpha_2} - P_{\beta_2} & \cdots & P_{\alpha_t} - P_{\beta_t} \end{bmatrix}.$$

Assume, without loss of generality (see Remark 8), that the indices in the cycle $C_t = (i_1, i_2, \ldots, i_t)$ are ordered such that $i_1 > i_2 > \cdots > i_t$. In this case, we can write

$$M = P(D \otimes I_d),$$

where $P := \begin{bmatrix} P_{i_1} & P_{i_2} & \ldots & P_{i_t} \end{bmatrix}$ and D is the incidence matrix of the directed cycle $\overrightarrow{C_t}$, which corresponds to C_t with the orientation $i_1 \to i_2 \to \ldots \to i_t$ and $i_1 \to i_t$. With this representation, we obtain

$$L = MM^\top = P\left((DD^\top) \otimes I_d\right) P^\top = P(\underbrace{L_{C_t} \otimes I_d}_{=:\tilde{L}_{C_t}})P^\top,$$

where L_{C_t} denotes the graph Laplacian of the cycle C_t. This decomposition enables us to view L as a multiplicative perturbation of $\tilde{L}_{C_t}$, whose spectrum is explicitly known (it coincides with that of L_{C_t}, up to multiplicity). Indeed, it is well known (see e.g., Spielman (2025)) that the eigenvalues of C_t correspond to the (unordered) values

$$2\left(1 - \cos\left(\frac{2\pi(t-k)}{t}\right)\right); \quad k = 1, \ldots, t.$$

Denote

$$s := \operatorname{rank}(L),$$

that is, L has s non-zero eigenvalues (later, we will determine s). The following general lemma helps us to lower-bound the quantity

$$\log\det\left(\frac{(1 - \|A^*\|_2)^3}{4\sigma^2} L + I_{i_1 d}\right).$$

Lemma 5. *Let $Z \in \mathbb{R}^{q \times q}$ be symmetric and p.s.d with rank $p \le q$, and let $\lambda_1(Z), \ldots, \lambda_p(Z)$ be its non-zero eigenvalues. Then, for any $\gamma > 0$,*

$$\log \prod_{k=1}^{p} (\gamma \lambda_k(Z) + 1) \ge p \log\left(\gamma \left(\prod_{k=1}^{p} \lambda_k(Z)\right)^{\frac{1}{p}} + 1\right).$$

To prove this, we need the following elementary classic result, whose proof can be found in Appendix C.1.

Lemma 6 (Super-additivity of geometric means)**.** *Let $(a_k)_{1 \le k \le n}, (b_k)_{1 \le k \le n}$ be two sequences of non-negative real numbers, then*

$$\left(\prod_{k=1}^{n} a_k\right)^{1/n} + \left(\prod_{k=1}^{n} b_k\right)^{1/n} \le \left(\prod_{k=1}^{n} (a_k + b_k)\right)^{1/n}.$$

Lemma 5 follows directly from Lemma 6, by considering the sequences $(a_k)_{1 \le k \le p}$ and $(b_k)_{1 \le k \le p}$ defined as

$$a_k = \gamma \lambda_k(Z), \quad b_k = 1, \quad \text{for } k \in \{1, \ldots, p\},$$

and then applying the logarithm. Applying Lemma 5 with $Z = L$ (so $p = s$) and $\gamma = \frac{(1 - \|A^*\|_2)^3}{4\sigma^2}$, noting that $\lambda_k(L) \ge 0$, for all $k \in [i_1 d]$, we get

$$\log\det\left(\frac{(1 - \|A^*\|_2)^3}{4\sigma^2} L + I_{i_1 d}\right) \ge s \log\left(\frac{(1 - \|A^*\|_2)^3}{4\sigma^2} \left(\prod_{k=1}^{s} \lambda_k(L)\right)^{\frac{1}{s}} + 1\right). \tag{3.13}$$

To complete the proof of the lower bound, two ingredients are required. First, we must determine the value of s. Second, we need to obtain a lower bound for

$$\det{}^*(L) := \prod_{k=1}^{s} \lambda_k(L),$$

known as the *pseudo determinant* of L.

The following lemma gives the value of s and the rank of P (which will be needed later).

Lemma 7. *Let P and L as defined above, and $s = rank(L)$. Then $s = (t-1)d$ and $\operatorname{rank}(P) = td$.*

Proof. The claim $\operatorname{rank}(P) = td$ follows directly from the structure of the matrices $P_{i_1}, \ldots, P_{i_t}$. Indeed, for each $k \in [t]$, the matrix P_{i_k} has column rank d, since it is a tall matrix containing I_d as one of its blocks. Moreover, the I_d blocks corresponding to different P_{i_k} are disjoint.

To see $s = (t-1)d$, recall $L = MM^\top$ where $M = P(D \otimes I_d)$. Since $P \in \mathbb{R}^{(i_1 d)\times(td)}$ is full column-rank (as $i_1 \geq t$) and $D \otimes I_d \in \mathbb{R}^{(td)\times(t-1)d}$ is also full column-rank, hence it readily follows that $s = (t-1)d$. □

From Lemma 7 and (3.13), it follows that

$$\log\det\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}L + I_{i_1 d}\right) \geq (t-1)d\log\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}\left(\prod_{k=1}^{(t-1)d}\lambda_k(L)\right)^{\frac{1}{(t-1)d}} + 1\right). \tag{3.14}$$

From the previous lemma, it also follows that $\operatorname{rank}(P^\top P) = td$, which will be used in the lower bound for $\det^*(L)$.

Bound for the pseudo determinant of L. To obtain a bound on $\det^*(L)$, we require a sequence of auxiliary lemmas. The first, stated in a more general form, shows that the pseudo determinant of $L = P(L_{C_t} \otimes I_d)P^\top$ can be factorized into the product of two terms, each depending exclusively on L_{C_t} or P, respectively. The proof can be found in Appendix C.2.

Lemma 8. *Let $Z \in \mathbb{R}^{q\times q}$ be a symmetric p.s.d matrix of rank $p \leq q$, and let $W \in \mathbb{R}^{q'\times q}$, with $q' \geq q$. Assume that W has rank q. Then,*

$$\det{}^*(WZW^\top) = \det{}^*(Z)\det\left((WU_p)^\top(WU_p)\right),$$

where $U_p \in \mathbb{R}^{q\times p}$ is the matrix whose columns are the eigenvectors of Z associated to its non-zero eigenvalues.

Given Lemma 7, the assumptions of Lemma 8 are satisfied with $Z = (L_{C_t} \otimes I_d)$, $W = P$, $p = (t-1)d$, $q = td$, $q' = i_1 d$. Consequently, by these lemmas, we get

$$\begin{aligned}\det{}^*\left(P(L_{C_t} \otimes I_d)P^\top\right) &= \det{}^*(L_{C_t} \otimes I_d)\det\left((PU_{(t-1)d})^\top(PU_{(t-1)d})\right)\\ &= \det{}^*(L_{C_t})^d\det\left((PU_{(t-1)d})^\top(PU_{(t-1)d})\right)\end{aligned}$$

By the Kirchoff's matrix tree theorem (Godsil & Royle, 2001, Lemma 13.2.4), we know that

$$\frac{1}{t}\det{}^*(L_{C_t}) = |\{\text{spanning trees in } C_t\}|.$$

Since the number of spanning trees in the t-cycle graph is equal to t, we obtain

$$\det{}^*(L_{C_t}) = t^2.$$

From this, we obtain,

$$\det{}^*(L) = \det{}^*\left(P(L_{C_t} \otimes I_d)P^\top\right) = t^{2d}\det\left((PU_{(t-1)d})^\top(PU_{(t-1)d})\right). \tag{3.15}$$

The following lemma helps us bound the right-hand side of the previous expression. Its proof is deferred to Appendix C.3.

Lemma 9. *We have*

$$\det\left((PU_{(t-1)d})^\top (PU_{(t-1)d})\right) \geq \frac{\det\left(P^\top P\right)}{\prod_{k=1}^d \lambda_k\left(P^\top P\right)}.$$

By the previous lemma, in order to obtain a lower bound on $\det^*(PLP^\top)$, there are two ingredients left: a lower bound on $\det(P^\top P)$, and an upper bound on $\prod_{k=1}^d \lambda_k(P^\top P)$. For the latter, we will use that

$$\prod_{k=1}^d \lambda_k(P^\top P) \leq \lambda_1(P^\top P)^d,$$

for which an upper bound on $\lambda_1(P^\top P)$ suffices.

Determinant of $P^\top P$. The following lemma, whose proof is given in Appendix C.4, provides a formula of the explicit value of $\det\left(P^\top P\right)$, which could be of independent interest. The proof relies on decomposing the matrix $P^\top P$ into the product of square block lower triangular matrices and a Gram matrix whose determinant is straightforward to compute.

Lemma 10. *For any cycle $C_t = (i_1, \ldots, i_t)$, where $t \geq 2$, it holds for $P = \begin{bmatrix} P_{i_1} & P_{i_2} & \ldots & P_{i_t} \end{bmatrix}$ that*

$$\det\left(P^\top P\right) = \left(\prod_{k=1}^{t-1} \det\left(\sum_{l=0}^{i_k - i_{k-1} - 1} ((A^*)^l)^\top (A^*)^l\right)\right)\left(\det\left(\sum_{l=0}^{i_t - 1} ((A^*)^l)^\top (A^*)^l\right)\right) \geq 1.$$

From the previous lemma, we obtain $\det(P^\top P) \geq 1$. Although simple, this bound is already nontrivial from the definition of P. Moreover, it allows us to identify the cases of equality: indeed, the bound is tight when $\|A^*\|_2 = 0$ or when the cycle is $C_t = (t, t-1, \ldots, 2, 1)$.

Upper bound on $\lambda_1\left(P^\top P\right)$. We will use a generalization of Gershogorin's theorem for block matrices (see (Tretter, 2008, Theorem 1.13.1)) to bound the largest eigenvalue of $P^\top P$. The proof of the next result is deferred to Appendix C.5.

Lemma 11. *For P defined as above, it holds*

$$\lambda_1\left(P^\top P\right) \leq \frac{1}{(1 - \|A^*\|_2)^2}.$$

Putting it together. From (3.15), and Lemmas 9, 10 and 11 we deduce

$$\det{}^*(L) \geq t^{2d}(1 - \|A^*\|_2)^{2d},$$

which together with (3.14) gives

$$\begin{aligned}
\log\det\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2} L + I_{i_1 d}\right)^{-\frac{1}{2}} &\leq -(t-1)\frac{d}{2}\log\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}\left(t^{2d}(1-\|A^*\|_2)^{2d}\right)^{\frac{1}{(t-1)d}} + 1\right) \\
&\leq -(t-1)\frac{d}{2}\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2} + 1\right).
\end{aligned}$$

In the last line we used that $t^{\frac{2}{t-1}} \geq 1$ for $t \geq 2$, and that $(1 - \|A^*\|_2)^{\frac{2}{t-1}} \geq (1 - \|A^*\|_2)^2$, for $t \geq 2$. From this, Proposition 2 follows.

Remark 8 (About the assumption $i_1 > i_2 > \ldots > i_t$)**.** *In the proof of Proposition 2, we assumed $i_1 > i_2 > \ldots > i_t$ for convenience. More generally, for any t-cycle $C_t = (i_1, i_2, \ldots, i_t)$ with i_1 being the largest element, we always have $(\alpha_1, \beta_1) = (i_1, i_2)$ and $(\alpha_t, \beta_t) = (i_1, i_t)$. This allows us to consistently orient the edges $i_1 \to i_2$ and $i_1 \to i_t$. The orientation of the remaining edges—namely, $\{i_2, i_3\}, \{i_3, i_4\}, \ldots, \{i_{t-1}, i_t\}$—depends on the relative ordering of $i_2, i_3, \ldots, i_t$. For the purposes of the analysis, this implies that we can express $M = P(D\Pi \otimes I_d)$, where Π is a permutation matrix. Clearly, the matrix $L = MM^\top$ remains unchanged compared to the case where $i_1 > i_2 > \ldots > i_t$.*

4 Algorithms for solving the MLE

Given the hardness of solving the joint optimization problem (2.2), we now describe an alternating minimization based heuristic for solving (2.2) consisting of the following two main steps, which will be iteratively applied.

- **Step 1: Estimation of A^* for fixed Π.** For a fixed Π, we can estimate A^* by solving (2.4). This has a closed form solution outlined in Lemma 2.
- **Step 2: Estimation of Π^* for fixed A.** For a fixed A, we consider the problem (2.3). Since the objective function is convex, we propose several relaxations of the permutation constraints, resulting in convex optimization problems.

We start by introducing algorithms for estimating Π^*, for a fixed A, in Section 4.1. Later, in Section 4.2 we introduce an iterative algorithm based on alternating optimization. Recall that we assume σ is known.

4.1 Relaxed MLE strategy for estimating Π^* given A

Recall the estimator for estimating Π^*, for a fixed A, introduced in (2.3) – the MLE

$$\widehat{\Pi}_{\mathrm{MLE}}(A) \in \operatorname*{argmin}_{\Pi \in \mathcal{P}_T} \|X^{\#}\Pi - AX^{\#}\Pi S - (X - AXS)\|_F^2,$$

where, S denotes the shift operator introduced in Section 2.1. Since a general quadratic program with permutaion constraints is hard in the worse case, it is not clear whether $\hat{\Pi}_{\mathrm{MLE}}$ can be found efficiently. For this reason, we study a two-step strategy: (1) first solving a *relaxed MLE* problem, and (2) then rounding the relaxed solution to a valid permutation.

Relaxed MLE. The objective in (2.3) is convex (the square norm of a linear function in Π), but the set of constraints is discrete. We will consider a convex set $\mathcal{K} \subseteq \mathbb{R}^{T\times T}$, containing the set of permutations matrices, to obtain the relaxed convex optimization problem

$$\min_{\Pi \in \mathcal{K}} \underbrace{\|X^{\#}\Pi - AX^{\#}\Pi S - (X - AXS)\|_F^2}_{=:f(\Pi; X, X^{\#}, A, S)}. \tag{4.1}$$

In what follows, we will consider specific relaxations induced by particular choices of $\mathcal{K}$. Although these relaxations can be solved using general-purpose convex optimization methods (e.g., interior point methods), for the sake of efficiency we will implement gradient-based algorithms in the numerical experiments (see Section 5 for details).

1. **Hyperplane**. This corresponds to the choice (where $\mathbb{1}$ denotes the all ones vector)

   $$\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z \mathbb{1} = T\},$$

   which is a hyperplane constraint in the space $\mathbb{R}^{T^2}$ (with the obvious identification of a matrix with a vector). The main motivation for this relaxation comes from its success in the graph matching problem, as studied in Fan et al. (2022) (with an additional regularization term) where performance guarantees were obtained under specific planted matching models.

2. **Simplex**. A tighter relaxation than the hyperplane constraint is given by the simplex

   $$\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z \mathbb{1} = T,\ Z \geq 0\}.$$

   The addition of positivity constraints, in comparison to the simplex relaxation, has proven beneficial in the context of graph matching, as demonstrated recently in Araya & Tyagi (2025). There, this relaxation was shown to outperform the hyperplane formulation experimentally.

3. **Birkhoff polytope.** The tightest convex relaxation of (2.3) is given by the Birkhoff polytope (the set of doubly stochastic matrices)

$$\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z = \mathbb{1}^\top, Z\mathbb{1} = \mathbb{1}, Z \geq 0\}.$$

Indeed, by the Birkhoff–von Neumann theorem, the Birkhoff polytope is precisely the convex hull of the set of permutation matrices.

Rounding procedure. The solutions of the relaxed problem (4.1) are, in general, not guaranteed to be permutation matrices. Therefore, an additional rounding step is required. This can be achieved by solving a linear assignment problem using the relaxed solution (denoted by $\widehat{\Pi}_{\text{rel}}$) as the cost matrix; that is, we solve

$$\max_{\Pi\in\mathcal{P}_T} \langle \Pi, \widehat{\Pi}_{\text{rel}}\rangle_F. \tag{4.2}$$

As previously mentioned, the linear assignment problem can, in general, be solved by the Hungarian algorithm with cubic complexity.

Algorithm 1 Relaxed MLE + LA rounding (`RelaxMLE-Round`)

Require: Time series matrices $X, X^\# \in \mathbb{R}^{d\times T}$, system matrix $A \in \mathbb{R}^{d\times d}$, a convex set $\mathcal{K} \supseteq \mathcal{P}_T$

1: Set $S = \begin{bmatrix} 0 & I_{T-1} \\ 0 & 0 \end{bmatrix}$.

2: Solve the relaxed MLE problem in (4.1)

$$\widehat{\Pi}_{\text{rel}} = \operatorname*{argmin}_{\Pi\in\mathcal{K}} f(\Pi; X, X^\#, A, S).$$

3: Round with linear assignment

$$\widehat{\Pi} = \operatorname*{argmax}_{\Pi\in\mathcal{P}_T} \langle \Pi, \widehat{\Pi}_{\text{rel}}\rangle_F.$$

4: **return** $\widehat{\Pi}$

Remark 9 (Other relaxations)**.** *We choose to present Algorithm 1 in a general form, which allows, in principle, the use of other convex relaxations. Our implementation in Section 5 will be based on the choices of $\mathcal{K}$ discussed above, which are motivated by their success in other quadratic optimization problems over the set of permutations, as well as by the availability of efficient algorithms based on gradient descent, mirror descent, and the alternating direction method of multipliers (ADMM). It is worth noting that in related problems, such as graph matching, non-convex relaxations have also been considered. For example, in the classic work of Umeyama (1991), relaxing the graph matching problem to the set of orthogonal matrices yields a closed-form solution based on spectral information. In the case of (4.1), however, it is not clear that a closed-form solution exists when $\mathcal{K}$ is taken to be the set of orthogonal matrices.*

4.2 Iterative algorithm for estimating Π^*

To recover Π^*, we propose an iterative algorithm that alternates between **Steps 1 and 2** discussed at the beginning of this section. In Algorithm 2, we present the proposed procedure in the general case, where any of the strategies based on convex relaxation for estimating Π given A, as discussed in Section 4.1, can be used. We choose to write it in this abstract form, using `RelaxMLE-Round`, which serves as a subroutine for estimating Π given A.

Initialization. Algorithm 2 requires an initial estimate $\Pi^{(0)}$ of the optimal matching. Several strategies can be used for this initialization. One option is to set the initial permutation to the linear assignment estimator, i.e., $\Pi^{(0)} = \widehat{\Pi}_{\text{LA}}$. In this case, Algorithm 2 can be seen as an iterative refinement of the linear assignment solution. This raises the natural question of how much improvement the algorithm provides over linear assignment—a question we explore experimentally in Section 5. Another approach is to initialize randomly, for example by drawing $\Pi^{(0)}$ uniformly from the set $\mathcal{P}_T$, which we also investigate empirically. A central question is the extent to which the initial estimate of the permutation influences the outcome.

Algorithm 2 Alternating minimization method for matching VAR time series

Require: Time series matrices $X, X^{\#} \in \mathbb{R}^{d\times T}$, noise parameter $\sigma > 0$, initial estimate $\Pi^{(0)} \in \mathcal{P}_T$, max iterations K, a convex set $\mathcal{K} \supseteq \mathcal{P}_T$.

1: Set $S = \begin{bmatrix} 0 & I_{T-1} \\ 0 & 0 \end{bmatrix}$.

2: **for** $k = 1$ to K **do**

3: **A-update:** Set

$$A^{(k)} = \left[X(XS)^\top + \frac{1}{\sigma^2}\left(X^{\#}\Pi^{(k-1)} - X\right)\left(X^{\#}\Pi^{(k-1)}S - XS\right)^\top\right] \times \left[(XS)(XS)^\top + \frac{1}{\sigma^2}(X^{\#}\Pi^{(k-1)}S - XS)(X^{\#}\Pi^{(k-1)}S - XS)^\top\right]^\dagger$$

4: **Π-update:** Use sub-routine `RelaxMLE-Round` to obtain

$$\Pi^{(k)} := \texttt{RelaxMLE-Round}(X, X^{\#}, A^{(k)}, \mathcal{K})$$

5: **end for**

6: **return** $\Pi^{(K)}$

Remark 10 (Another strategy: estimate A^* first)**.** *Notice that Algorithm 2 does not require prior knowledge of the system matrix A^*, as it is updated iteratively (specifically in line 4). If enough data is available—that is, sufficiently long time series—an alternative strategy is to first estimate the system matrix using only the time series X, and then address the problem of estimating the permutation Π, without further updating the estimate of A^*. The intuition is that the estimate of A^* would not change significantly, since the amount of information about A^* contained in $(X, \widehat{X})$ should be asymptotically similar to that contained in X alone. On the other hand, an interesting question is wether one can consistently estimate the matching (or achieve non-trivial recovery), even if the system matrix cannot be consistently estimated.*

Remark 11 (Complexity)**.** *The computational complexity of the A-update step in Algorithm 2 is dominated by matrix multiplications involving matrices of sizes $d \times d$, $d \times T$, and $T \times T$. In the worst case, this yields a complexity of $O(\max\{d, T\}^\omega)$, where $\omega \leq 3$ (see e.g. Le Gall (2014)) denotes the matrix multiplication exponent. The complexity of the Π-update step depends on the specific optimization method employed. For instance, when using a gradient-based method for the hyperplane relaxation with $\log T$ iterations, the cost is approximately $O(\max\{d, T\}^\omega \log T)$, as the most expensive operation—gradient computation—also reduces to matrix multiplications. The rounding step incurs an additional $O(T^3)$ cost in the worst case when solved via the Hungarian algorithm (Kuhn, 1955). Therefore, the overall computational complexity of the alternating scheme is $O(\max\{d, T\}^\omega K \log T)$, where K denotes the number of outer iterations. This complexity can be reduced in practice under structural assumptions. For example, if some of the matrices involved are sparse, matrix multiplications become more efficient. Alternatively, replacing the linear assignment rounding step with a greedy method reduces the rounding cost to $O(T^2)$, see (Araya & Tyagi, 2025, Algorithm 1), for example.*

5 Numerical experiments

In this section, we empirically test the recovery algorithms discussed in Section 4, and the linear assignment estimator analyzed in Section 3. We focus on synthetic data generated under the **CVAR** model, introduced in Section 2. In Section 5.1, we provide some details about the implementation of the relaxed-MLE strategy (Algorithm 1) for different choices of $\mathcal{K}$. In Section 5.2, we test the relaxed MLE algorithms under the assumption that A^* is known. In Section 5.3, we assume that A^* is unknown and evaluate the alternating optimization approach of Algorithm 2, as well as the linear assignment (LA) estimator.

5.1 Algorithmic implementation details

To solve the relaxed MLE problem (4.1), we use different approaches depending on the convex set $\mathcal{K}$ considered. Although this problem could be solved with general purpose convex optimization algorithms, for the sake of efficiency, we focus on first-order methods as described below.

- **Hyperplane.** In the case $\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z \mathbb{1} = T\}$, we will use a *projected gradient descent* (PGD) strategy to optimize. More specifically, we consider a learning rate $\gamma_k > 0$, and an initial vector $\widehat{\Pi}^{(0)}_{\text{rel}} = \frac{1}{T} J_T$, where J_T is the $T \times T$ all-ones matrix. This choice can be considered agnostic, since J_T is at the same distance with respect to all the permutation matrices. Each iteration is described by
$$\widehat{\Pi}^{(k)}_{\text{rel}} = \mathscr{P}_{\mathcal{K}}\left(\widehat{\Pi}^{(k-1)}_{\text{rel}} - \gamma_k \nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}; X, X^{\#}, A, S\right)\right), \text{ for } k \geq 1,$$
where f is the relaxed MLE objective defined in (4.1), S is the shift matrix defined in Section 2, and $\mathscr{P}_{\mathcal{K}}$ is the Euclidean projection onto $\mathcal{K}$. We use an adaptive learning rate strategy, given by
$$\gamma_k = \gamma \frac{\log(k+1)}{\left(\left\|\nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}\right)\right\|_2 \vee 10^{-4}\right)\sqrt{k+1}}, \tag{5.1}$$
where $\gamma > 0$ is a user-specified constant, and the small term 10^{-4} is included arbitrarily to prevent numerical blow-up. This strategy is commonly used in practice; see (Beck, 2017, Chapter 8) for this and other related learning rate schemes.

- **Simplex.** When $\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z \mathbb{1} = T,\ Z \geq 0\}$, we use the *Entropic Mirror Descent* algorithm (see (Beck, 2017, Chapter 9)), which results in a multiplicative weights update algorithm. As in the case of the simplex, we use a learning rate $\gamma_k > 0$ and $\widehat{\Pi}^{(0)}_{\text{rel}} = \frac{1}{T} J_T$ as the initial point. The iterative step is, for each $k \geq 1$,
$$\widehat{\Pi}^{(k)}_{\text{rel}} = T \frac{\widehat{\Pi}^{(k-1)}_{\text{rel}} \odot \exp\left(-\gamma_k \nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}; X, X^{\#}, A, S\right)\right)}{\left\|\widehat{\Pi}^{(k-1)}_{\text{rel}} \odot \exp\left(-\gamma_k \nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}; X, X^{\#}, A, S\right)\right)\right\|_1},$$
where, for a matrix $A \in \mathbb{R}^{N\times N}$, $\|A\|_1 := \sum_{i,j\in[N]} |A_{ij}|$. The learning rate is chosen as follows,
$$\gamma_k = \gamma \frac{\log(k+1)}{\left(\left\|\nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}\right)\right\|_\infty \vee 10^{-4}\right)\sqrt{k+1}}. \tag{5.2}$$

- **Birkhoff Polytope.** Consider $\mathcal{K} = \{Z \in \mathbb{R}^{T\times T} : \mathbb{1}^\top Z = \mathbb{1}^\top, Z\mathbb{1} = \mathbb{1}, Z \geq 0\}$. To solve the relaxed MLE problem on the Birkhoff polytope, we employ a projected gradient descent strategy (which we call Birkhoff PGD). We consider the same initialization as the previous algorithms $\widehat{\Pi}^{(0)}_{\text{rel}} = \frac{1}{T} J_T$, and the iterations are of the form
$$\widehat{\Pi}^{(k)}_{\text{rel}} = \mathscr{P}_{\mathcal{K}}\left(\widehat{\Pi}^{(k-1)}_{\text{rel}} - \gamma_k \nabla f\left(\widehat{\Pi}^{(k-1)}_{\text{rel}}; X, X^{\#}, A, S\right)\right), \text{ for } k \geq 1.$$
Similar as before, here f is the relaxed MLE objective defined in (4.1), S is the shift matrix defined in Section 2, and $\mathscr{P}_{\mathcal{K}}$ is the Euclidean projection onto $\mathcal{K}$ (the Birkhoff polytope). We use the Dykstra algorithm (Dykstra, 1983) to approximate this projection. The learning rate follows (5.1), with the constant γ adjusted as needed. We refer to this algorithm as Birkhoff PGD.

Error metrics. To quantify the success of the proposed methods in the recovery of $\Pi^* \in \mathcal{P}_T$ we consider the recovery fraction, defined for any matrix $\Pi \in \mathcal{P}_T$ as
$$\text{Recovery fraction}(\Pi) := \langle \Pi, \Pi^* \rangle_F / T.$$

Notice that in the previous definition the ground truth Π^* is implicit. Even if our goal is mainly recovery of Π^*, the alternating optimization method proposed in Algorithm 2 allows us to jointly recover Π^* and A^*. We measure the error for the estimation of $A^* \in \mathbb{R}^{d\times d}$ with the MSE, defined for a matrix $A \in \mathbb{R}^{d\times d}$ as follows,

$$\mathrm{MSE}(A) := \|A - A^*\|_F^2/d.$$

Parametric assumption on A^*. Throughout our experiments, we consider a random A^*, constructed as follows. First, we sample a A' with iid standard Gaussian entries. Then we set

$$A^* = \theta \frac{A'}{\|A'\|_2}\,, \tag{5.3}$$

where the parameter $\theta \in \mathbb{R}$ helps to control $\|A^*\|_2$.

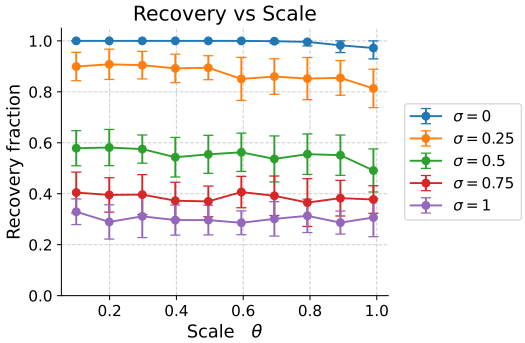
5.2 Relaxed MLE with known A^*

We begin by evaluating the performance of the `RelaxMLE-Round` subroutine, see Algorithm 1, when the system matrix A^* is known. We generate time series following the $\mathbf{CVAR}(1, d, T; A^*, \pi^*, \sigma)$ model under different noise levels σ, with A^* satisfying the parametric assumption in (5.3). To unravel the influence of θ and σ, we fix one parameter and vary the other, reporting the recovery accuracy of Π^*. We choose $\Pi^* = I_T$, and for all methods we initialize with $\widehat{\Pi}^{(0)}_{\mathrm{rel}} = \frac{1}{T}J_T$. In addition, we consider two contrasting regimes: $(d, T) = (5, 50)$, where the ambient dimension is small relative to the number of observations, and $(d, T) = (50, 5)$, where the opposite holds.

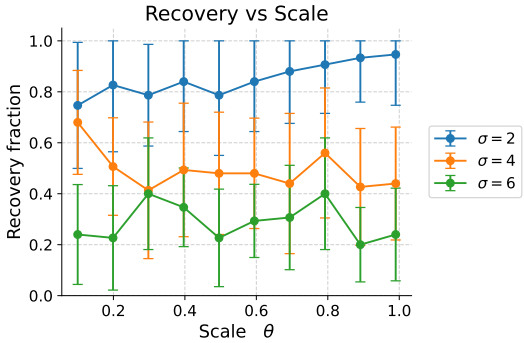
Figure 1 illustrates the effect of the scale parameter θ on recovery. Interestingly, in both regimes considered for d, T the estimation error remains essentially unaffected by θ, suggesting that for these algorithms the problem does not become more difficult at larger scales. It should be noted that, in the case $d = 50$, $T = 5$, the considered σ is higher, since for smaller values of σ perfect recovery is achieved for all the algorithms for most random realizations. Intuitively, the problem becomes easier in this regime, as discussed in more detail below. In terms of performance, the relaxation to the Birkhoff polytope generally outperforms the simplex and hyperplane relaxations, particularly in high-noise settings (e.g., in the case $d = 50, T = 5$ considered here). Interestingly, the overall differences between the relaxations are not drastic—the simplex relaxation performs comparably to the Birkhoff relaxation across most conditions. One advantage of the hyperplane and simplex relaxations is their comparatively lower computational complexity relative to the Birkhoff relaxation.

In Figure 2 we plot the recovery fraction versus the noise level to highlight the effect of σ on the recovery level for the same pairs (d, T). This is a complementary plot to Figure 1, which allows us to read how the recovery decays with the noise level. We obtain similar conclusions: the scale does not seem to affect the recovery fraction and while Birkhoff PGD performs slightly better, the difference in performance remains moderate (expect for high noise scenarios in $d = 50\,, T = 5$). In addition, we notice that the performance of Birkhoff PGD is very close to the LA estimator. This is expected for small scales, since for $A = 0$ solving (4.1) on the Birkhoff polytope is equivalent to the linear assignment solution in (2.6). On the other hand, for larger scales it is not obvious from (4.1) that both approaches would have a similar performance.

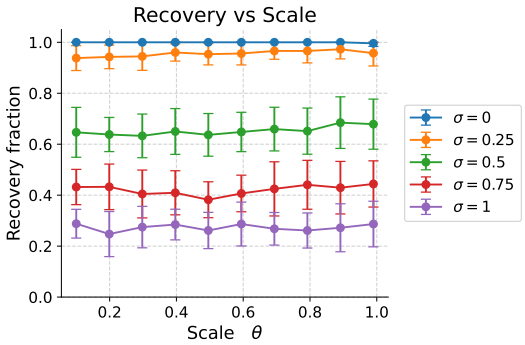
These plots suggest that case $d = 50, T = 5$ is easier for all algorithms considered, compared to case $d = 5, T = 50$ (notice that we considered higher levels of noise in case $d = 50, T = 5$). This is expected since in the case $d = 50, T = 5$ we have few points in a high dimension, which implies a higher separation between them. More surprising is the fact that the LA estimator performs on par with the best-performing MLE relaxation: the one on the Birkhoff polytope. It seems that the model information, used by the MLE relaxations, does not lead to a noticeable advantage in terms of the matching performance. A natural question is whether LA is optimal in terms of recovery for time series matching, at least in the regime $\|A^*\|_2 < 1$. We explore the case $\|A^*\|_2 \geq 1$ below.



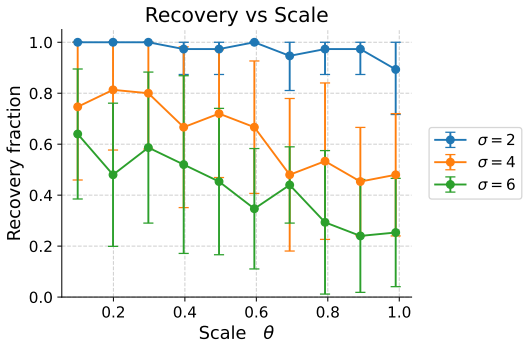
(a) Hyperplane relaxed MLE $d = 5$, $T = 50$.



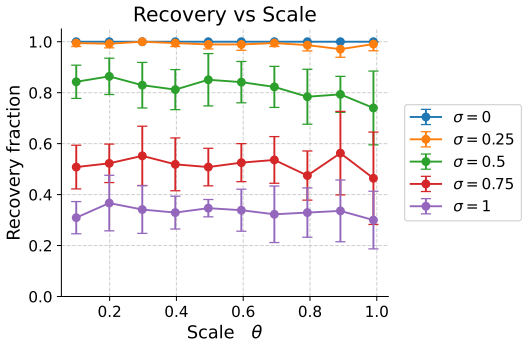
(b) Hyperplane relaxed MLE $d = 50$, $T = 5$.



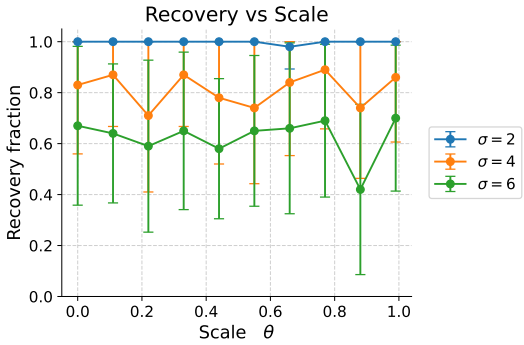
(c) Simplex relaxed MLE $d = 5$, $T = 50$.



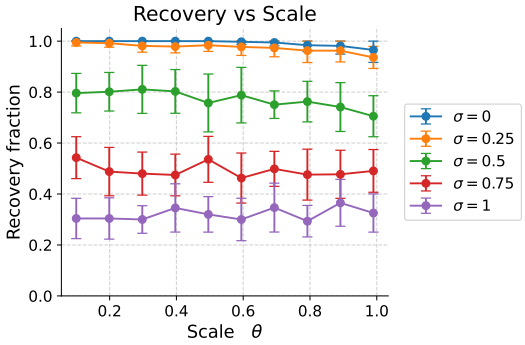
(d) Simplex relaxed MLE $d = 50$, $T = 5$.



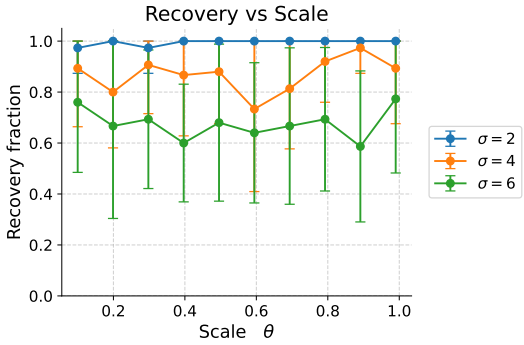
(e) Linear assignment estimator $d = 5$, $T = 50$.



(f) Linear assignment estimator $d = 50$, $T = 5$.

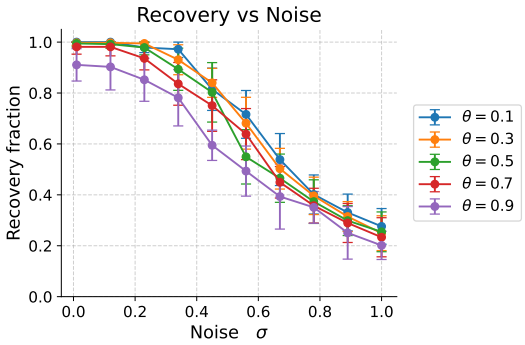


(g) Birkhoff relaxed MLE $d = 5$, $T = 50$.

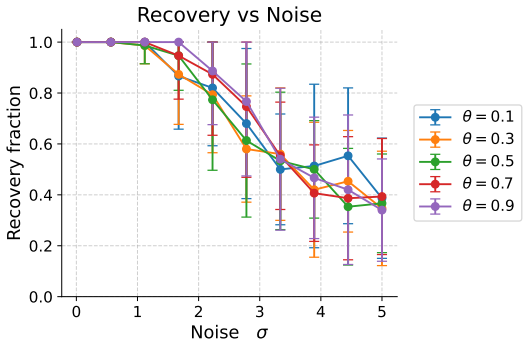


(h) Birkhoff relaxed MLE $d = 50$, $T = 5$.

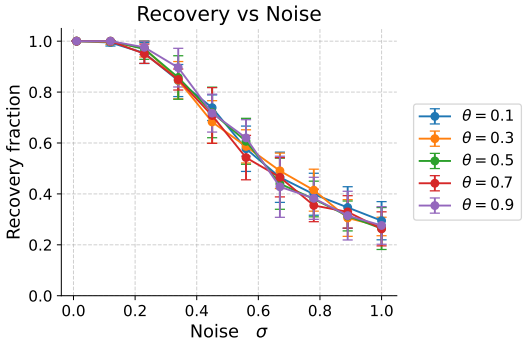
Figure 1: Recovery fraction vs. scale θ using Algorithm 1. We assume known A^* of the form (5.3), and consider different values for θ. For each (θ, σ) pair, the plotted value corresponds to the average over 30 Monte Carlo samples of the **CVAR**$(1, d, T; A^*, \pi^*, \sigma)$ model. The error bars reflects one standard deviation above and below the mean.



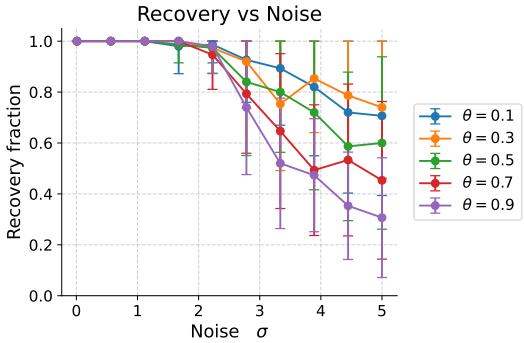
(a) Hyperplane relaxed MLE $d = 5$, $T = 50$.



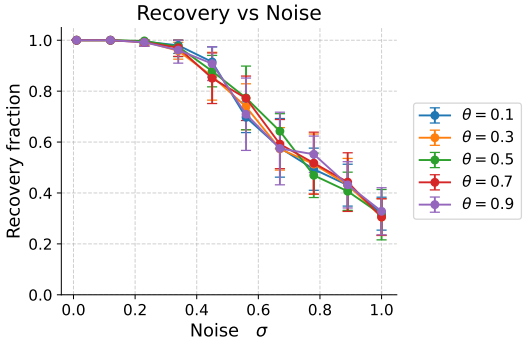
(b) Hyperplane relaxed MLE $d = 50$, $T = 5$.



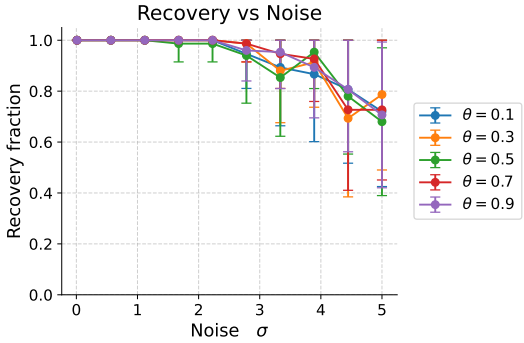
(c) Simplex relaxed MLE $d = 5$, $T = 50$.



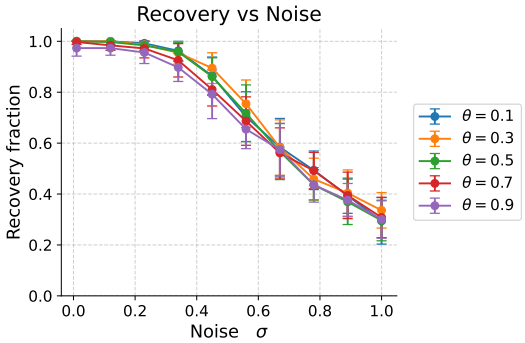
(d) Simplex relaxed MLE $d = 50$, $T = 5$.



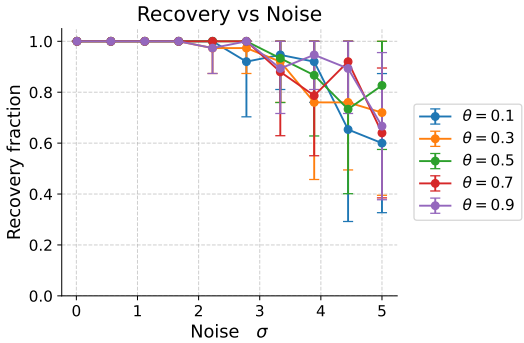
(e) Linear assignment estimator $d = 5$, $T = 50$.



(f) Linear assignment estimator $d = 50$, $T = 5$.



(g) Birkhoff relaxed MLE $d = 5$, $T = 50$.



(h) Birkhoff relaxed MLE $d = 50$, $T = 5$.

Figure 2: Recovery fraction vs. noise σ using Algorithm 1. The setting is analogous to Fig.1.

5.3 Algorithms with unknown A^*

We now evaluate the proposed alternating optimization procedure in Algorithm 2 and the linear assignment estimator in (2.6), in terms of Π^* recovery. In these experiments, we assume that σ is known, while A^* remains unknown. We initialize $\Pi^{(0)}$ uniformly at random and use $K = 5$ alternating minimization steps. In Appendix E we include some experiments for the strategy of estimating A^* first, discussed in Remark 10.

In Figure 3, we present the recovery fraction as a function of the scale parameter θ for $(d, T) \in \{(5, 50), (50, 5)\}$. As in the case where A^* is known, the recovery fraction remains fairly stable across scales. In this case, all algorithms based on MLE relaxations have a similar performance in the case $d = 5$, $T = 50$, while the LA estimator slightly outperform them. On the other hand, linear assignment outperforms relaxed MLE-based algorithms in the case $d = 50$, $T = 5$, achieving perfect recovery for the considered values of σ. This suggests that random initialization performs poorly for estimating A^*, particularly with small T. It should be noted that the variance, reflected in the error bars, is relatively high for this choice of parameters d, T. Interestingly, for $d = 5$ and $T = 50$, the Birkhoff relaxation achieves an average recovery performance comparable to linear assignment. This is somewhat surprising, as one might expect that an initial random permutation would adversely affect the A-update step in Algorithm 2. Similar conclusions can be obtained from the complementary plots, in Figure 4, where the recovery fraction is plotted against the noise level.

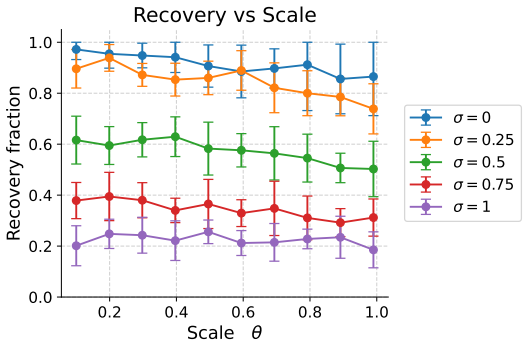
Remark 12 (Number of alternating optimization iterations)**.** *We observe that often after one or two alternating optimization iterations in Algorithm 2, the estimator for the permutation converges. This suggest that estimating A^* first, as discussed in Remark 10, is a viable alternative. We evaluate this in Appendix E.1.*

Estimation of A^*. Although the estimation of A^* is not our primary objective, Algorithm 2 jointly estimates both Π^* and A^*. To assess the performance of the algorithm for the estimation of A^*, we fix $d = 5$, $\theta = 0.5$, and vary the time horizon $T \in \{10, 20, 30, 50, 100\}$. We plot the performance of the Birkhoff-based relaxation, but in our experiments the hyperplane and the simplex performs similarly for this task. We report the estimation error for $\sigma = 0.5$, since other values produce qualitatively similar results. In Figure 5a, we plot the MSE for the estimation of A^* for different values of T. As expected, $\mathrm{MSE}(A)$ decreases as T increases, indicating that Algorithm 2 successfully estimates A^*. Figure 5b shows a scatterplot of the estimation errors of A^* and Π^* (in terms of the recovery fraction) over 50 samples with $T = 100$. Although there is a tendency for better estimation of A^* to correspond to improved recovery of Π^*, similar estimation errors for A^* can still lead to markedly different recovery fractions.

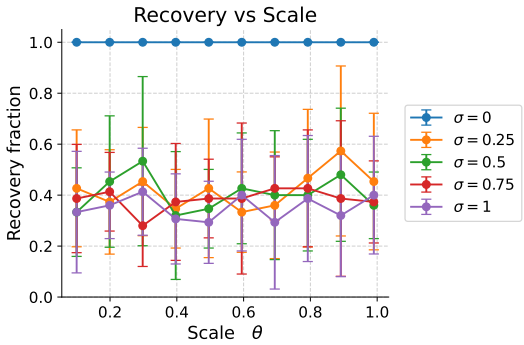
Recovery vs. T. We examine the recovery performance of the LA estimator and Birkhoff PGD in function of the time horizon T. In Figure 6 we show the average recovery of these estimators for the scale $\theta = 0.5$ and $d \in \{5, 25\}$. We observe that the performance of both estimators is very similar for both dimensions. It should be noted that while we report only the scale $\theta = 0.5$, similar results were obtained for other values $\theta < 1$. This supports the hypothesis that LA is already optimal, or near optimal, in this regime. In addition, we observe that the recovery worsens as T grows, as predicted by Theorem 1. As we will see next, the situation is slightly different for $\theta > 1$.

The case $\theta \geq 1$. This case is interesting, since the processes $(x_t)_{t\in[T]}, (x^{\#})_{t\in[T]}$ become unstable, from the dynamical systems perspective. A natural question is if its possible to non-trivially recover the hidden permutation Π^* in this regime, and does the problem becomes easier from the matching perspective. Intuitively, the problem might become easier if the individual points are more separated. We evaluate this by considering the scales $\theta \in \{1.5, 2, 2.5, 3\}$.

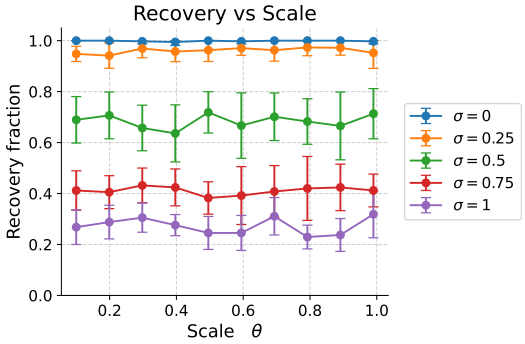
In Figure 7, we plot the average recovery fraction for the estimated permutation as a function of T. The comparison includes the performance of the LA estimator and the Birkhoff PGD method. We observe a transition in recovery performance as the scale parameter θ increases: for $\theta = 1.5$, the Birkhoff-based estimator slightly outperforms LA, whereas for $\theta = 3$, LA achieves better recovery. A similar trend appears when varying T: for smaller T, Birkhoff PGD performs slightly better, while for larger T, LA tends to outperform it in average. It should be noted that the LA estimator has larger variance at the considered scales. Intuitively, as T grows, the underlying unstable processes evolve for longer periods, leading to more



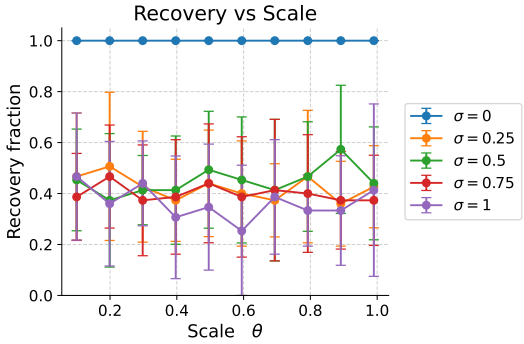
(a) Hyperplane relaxed MLE $d = 5$, $T = 50$.



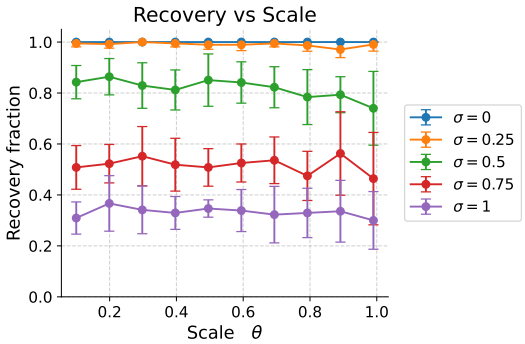
(b) Hyperplane relaxed MLE $d = 50$, $T = 5$.



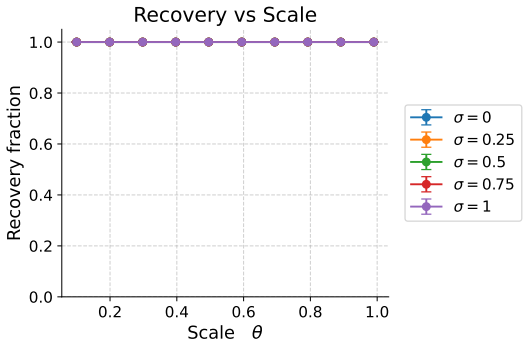
(c) Simplex relaxed MLE $d = 5$, $T = 50$.



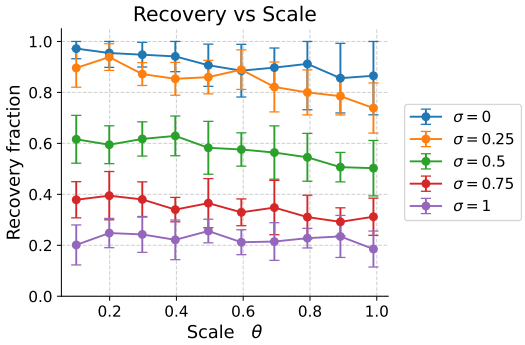
(d) Simplex relaxed MLE $d = 50$, $T = 5$.



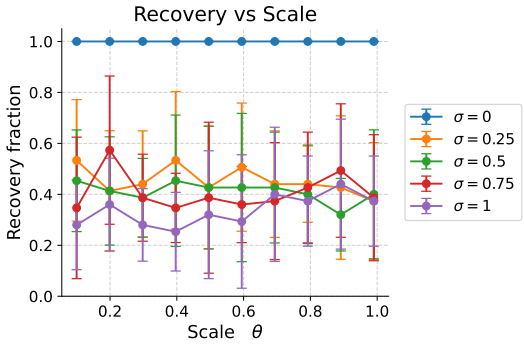
(e) Linear assignment estimator $d = 5$, $T = 50$.



(f) Linear assignment estimator $d = 50$, $T = 5$.

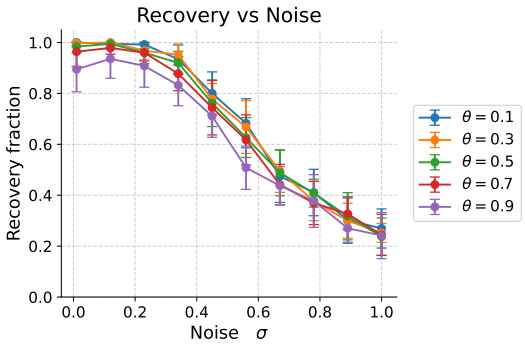


(g) Birkhoff relaxed MLE $d = 5$, $T = 50$.

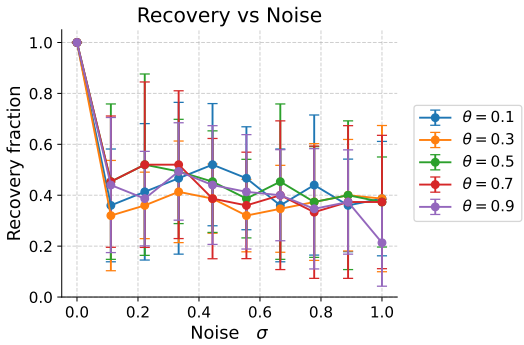


(h) Birkhoff relaxed MLE $d = 50$, $T = 50$.

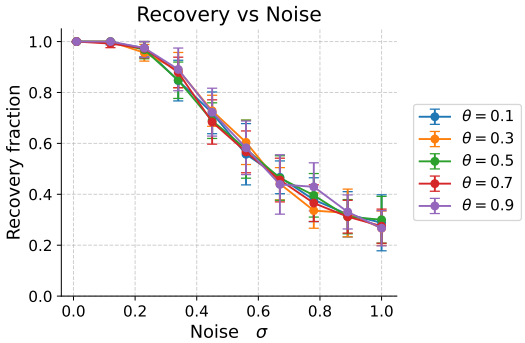
Figure 3: Recovery fraction vs. scale θ using Algorithm 2 with $K = 5$. A^* (unknown) is of the form (5.3). We average 30 Monte Carlo samples of the **CVAR**$(1, d, T; A^*, \pi^*, \sigma)$ model. The error bars reflects one standard deviation above and below the mean.



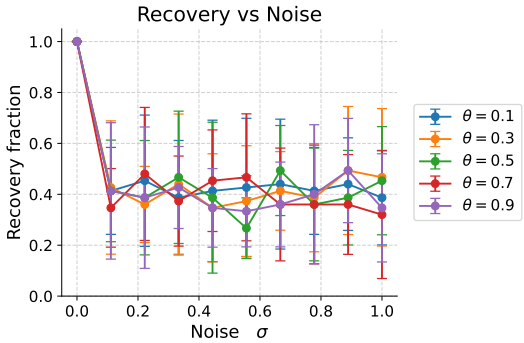
(a) Hyperplane relaxed MLE $d = 5$, $T = 50$.



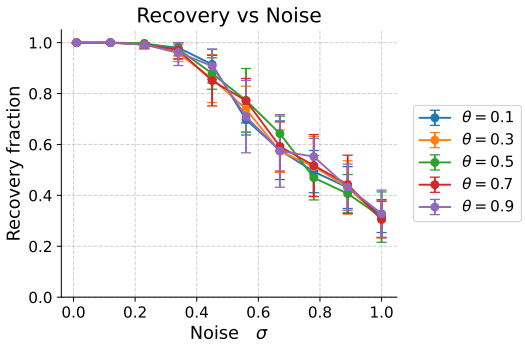
(b) Hyperplane relaxed MLE $d = 50$, $T = 5$.



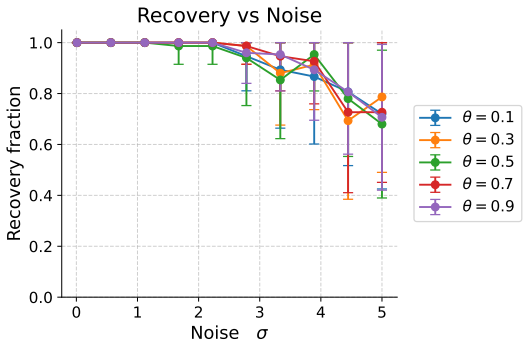
(c) Simplex relaxed MLE $d = 5$, $T = 50$.



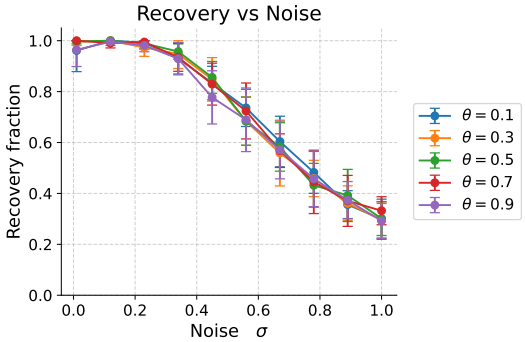
(d) Simplex relaxed MLE $d = 50$, $T = 5$.



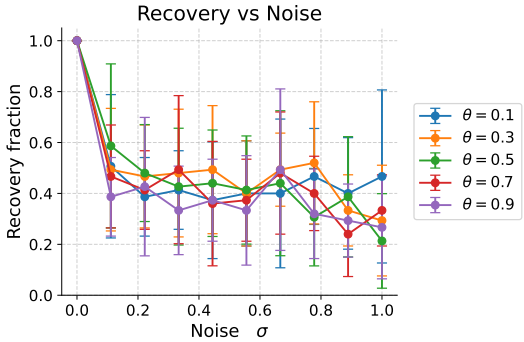
(e) Linear assignment estimator $d = 5$, $T = 50$.



(f) Linear assignment estimator $d = 50$, $T = 5$.

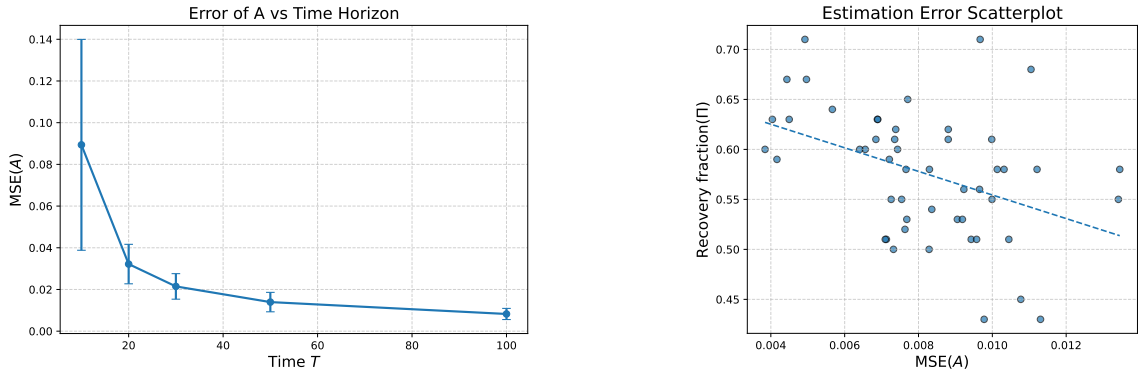


(g) Birkhoff relaxed MLE $d = 5$, $T = 50$.



(h) Birkhoff relaxed MLE $d = 50$, $T = 5$.

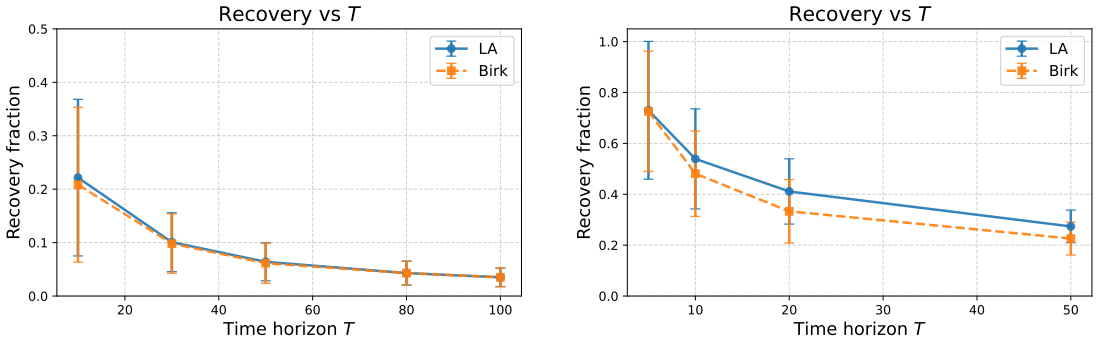
Figure 4: Recovery fraction vs. noise σ with A^* unknown. This figure is complementary to Fig. 3, under an analogous setting.



(a) Estimation MSE for A^*

(b) Error for estimating Π^* and A^*

Figure 5: Estimation error for A^*. We fix $d = 5$, $\sigma = 0.5$, $\theta = 0.5$. In Fig.5a we plot MSE(A) for $T \in \{10, 20, 30, 50, 100\}$ averaged over 30 Monte Carlo samples (the error bars reflect one standard deviation above and below the mean). Fig.5b is a scatter plot, over 50 samples, for the error of estimating Π^* and A^*. The dashed line represent the linear trend.



(a) $\theta = 0.5$, $d = 5$

(b) $\theta = 0.5$, $d = 25$

Figure 6: Comparison of recovery performance for the LA and Birkhoff PGD estimators for different time horizons. The recovery fraction is the average over 30 samples, and the error bars reflect one standard deviation above and below the average.

separated trajectories and hence an easier matching task. The same argument applies for larger scales, as the separation between points increases quickly.

6 Conclusion and open questions

In this paper, we studied the problem of matching correlated VAR time series, extending the point-cloud matching framework of Kunisky & Niles-Weed (2022) to a temporal setting. We introduced a model in which a base VAR process $(x_t)_{t\in[T]}$ is perturbed by a σ-scaled independent copy and then permuted by an unknown π^*, yielding $(x_t^{\#})_{t\in[T]}$.

We derived the MLE for recovering π^*, from the observation of $\left((x_t)_{t\in[T]}, (x^{\#}_{t\in[T]})\right)$, and showed that it leads to a quadratic assignment problem, which is NP-hard in general. To obtain tractable alternatives, we theoretically analyzed the linear assignment estimator and established conditions—expressed as thresholds on σ—under which perfect or partial recovery is guaranteed. We also developed an alternating minimization based framework for solving the MLE, where the latent permutation is iteratively estimated by solving a suitable convex relaxation of the set of permutation matrices, thus enabling efficient first-order algorithms.

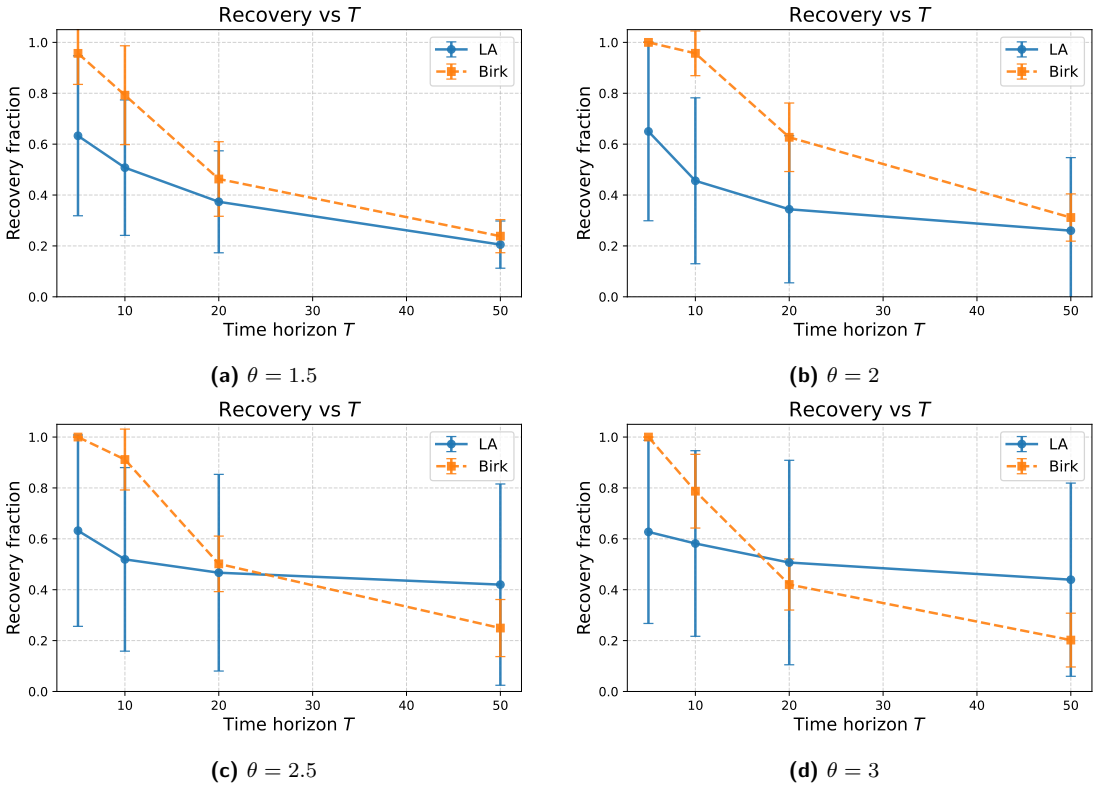


Figure 7: Comparison of recovery performance across different values of θ. Here $d = 25$ and the recovery fraction is the average over 30 samples. The error bars reflect one standard deviation above and below the average.

Finally, we evaluated both the linear assignment estimator and the MLE relaxations on synthetic datasets, demonstrating their practical performance.

There are several promising directions for future work.

- **Information-theoretic limits.** A natural open question is to determine the fundamental limits for recovering π^*. While related bounds are known for point-cloud matching (da Rocha Schwengber & Oliveira, 2024), it remains unclear whether those techniques extend to settings with temporal correlations.

- **Extending the analysis.** It would be interesting to extend our results to the regime $\rho(A^*) \leq 1$, or even $\rho(A^*) > 1$, where the spectral radius $\rho(\cdot)$ characterizes the stability of linear dynamical systems. In the system identification literature for learning VAR models from a single trajectory, these two regimes have been studied recently for the setting $\rho(A^*) \leq 1$ (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019) where non-asymptotic error bounds for recovering A^* were obtained. These were also studied for the case $\rho(A^*) > 1$ in Sarkar & Rakhlin (2019) where the results hold for "regular" matrices[5] A^*.

- **Alternative problem formulations.** Our analysis focused on permutations of time indices, i.e., column permutations of the $d \times T$ matrix X' (whose columns are the elements of $(x'_t)_{t\in[T]}$, defined in (1.3)). Another meaningful variant permutes the rows of X', effectively shuffling the coordinates of the time series. This has potential applications in problems such as dynamic time warping and time-series alignment, and may require different analytical tools.

[5]Matrices for which the geometric multiplicity of eigenvalues lying outside the unit circle, is one.

References

John Aach and George M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

Ernesto Araya and Hemant Tyagi. Graph matching via convex relaxation to the simplex. *Foundations of Data Science*, 7(2):464–501, 2025.

Amir Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. SIAM, 2017.

Jan Beutel, Stephan Gruber, Andreas Hasler, Roman Lim, Andreas Meier, Christian Plessl, Igor Talzi, Lothar Thiele, Christian Tschudin, Matthias Woehrle, and Mustafa Yuecel. Permadaq: A scientific instrument for precision sensing and data recovery in environmental extremes. In *2009 International Conference on Information Processing in Sensor Networks*, pp. 265–276, 2009.

A. Bottcher and B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices*. Springer New York, NY, 1999.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Olivier Collier and Arnak S. Dalalyan. Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.*, 17(1):162–192, January 2016.

Lucas da Rocha Schwengber and Roberto Imbuzeiro Oliveira. Geometric planted matchings beyond the gaussian model. *arXiv:2403.17469*, 2024.

Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1-2):29–115, 2021a.

Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. The planted matching problem: sharp threshold and infinite-order phase transition. *Probability Theory and Related Fields*, 187:1–71, 2021b.

Richard L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346-347:180–197, 2016. ISSN 0020-0255.

Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations i: Algorithm and gaussian analysis. *Foundations of Computational Mathematics*, 23(5):1511–1565, June 2022.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Tigran Galstyan, Arshak Minasyan, and Arnak S. Dalalyan. Optimal detection of the feature matching map in presence of noise and outliers. *Electronic Journal of Statistics*, 16(2):5720 – 5750, 2022.

Chris Godsil and Gordon Royle. *Algebraic Graph Theory*, volume 207 of *Graduate Texts in Mathematics*. Springer, 2001. ISBN 978-0-387-95220-9.

Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16(97):3115–3150, 2015.

Yanjun Han, Philippe Rigollet, and George Stepaniants. Covariance alignment: From maximum likelihood estimation to gromov–wasserstein. *SIAM Journal on Mathematics of Data Science*, 7(3):1491–1513, 2025.

Nicholas J. Higham and Sheung Hun Cheng. Modifying the inertia of matrices arising in optimization. *Linear Algebra and its Applications*, 275-276:261–279, 1998. Proceedings of the Sixth Conference of the International Linear Algebra Society.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2012. ISBN 9780521839402.

Tzu-Kuo Huang and Jeff Schneider. Learning linear dynamical systems without sequence information. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 425–432, 2009.

Tzu-Kuo Huang and Jeff Schneider. Learning auto-regressive models from sequence and non-sequence data. *Advances in Neural Information Processing Systems*, 24, 2011.

Tzu-Kuo Huang and Jeff Schneider. Learning hidden markov models from non-sequence data via tensor decomposition. *Advances in Neural Information Processing Systems*, 26, 2013.

Tzu-Kuo Huang, Le Song, and Jeff Schneider. Learning nonlinear dynamic models from non-sequenced data. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 350–357, 2010.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 996–1001, 2020a.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems: Optimality of the least-squares estimator. *arxiv:2003.07937*, 2020b.

Keisuke Kawano, Takuro Kutsuna, and Satoshi Koide. Neural time warping for multiple sequence alignment. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3837–3841, 2020.

Matthias Keller, Lothar Thiele, and Jan Beutel. Reconstruction of the correct temporal order of sensor network data. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 282–293, 2011.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Dmitriy Kunisky and Jonathan Niles-Weed. Strong recovery of geometric planted matchings. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 834–876, 2022.

François Le Gall. Algebraic complexity theory and matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, pp. 23, 2014.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637 – 1664, 2012.

Vince Lyzinski, Donniell E. Fishkind, Marcelo Fiori, Joshua T. Vogelstein, Carey E. Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):60–73, 2016.

Mehrdad Moharrami, Cristopher Moore, and Jiaming Xu. The planted matching problem: Phase transitions and exact results. *The Annals of Applied Probability*, 31(6):2663 – 2720, 2021.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008.

Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 1235–1243, 2011.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, pp. 5610–5618, 2019.

Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. Recovery thresholds in the sparse planted matching problem. *Phys. Rev. E*, 102:022304, 2020.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 439–473, 2018.

D.A Spielman. Spectral and algebraic graph theory. incomplete draft (2025), available at `http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf`, 2025.

Eric Chu Borja Peleato Stephen Boyd, Neal Parikh and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3: 1–122, 2011.

Christiane Tretter. *Spectral Theory of Block Operator Matrices and Applications*. Imperial College Press, London, 2008. ISBN 9781860947681.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pp. 3441–3488, 2022.

Qingqing Ye, Haibo Hu, Ninghui Li, Xiaofeng Meng, Huadi Zheng, and Haotian Yan. Beyond value perturbation: Local differential privacy in the temporal setting. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–10, 2021.

Qingqing Ye, Haibo Hu, Kai Huang, Man Ho Au, and Qiao Xue. Stateful switch: Optimized time series release with local differential privacy. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pp. 1–10, 2023.

Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–i267, June 2009.

A Proof of Lemma 2

The solution to (2.4) can be found in closed form. Indeed, note that

$$\|X - AXS\|_F^2 = \|\operatorname{vec}(X) - \big((XS)^\top \otimes I_d\big)\operatorname{vec}(A)\|_2^2,$$
$$\|X^\#\Pi - AX^\#\Pi S - (X - AXS)\|_F^2 = \|\operatorname{vec}(X^\#\Pi - X) - \big((X^\#\Pi S)^\top \otimes I_d\big)\operatorname{vec}(A) + \big((XS)^\top \otimes I_d\big)\operatorname{vec}(A)\|_2^2$$

Define,

$$\begin{aligned} v &:= \operatorname{vec}(X), \\ V &:= (XS)^\top \otimes I_d, \\ u &:= \operatorname{vec}(X^\#\Pi - X), \\ U &:= (X^\#\Pi S)^\top \otimes I_d + (XS)^\top \otimes I_d, \\ a &:= \operatorname{vec}(A). \end{aligned}$$

So (2.4) can be rewritten as

$$\operatorname{vec}\left(\widehat{A}_{\mathrm{MLE}}(\Pi)\right) = \underset{a\in\mathbb{R}^{d^2}}{\operatorname{argmin}} \left\{ \|v - Va\|_2^2 + \frac{1}{\bar{\sigma}^2}\|u - Ua\|_2^2 \right\}.$$

Define $g(a) := \|v - Va\|_2^2 + \frac{1}{\bar{\sigma}^2}\|u - Ua\|_2^2$. Then,

$$\begin{aligned}\nabla g(a) &= 2V^\top Va - 2V^\top v + \frac{1}{\bar{\sigma}^2}\left(2U^\top Ua - 2U^\top u\right)\\ &= 2\left((V^\top V + \frac{1}{\bar{\sigma}^2}U^\top U)a - (V^\top v + \frac{1}{\bar{\sigma}^2}U^\top u)\right).\end{aligned}$$

Hence, the condition $\nabla g(a) = 0$ is equivalent to

$$\left(V^\top V + \frac{1}{\bar{\sigma}^2}U^\top U\right)a = V^\top v + \frac{1}{\bar{\sigma}^2}U^\top u,$$

Now,

$$V^\top V + \frac{1}{\bar{\sigma}^2}U^\top U = \left((XS)(XS)^\top + \frac{1}{\bar{\sigma}^2}(X^\#\Pi S - XS)(X^\#\Pi S - XS)^\top\right)\otimes I_d.$$

Also,

$$\begin{aligned}V^\top v &= ((XS)\otimes I_d)\operatorname{vec}(X)\\ U^\top u &= \left((X^\#\Pi S - XS)\otimes I_d\right)\operatorname{vec}(X^\#\Pi - X).\end{aligned}$$

This translates to

$$\left(V^\top V + \frac{1}{\bar{\sigma}^2}U^\top U\right)\operatorname{vec}(A) = A\left((XS)(XS)^\top + \frac{1}{\bar{\sigma}^2}(X^\#\Pi S - XS)(X^\#\Pi S - XS)^\top\right),$$

and

$$V^\top v + \frac{1}{\bar{\sigma}^2}U^\top u = X(XS)^\top + \frac{1}{\bar{\sigma}^2}\left(X^\#\Pi - X\right)\left(X^\#\Pi S - XS\right)^\top.$$

So, if $\widehat{A}$ satisfies (2.4), then

$$\widehat{A}\left((XS)(XS)^\top + \frac{1}{\bar{\sigma}^2}(X^\#\Pi S - XS)(X^\#\Pi S - XS)^\top\right) = X(XS)^\top + \frac{1}{\bar{\sigma}^2}\left(X^\#\Pi - X\right)\left(X^\#\Pi S - XS\right)^\top.$$

From this, the result follows.

B Proof of Lemma 3

To ease notation, we use A instead of A^*. Note that, for $a, b \in \mathbb{N}$, with $a > b$, we have

$$\begin{aligned}\tilde{x}_a &= A^{a-1}\tilde{\xi}_1 + A^{a-2}\tilde{\xi}_2 + \ldots + A^{a-b}\tilde{\xi}_b + A^{a-b-1}\tilde{\xi}_{b+1} + \ldots + \tilde{\xi}_a,\\ \tilde{x}_b &= A^{b-1}\tilde{\xi}_1 + A^{b-2}\tilde{\xi}_2 + \ldots + \tilde{\xi}_b,\end{aligned}$$

from which we obtain

$$\begin{aligned}A(\tilde{x}_{a-1} - \tilde{x}_{b-1}) =& A^{b-1}\left(A^{a-b} - I_d\right)\tilde{\xi}_1 + A^{b-2}\left(A^{a-b} - I_d\right)\tilde{\xi}_2 + \ldots + A\left(A^{a-b} - I_d\right)\tilde{\xi}_{b-1}\\ &+ \left(A^{a-b}\tilde{\xi}_b + \ldots + A\tilde{\xi}_{a-1}\right).\end{aligned}$$

With some algebra, we get

$$\tilde{y}_{ab} := A(\tilde{x}_{a-1} - \tilde{x}_{b-1}) + \tilde{\xi}_a - \tilde{\xi}_b = \sum_{i=0}^{b-1} A^i\left(A^{a-b} - I_d\right)\tilde{\xi}_{b-i} + \sum_{i=1}^{a-b} A^{a-b-i}\tilde{\xi}_{b+i}.$$

Recalling the definition of $y_{ab}, \tilde{y}_{ab}$, it is clear that

$$\langle \tilde{y}_{ab}, y_{ab} \rangle = \underbrace{\left\langle y_{ab}, \sum_{i=0}^{b-1} A^i \left(A^{a-b} - I_d\right) \tilde{\xi}_{b-i} \right\rangle}_{=:\zeta_1} + \underbrace{\left\langle y_{ab}, \sum_{i=1}^{a-b} A^{a-b-i} \tilde{\xi}_{b+i} \right\rangle}_{=:\zeta_2}.$$

Conditioned on $\xi_1, \ldots, \xi_a$, note that ζ_1 and ζ_2 are (independent) Gaussians (being linear combination of jointly Gaussian variables) of the form $\zeta_1 \sim \mathcal{N}(0, \tilde{\sigma}_1^2)$, $\zeta_2 \sim \mathcal{N}(0, \tilde{\sigma}_2^2)$, with

$$\tilde{\sigma}_1^2 = y_{ab}^\top \left(\sum_{i=0}^{b-1} A^i \left(A^{a-b} - I_d\right) \left(A^{a-b} - I_d\right)^\top {A^i}^\top \right) y_{ab}$$
$$\tilde{\sigma}_2^2 = y_{ab}^\top \left(\sum_{i=1}^{a-b} A^{a-b-i} {A^{a-b-i}}^\top \right) y_{ab}.$$

Furthermore, it holds

$$\begin{aligned} \tilde{\sigma}_1^2 \;\le\; \|y_{ab}\|_2^2 \|A^{a-b} - I_d\|_2^2 \sum_{i=0}^{b-1} \|A^i\|_2^2 \;&\le\; \|y_{ab}\|_2^2 \|A^{a-b} - I_d\|_2^2 \sum_{i=0}^{b-1} \|A\|_2^{2i} \\ &\le \frac{\|y_{ab}\|_2^2 \|A^{a-b} - I_d\|_2^2}{1 - \|A\|_2^2}, \end{aligned}$$

and also

$$\begin{aligned} \tilde{\sigma}_2^2 \;\le\; \|y_{ab}\|_2^2 \left\| \sum_{i=1}^{a-b} A^{a-b-i} {A^{a-b-i}}^\top \right\|_2 \;&\le\; \|y_{ab}\|_2^2 \sum_{i=1}^{a-b} \|A\|_2^{2(a-b-i)} \\ &= \|y_{ab}\|_2^2 \frac{1}{1 - \|A\|_2^2}. \end{aligned}$$

From this, we deduce that if $\left\|A\right\|_2 < 1$, then

$$\sigma^2(\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2) \;\le\; \frac{\sigma^2 \|y_{ab}\|_2^2}{1 - \|A\|_2^2} \left(1 + \underbrace{\|A^{a-b} - I_d\|_2^2}_{\le 4} \right) \;\le\; \frac{5\sigma^2 \|y_{ab}\|_2^2}{1 - \|A\|_2^2}.$$

This completes the proof.

C Auxiliary lemmas for proof of Proposition 2

C.1 Proof of Lemma 6

We have

$$\begin{aligned} \left(\frac{\prod_{k=1}^n a_k}{\prod_{k=1}^n (a_k + b_k)} \right)^{1/n} + \left(\frac{\prod_{k=1}^n b_k}{\prod_{k=1}^n (a_k + b_k)} \right)^{1/n} &= \left(\prod_{k=1}^n \frac{a_k}{a_k + b_k} \right)^{1/n} + \left(\prod_{k=1}^n \frac{b_k}{a_k + b_k} \right)^{1/n} \\ &\le \frac{1}{n} \sum_{k=1}^n \left(\frac{a_k}{a_k + b_k} \right) + \frac{1}{n} \sum_{k=1}^n \left(\frac{b_k}{a_k + b_k} \right) \\ &= 1, \end{aligned}$$

where we used the AM-GM inequality in the second line. Multiplying both sides by $\left(\prod_{k=1}^n (a_k + b_k)\right)^{1/n}$ gives the result.

C.2 Proof of Lemma 8

Consider the eigenvalue decomposition of Z,

$$Z = U \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} U^\top,$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive entries. Let $U_p \in \mathbb{R}^{q\times p}$ be the matrix whose columns are the first p columns of U (i.e., the columns are the eigenvectors associated with non-zero eigenvalues). Then,

$$Z = U_p \Lambda U_p,$$

and

$$WZW^\top = (WU_p)\Lambda(WU_p)^\top.$$

This implies that,

$$\begin{aligned}
\det{}^*(WZW^\top) &= \det{}^*\left((WU_p)\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}(WU_p)^\top\right) \\
&\overset{(1)}{=} \det{}^*\left(\Lambda^{\frac{1}{2}}(WU_p)^\top(WU_p)\Lambda^{\frac{1}{2}}\right) \\
&\overset{(2)}{=} \det\left(\Lambda^{\frac{1}{2}}(WU_p)^\top(WU_p)\Lambda^{\frac{1}{2}}\right) \\
&= \det(\Lambda)\det\left((WU_p)^\top(WU_p)\right) \\
&= \det{}^*(L)\det\left((WU_p)^\top(WU_p)\right).
\end{aligned}$$

In (1) we used that $\det^*(AB) = \det^*(BA)$, which holds because AB and BA have the same nonzero eigenvalues, counting multiplicities, for any matrices A, B such that the products AB, BA are well defined. To obtain (2), we used that the $p \times p$ matrix $\Lambda^{\frac{1}{2}}(WU_p)^\top(WU_p)\Lambda^{\frac{1}{2}}$ has full rank p, which holds because $W^\top W$ has rank p (i.e., it is full rank) by assumption. In the last two lines we used well-known properties of the determinant and the fact that $\det^*(L) = \det(\Lambda)$, by definition.

C.3 Proof of Lemma 9

To prove this lemma, we use a generalized version of Ostrowski's inequality (Higham & Cheng, 1998, Thm.3.2) to rectangular matrices. By that result, we get

$$\lambda_k\left(U_{(t-1)d}^\top P^\top P U_{(t-1)d}\right) = \eta_k \mu_k, \text{ for } k \in [(t-1)d],$$

where

$$\lambda_{td-k+1}\left(P^\top P\right) \le \mu_{(t-1)d-k+1} \le \lambda_{(t-1)d-k+1}\left(P^\top P\right),$$

and

$$\lambda_{(t-1)d}\left(U_{(t-1)d}^\top U_{(t-1)d}\right) \le \eta_k \le \lambda_1\left(U_{(t-1)d}^\top U_{(t-1)d}\right).$$

But, $\lambda_1\left(U_{(t-1)d}^\top U_{(t-1)d}\right) = \lambda_{(t-1)d}\left(U_{(t-1)d}^\top U_{(t-1)d}\right) = 1$, which implies that $\eta_k = 1$, for all $k \in [(t-1)d]$. From this, we deduce that

$$\begin{aligned}
\det\left((PU_{(t-1)d})^\top(PU_{(t-1)d})\right) &= \prod_{k=1}^{(t-1)d} \lambda_k\left((PU_{(t-1)d})^\top(PU_{(t-1)d})\right) \\
&= \prod_{k=1}^{(t-1)d} \mu_k \\
&\ge \prod_{k=1}^{(t-1)d} \lambda_{td-k+1}\left(P^\top P\right) \\
&= \prod_{k=d+1}^{td} \lambda_k\left(P^\top P\right) = \frac{\det\left(P^\top P\right)}{\prod_{k=1}^{d}\lambda_k\left(P^\top P\right)}.
\end{aligned}$$

C.4 Proof of Lemma 10

We will use A instead of A^* to ease notation. Note that, by definition,

$$P = \begin{bmatrix} A^{i_1-1} & A^{i_2-1} & \cdots & \cdots & A^{i_t-1} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & I_d \\ \vdots & A & \cdots & \cdots & 0 \\ A^{i_1-i_2} & I_d & \cdots & \cdots & \vdots \\ A^{i_1-i_2-1} & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ A & \vdots & \cdots & \cdots & \vdots \\ I_d & 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

We now define the following matrices in $\mathbb{R}^{i_1 d \times d}$

$$V_1 = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ 0 \\ A^{i_1-i_2-1} \\ \vdots \\ A \\ I_d \end{bmatrix} \left.\vphantom{\begin{matrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{matrix}}\right\} \begin{matrix} i_2\, d\times d \\ \text{zero blocks} \end{matrix}, V_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ A^{i_2-i_3-1} \\ \vdots \\ A \\ I_d \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}0\\0\\0\\0\end{matrix}}\right\} \begin{matrix} i_3\, d\times d \\ \text{zero blocks} \end{matrix} \\ \\ \\ \\ \\ \\ \left.\vphantom{\begin{matrix}0\\0\\0\\0\end{matrix}}\right\} \begin{matrix} (i_1-i_2)\, d\times d \\ \text{zero blocks} \end{matrix} \end{matrix}, \cdots, V_t = \begin{bmatrix} A^{i_t-1} \\ \vdots \\ I_d \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \left.\vphantom{\begin{matrix}0\\0\\0\\0\\0\\0\\0\\0\\0\end{matrix}}\right\} \begin{matrix} (i_1-i_t)\, d\times d \\ \text{zero blocks} \end{matrix}.$$

In other words, for $l \in [i_1 d]$, and $j \in [d]$,

$$(V_1)_{lj} = \begin{cases} 0 & \text{if } 1 \le l \le i_2 d \\ A_{lj}^{i_1 - \lceil l/d \rceil} & \text{if } i_2 d + 1 \le l \le i_1 d, \end{cases},$$

$$(V_k)_{lj} = \begin{cases} 0 & \text{if } 1 \le l \le i_{k+1} d \\ A_{lj}^{i_k - \lceil l/d \rceil} & \text{if } i_{k+1} d + 1 \le l \le i_k d \\ 0 & \text{if } i_k d + 1 \le l \le i_1 d \end{cases}, \quad \text{for } k \in \{2, \ldots, t-1\},$$

$$(V_t)_{lj} = \begin{cases} A_{lj}^{i_t - \lceil l/d \rceil} & \text{if } 1 \le i \le i_t \\ 0 & \text{if } i_t + 1 \le i \le i_1 \end{cases}.$$

From this definition, it is clear that the matrices $V_1, \ldots, V_t$ are pairwise orthogonal, i.e., $\langle V_k, V_{k'} \rangle_F = 0$, for $k, k' \in [t]$. We define the following shifting matrices, $\{S_{k,j-1}\}_{2 \le k \le t, 2 \le j \le t}$ in $\mathbb{R}^{t \times t}$,

$$(S_{k,j-1})_{ll'} = \begin{cases} 1 & \text{for } l = k, l' = j-1 \\ 0 & \text{otherwise} \end{cases}.$$

To see the effect of post-multiplying by this matrices, consider $U = \begin{bmatrix} u_1 & u_2 & \ldots & u_t \end{bmatrix} \in \mathbb{R}^{i_1 \times t}$. Then

$$US_{k,j-1} = \begin{bmatrix} 0 & \cdots & 0 & \underbrace{u_k}_{j-1 \text{ position}} & 0 & \cdots & 0 \end{bmatrix}.$$

In words, post-multiplying U by $S_{k,j-1}$ forms a new matrix with the same dimensions of U, with its $(j-1)$-th columns equal to the k-th column of U, and the rest of the columns are zero. With this, we express P as follows,

$$\begin{aligned}P =& \begin{bmatrix}V_1 & V_2 & \cdots & V_t\end{bmatrix} + \sum_{k=2}^{t}(I_{i_1} \otimes A^{i_1-i_k}) \begin{bmatrix}V_k & 0 & \cdots & 0\end{bmatrix} + \sum_{k=3}^{t}(I_{i_1} \otimes A^{i_2-i_k}) \begin{bmatrix}0 & V_k & 0 & \cdots & 0\end{bmatrix} \\ &+ \ldots + (I_{i_1} \otimes A^{i_{t-1}-i_t}) \begin{bmatrix}0 & \cdots & 0 & V_t & 0\end{bmatrix} \\ =& V + \sum_{j=2}^{t}\sum_{k=j}^{t}(I_{i_1} \otimes A^{i_{j-1}-i_k})V(S_{k,j-1} \otimes I_d),\end{aligned}$$

where $V := \begin{bmatrix}V_1 & V_2 & \cdots & V_t\end{bmatrix}$. On the other hand, for all $k, j \in [t]$,

$$(I_{i_1} \otimes A^{i_{j-1}-i_k})V(S_{k,j-1} \otimes I_d) = V(S_{k,j-1} \otimes I_d)(I_t \otimes A^{i_{j-1}-i_k}),$$

which implies,

$$\begin{aligned}P &= V + \sum_{j=2}^{t}\sum_{k=j}^{t} V(S_{k,j-1} \otimes I_d)(I_t \otimes A^{i_{j-1}-i_k}) \\ &= V + \sum_{j=2}^{t}\sum_{k=j}^{t} V\left(S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right) \\ &= V\left(I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right).\end{aligned}$$

With this, we have

$$P^\top P = \left(I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right)^\top V^\top V \left(I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right).$$

Given that $I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}$ and $V^\top V$ are square matrices, we have

$$\det\left(P^\top P\right) = \det(V^\top V)\det\left(I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right)^2.$$

The term $\sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}$ only contains matrices of the form $S_{k,j-1}$, with $k \geq j$, which are all strictly lower triangular. Then it is easy to see that $I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}$ is a block-lower triangular matrix with I_d's on its main block-diagonal. On the other hand, the determinant of a block-lower triangular matrix equals the determinant of the block-diagonal matrix formed by its diagonal blocks (see (Horn & Johnson, 2012, Section 0.9.4)). This implies that

$$\det\left(I_{td} + \sum_{j=2}^{t}\sum_{k=j}^{t} S_{k,j-1} \otimes A^{i_{j-1}-i_k}\right) = \det\left(I_{td}\right) = 1.$$

On other hand, by the orthogonality of the blocks $V_1, \ldots, V_t$ that form V, it is easy to see that $V^\top V$ is block diagonal, of the form,

$$V^\top V = \text{blkdiag}\left(V_1^\top V_1, V_2^\top V_2, \ldots, V_t^\top V_t\right) = \begin{bmatrix} V_1^\top V_1 & 0 & \cdots & \cdots & 0 \\ 0 & V_2^\top V_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & V_t^\top V_t \end{bmatrix}.$$

Given that the determinant of a block diagonal matrix is the product of the determinant of its blocks, we obtain

$$\det\left(P^\top P\right) = \det(V^\top V) = \prod_{k=1}^{t} \det(V_k^\top V_k),$$

where we note that

$$V_k^\top V_k = \begin{cases} \sum_{l=0}^{i_k - i_{k-1} - 1} (A^l)^\top A^l, & \text{for } k \in [t-1] \\ \sum_{l=0}^{i_t - 1} (A^l)^\top A^l, & \text{for } k = t. \end{cases}$$

In particular, since $V_k^\top V_k \succeq I_d$ for each k, this implies $\det\left(P^\top P\right) \geq 1$.

C.5 Proof of Lemma 11

We will again use A instead of A^* to ease notation. Notice that $P^\top P$ has the following block structure

$$P^\top P = \begin{bmatrix} P_{i_1}^\top P_{i_1} & P_{i_1}^\top P_{i_2} & \cdots & P_{i_1}^\top P_{i_t} \\ P_{i_2}^\top P_{i_1} & P_{i_2}^\top P_{i_2} & \cdots & P_{i_2}^\top P_{i_t} \\ \vdots & \ddots & \cdots & \vdots \\ P_{i_t}^\top P_{i_1} & P_{i_t}^\top P_{i_2} & \cdots & P_{i_t}^\top P_{i_t} \end{bmatrix}.$$

In order to bound the eigenvalues of $P^\top P$ we use (Tretter, 2008, Theorem 1.13.1) which generalizes the Gershgorin disk theorem to the block matrix case. In particular, it says that

$$\text{spec}(P^\top P) \in \cup_{k=1}^{t} \mathcal{G}_k,$$

where

$$\mathcal{G}_k := \text{spec}(P_{i_k}^\top P_{i_k}) \cup \left\{ \lambda \notin \text{spec}(P_{i_k}^\top P_{i_k}) : dist\big(\lambda, \text{spec}(P_{i_k}^\top P_{i_k})\big) \leq \sum_{\substack{k'=1 \\ k' \neq k}}^{t} \|P_{i_k}^\top P_{i_k}\|_2 \right\}.$$

Using the above result, we have the following estimates (we use A instead of A^* to ease notation).

(a) For $k \in [t]$, we have

$$P_{i_k}^\top P_{i_k} = \sum_{j=0}^{i_k - 1} (A^j)^\top A^j = I_d + \sum_{j=1}^{i_k - 1} (A^j)^\top A^j, \text{ and}$$

$$\left\| \sum_{j=1}^{i_k - 1} (A^j)^\top A^j \right\|_2 \leq \sum_{j=1}^{i_k - 1} \|A\|_2^{2j} = \frac{\|A\|_2^2}{1 - \|A\|_2^2} =: \delta(A).$$

Then, $\text{spec}(P_{i_k}^\top P_{i_k}) \in [1 - \delta(A), 1 + \delta(A)]$.

(b) For $k', k \in [t]$, with $k \neq k'$, we distinguish the following two cases.

(b.1) For k' such that $i_{k'} < i_k$, we have

$$P_{i_k}^\top P_{i_{k'}} = A^{i_k - i_{k'}} \left(I + A^2 + A^4 + \ldots + A^{2(i_{k'} - 1)}\right),$$

which implies that

$$\|P_{i_k}^\top P_{i_{k'}}\|_2 \leq \|A\|_2^{i_k - i_{k'}} \left(\frac{1 - \|A\|_2^{2 i_{k'}}}{1 - \|A\|_2^2} \right) \leq \frac{\|A\|_2^{i_k - i_{k'}}}{1 - \|A\|_2^2}.$$

From above, we obtain

$$\sum_{k' : i_{k'} < i_k} \|P_{i_k}^\top P_{i_{k'}}\|_2 \leq \frac{\|A\|_2}{(1 - \|A\|_2)(1 - \|A\|_2^2)}.$$

(b.2) For k' such that $i_{k'} > i_k$, we have

$$P_{i_k}^\top P_{i_{k'}} = A^{i_{k'}-i_k}\big(I + A^2 + A^4 + \ldots + A^{2(i_k-1)}\big),$$

from which we obtain

$$\|P_{i_k}^\top P_{i_{k'}}\|_2 \le \|A\|_2^{i_{k'}-i_k}\left(\frac{1-\|A\|_2^{2i_k}}{1-\|A\|_2^2}\right) \le \frac{\|A\|_2^{i_k-i_{k'}}}{1-\|A\|_2^2}.$$

Hence,

$$\sum_{k':i_{k'}>i_k} \|P_{i_k}^\top P_{i_{k'}}\|_2 \le \frac{\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)}.$$

Combining the estimates in (b.1) and (b.2), we obtain

$$\sum_{k':i_{k'}\neq i_k} \|P_{i_k}^\top P_{i_{k'}}\|_2 \le \frac{2\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)} := \kappa(A).$$

From the above calculations, we see that (let $\lambda_j^{(k)}$ lie in $\mathrm{spec}(P_{i_k}^\top P_{i_k})$)

$$\begin{aligned}
\mathcal{G}_k &\subseteq [1-\delta(A), 1+\delta(A)] \cup \left\{\cup_{j=1}^d \{\lambda \neq \lambda_j^{(k)} \in \mathrm{spec}(P_{i_k}^\top P_{i_k}) : |\lambda - \lambda_j^{(k)}| \le \kappa(A)\}\right\} \\
&\subseteq [1-\delta(A)-\kappa(A), 1+\delta(A)+\kappa(A)] \\
&= \left[1 - \frac{\|A\|_2^2}{1-\|A\|_2^2} - \frac{2\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)}, 1 + \frac{\|A\|_2^2}{1-\|A\|_2^2} + \frac{2\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)}\right]
\end{aligned}$$

Now,

$$\begin{aligned}
1 + \frac{\|A\|_2^2}{1-\|A\|_2^2} + \frac{2\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)} &\le \frac{1}{1-\|A\|_2^2} + \frac{2\|A\|_2}{(1-\|A\|_2)(1-\|A\|_2^2)} \\
&\le \frac{1-\|A\|_2^2}{(1-\|A\|_2^2)(1-\|A\|_2)} \\
&\le \frac{1}{(1-\|A\|_2)^2}.
\end{aligned}$$

Then,

$$\lambda_1\left(P^\top P\right) \le \frac{1}{(1-\|A\|_2)^2}.$$

D Proof of Theorem 1

The proof of Theorem 1 follows directly from the next three lemmas.

Lemma 12. *Let $s_0 := 2^{1/d}$, and assume A^* satisfies $\|A^*\|_2 < 1$. Suppose that*

$$\sigma^2 \le \frac{(1-\|A^*\|_2)^5}{4(s_0^{\omega(1)}T^{4/d}-1)}.$$

Then we have $\mathbb{E}[|\mathcal{E}|] \to 0$, when $T \to \infty$. In particular, $|\mathcal{E}| = 0$ with high probability.

Proof. From Proposition 2, we get

$$\begin{aligned}
\log\det\left(\frac{(1-\|A^*\|_2)^3}{4\sigma^2}L + I_{i_1 d}\right)^{-\frac{1}{2}} &\le -\frac{d}{2}(t-1)\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right) \\
&\le -\frac{d}{4}t\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right),
\end{aligned}$$

where the last inequality follows from the fact that $2(t-1) \geq t$, for $t \geq 2$. From this, we deduce,

$$\begin{aligned} t\log T + \log\det\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}L + I_{i_1 d}\right)^{-\frac{1}{2}} &\leq (t\log T)\left(1 - \frac{d}{4\log T}\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)\right) \\ &\leq (t\log T)\left(1 - \frac{d}{4\log T}\log s_0^{\omega(1)}T^{\frac{4}{d}}\right) \\ &\leq (t\log T)\left(-\omega(1)\frac{\log 2}{\log T}\right) \\ &\leq -t\omega(1). \end{aligned} \tag{D.1}$$

From (3.2), (3.9) and the fact that number of t-cycles in $[T]$ is bounded by T^t/t, we obtain,

$$\begin{aligned} \mathbb{E}[|\mathcal{E}|] &\leq \sum_{t=2}^{T}\exp\left(t\log T + \log\det\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}L + I_{i_1 d}\right)^{-\frac{1}{2}}\right) \\ &\leq \sum_{t=2}^{T}(e^{-\omega(1)})^t = o(1), \end{aligned}$$

where in the second line we used (D.1). The high probability statement follows directly from Markov's inequality. □

When $\|A^*\|_2 = 0$, our result recovers up-to-constants the results in Kunisky & Niles-Weed (2022). Similarly, we have an analogous result to (Kunisky & Niles-Weed, 2022, Lemma 3.4), which leads to a constant error upper bound.

Lemma 13. *Let $s_0 = 2^{1/d}$, and assume A^* satisfies $\|A^*\|_2 < 1$. Suppose that*

$$\sigma^2 \leq \frac{(1-\|A^*\|_2)^5}{4(s_0^{O(1)}T^{4/d}-1)}.$$

Then, $\mathbb{E}[|\mathcal{E}|] = O(1)$. In particular, for any function $f(T) = w(1)$, we have $|\mathcal{E}| \leq f(T)$ with high probability.

Proof. The proof mimics the argument in the proof of Lemma 12, but exchanging $\omega(\cdot)$ by $\mathcal{O}(\cdot)$. This shows that $\mathbb{E}[|\mathcal{E}|] = \mathcal{O}(1)$, and the rest of the argument follows from an application of Markov's inequality. □

We now establish an analogue of (Kunisky & Niles-Weed, 2022, Lemma 3.5). Our proof proceeds in the same manner as the previous lemmas. Interestingly, in Kunisky & Niles-Weed (2022) the argument for this lemma differs slightly from the others, whereas in our case no such modifications are required.

Lemma 14. *Let $s_0 = 2^{1/d}$, and assume A^* satisfies $\|A^*\|_2 < 1$. Suppose that*

$$\sigma^2 \leq \frac{(1-\|A^*\|_2)^5}{4(s_0^{\omega(1)}T^{2/d}-1)}.$$

Then,

$$\mathbb{E}[|\mathcal{E}|] = \mathcal{O}\left(\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)^{-\frac{d}{2}}T^2\right).$$

Proof. Recall that by Proposition 2, we have

$$\begin{aligned} t\log T + \log\det\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}L + I_{i_1 d}\right)^{-\frac{1}{2}} &\leq t\log T - \frac{d}{2}(t-1)\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right) \\ &= 2\log T + (t-2)\log T - \frac{d}{2}(t-2)\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right) \\ &\quad - \frac{d}{2}\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right) \\ &\leq 2\log T - \frac{d}{2}\log\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right) - \omega(1)(t-2). \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E}[|\mathcal{E}|] &\leq \sum_{t=2}^{T}\exp\left(t\log T + \log\det\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}L + I_{i_1 d}\right)^{-\frac{1}{2}}\right) \\ &\leq \sum_{t=2}^{T}\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)^{-\frac{d}{2}} T^2 (e^{-\omega(1)})^{t-2} \\ &= \mathcal{O}\left(\left(\frac{(1-\|A^*\|_2)^5}{4\sigma^2}+1\right)^{-\frac{d}{2}} T^2\right). \end{aligned}$$

□

E Additional experiments

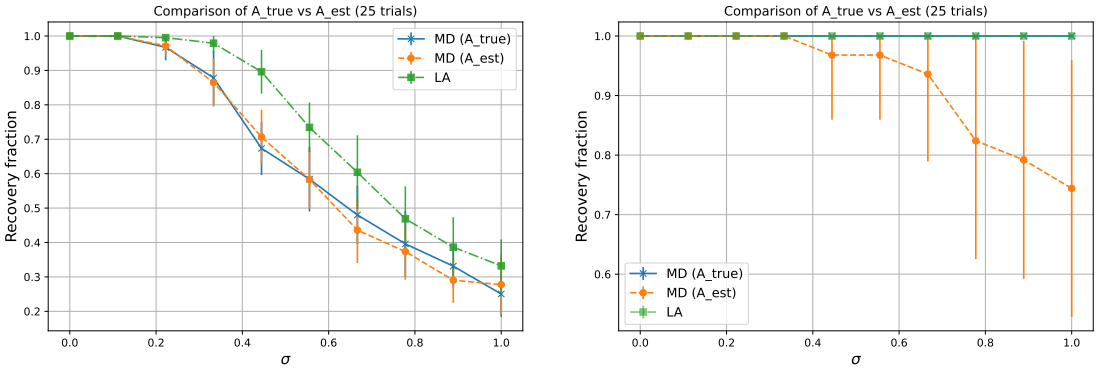
In this section, we include additional experiments that complement the experiments in the main paper. In Section E.1 we show the performance of the estimation method described in Remark 10. Finally, in Section E.2 we implement other algorithms for the Birkhoff polytope relaxation and show that they all have similar performance.

E.1 Estimate A^* first

One natural approach for estimating Π consists in the two-step strategy described in Remark 10. Here, first A^* is estimated using only the time series $(x_t)_{t\in[T]}$, for which the right order is known, via least-squares (the MLE for estimating A^* under the **VAR** model). Call this estimator $\widehat{A}_{\text{MLE}}$. Then we solve (4.1) replacing A with $\widehat{A}_{\text{MLE}}$. We show the results in the regimes $d = 5, T = 50$ and $d = 50, T = 5$, in Figure 8 using the Mirror Descent (MD) algorithm for the simplex relaxation. We observe that in the regime $d = 5$, $T = 50$, the performance under the true and the estimated A^* are very similar, likely because the time series is long enough so that $\widehat{A}_{\text{MLE}}$ is already close to A^*. On the other hand, for $d = 50$, $T = 5$, the performance of `RelaxMLE-Round` with $\widehat{A}_{\text{MLE}}$ degrades as the noise increases. In the considered regime, the LA estimator and `RelaxMLE-Round` with A^* both achieve perfect recovery.

E.2 Other algorithms for the Birkhoff relaxation

We consider different optimization schemes for the Birkhoff relaxation, including a Frank–Wolfe method, an ADMM-based approach, and a quasi-Newton variant. The Frank–Wolfe algorithm (Frank & Wolfe, 1956) performs iterative linear minimization over the Birkhoff polytope using the gradient of the objective and a line search step, offering a projection-free alternative particularly suited for large-scale problems. The ADMM algorithm (Stephen Boyd & Eckstein, 2011) enforces the Birkhoff constraints via alternating updates of primal and dual variables, with efficient projections implemented through Dykstra's algorithm. Finally, the quasi-Newton method (Byrd et al., 1995) applies an L-BFGS step on the flattened permutation matrix followed by projection onto the Birkhoff polytope, providing a curvature-aware but more computationally



(a) $d = 5$, $T = 50$ **(b)** $d = 50$, $T = 5$

Figure 8: Recovery fraction vs. σ for the MD algorithm, which solves the simplex relaxation. Here $\theta = 0.5$ and the average is computed over 30 Monte Carlo runs. In Fig. 8b both the MD estimator with access to A^* and the LA estimator achieve perfect recovery.

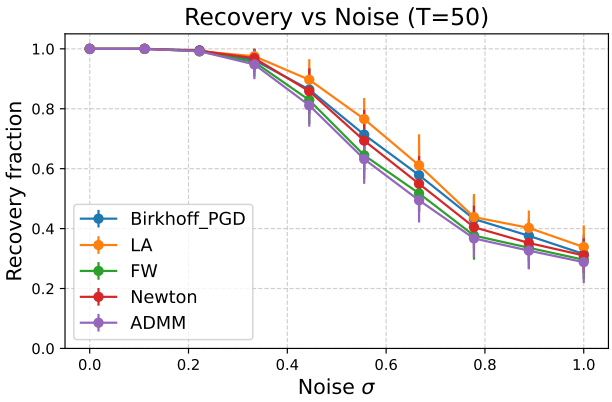


Figure 9: Comparison between different algorithms to solve (4.1) on the Birkhoff polytope. We evaluate them in the setting of Algorithm 1. Here we consider the setting of known $d = 5, T = 50$, $\theta = 0.5$ and different levels of noise σ.

demanding alternative. In Figure 9, we compare all three methods for $d = 5$, $T = 50$, and $\theta = 0.5$. We run algorithm 1, assuming a known A^*, which is sampled at random according to (5.3). In this setting, all algorithms achieve similar recovery performance, and this pattern remains consistent across the different parameter combinations we tested.