HAL open science

Matrix Factorization for Multivariate Time Series Analysis

Pierre Alquier, Nicolas Marie

To cite this version:

Pierre Alquier, Nicolas Marie. Matrix Factorization for Multivariate Time Series Analysis. Electronic Journal of Statistics , 2019, 13 (2), pp.4346-4366. 10.1214/19-EJS1630 . hal-02068455v2

HAL Id: hal-02068455

https://hal.science/hal-02068455v2

Submitted on 6 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MATRIX FACTORIZATION FOR MULTIVARIATE TIME SERIES ANALYSIS

PIERRE ALQUIER* AND NICOLAS MARIE†

Abstract. Matrix factorization is a powerful data analysis tool. It has been used in multivariate time series analysis, leading to the decomposition of the series in a small set of latent factors. However, little is known on the statistical performances of matrix factorization for time series. In this paper, we extend the results known for matrix estimation in the i.i.d setting to time series. Moreover, we prove that when the series exhibit some additional structure like periodicity or smoothness, it is possible to improve on the classical rates of convergence.

Acknowledgements. The authors gratefully acknowledge Maxime Ossonce for discussions which helped to improve the paper. The first author was working at CREST, ENSAE Paris when this paper was written; he gratefully acknowledges financial support from Labex ECODEC (ANR-11-LABEX-0047).

1. Introduction

Matrix factorization is a very powerful tool in statistics and data analysis. It was used as early as in the 70's in econometrics in reduced-rank regression [27, 22, 29]. There, matrix factorization is mainly a tool to estimate a coefficient matrix under a low-rank constraint. There was recently a renewed interest in matrix factorization as a data analysis tool for huge datasets. Nonnegative matrix factorization (NMF) was introduced by [35] as a tool to represent a huge number of objects as linear combinations of elements of "parts" of objects. The method was indeed applied to large facial image datasets and the dictionary indeed contained typical parts of faces. Since then, various methods of matrix factorization were successfully applied such various fields as collaborative filtering and recommender systems on the Web [34, 51], document clustering [46], separation of sources in audio processing [42], missing data imputation [26], quantum tomography [23, 25, 53, 7, 39], medical image processing [21] topics extraction in texts [43] or transports data analysis [12]. Very often, matrix factorization provides interpretable and accurate representations of the data matrix as the product of two much smaller matrices. The theoretical performances of matrix completion were studied in a series of papers by Candès with many co-authors [10, 11, 9]. Minimax rates for matrix completion and more general matrix estimation problems were derived in [32, 8, 30, 31, 41]. Bayesian estimators and aggregation procedures were studied in [1, 47, 38, 2, 37, 4, 17, 36, 16, 15].

To apply matrix factorization techniques to multivariate time series is a very natural idea. First, the low-rank structure induced by the factorization leads to high correlations that are indeed observed in some applications (this structure is actually at the core of cointegration models in econometrics [20, 28, 5]). Moreover, the factorization provides a decomposition of each series in a dictionary which member that can be interpreted as latent factors used for example in state-space models, see e.g. Chapter 3 in [33]. For this reasons, matrix factorization was used in multivariate time series analysis beyond econometrics: electricity consumptions

forecasting [18, 40], failure detection in transports systems [48], collaborative filtering [24], social media analysis [45] to name a few.

It is likely that the temporal structure in the data can be exploited to obtain an accurate and sensible factorization: autocorrelation, smoothness, periodicity... Indeed, while some authors use matrix factorization as a black box for data analysis, others propose in a way or another to adapt the algorithm to the temporal structure of the data [54, 45, 14, 24]. However, there is no theoretical guarantee that this leads to better predictions or better rates of convergence. Moreover, the aforementionned theoretical studies [10, 11, 9, 32, 8, 30, 31, 41] all assumed i.i.d noise, strongly limiting their applicability to study algorithms designed for time series such as in [54]. The objective of this paper is to address both issues.

Consider for example that one observes a d series $(x_{i,t})_{t=1}^T = \mathbf{X}$ and assume that $\mathbf{X} = \mathbf{M} + \varepsilon$ where $\mathbf{M}$ is a rank k matrix and ε is some noise. In a first time, assume that entries $\varepsilon_{i,t}$ of ε are i.i.d with variance σ^2. Theorem 3 in [32] implies that there is an estimator $\hat{\mathbf{M}}_1$ of $\mathbf{M}$, such that $\frac{1}{dT}\|\hat{\mathbf{M}}_1 - \mathbf{M}\|_F^2 = \mathcal{O}(\sigma^2 \frac{k(d+T)}{dT})$, up to log terms. Moreover, Theorem 5 in the same paper shows that this rate cannot be improved. Here, we propose an estimator $\hat{\mathbf{M}} = \hat{\mathbf{U}}\hat{\mathbf{V}}$, where $\hat{\mathbf{U}}$ is a $d \times k$ matrix and $\hat{\mathbf{V}}$ is $k \times T$. We study this estimator under the assumption that the rows $\varepsilon_{i,\cdot}$ of ε are independent, centered, with covariance matrix Σ_ε, allowing a temporal dependence in the noise. We prove that $\frac{1}{dT}\|\hat{\mathbf{M}} - \mathbf{M}\|_F^2 = \mathcal{O}(\|\Sigma_\varepsilon\|_{\text{op}} \frac{k(d+T)}{dT})$ where $\|\Sigma_\varepsilon\|_{\text{op}}$ is the operator norm of Σ_ε. Note that in the i.i.d case $\Sigma_\varepsilon = \sigma^2 \mathbb{I}_T$, we recover the rate of [32] as $\|\Sigma_\varepsilon\|_{\text{op}} = \sigma^2$. However, our result is more general: we provide examples where the noise is non i.i.d and we still have a control on $\|\Sigma_\varepsilon\|_{\text{op}}$. For example, when the noise is row-wise AR(1), that is $\varepsilon_{i,t+1} = \rho\varepsilon_{i,t} + \eta_{i,t}$ where the $\eta_{i,t}$ are i.i.d with variance σ^2 and $|\rho| < 1$, we have $\|\Sigma_\varepsilon\|_{\text{op}} \leqslant \sigma^2 \frac{1+|\rho|}{1-|\rho|}$. Moreover, our estimator can be tuned to take into account a possible periodicity or smoothness of the series. This is done by rewriting $\hat{\mathbf{W}} = \hat{\mathbf{V}}\mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is a $\tau \times T$ matrix encoding the temporal structure, and $\tau \leqslant T$. In this case, we always improve on the rate $\mathcal{O}(\|\Sigma_\varepsilon\|_{\text{op}} \frac{k(d+T)}{dT})$.

We obtain the following rates, for some constant $C(\beta, L)$:

	no structure	τ-periodic case	β-smooth case
order of $\frac{1}{dT}\|\hat{\mathbf{M}} - \mathbf{M}\|_F^2$	$\frac{\|\Sigma_\varepsilon\|_{\text{op}} k(d+T)}{dT}$	$\frac{\|\Sigma_\varepsilon\|_{\text{op}} k(d+\tau)}{dT}$	$\frac{\|\Sigma_\varepsilon\|_{\text{op}} kd}{dT} + \left(\frac{\|\Sigma_\varepsilon\|_{\text{op}} k}{dT}\right)^{\frac{2\beta}{2\beta+1}}$

All the results are first stated under a known structure, that is, we assume that we know the rank k, the period τ or the smoothness β of the series. We provide at the end of the paper a model selection procedure that allows to obtain the same rates of convergence without assuming this prior knowledge.

Finally, we should mention the nice paper [44] where the authors studied time-evolving adjacency matrices for graphs with autoregressive features. However, the rows of an adjacency matrix are not interpreted as time series, so the objective of this work is quite different from ours.

The paper is organized as follows. In Section 2 we introduce the notations that will be used in all the paper. In Subection 3.1 we study matrix factorization without additional temporal structure. In Section 3.2, we study the estimator $\hat{\mathbf{M}} = \hat{\mathbf{U}}\hat{\mathbf{V}}\mathbf{\Lambda}$ in the general case, and show how it improves the rates of convergence for a well chosen matrix $\mathbf{\Lambda}$ for periodic and/or smooth series. Finally, adaptation to unknown rank, periodicity and/or smoothness is tackled in Section 4. The proofs are given in Section 5.

2. Setting of the problem and notation

Assume that we observe a multivariate series

$$\mathbf{X} = (x_{i,t})_{(i,t)\in[\![1,d]\!]\times[\![1,T]\!]}.$$

where $d \in \mathbb{N}^*$ and $T \in \mathbb{N}\backslash\{0,1\}$. This multivariate series is modelled as a stochastic process. We actually assume that

$$\mathbf{X} = \mathbf{M} + \varepsilon, \tag{1}$$

where ε is a noise and $\mathbf{M}$ is a matrix of rank $k \in [\![1,T]\!]$. Then, there exist $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{W} \in \mathcal{M}_{k,T}(\mathbb{R})$ such that $\mathbf{M} = \mathbf{UW}$. We will refer to $\mathbf{W}$ as the *dictionary* or as the *latent series*.

We also want to model more structure in $\mathbf{M}$. This is done by rewritting $\mathbf{W} = \mathbf{V\Lambda}$, where $\tau \in \mathbb{N} \setminus \{0\}$, $\mathbf{V} \in \mathcal{M}_{k,\tau}$ and $\mathbf{\Lambda} \in \mathcal{M}_{\tau,T}$, where $\mathbf{\Lambda}$ is a *known* matrix. The matrix Λ depends on the structure assumed on $\mathbf{M}$.

Example 2.1 (Periodic series)**.** *Assume that $T = p\tau$ with $p \in \mathbb{N}^*$ for the sake of simplicity. To assume that the latent series in the dictionary $\mathbf{W}$ are τ-periodic is exactly equivalent to writing $\mathbf{W} = \mathbf{V\Lambda}$ where $\mathbf{V} \in \mathcal{M}_{k,\tau}(\mathbb{R})$ and $\mathbf{\Lambda} = (\mathbf{I}_\tau|\dots|\mathbf{I}_\tau) \in \mathcal{M}_{\tau,T}(\mathbb{R})$ is defined by blocks, $\mathbf{I}_\tau$ being the indentity matrix in $\mathcal{M}_{\tau,\tau}(\mathbb{R})$.*

Example 2.2 (Smooth series)**.** *We can assume that the series in $\mathbf{W}$ are smooth. For example, say that they belong a a Sobolev space with smoothness β, we have*

$$\mathbf{W}_{i,t} = \sum_{n=0}^{\infty} \mathbf{U}_{i,n}\mathbf{e}_n\left(\frac{t}{T}\right)$$

where $(\mathbf{e}_n)_{n\in\mathbb{N}}$ is the Fourier basis (the definition of a Sobolev space is reminded is Section 3.2 below). Of course, there are infinitely many coefficients $\mathbf{U}_{i,n}$ and to estimate them all is not feasible, however, for τ large enough, the approximation

$$\mathbf{W}_{i,t} \simeq \sum_{n=0}^{\tau-1} \mathbf{U}_{i,n}\mathbf{e}_n\left(\frac{t}{T}\right)$$

will be suitable, and can be rewritten as $\mathbf{W} = \mathbf{U\Lambda}$ where $\mathbf{\Lambda}_{i,t} = e_i(t/T)$. More details will be given in Section 3.2, where we actually cover more general basis of functions.

So our complete model will finally be written as

$$\mathbf{X} = \mathbf{M} + \varepsilon = \mathbf{UV\Lambda} + \varepsilon, \tag{2}$$

where $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{V} \in \mathcal{M}_{k,\tau}(\mathbb{R})$ are unknown, but $\tau \leqslant T$ and $\mathbf{\Lambda} \in \mathcal{M}_{\tau,T}(\mathbb{C})$ such that $\mathrm{rank}(\mathbf{\Lambda}) = \tau$ are known (note that the unstructured case corresponds to $\tau = T$ and $\mathbf{\Lambda} = \mathbf{I}_T$).

Note that more constraint can be imposed on the estimator. For example, in nonnegative matrix factorization [35], one imposes that all the entries in $\hat{\mathbf{U}}$ and $\hat{\mathbf{W}}$ are nonnegative. Here, we will more generally assume that $\hat{\mathbf{U}}\hat{\mathbf{V}}$ belongs to some prescribed subset $\mathcal{S} \subseteq \mathcal{M}_{d,T}(\mathbb{R})$.

In what follows, we will consider two norms on $\mathcal{M}_{d,T}$. For a matrix A, the Frobenius norms is given by

$$\|\mathbf{A}\|_F = \mathrm{trace}(\mathbf{AA}^*)^{1/2}.$$

and the operator norm by

$$\|\mathbf{A}\|_{\mathrm{op}} = \sup_{\|x\|=1} \|\mathbf{A}x\|$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^T$.

2.1. **Estimation by empirical risk minimization.** By multiplying both sides in (2) by the pseudo-inverse $\mathbf{\Lambda}^+ = \mathbf{\Lambda}^*(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{-1}$, we obtain the "simplified model"

$$\widetilde{\mathbf{X}} = \widetilde{\mathbf{M}} + \widetilde{\varepsilon}$$

with $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda}^+$, $\widetilde{\mathbf{M}} = \mathbf{U}\mathbf{V}$ and $\widetilde{\varepsilon} = \varepsilon\mathbf{\Lambda}^+$. In this model, the estimation of $\widetilde{\mathbf{M}}$ can be done by empirical risk minimization:

$$\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} \in \arg\min_{\mathbf{A}\in\mathcal{S}} \widetilde{r}(\mathbf{A}) \tag{3}$$

where

$$\widetilde{r}(\mathbf{A}) = \|\mathbf{A} - \widetilde{\mathbf{X}}\|_F^2 \ ; \forall \mathbf{A} \in \mathcal{M}_{d,\tau}(\mathbb{R}).$$

Therefore, we can define the estimator $\widehat{\mathbf{M}}_{\mathcal{S}} = \widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}}\mathbf{\Lambda}$ of $\mathbf{M}$.

In Section 3, we study the statistical performances of this estimator. The first step is done in Subsection 3.1, where we derive upper bounds on

$$\left\|\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} - \widetilde{\mathbf{M}}\right\|_F^2.$$

The corresponding upper bounds on

$$\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2$$

are derived in Subsection 3.2.

3. Oracle inequalities

Throughout this section, assume that ε fulfills the following..

Assumption 3.1. *The rows of ε are independent and have the same T-dimensional sub-Gaussian distribution, with second moment matrix Σ_ε. Moreover, $\varepsilon_{1,.}\Sigma_\varepsilon^{-1/2}$ is isotropic and has a finite sub-Gaussian norm*

$$\mathfrak{K}_\varepsilon := \sup_{\|x\|=1} \sup_{p\in[1,\infty[} p^{-1/2}\mathbb{E}(|\langle \varepsilon_{1,.}\Sigma_\varepsilon^{-1/2}, x\rangle|^p)^{1/p} < \infty.$$

In the sequel, we also consider $\widetilde{\mathfrak{K}}_\varepsilon := \mathfrak{K}_\varepsilon^2 \vee \mathfrak{K}_\varepsilon^4$.

We remind (see e.g Chapter 1 in [13]) that when $X \sim \mathcal{N}(0, \mathbf{I}_n)$,

$$\sup_{\|x\|=1} \sup_{p\in[1,\infty[} p^{-1/2}\mathbb{E}(|\langle X, x\rangle|^p)^{1/p} = C \tag{4}$$

for some universal constant $C > 0$ (that is, C does not depend on n). Thus, for Gaussian noise, Assumption 3.1 is satisfied and $\mathfrak{K}_\varepsilon = C$ does not depend on the dimension T.

3.1. **The case $\mathbf{\Lambda} = \mathbf{I}_T$.** In this subsection only, we assume that $\mathbf{\Lambda} = \mathbf{I}_T$ (and thus $\tau = T$). So the simplified model is actually the original model $\widetilde{\mathbf{X}} = \mathbf{X}$, $\widetilde{\mathbf{M}} = \mathbf{M}$, $\widetilde{\varepsilon} = \varepsilon$ and $\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} = \widehat{\mathbf{M}}_{\mathcal{S}}$.

Theorem 3.2. *Under Assumption 3.1, for every $\lambda \in]0,1[$ and $s \in \mathbb{R}_+$,*

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} \frac{1}{dT}\|\mathbf{A} - \mathbf{M}\|_F^2 + \frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)} \cdot \frac{k(d+T+s)}{dT}$$

with probability larger than $1 - 2e^{-s}$.

As a consequence, if we have indeed $\mathbf{M} \in \mathcal{S}$, then with large probability,

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 \leqslant \frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)} \cdot \frac{k(d+T+s)}{dT}.$$

Thus, we recover the rate $\mathcal{O}(\|\Sigma_\varepsilon\|_{\mathrm{op}} \frac{k(d+T)}{dT})$ claimed in the introduction.

Remark 3.3. *Since the bound relies on the constant $\|\Sigma_\varepsilon\|_{\mathrm{op}}$, let us provide its value in some special cases:*

(1) *If $\mathrm{cov}(\varepsilon_{1,t}, \varepsilon_{1,t'}) = \sigma^2 \mathbf{1}_{\{t=t'\}}$ then*

$$\|\Sigma_\varepsilon\|_{op} = \sigma^2.$$

More generally, when $\varepsilon_{1,1}, \dots, \varepsilon_{1,T}$ are uncorrelated,

$$\|\Sigma_\varepsilon\|_{op} = \max_{t \in [\![1,T]\!]} \mathrm{var}(\varepsilon_{1,t}).$$

(2) *Let $(\eta_t)_{t\in\mathbb{Z}}$ be a white noise of standard deviation $\sigma > 0$ and assume that there exists $\theta \in \mathbb{R}^*$ such that $\varepsilon_{1,t} = \eta_t - \theta\eta_{t-1}$ for every $t \in [\![1,T]\!]$. In other words, $(\varepsilon_{1,t})_{t=1,\dots,T}$ is the restriction of a MA(1) process to $[\![1,T]\!]$. So,*

$$\Sigma_\varepsilon = \sigma^2 \begin{pmatrix} 1+\theta^2 & -\theta & 0 & \dots & 0 \\ -\theta & 1+\theta^2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 1+\theta^2 & -\theta \\ 0 & \dots & 0 & -\theta & 1+\theta^2 \end{pmatrix}$$

and then

$$\|\Sigma_\varepsilon\|_{op} = \sigma^2 \left[1 + \theta^2 - 2\theta \min_{\ell \in [\![1,T]\!]} \cos\left(\frac{\ell\pi}{1+T}\right)\right] \leqslant \sigma^2(1+\theta)^2.$$

(3) *Let $(\eta_t)_{t\in\mathbb{Z}}$ be a white noise of standard deviation $\sigma > 0$ and assume that there is a ρ with $|\rho| < 1$ such that $\varepsilon_{1,t} = \rho\varepsilon_{1,t-1} + \eta_t$. So $(\varepsilon_{1,t})_{t=1,\dots,T}$ is the restriction of a AR(1) process to $[\![1,T]\!]$. So,*

$$\Sigma_\varepsilon = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{T-2} & \dots & \rho & 1 & \rho \\ \rho^{T-1} & \dots & \rho^2 & \rho & 1 \end{pmatrix} = \sigma^2 \left[\mathbf{I}_T + \sum_{t=1}^{T-1} \rho^t \left(\mathbf{J}_T^t + (\mathbf{J}_T^*)^t\right)\right].$$

where

$$\mathbf{J}_T = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 1 \\ 0 & \dots & 0 & 0 & 0 \end{pmatrix}.$$

As $\|\mathbf{J}_T\|_{\mathrm{op}} = 1$, we have

$$\|\Sigma_\varepsilon\|_{\mathrm{op}} \leqslant \sigma^2 \left(1 + 2\sum_{t=1}^{T} |\rho|^t\right) \leqslant \sigma^2 \left(1 + \frac{2|\rho|}{1-|\rho|}\right) = \sigma^2 \frac{1+|\rho|}{1-|\rho|}.$$

3.2. **The general case.** Let us now come back to the general case. An application of Theorem 3.2 to the "simplified model" (2.1) shows that for any $\lambda \in]0,1[$ and $s \in \mathbb{R}_+$,

$$\left\|\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} - \widetilde{\mathbf{M}}\right\|_F^2 \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} \left\|\mathbf{A}\mathbf{\Lambda} - \widetilde{\mathbf{M}}\right\|_F^2 + \frac{4\mathfrak{c}k}{\lambda(1-\lambda)}(d+\tau+s)\widetilde{\mathfrak{K}}_{\widetilde{\varepsilon}}\|\Sigma_{\varepsilon\mathbf{\Lambda}^+}\|_{\mathrm{op}} \tag{5}$$

with probability larger than $1-2e^{-s}$.

In order to obtain the desired bound on $\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\|_F^2$, we must now understand the behaviour of $\|\Sigma_{\varepsilon\mathbf{\Lambda}^+}\|_{\mathrm{op}}$ and $\widetilde{\mathfrak{K}}_{\widetilde{\varepsilon}}$.

Lemma 3.4. *For any matrix* $\mathbf{C} \in \mathcal{M}_{T,\tau}(\mathbb{C})$,

$$\|\Sigma_{\varepsilon_{1,.}\mathbf{C}}\|_{\mathrm{op}} \leqslant \|\Sigma_\varepsilon\|_{\mathrm{op}}\|\mathbf{C}^*\mathbf{C}\|_{\mathrm{op}}.$$

The situation regarding $\widetilde{\mathfrak{K}}_{\widetilde{\varepsilon}} = \mathfrak{K}_{\widetilde{\varepsilon}}^2 \vee \mathfrak{K}_{\widetilde{\varepsilon}}^4$ is different, we are not aware of a general simple upper bound on $\mathfrak{K}_{\widetilde{\varepsilon}} = \mathfrak{K}_{\varepsilon\mathbf{\Lambda}^+}$ in terms of $\mathfrak{K}_\varepsilon$ and $\mathbf{\Lambda}^+$. Still, there are two cases where we actually have $\mathfrak{K}_{\widetilde{\varepsilon}} = \mathfrak{K}_\varepsilon$. Indeed, in the Gaussian case, $\mathfrak{K}_{\widetilde{\varepsilon}} = \mathfrak{K}_\varepsilon = C$, see (4) above. For non Gaussian noise, we have the following result.

Lemma 3.5. *Assume that there is* $c(\tau,T) > 0$ *such that* $\mathbf{\Lambda}\mathbf{\Lambda}^* = c(\tau,T)\mathbf{I}_\tau$. *If* $\Sigma_\varepsilon = \sigma^2\mathbf{I}_T$ *with* $\sigma > 0$, *then* $\mathfrak{K}_{\widetilde{\varepsilon}} = \mathfrak{K}_\varepsilon$.

Note that the assumption on $\mathbf{\Lambda}$ is fullfilled by the examples covered in Subsections 3.3 and (3.4).

The previous discussion legitimates the following assumption.

Assumption 3.6. $\mathfrak{K}_{\widetilde{\varepsilon}} \leqslant \mathfrak{K}_\varepsilon$.

Finally, note that

$$\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 = \left\|(\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} - \widetilde{\mathbf{M}})\mathbf{\Lambda}\right\|_F^2 \leqslant \left\|\widehat{\widetilde{\mathbf{M}}}_{\mathcal{S}} - \widetilde{\mathbf{M}}\right\|_F^2 \|\mathbf{\Lambda}\mathbf{\Lambda}^*\|_{\mathrm{op}}$$

and in the same way

$$\left\|\mathbf{A} - \widetilde{\mathbf{M}}\right\|_F^2 = \left\|(\mathbf{A}\mathbf{\Lambda} - \mathbf{M})\mathbf{\Lambda}^+\right\|_F^2 \leqslant \|\mathbf{A}\mathbf{\Lambda} - \mathbf{M}\|_F^2 \, \|\mathbf{\Lambda}\mathbf{\Lambda}^+\|(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{-1}\|_{\mathrm{op}}.$$

By Inequality (5) together with Lemmas 3.4 and 3.5, we obtain the following result.

Corollary 3.7. *Fix* $\lambda \in]0,1[$ *and* $s \in \mathbb{R}_+$. *Under Assumption 3.1 and Assumption 3.6,*

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} \frac{1}{dT}\|\mathbf{A}\mathbf{\Lambda} - \mathbf{M}\|_F^2 + \frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)} \cdot \frac{k(d+\tau+s)}{dT}$$

with probability larger than $1-2e^{-s}$.

Corollary 3.7 provides an oracle inequality: it says that our estimator provides the optimal tradeoff between a variance term in $\|\Sigma_\varepsilon\|_{\mathrm{op}}k(d+\tau)/(dT)$, and a bias term. The bias term is the distance of $\mathbf{M}$ to its best approximation by a matrix of the form $\mathbf{A}\mathbf{\Lambda}$. In order to explicit the rates of convergence, assumptions can be made to upper-bound the bias term. We now apply Corollary 3.7 in the case of periodic time series, and then in the case of smooth time series. In each case, we explicit the bias term and the rate of convergence.

3.3. **Application: periodic time series.** In the case of τ-periodic time series, remind that we assumed for simplicity that there is an integer p such that $\tau p = T$ and we defined

$$\mathbf{\Lambda} = (\mathbf{I}_\tau | \dots | \mathbf{I}_\tau) \in \mathcal{M}_{\tau,T}(\mathbb{R}).$$

Then

$$\mathbf{\Lambda}\mathbf{\Lambda}^* = \frac{T}{\tau}\mathbf{I}_\tau \Rightarrow \|\mathbf{\Lambda}\mathbf{\Lambda}^*\|_{\mathrm{op}} = \frac{T}{\tau} \text{ and } \|(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{-1}\|_{\mathrm{op}} = \frac{\tau}{T}.$$

Therefore, by Corollary 3.7, for every $\lambda \in]0,1[$ and $s \in \mathbb{R}_+$, under Assumptions 3.1 and 3.6,

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} \frac{1}{dT}\|\mathbf{A}\mathbf{\Lambda} - \mathbf{M}\|_F^2 + \frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)} \cdot \frac{k(d+\tau+s)}{dT}$$

with probability larger than $1-2e^{-s}$. Now, define

$$\mathcal{S} = \{\mathbf{A} \in \mathcal{M}_{n,T}(\mathbb{R})\colon \operatorname{rank}(\mathbf{A}) \leqslant k \text{ and } \forall i, \forall t, \mathbf{A}_{i,t+\tau} = A_{i,t}\}$$

and assume that $\mathbf{M} \in \mathcal{S}$. Then,

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 = \mathcal{O}\left(\|\Sigma_\varepsilon\|_{\mathrm{op}} \frac{k(d+\tau)}{dT}\right)$$

which is indeed an improvement with respect to the rate obtained without taking the periodicity into account, that is $\mathcal{O}(\|\Sigma_\varepsilon\|_{\mathrm{op}} \frac{k(d+T)}{dT})$.

3.4. **Application: time series with smooth trend.** Assume we are given a dictionary of functions $(\mathbf{e}_n)_{|n|\leqslant N}$ for some finite $N \in \mathbb{N}$. This dictionary can for example be a finite subset of a basis of an Hilbert space $(\mathbf{e}_n)_{n\in\mathbb{Z}}$, like the Fourier basis or a wavelet basis.

Define

$$\mathbf{\Lambda}_N = \left(\mathbf{e}_n\left(\frac{t}{T}\right)\right)_{(n,t)\in[\![-N,N]\!]\times[\![1,T]\!]}.$$

Note that $\mathbf{\Lambda}_N$ is a $\tau \times T$ matrix where $\tau = 2N+1$.

Assume that

$$\mathbf{\Lambda}_N\mathbf{\Lambda}_N^* = T\mathbf{I}_\tau.$$

This implies that $\|(\mathbf{\Lambda}_N\mathbf{\Lambda}_N^*)^{-1}\|_{\mathrm{op}} = 1/T$ and $\|\mathbf{\Lambda}_N\mathbf{\Lambda}_N^*\|_{\mathrm{op}} = T$. This can be the case for a well-chosen basis, otherwise, we can apply the Gram-Schmidt to the dictionary of functions.

Example 3.8. *(Fourier's basis) Consider the Fourier basis $(\mathbf{e}_n)_{n\in\mathbb{Z}}$ defined by*

$$\mathbf{e}_n(x) = e^{2i\pi nx}\ ;\ \forall n \in \mathbb{Z},\ \forall x \in \mathbb{R}.$$

On the one hand, for every $n \in [\![-N,N]\!]$ and $t \in [\![1,T]\!]$, $|e_n(t/T)| = 1$. On the other hand, for every $m,n \in [\![-N,N]\!]$ such that $m \neq n$,

$$\begin{aligned}\sum_{t=1}^{T}\mathbf{e}_n\left(\frac{t}{T}\right)\overline{\mathbf{e}_m\left(\frac{t}{T}\right)} &= \sum_{t=1}^{T} e^{2i\pi(n-m)t/T} \\ &= \frac{e^{2i\pi(n-m)/T}(1-e^{2i\pi(n-m)})}{1-e^{2i\pi(n-m)/T}} = 0.\end{aligned}$$

Therefore, by Corollary 3.7, for every $\lambda \in]0,1[$ and $s \in \mathbb{R}_+$, under Assumptions 3.1 and 3.6,

$$\left\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\right\|_F^2 \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} \|\mathbf{A}\mathbf{\Lambda}_N - \mathbf{M}\|_F^2 + \frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)} \cdot \frac{k(d+(2N+1)+s)}{dT} \tag{6}$$

with probability larger than $1-2e^{-s}$. We will now show the consequences of these results when the rows of $\mathbf{M}$ are smooth in the sense that they belong to a given Sobolev ellipsoid. In this case, we will not have a $\mathbf{A}$ such that $\|\mathbf{A}\mathbf{\Lambda}-\mathbf{M}\|_F^2=0$, but this quantity will be small and can be controlled as a function of N. We introduce a few definitions.

Definition 3.9. *The Sobolev ellipsoid $W(\beta,L)$ is the set of functions $f:[0,1]\to\mathbb{R}$ such that f is $\beta-1$ times differentiable, $f^{(\beta-1)}$ is absolutely continuous and*

$$\int_0^1 f^{(\beta)}(x)dx \leqslant L^2.$$

From now, we assume that $\mathbf{e}_n(x)=e^{2i\pi nx}$ is the Fourier basis. It is well-known from Chapter 1 in [50] that any $f\in W(\beta,L)$ and $x\in[0,1]$,

$$f(x)=\sum_{n=-\infty}^{\infty} c_n(f)\mathbf{e}_n(x)$$

and that there is a (known) constant $C(\beta,L)>0$ such that

$$\frac{1}{T}\sum_{t=1}^{T}\left[f\left(\frac{t}{T}\right)-\sum_{n\leqslant|N|}c_n\mathbf{e}_n\left(\frac{t}{T}\right)\right]^2\leqslant C(\beta,L)N^{-2\beta}. \tag{7}$$

Definition 3.10. *We define $\mathcal{S}(k,\beta,L)\subset\mathcal{M}_{d,T}(\mathbb{R})$ as the set of matrices $\mathbf{M}$ such that $\mathbf{M}=\mathbf{U}\mathbf{W}$, $\mathbf{U}\in\mathcal{M}_{k,T}(\mathbb{R})$, $\mathbf{W}\in\mathcal{M}_{d,k}(\mathbb{R})$ and*

1. *For any $i\in[\![1,d]\!]$, $\|\mathbf{U}_{i,\cdot}\|^2\leqslant 1$,*
2. *For any $\ell\in[\![1,k]\!]$ and $t\in[\![1,T]\!]$, $\mathbf{W}_{\ell,t}=f_\ell\left(\frac{t}{T}\right)$ for some $f_\ell\in W(\beta,L)$.*

Denote $\mathbf{V}_{N,\mathbf{W}}=(c_n(f_\ell))_{\ell\leqslant k,|n|\leqslant N}$.

Then (7) implies

$$\frac{1}{dT}\|\mathbf{M}-\mathbf{U}\mathbf{V}_{N,\mathbf{W}}\mathbf{\Lambda}_N\|_F^2\leqslant C(\beta,L)N^{-2\beta}.$$

Pluging this into (6) gives

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}(k,\beta,L)}-\mathbf{M}\right\|_F^2\leqslant\frac{1+\lambda}{1-\lambda}\cdot C(\beta,L)N^{-2\beta}+\frac{4\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)}\cdot\frac{k(d+\tau+s)}{dT}.$$

If β is known, an adequate optimization with respect to N gives the following result.

Corollary 3.11. *Assume that $\mathbf{M}\in\mathcal{S}(k,\beta,L)$. Under Assumptions 3.1 and 3.6, the choice $N=\lfloor(dTC(\beta,L)/(\|\Sigma_\varepsilon\|_{\mathrm{op}}k))^{1/(2\beta+1)}\rfloor$ ensures*

$$\frac{1}{dT}\left\|\widehat{\mathbf{M}}_{\mathcal{S}(k,\beta,L)}-\mathbf{M}\right\|_F^2\leqslant\mathcal{C}\left[\|\Sigma_\varepsilon\|_{\mathrm{op}}\frac{kd+s}{dT}+C(\beta,L)^{\frac{1}{2\beta+1}}\left(\|\Sigma_\varepsilon\|_{\mathrm{op}}\frac{k}{dT}\right)^{\frac{2\beta}{2\beta+1}}\right]$$

with probability larger than $1-2e^{-s}$, where $\mathcal{C}>0$ is some constant depending on λ, $\mathfrak{c}$ and $\mathfrak{K}_\varepsilon$.

However, in practice, β is not known - nor the rank k. This problem is tackled in the next section.

4. Model selection

Assume that we have many possible matrices $\mathbf{\Lambda}_\tau$, for $\tau\in\mathcal{T}\subset\{1,\dots,T\}$ and for each τ, many possible $\mathcal{S}_{\tau,k}$ for different possible ranks $k\in\mathcal{K}\subset\{1,\dots,d\wedge T\}$.

Consider $s \in \mathbb{R}_+$ and the penalized estimator $\widehat{\mathbf{M}}_s = \widehat{\mathbf{M}}_{\mathcal{S}_{\hat{\tau}_s, \hat{k}_s}}$ with

$$(\widehat{\tau}_s, \widehat{k}_s) \in \arg \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}} \left\{ \left\| \widehat{\mathbf{M}}_{\mathcal{S}_{\tau,k}} - \mathbf{X} \right\|_F^2 + \mathrm{pen}_{s+\tau+k}(\tau, k) \right\},$$

where

$$\mathrm{pen}_s(\tau, k) = \frac{2\mathfrak{c}k}{\lambda}(d + \tau + s)\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}.$$

Theorem 4.1. *Under Assumptions 3.1 and 3.6, for every $\lambda \in]0, 1[$,*

$$\frac{1}{dT} \left\| \widehat{\mathbf{M}}_{\mathcal{S}_{\hat{\tau}_s, \hat{k}_s}} - \mathbf{M} \right\|_F^2 \leqslant \min_{\substack{(\tau,\, k) \,\in\, \mathcal{T} \,\times\, \mathcal{K} \\ \mathbf{A} \,\in\, \mathcal{S}_{\tau,k}}} \left\{ \left(\frac{1+\lambda}{1-\lambda} \right)^2 \frac{1}{dT} \|\mathbf{A}\mathbf{\Lambda}_\tau - \mathbf{M}\|_F^2 \right.$$
$$\left. + \frac{16\mathfrak{c}\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}}{\lambda(1-\lambda)^2} \cdot \frac{k(d+\tau+s)}{dT} \right\}.$$

with probability larger than $1 - 2e^{-s}$.

Remark 4.2. *The reader might feel uncomfortable with the fact that the model selection procedure leads to $\hat{k}_s$ and $\hat{\tau}_s$ that depend on the prescribed confidence level s. Note that if $k = k_0$ is known, that is $\mathcal{K} = \{k_0\}$, then it is clear from the definition that $\hat{\tau}_s$ actually does not depend on s.*

As an application, assume that $\mathbf{M} \in \mathcal{S}(k, \beta, L)$ where k is known, but β is unknown. Then the model selection procedure is feasible as it does not depend on β, and it satisfies exactly the same rate as $\hat{\mathbf{M}}_{\mathcal{S}(k,\beta,L)}$ in Corollary 3.11.

References

[1] P. Alquier. Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In *International Conference on Algorithmic Learning Theory*, pages 309–323. Springer, 2013.

[2] P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *arXiv preprint, to appear in the Annals of Statistics*, 2017.

[3] P. Alquier and P. Doukhan. Sparsity considerations for dependent variables. *Electronic journal of statistics*, 5:750–774, 2011.

[4] P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017.

[5] L. Bauwens and M. Lubrano. Identification restriction and posterior densities in cointegrated Gaussian VAR systems. In T. M. Fomby and R. Carter Hill, editors, *Advances in econometrics, vol. 11(B)*. JAI Press, Greenwich, 1993.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[7] T. Cai, D. Kim, Y. Wang, M. Yuan, and H. Zhou. Optimal large-scale quantum state tomography with Pauli measurements. *The Annals of Statistics*, 44(2):682–712, 2016.

[8] T. Cai and A. Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[9] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[10] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[11] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[12] L. Carel and P. Alquier. Non-negative matrix factorization as a pre-processing tool for travelers temporal profiles clustering. In *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 417–422, 2017.

[13] D. Chafaï, O. Guédon, G. Lecué, and A. Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*. Société Mathématique de France, 2012.

[14] V. Cheung, K. Devarajan, G. Severini, A. Turolla, and P. Bonato. Decomposing time series data by a non-negative matrix factorization algorithm with temporally constrained coefficients. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3496–3499. IEEE, 2015.

[15] S. Chrétien and B. Guedj. Revisiting clustering as matrix factorisation on the Stiefel manifold. *arXiv preprint arXiv:1903.04479*, 2019.

[16] A. S. Dalalyan. Exponential weights in multivariate regression and a low-rankness favoring prior. *arXiv preprint arXiv:1806.09405*, 2018.

[17] A. S. Dalalyan, E. Grappin, and Q. Paris. On the exponentially weighted aggregate with the Laplace prior. *The Annals of Statistics*, 46(5):2452–2478, 2018.

[18] Y. De Castro, Y. Goude, G. Hébrail, and J. Mei. Recovering multiple nonnegative time series from a few temporal aggregates. In *ICML 2017-34th International Conference on Machine Learning*, pages 1–9, 2017.

[19] J. Dedecker, P. Doukhan, G. Lang, L. R. J. Rafael, S. Louhichi, and C. Prieur. Weak dependence. In *Weak dependence: With examples and applications*, pages 9–20. Springer, 2007.

[20] R. F. Engle and C. W. J. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

[21] I. A. Genevera, L. Grosenick, and J. Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159, 2014.

[22] J. Geweke. Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75:121–146, 1996.

[23] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[24] S. Gultekin and J. Paisley. Online forecasting matrix factorization. *arXiv preprint arXiv:1712.08734*, 2017.

[25] M. Guţă, T. Kypraios, and I. Dryden. Rank-based model selection for multiple ions quantum tomography. *New Journal of Physics*, 14(10):105002, 2012.

[26] F; Husson, J. Josse, B. Narasimhan, and G. Robin. Imputation of mixed data with multilevel singular value decomposition. *arXiv preprint arXiv:1804.11087*, 2018.

[27] A. Izenman. Reduced rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

[28] F. Kleibergen and H. K. van Dijk. On the shape of the likelihood-posterior in cointegration models. *Econometric theory*, 10:514–551, 1994.

[29] F. Kleibergen and H. K. van Dijk. Bayesian simultaneous equation analysis using reduced rank structures. *Econometric theory*, 14:699–744, 1998.

[30] O. Klopp, K. Lounici, and A. B. Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017.

[31] O. Klopp, Y. Lu, A. B. Tsybakov, and H. H. Zhou. Structured matrix estimation and completion. *arXiv preprint arXiv:1707.02090*, 2017.

[32] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[33] G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4):267–358, 2010.

[34] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[35] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[36] A. Lumbreras, L. Filstroff, and C. Févotte. Bayesian mean-parameterized nonnegative binary matrix factorization. *arXiv preprint arXiv:1812.06866*, 2018.

[37] T. D. Luu, J. Fadili, and C. Chesneau. Sharp oracle inequalities for low-complexity priors. *arXiv preprint arXiv:1702.03166*, 2017.

[38] T. T. Mai and P. Alquier. A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1):823–841, 2015.

[39] T. T. Mai and P. Alquier. Pseudo-bayesian quantum tomography with rank-adaptation. *Journal of Statistical Planning and Inference*, 184:62–76, 2017.

[40] J. Mei, Y. De Castro, Y. Goude, J.-M. Azaïs, and G. Hébrail. Nonnegative matrix factorization with side information for time series recovery and prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[41] K. Moridomi, K. Hatano, and E. Takimoto. Tighter generalization bounds for matrix completion via factorization into constrained matrices. *IEICE Transactions on Information and Systems*, 101(8):1997–2004, 2018.

[42] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.
[43] J. Paisley, D. Blei, and M. I. Jordan. *Bayesian nonnegative matrix factorization with stochastic variational inference*, volume Handbook of Mixed Membership Models and Their Applications, chapter 11. Chapman and Hall/CRC, 2015.
[44] E. Richard, S. Gaïffas, and N. Vayatis. Link prediction in graphs with autoregressive features. *The Journal of Machine Learning Research*, 15(1):565–593, 2014.
[45] A. Saha and V. Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 693–702. ACM, 2012.
[46] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.
[47] T. Suzuki. Convergence rate of Bayesian tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, pages 1273–1282, 2015.
[48] E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Anomaly detection in smart card logs and distant evaluation with twitter: a robust framework. *Neurocomputing*, 298:109–121, 2018.
[49] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
[50] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. 2009.
[51] C. Vernade and O. Cappé. Learning from missing data using selection bias in movie recommendation. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE, 2015.
[52] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
[53] D. Xia and V. Koltchinskii. Estimation of low rank density matrices: bounds in schatten norms and other distances. *Electronic Journal of Statistics*, 10(2):2717–2745, 2016.
[54] H.-F. Yu, N. Rao, and I. S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 847–855. Curran Associates, Inc., 2016.

5. Proofs

5.1. **Additional notations.** Let us first introduce a few additional notations.

First, for the sake of shortness, we introduce the estimation risk R and the empirical risk r. These notations also make clear the fact that our estimator can be seen as an empirical risk minimizer.

$$R(\mathbf{A}) = \|\mathbf{A} - \mathbf{M}\|_F^2 \text{ and } r(\mathbf{A}) = \|\mathbf{A} - \mathbf{X}\|_F^2 \ ; \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}).$$

Let $\Delta(\mathcal{S}) = \{\mathbf{A} - \mathbf{B} \ ; \ \mathbf{A}, \mathbf{B} \in \mathcal{S}\}$.

For any $\mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{C})$, the spectral radius of $\mathbf{A}$ is given by

$$\rho(\mathbf{A}) := \max\{|\lambda| \ ; \ \lambda \in \mathrm{sp}(\mathbf{A})\}.$$

Note that $\|\mathbf{A}\|_{\mathrm{op}}^2 = \rho(\mathbf{A}\mathbf{A}^*) = \rho(\mathbf{A}^*\mathbf{A})$.

For any subset $\mathcal{K}$ of $\mathcal{M}_{d,T}(\mathbb{C})$,

$$\mathrm{rk}(\mathcal{K}) = \max\{\mathrm{rank}(\mathbf{A}) \ ; \ \mathbf{A} \in \mathcal{K}\}$$

and

$$\mathcal{K}^1 = \{\mathbf{A} \in \mathcal{K} : \|\mathbf{A}\|_F \leqslant 1\}.$$

5.2. **Some lemmas.** Let us now state the key lemmas for the proof of our results. The first one will be used to estimate how far from the minimizer of R is the minimizer of r.

Lemma 5.1. *For any* $\mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R})$,

$$R(\mathbf{A}) - r(\mathbf{A}) + \|\varepsilon\|_F^2 = 2\langle \varepsilon, \mathbf{A} - \mathbf{M}\rangle_F.$$

Moreover, for every $\lambda \in]0,1[$,

$$R(\mathbf{A}) \leqslant \frac{r(\mathbf{A}) - \|\varepsilon\|_F^2}{1-\lambda} + \frac{1}{\lambda(1-\lambda)} \left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\|\mathbf{A}-\mathbf{M}\|_F} \right\rangle_F^2 \tag{8}$$

and

$$r(\mathbf{A}) - \|\varepsilon\|_F^2 \leqslant (1+\lambda) R(\mathbf{A}) + \frac{1}{\lambda} \left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\|\mathbf{A}-\mathbf{M}\|_F} \right\rangle_F^2. \tag{9}$$

Proof. Consider $\mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R})$. First of all,

$$\begin{aligned} R(\mathbf{A}) - r(\mathbf{A}) &= \|\mathbf{A}-\mathbf{X}\|_F^2 - \|\mathbf{A}-\mathbf{M}\|_F^2 \\ &= \langle \mathbf{X}-\mathbf{M}, 2\mathbf{A}-\mathbf{X}-\mathbf{M}\rangle_F^2 \\ &= -\|\varepsilon\|_F^2 + 2\langle \varepsilon, \mathbf{A}-\mathbf{M}\rangle_F. \end{aligned}$$

Then, for any $\lambda \in]0,1[$,

$$R(\mathbf{A}) - r(\mathbf{A}) + \|\varepsilon\|_F^2 = 2\sqrt{\lambda R(\mathbf{A})} \left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\sqrt{\lambda}\cdot\|\mathbf{A}-\mathbf{M}\|_F} \right\rangle_F. \tag{10}$$

On the one hand, by Equation (10) together with the classic inequality $2ab \leqslant a^2+b^2$ for every $a, b \in \mathbb{R}$,

$$R(\mathbf{A}) - r(\mathbf{A}) + \|\varepsilon\|_F^2 \leqslant \lambda R(\mathbf{A}) + \frac{1}{\lambda} \left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\|\mathbf{A}-\mathbf{M}\|_F} \right\rangle_F^2.$$

So, Inequality (8) it true.

On the other hand, by Equation (10) together with the classic inequality $-2ab \leqslant a^2+b^2$ for every $a, b \in \mathbb{R}$,

$$r(\mathbf{A}) - R(\mathbf{A}) - \|\varepsilon\|_F^2 \leqslant \lambda R(\mathbf{A}) + \frac{1}{\lambda} \left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\|\mathbf{A}-\mathbf{M}\|_F} \right\rangle_F^2.$$

So, Inequality (9) it true. □

In the proof of the theorems, $\mathbf{A}$ will be replaced by an estimator of $\mathbf{M}$ that will be data dependent. Thus, it is now crucial to obtain uniform bounds on the scalar product in Lemma 5.1. In machine learning theory, concentration inequalities are the standard tools to derive such a uniform bound, see [6] for a comprehensive introduction to concentration inequalities for independent observations, and their applications to statistics. Some inequalities for time series can be found for example in [19], and were applied to machine learning in [3]. Here, we require more specifically a concentration inequality on random matrices. Such inequalities can be found in [49, 52]. We will actually use the following result (Theorem 5.39 and Remark 5.40.2 from [52]). As the proof can be found in [52], we don't reproduce it here.

Proposition 5.2. *Under Assumption 3.1, there exists a deterministic constant* $\mathfrak{m} > 1$, *not depending on* ε, d *and* T, *such that for every* $s \in \mathbb{R}_+$,

$$\left\| \frac{1}{d}\varepsilon^*\varepsilon - \Sigma_\varepsilon \right\|_{\mathrm{op}} \leqslant \mathfrak{m} \max\left\{ \sqrt{\frac{T}{d}} + \sqrt{\frac{s}{d}}; \left(\sqrt{\frac{T}{d}} + \sqrt{\frac{s}{d}}\right)^2 \right\} \widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}$$

with probability larger than $1 - 2e^{-s}$, *where* $\widetilde{\mathfrak{K}}_\varepsilon := \mathfrak{K}_\varepsilon^2 \vee \mathfrak{K}_\varepsilon^4$.

We are now in position to provide a uniform bound on the scalar product in Lemma 5.1.

Lemma 5.3. *Under Assumption 3.1, there exists a constant $\mathfrak{c} > 1$, not depending on ε, d and T, such that for every $s \in \mathbb{R}_+$ and $\mathcal{K} \subset \mathcal{M}_{d,T}(\mathbb{R})$,*

$$\sup_{\mathbf{A}\in\mathcal{K}^1} \langle \varepsilon, \mathbf{A}\rangle_F^2 \leqslant \mathfrak{c} \cdot \mathrm{rk}(\mathcal{K}^1)(d+T+s)\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}$$

with probability larger than $1-2e^{-s}$.

Proof. Consider a subset $\mathcal{K}$ of $\mathcal{M}_{d,T}(\mathbb{R})$ and $s \in \mathbb{R}_+$. Let $\sigma_1(\varepsilon) \geqslant \cdots \geqslant \sigma_d(\varepsilon)$ be the singular values of ε. On the one hand, consider a matrix $\mathbf{A} \in \mathcal{K}^1$ with singular values $\sigma_1(\mathbf{A}) \geqslant \cdots \geqslant \sigma_{\mathrm{rk}(\mathcal{K}^1)}(\mathbf{A})$. By Cauchy-Schwarz's inequality:

$$\begin{aligned}
|\langle \varepsilon, \mathbf{A}\rangle_F| &\leqslant \sum_{i=1}^{\mathrm{rk}(\mathcal{K}^1)} \sigma_i(\varepsilon)\sigma_i(\mathbf{A}) \\
&\leqslant \left| \sum_{i=1}^{\mathrm{rk}(\mathcal{K}^1)} \sigma_i(\varepsilon)^2 \right|^{1/2} \|\mathbf{A}\|_F \leqslant \mathrm{rk}(\mathcal{K}^1)^{1/2}\sigma_1(\varepsilon).
\end{aligned}$$

Then,

$$\sup_{\mathbf{A}\in\mathcal{K}^1} \langle \varepsilon, \mathbf{A}\rangle_F^2 \leqslant \mathrm{rk}(\mathcal{K}^1)\sigma_1(\varepsilon)^2. \tag{11}$$

On the other hand, consider

$$\omega \in \left\{ \left\| \frac{1}{d}\varepsilon^*\varepsilon - \Sigma_\varepsilon \right\|_{\mathrm{op}} \leqslant \mathfrak{m} \max\left\{ \sqrt{\frac{T}{d}} + \sqrt{\frac{s}{d}}; \left(\sqrt{\frac{T}{d}} + \sqrt{\frac{s}{d}}\right)^2 \right\} \widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}} \right\}.$$

Then,

$$\begin{aligned}
\left| \frac{1}{d}\sigma_1(\varepsilon(\omega))^2 - \|\Sigma_\varepsilon\|_{\mathrm{op}} \right| &= \left| \frac{1}{d}\|\varepsilon(\omega)\|_{\mathrm{op}}^2 - \|\Sigma_\varepsilon\|_{\mathrm{op}} \right| \\
&= \left| \frac{1}{d}\|\varepsilon(\omega)^*\varepsilon(\omega)\|_{\mathrm{op}} - \|\Sigma_\varepsilon\|_{\mathrm{op}} \right| \\
&\leqslant \left\| \frac{1}{d}\varepsilon(\omega)^*\varepsilon(\omega) - \Sigma_\varepsilon \right\|_{\mathrm{op}} \\
&\leqslant \mathfrak{m}\left(\sqrt{\frac{T}{d}} + \sqrt{\frac{s}{d}} + \frac{2T}{d} + \frac{2s}{d} \right) \widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}.
\end{aligned}$$

In particular,

$$\begin{aligned}
\sigma_1(\varepsilon(\omega))^2 &\leqslant \mathfrak{m}(\sqrt{Td} + \sqrt{sd} + 2T + 2s + d)\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}} \\
&\leqslant \mathfrak{m}\left(2d + \frac{5}{2}T + \frac{5}{2}s\right) \widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}} \\
&\leqslant \mathfrak{c}\,(d+T+s)\,\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}} \qquad (12)
\end{aligned}$$

with $\mathfrak{c} = 5\mathfrak{m}/2$. Therefore, by Inequalities (11) and (12) together with Proposition 5.2,

$$\sup_{\mathbf{A}\in\mathcal{K}^1} \langle \varepsilon, \mathbf{A}\rangle_F^2 \leqslant \mathfrak{c} \cdot \mathrm{rk}(\mathcal{K}^1)(d+T+s)\widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}$$

with probability larger than $1-2e^{-s}$. □

5.3. **Proof of Theorem 3.2.** Consider $\lambda \in]0,1[$ and $s \in \mathbb{R}_+$. By applying the Inequalities (8) and (9) of Lemma 5.1 successively:

$$\begin{aligned} R(\widehat{\mathbf{M}}_{\mathcal{S}}) &\leqslant \frac{r(\widehat{\mathbf{M}}_{\mathcal{S}}) - \|\varepsilon\|_F^2}{1-\lambda} + \frac{1}{\lambda(1-\lambda)} \left\langle \varepsilon, \frac{\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}}{\|\widehat{\mathbf{M}}_{\mathcal{S}} - \mathbf{M}\|_F} \right\rangle_F^2 \\ &\leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} R(\mathbf{A}) + \frac{2}{\lambda(1-\lambda)} \sup_{\mathbf{A}\in\Delta(\mathcal{S})^1} \langle \varepsilon, \mathbf{A}\rangle_F^2. \end{aligned}$$

By Lemma 5.3:

$$R(\widehat{\mathbf{M}}_{\mathcal{S}}) \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}} R(\mathbf{A}) + \frac{4\mathfrak{c}k}{\lambda(1-\lambda)} (d+T+s) \widetilde{\mathfrak{K}}_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}$$

with probability larger than $1 - 2e^{-s}$.

5.4. **Proof of Lemma 3.4.** First of all,

$$\Sigma_{\varepsilon_{1,.}\mathbf{C}} = \mathbb{E}(\mathbf{C}^* \varepsilon_{1,.}^* \varepsilon_{1,.} \mathbf{C}) = \mathbf{C}^* \mathbb{E}(\varepsilon_{1,.}^* \varepsilon_{1,.}) \mathbf{C} = \mathbf{C}^* \Sigma_\varepsilon \mathbf{C}.$$

Then, since the matrix $\Sigma_{\varepsilon\mathbf{C}}$ is Hermitian,

$$\begin{aligned} \|\Sigma_{\varepsilon_{1,.}\mathbf{C}}\|_{\mathrm{op}} &= \sup_{x\in\mathbb{C}^\tau\setminus\{0\}} \frac{\|\mathbf{C}^*\Sigma_\varepsilon \mathbf{C} x\|}{\|x\|} = \sup_{x\in\mathbb{C}^\tau\setminus\{0\}} \frac{x^*\mathbf{C}^*\Sigma_\varepsilon\mathbf{C}x}{\|x\|^2} \\ &= \sup_{x\in\mathbb{C}^\tau\setminus\{0\}} \frac{x^*\mathbf{C}^*\Sigma_\varepsilon\mathbf{C}x}{\|\mathbf{C}x\|^2} \times \frac{\|\mathbf{C}x\|^2}{\|x\|^2} \\ &\leqslant \left(\sup_{y\in\mathbb{C}^T\setminus\{0\}} \frac{y^*\Sigma_\varepsilon y}{\|y\|^2} \right) \left(\sup_{x\in\mathbb{C}^\tau\setminus\{0\}} \frac{\|\mathbf{C}x\|^2}{\|x\|^2} \right) = \|\Sigma_\varepsilon\|_{\mathrm{op}} \|\mathbf{C}^*\mathbf{C}\|_{\mathrm{op}}. \end{aligned}$$

5.5. **Proof of Lemma 3.5.** Since $\widetilde{\varepsilon} = \varepsilon\mathbf{\Lambda}^+$, $\Sigma_\varepsilon = \sigma^2\mathbf{I}_T$ and $\mathbf{\Lambda}\mathbf{\Lambda}^* = c(\tau,T)\mathbf{I}_\tau$,

$$\begin{aligned} \Sigma_{\widetilde{\varepsilon}}^{-1/2} &= ((\mathbf{\Lambda}^+)^*\Sigma_\varepsilon\mathbf{\Lambda}^+)^{-1/2} \\ &= \sigma^{-1}((\mathbf{\Lambda}^+)^*\mathbf{\Lambda}^+)^{-1/2} \\ &= \sigma^{-1}(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{1/2} \\ &= \sigma^{-1}c(\tau,T)^{1/2}\mathbf{I}_\tau. \end{aligned}$$

Then, for any x with $\|x\| = 1$,

$$\begin{aligned} \langle \widetilde{\varepsilon}_{1,.}\Sigma_{\widetilde{\varepsilon}}^{-1/2}, x\rangle &= \sigma^{-1}c(\tau,T)^{-1/2}\langle \varepsilon_{1,.}\mathbf{\Lambda}^*, x\rangle \\ &= c(\tau,T)^{-1/2}\|x\mathbf{\Lambda}\| \cdot \left\langle \varepsilon_{1,.}\Sigma_\varepsilon^{-1/2}, \frac{x\mathbf{\Lambda}}{\|x\mathbf{\Lambda}\|} \right\rangle. \end{aligned}$$

Moreover,

$$\|x\mathbf{\Lambda}\|^2 = x\mathbf{\Lambda}\mathbf{\Lambda}^*x^* = c(\tau,T)xx^* = c(\tau,T).$$

Therefore,

$$\begin{aligned} \mathfrak{K}_{\widetilde{\varepsilon}} &= \sup_{\|x\|=1} \sup_{p\in[1,\infty[} p^{-1/2}\mathbb{E}(|\langle \widetilde{\varepsilon}_{1,.}\Sigma_{\widetilde{\varepsilon}}^{-1/2}, x\rangle|^p)^{1/p} \\ &= c(\tau,T)^{-1/2} \\ &\quad \times \sup_{\|x\|=1} \left\{ \|x\mathbf{\Lambda}\| \sup_{p\in[1,\infty[} p^{-1/2}\mathbb{E}\left(\left| \left\langle \varepsilon_{1,.}\Sigma_\varepsilon^{-1/2}, \frac{x\mathbf{\Lambda}}{\|x\mathbf{\Lambda}\|} \right\rangle \right|^p \right)^{1/p} \right\} = \mathfrak{K}_\varepsilon \end{aligned}$$

and finally, $\widetilde{\mathfrak{K}}_{\widetilde{\varepsilon}} = \widetilde{\mathfrak{K}}_\varepsilon$.

5.6. **Proof of Theorem 4.1.** For short, let us denote

$$\widehat{\mathbf{M}}_s := \widehat{\mathbf{M}}_{\mathcal{S}_{\widehat{\tau}_s,\widehat{k}_s}}.$$

Consider $\lambda \in]0,1[$. On the one hand, by applying the Inequalities (8) and (9) of Lemma 5.1 successively:

$$\begin{aligned}
R(\widehat{\mathbf{M}}_s) &\leqslant \frac{r(\widehat{\mathbf{M}}_s) - \|\varepsilon\|_F^2}{1-\lambda} + \frac{1}{\lambda(1-\lambda)}\left\langle \varepsilon, \frac{\widehat{\mathbf{M}}_s - \mathbf{M}}{\|\widehat{\mathbf{M}}_s - \mathbf{M}\|_F}\right\rangle_F^2 \\
&= \frac{1}{1-\lambda}\cdot \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}}\{r(\widehat{\mathbf{M}}_{\tau,k}) + \mathrm{pen}_{s+\tau+k}(\tau,k) - \|\varepsilon\|_F^2\} \\
&\quad + \frac{1}{1-\lambda}\left(-\mathrm{pen}_{s+\tau+k}(\widehat{\tau}_s,\widehat{k}_s) + \frac{1}{\lambda}\left\langle \varepsilon, \frac{\widehat{\mathbf{M}}_s - \mathbf{M}}{\|\widehat{\mathbf{M}}_s - \mathbf{M}\|_F}\right\rangle_F^2\right) \\
&\leqslant \frac{1}{1-\lambda}\cdot \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}}\{(1+\lambda)R(\widehat{\mathbf{M}}_{\tau,k}) + \mathrm{pen}_{s+\tau+k}(\tau,k) + \psi_\varepsilon(\widehat{\mathbf{M}}_{\tau,k})\} \\
&\quad + \frac{1}{1-\lambda}(-\mathrm{pen}_{s+\tau+k}(\widehat{\tau}_s,\widehat{k}_s) + \psi_\varepsilon(\widehat{\mathbf{M}}_s)),
\end{aligned} \tag{13}$$

where

$$\psi_\varepsilon(\mathbf{A}) = \frac{1}{\lambda}\left\langle \varepsilon, \frac{\mathbf{A}-\mathbf{M}}{\|\mathbf{A}-\mathbf{M}\|_F}\right\rangle_F^2 ;\ \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}).$$

On the other hand, consider $(\tau,k) \in \mathcal{T}\times\mathcal{K}$. Since $\widetilde{\varepsilon} = \varepsilon\mathbf{\Lambda}_\tau^+$,

$$\begin{aligned}
\psi_\varepsilon(\widehat{\mathbf{M}}_{\tau,k}) &= \frac{1}{\lambda}\left\langle \varepsilon, \frac{(\widehat{\widetilde{\mathbf{M}}}_{\tau,k} - \widetilde{\mathbf{M}})\mathbf{\Lambda}_\tau}{\|\widehat{\mathbf{M}}_{\tau,k} - \mathbf{M}\|_F}\right\rangle_F^2 = \frac{1}{\lambda}\left\langle \varepsilon\mathbf{\Lambda}_\tau^+\mathbf{\Lambda}_\tau\mathbf{\Lambda}_\tau^*, \frac{\widehat{\widetilde{\mathbf{M}}}_{\tau,k} - \widetilde{\mathbf{M}}}{\|\widehat{\mathbf{M}}_{\tau,k} - \mathbf{M}\|_F}\right\rangle_F^2 \\
&\leqslant \frac{1}{\lambda}\cdot \sup_{\mathbf{A}\in\Delta(\mathcal{S}_{\tau,k})^1}\langle \widetilde{\varepsilon}, \mathbf{A}\mathbf{\Lambda}_\tau^*\rangle_F^2 \leqslant \frac{1}{\lambda}\|\mathbf{\Lambda}_\tau^*\|_{\mathrm{op}}^2 \cdot \mathrm{rk}(\Delta(\mathcal{S}_{\tau,k})^1)\cdot \sigma_1(\widetilde{\varepsilon})^2.
\end{aligned}$$

As in the proof of Proposition 5.3, by Lemma 5.2 and since

$$\|\Sigma_{\varepsilon\mathbf{\Lambda}_\tau^+}\|_{\mathrm{op}}\|\mathbf{\Lambda}_\tau^*\|_{\mathrm{op}}^2 \leqslant \|\Sigma_\varepsilon\|_{\mathrm{op}}\|(\mathbf{\Lambda}_\tau\mathbf{\Lambda}_\tau^*)^{-1}\|_{\mathrm{op}}\rho(\mathbf{\Lambda}_\tau\mathbf{\Lambda}_\tau^*) = \|\Sigma_\varepsilon\|_{\mathrm{op}},$$

with probability larger than $1-2e^{-u}$,

$$\psi_\varepsilon(\widehat{\mathbf{M}}_{\tau,k}) \leqslant \frac{2\mathfrak{c}k}{\lambda}(d+\tau+u)\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}} = \mathrm{pen}_u(\tau,k).$$

Take $u = s+\tau+k$, we obtain that with probability at least $1-2e^{-s-\tau-k}$,

$$\psi_\varepsilon(\widehat{\mathbf{M}}_{\tau,k}) \leqslant \frac{2\mathfrak{c}k}{\lambda}(d+\tau+(s+2\tau+2k))\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}} = \mathrm{pen}_{s+\tau+k}(\tau,k).$$

Then, by a union bound,

$$\begin{aligned}
\mathbb{P}(\forall k, \forall\tau\colon\ \psi_\varepsilon(\widehat{\mathbf{M}}_{\mathcal{S}_{\tau,k}}) \leqslant \mathrm{pen}_{s+2\tau+2k}(\tau,k)) &\geqslant 1 - 2\sum_{(\tau,k)\in\mathcal{T}\times\mathcal{K}} e^{-s-\tau-k} \\
&\geqslant 1 - 2e^{-s}\left(\sum_{\tau\geq 1} e^{-\tau}\right)\left(\sum_{k\geq 1} e^{-k}\right) \\
&\geqslant 1 - 2e^{-s}.
\end{aligned}$$

Together with Inequality (13), this gives, with probability at least $1-2e^{-s}$,

$$R(\widehat{\mathbf{M}}_s) \leqslant \frac{1}{1-\lambda}\cdot \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}}\left\{(1+\lambda)R(\widehat{\mathbf{M}}_{\tau,k}) + 2\mathrm{pen}_{s+\tau+k}(\tau,k)\right\}. \tag{14}$$

Finally, follow the proof of Corollary 3.7 to obtain, on the same event with probability at least $1-2e^{-s}$, for any τ and k,

$$R(\widehat{\mathbf{M}}_{\mathcal{S}_{\tau,k}}) \leqslant \frac{1+\lambda}{1-\lambda} \cdot \min_{\mathbf{A}\in\mathcal{S}_{\tau,k}} \|\mathbf{A}\mathbf{\Lambda}_\tau - \mathbf{M}\|_F^2 + \frac{2}{1-\lambda}\mathrm{pen}_{s+\tau+k}(\tau,k).$$

Plugging this into (14) gives, with probability at least $1-2e^{-s}$,

$$R(\widehat{\mathbf{M}}_s) \leqslant \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}} \left\{ \left(\frac{1+\lambda}{1-\lambda}\right)^2 \min_{\mathbf{A}\in\mathcal{S}_{\tau,k}} \|\mathbf{A}\mathbf{\Lambda}_\tau - \mathbf{M}\|_F^2 + \frac{4}{(1-\lambda)^2}\mathrm{pen}_{s+\tau+k}(\tau,k) \right\}.$$

Finally, note that $k \leqslant d$ so

$$\begin{aligned}\mathrm{pen}_{s+\tau+k}(\tau,k) &= \frac{2\mathfrak{c}k}{\lambda}(d+\tau+(s+\tau+k))\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}} \\ &\leqslant \frac{4\mathfrak{c}k}{\lambda}(d+\tau+s)\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}}\end{aligned}$$

and so, with probability at least $1-2e^{-s}$,

$$R(\widehat{\mathbf{M}}_s) \leqslant \min_{(\tau,k)\in\mathcal{T}\times\mathcal{K}} \left\{ \left(\frac{1+\lambda}{1-\lambda}\right)^2 \min_{\mathbf{A}\in\mathcal{S}_{\tau,k}} \|\mathbf{A}\mathbf{\Lambda}_\tau - \mathbf{M}\|_F^2 + \frac{16\mathfrak{c}k}{\lambda(1-\lambda)^2}(d+\tau+s)\widetilde{\mathfrak{K}}_\varepsilon\|\Sigma_\varepsilon\|_{\mathrm{op}} \right\}.$$

This ends the proof.

*RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
Email address: pierre.alquier.stat@gmail.com

†Laboratoire Modal'X, Université Paris Nanterre, Nanterre, France
Email address: nmarie@parisnanterre.fr

†ESME Sudria, Ivry-sur-Seine, France
Email address: nicolas.marie@esme.fr