INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAG002

Thèse de doctorat

ENSAE

IP PARIS

Matrix-valued Time Series in High Dimension

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 5 juillet 2024, par

NAYEL BETTACHE

Composition du Jury :

Alexandre Tsybakov Professeur, CREST, ENSAE, IP Paris	Président
Florentina Bunea Professeur, Cornell University (Unité de recherche)	Rapporteur
Maxim Panov Professeur, MBZUAI	Rapporteur
Catherine Matias Directrice de Recherche, CNRS	Examinateur
Christophe Pouet Professeur, Ecole centrale de Marseille	Examinateur
Farida Enikeeva Maître de Conférences, Université de Poitiers	Examinateur
Cristina BUTUCEA Professeur, CREST, ENSAE, IP Paris	Directeur de thèse

"In God we trust. All others must bring data."
W. Edwards Deming

Remerciements

Je dois infiniment à l'accompagnement et au soutien indéfectibles de personnes exceptionnelles qui m'ont accompagné au long de mon parcours. Je souhaite ici les remercier. Mes premières pensées vont d'abord à Cristina. Cristina, de mon premier stage en 2018 sous ta direction jusqu'à ce manuscrit en 2024 et sa rédaction, je sais tout ce que je te dois. Tu m'as proposé des sujets por- teurs et intéressants, m'ouvrant à des perspectives de recherche nouvelles qui m'ont passionné. Grâce à toi, j'ai beaucoup voyager pendant ces trois années : à Londres, Boston, Ithaca, Marseille, Varsovie et Oberwolfach. Tu m'as formé scientifiquement, m'apprenant notamment à être très rigoureux dans la rédaction d'un papier et à présenter mes idées de manière claire et concise.Tu m'as également poussé à faire preuve d'esprit critique et à sans cesse me questionner. Toutes ces qualités m'ont permis de faire mes preuves auprès d'autres scientifiques dont je te dois la rencontre. Je pense en particulier à Tracy et Flori. Je quitte ta direction avec une plus grande maturité scientifique, un recul plus important sur mes centres d'interêt mathématique et surtout, humainement grandi. Pour tout, Merci. Secondly, I want to thank you Flori. It is not possible to imagine a warmer welcome than the one you offered me. Professionally, thank you for having accepted to review this thesis with a very short notice. It is an honour to have your name on the front page of my thesis. Personally, I have to tell that your energy and enthusiasm for research are deeply inspiring. I keep them in mind whenever my research flame flickers. In addition to be a brilliant researcher, you opened me your door as spontaneously as if we were family. I felt home the moment I arrived at Ithaca. It pushed me to give you my best. I'll do everything I can to honour your trust. Maxim, thank you for having taken the time to review my thesis. I enjoyed our e-discussion and the interest you expressed on my work. Je tiens à remercier chaleureusement les autres membres de mon jury. C'est un honneur de présenter mes travaux devant vous. Merci Sacha, tout d'abord pour tes cours d'une immense qualité que j'ai suivi à l'ENSAE. Merci également pour tes travaux qui ont grandement contribué à ma formation scientifique. J'ai beaucoup appris en te lisant. La qualité de tes travaux, ton humilité, ta gentillesse sont pour moi sources d'inspiration. Catherine, j'ai eu la chance de te rencontrer à Varsovie l'an dernier. Merci pour les discussions bienveillantes que nous avons eues là-bas. Je suis heureux de te compter parmi les membres de mon jury. Christophe, ta gentillesse et ta disponibilité à Luminy m'ont marquées. Merci d'avoir accepté de faire partie de mon jury. Farida, je n'ai pas encore eu l'opportunité de te rencontrer. Merci de prendre de ton temps pour faire parti de mon jury. Ces trois dernières années forment une étape importante de ma construction professionnelle et person- nelle. J'ai vécu une aventure scientifique et humaine riche. J'ai eu la chance de côtoyer des personnes à la fois brillantes et amicales. Après trois années en tant qu'élève à l'ENSAE, je me suis rapidement senti à ma place au CREST. Je vous suis reconnaissant de m'avoir aussi bien accueilli. Arnak, merci pour tous nos échanges, qu'ils soient autour d'un café ou d'un ballon. Tu as répondu à beaucoup de mes interrogations pendant ces trois années, tu m'as beaucoup aiguillé et un nombre non négligeable de mes décisions ont été prises à la lumière de tes conseils. Victor, merci pour ta gentillesse, ta bonne humeur permanente et le travail administratif que tu effectues pour nous. Ca a

été un plaisir d'assurer les travaux dirigés de tes cours. Guillaume, merci pour nos discussions sur la longévité. Tu es la seule personne que j'ai rencontré à partager un interêt scientifique à ce sujet. J'espère que nous continuerons à échanger à ce propos. Matthieu, Nicolas, merci pour les déjeuners, les cafés et les discussions que nous avons partagés. Anna, Jaouad, merci pour votre sympathie, votre humour et votre disponibilité. J'espère vous croiser encore régulièrement à l'avenir. Katia, Mohamed, Evgenii, Arshak, Badr, j'ai trouvé chez vous des personnes en qui je me reconnais et dont je partage les valeurs. Je suis heureux de vous avoir dans mon entourage. Nicolas, merci pour ta présence durant mon aventure doctorale. Tu es devenu un ami. Enfin merci à tous les doctorants. En premier lieu à vous, Hugo et Théo. Vous avez été des lumières dans cette aventure parfois souterraine. Puisse notre amitié perdurer. Clara, Nina, Flore, Julien, Younes, Ziyad, Etienne, Clémentine, Meyer, merci pour tout le temps qu'on a partagé. Leyla, Djamila merci pour votre bonne humeur et votre accompagnement dans les démarches administratives. Je souhaite ensuite remercier mes amis de longue date. Merci à Gary, Vincent, David, que j'ai à mes côtés depuis plus de dix ans maintenant. Merci à Charles, Guillaume, Stan, Tom, Tam, Elias, Léa, Agathe, Emma. Votre amitié est rare et je mesure la chance de vous avoir dans ma vie. Merci également à ma belle famille, Cathy, Thibaud, Juliette, Dorian, Denis, de m'avoir si bien accueilli. Merci à mon frère, Kenzi, qui a toujours partagé mes joies et mes peines. La Volvic citron et les petites pierres qui tombent du lit sont des bouées auxquelles je pourrai à jamais me rattacher. Merci à mes parents, Odile et Abel, de m'avoir transmis des valeurs d'amour, de générosité et d'abnégation. Merci de nous avoir offert, avec Kenzi, la vie que nous avons. Merci d'avoir été exigeant avec nous, exigence sans laquelle je n'écrirais pas ces lignes aujourd'hui. Merci d'avoir mis autant d'énergie, de souci, d'investissement, dans notre éducation. Merci de nous avoir toujours soutenu, poussé, encouragé. Merci d'avoir fait autant de sacrifices pour que nous puissions, un jour, avoir mieux que ce que vous avez eu. Je suis fier d'être votre fils. Enfin, merci à toi, Laure. Merci de partager ma vie depuis cinq ans maintenant. Merci pour ton amour, ton affection, ton attention, ta douceur (pas toujours) et ton engagement. Merci pour ta patience, tu vois maintenant l'aboutissement de tous ces moments où je suis physiquement là mais mentalement ailleurs. J'espère que mes longs moments passés seul dans le silence à la maison te semblent aujourd'hui plus concrets. Merci pour ton soutien permanent pendant ces trois dernières années. Tu m'as bien remis sur pieds dans mes moments de doute. Merci de m'avoir toujours écouté avec attention, même si tu ne comprenais pas forcément tout, lorsque j'avais besoin de partager une idée. L'effort de vulgarisation nécessaire m'a souvent permis de pointer ce que je n'avais pas compris moi-même. Pour finir, merci de soutenir mon départ pour Ithaca. Je sais ce que mon choix te coûte et les concessions qu'il implique pour nous. Merci.

Notations

$A := B$	A is defined as being equal to B,
$u_n \overset{n\to\infty}{\sim} v_n$	u_n is equivalent to v_n,
$X \sim \mathbb{P}_X$	X follows the distribution $\mathbb{P}_X$,
$[n] = [\![1, n]\!]$	Set of the firsts n integers,
$\lvert E\rvert = \#E$	Cardinal of the set E,
$\lvert x\rvert$	Absolute value of the real number x,
$n \wedge m = \min(n, m)$	Minimum between n and m,
$n \vee m = \max(n, m)$	Maximum between n and m,
$\mathbb{R}_+$	Set of non negative real numbers,
$\mathbb{R}_+^*$	Set of positive real numbers,
$\mathbb{R}^p$	Set of real-valued vectors of size p,
$\mathbb{R}^{p\times q}$	Set of real-valued matrices of size $p \times q$,
$x(k)$	k^{th} coefficient of the vector x,
$\mathcal{S}_p$	Set of symmetric matrices of size p,
$\mathcal{S}_p^+$	Set of symmetric positive matrices of size p,
$\mathcal{S}_p^{++}$	Set of symmetric positive definite matrices of size p,
$\mathcal{T}_p$	Set of Toeplitz matrices of size p,
$\mathcal{O}_p$	Set of orthogonal matrices of size p,
$Diag_{n,m}(\alpha_k,\, 1 \leq k \leq r)$	Matrix of size $n \times m$ with diagonal entries in the list and zero elsewhere,
$[M]_{i.}$	i^{th} row of the matrix M,
$[M]_{.j}$	j^{th} column of the matrix M,
$[M]_{ij}$	Coefficient on the i^{th} row and j^{th} column of the matrix M,
$M^+ := M^\dagger$	Moore-Penrose pseudo inverse of M,
$U_M \Sigma_M V_M^\top$	Singular Value Decomposition of M,
$\lVert M\rVert_F$	Frobenius norm of M,
$\lVert M\rVert_*$	Nuclear norm of M,
$\lVert M\rVert_{op}$	Operator norm of M,
$\mathsf{Tr}(M)$	Trace of the matrix M,
$\mathbb{1}$	Indicator function,
$\mathcal{N}\left(\mu, \sigma^2\right)$	Univariate Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+^*$,
$(\Omega, \mathcal{A}, \mathbb{P})$	Probability space with a sample space $\Omega,$ a σ-algebra $\mathcal{A}$ and a probability function $\mathbb{P}$,
$\mathbb{P}_\Sigma(A)$	Probability of the event A when the parameter takes the value Σ,
$\mathbb{E}_\Sigma[X]$	Expectation of the random variable X when the parameter takes the value Σ,
$\asymp$	Equality up to constants.

Table des matières

1 Introduction **1**
- 1.1 A non-asymptotic viewpoint . . . 3
  - 1.1.1 From Chebyshev to McDiarmid . . . 3
  - 1.1.2 High-dimensional covariance matrix estimation . . . 9
  - 1.1.3 Random matrices with independent entries or rows . . . 10
  - 1.1.4 Stochastically dependent data . . . 13
  - 1.1.5 Time series analysis . . . 15
- 1.2 Problems and contributions . . . 20
  - 1.2.1 Hypothesis Testing : deciding where lives a covariance matrix . . . 20
  - 1.2.2 Regression framework . . . 24
  - 1.2.3 Topic Modeling . . . 28
- 1.3 List of publications . . . 31

2 Covariance matrix testing and support recovery **33**
- 2.1 Introduction . . . 33
- 2.2 Linear functionals of the covariance matrix . . . 35
- 2.3 Non-parametric testing for stationary time series . . . 38
  - 2.3.1 Moderately sparse covariance structure . . . 38
  - 2.3.2 Highly sparse covariance structure . . . 39
- 2.4 Lag-selection for stationary time-series . . . 42
- 2.5 Proofs . . . 43
- 2.6 Supplementary material . . . 48
  - 2.6.1 Power curves of the test procedures . . . 49
  - 2.6.2 Effect of non null entries . . . 50
  - 2.6.3 Comparison between Δ_n^{MS} and Δ_n^{HS} . . . 55
  - 2.6.4 A moderately sparse high-dimensional MA series . . . 56
  - 2.6.5 Comparison to other test procedures . . . 57
  - 2.6.6 Application to real data . . . 61

3 Two-sided Matrix Regression **65**
- 3.1 Introduction . . . 65
- 3.2 Rank penalized learning . . . 67
  - 3.2.1 Prediction for given ranks . . . 67
  - 3.2.2 Rank-adaptive prediction . . . 69
  - 3.2.3 Consistent rank selection . . . 70

3.2.4 Data-driven rank-adaptive prediction . . . 71
3.3 Nuclear norm penalized learning . . . 72
3.4 Numerical Results . . . 73
3.5 Proofs . . . 74
3.5.1 Proof of Theorem 3.2.1 . . . 75
3.5.2 Proof of Corollary 3.2.2 . . . 76
3.5.3 Proof of Theorem 3.2.3 . . . 77
3.5.4 Proofs of results in Section 3.2.3 . . . 78
3.5.5 Proof of Theorem 3.2.7 . . . 78
3.5.6 Proof of Theorem 3.3.1 . . . 80
3.6 Auxiliary results . . . 82

4 Dynamic Expected Topic Models 83
4.1 Introduction . . . 83
4.2 Dynamic topic model framework . . . 84
4.3 Recovery of the word-topic matrix . . . 87
4.4 Estimation of the autoregressive model . . . 91
4.5 Proofs . . . 93
4.5.1 Proof of Theorem 4.3.3 . . . 93
4.5.2 Proof of Proposition 4.3.4 . . . 94
4.5.3 Proof of Theorem 4.4.1 . . . 94
4.5.4 Proof of Theorem 4.4.2 . . . 97
4.5.5 Proof of Theorem 4.4.3 . . . 104

5 Dynamic topic model 107
5.1 Introduction . . . 107
5.2 Estimation of the word-topic matrix A^* . . . 108
5.3 Estimation of the topic-document matrix . . . 117
5.4 Estimation of the underlying parameters of the autoregressive model . . . 119
5.5 Proofs . . . 123
5.5.1 Proof of Proposition 5.2.1 and its Corollary . . . 123
5.5.2 Proof of Proposition 5.2.3 and its Corollaries . . . 127
5.5.3 Proof of Proposition 5.2.6 . . . 130
5.5.4 Proof of Proposition 5.2.7 . . . 134
5.5.5 Proof of Proposition 5.2.8 . . . 139
5.5.6 Proof of Proposition 5.2.9 . . . 139
5.5.7 Proof of Proposition 5.2.10 . . . 141
5.5.8 Proof of Theorem 5.2.11 . . . 141
5.5.9 Proof of Theorem 5.2.12 . . . 152
5.5.10 Proof of Theorem 5.2.13 . . . 157
5.5.11 Proof of Proposition 5.2.14 . . . 161
5.5.12 Proof of Theorem 5.2.15 . . . 162
5.5.13 Proof of Theorem 5.2.16 . . . 169
5.5.14 Proof of Theorem 5.2.17 . . . 180
5.5.15 Proof of Theorem 5.3.1 . . . 182
5.5.16 Proof of Theorem 5.4.2 . . . 192

5.5.17 Proof of Theorem 5.4.3 196
5.6 Auxiliary results 199

6 Introduction en français **205**

Chapitre 1

Introduction

The main motivation of this manuscript is to deepen our understanding of phenomena with a temporal component. Most machine learning algorithms and high-dimensional statistical models are largely studied under assumptions of independence of observations. Indeed, there are fewer and technically more demanding tools for measure concentration under this setting. This leads to non-asymptotic control of deviations being more challenging in the more realistic setup of dependence between observations. Very frequently, an evolution with time is obvious in the underlying model but not always taken into account in the proposed methods and the inference results.

This thesis investigates various non-parametric and high-dimensional inference problems including hypothesis testing, support recovery, prediction in matrix regression and estimation of dynamic topic models that combine matrix factorization and auto-regression. Although they share a common motivation, the chapters presented in this thesis can be read and understood separately as they are focusing on specific problems.

Assessing the quality of forecasting algorithms is crucial across diverse applications, from natural phenomena like weather patterns and seismic events to economic variables such as stock prices and energy demand. A key indicator of algorithm performance is the quality of residuals, representing the difference between observed and predicted values. More precisely, the closer the residuals are to a white noise distribution, the less information was lost by the predictor or the model at hand. In chapter 2 we study the testing and support recovery problems of a high-dimensional covariance matrix of a stationary time series. Specifically, we consider $X_1, \dots, X_n$ independent p-dimensional Gaussian vectors with a covariance matrix Σ. When the vectors X_i are issued from a stationary process, the covariance matrix Σ has a Toeplitz structure, that is its diagonal elements are all constants. As mentioned in [46], stationary time series are used as approximations of geometrically ergodic time series. This setting is motivated by the following observation : given a time series of length T with $T \gg p$, it is possible to consider vectors of length p sufficiently far apart to assume they are independent vectors of dimension p. The aim is then to test whether the distribution is close to a white noise. To do so we test if the covariance matrix Σ is the identity matrix I_p or there exists a number s of covariance elements that are significantly positive or significantly different from zero. We provide testing procedures with non asymptotic upper bounds on the maximal testing risks both for moderately sparse and highly sparse covariance structures. If the test is rejected, it is of interest to select the non-null entries in Σ, pinpointing where information may be lost in the modelling process. We then define a lag-selection procedure and provide a non asymptotic upper bound on its risk.

Next, we introduce a new matrix regression model where the correlations in the output matrix are

explained by two matrix parameters that multiply the design matrix from the left and from the right, respectively. We assume that the noise matrix has independent σ^2-subGaussian entries. This general matrix regression model is highly non-identifiable without additional stringent assumptions, thus only prediction results were provided. The predictors are first defined as solutions of the minimization problem of the squared Frobenius prediction risk under a maximal fixed rank constraint. By using the SVD of the target and design matrices we provide solutions to this optimization problem together with a non asymptotic upper bound on the prediction risk. We show that this upper bound can be decomposed as the sum of a bias term and a stochastic term. We then derive a model selection procedure for estimating the true common rank of the parameter matrices, first under the assumption that the noise parameter σ is available. We examine the non asymptotic performance of this procedure and we adapt the initial minimization problem by fixing the rank constraint to this estimated rank. This leads to new rank-adaptive predictors. We provide again a non asymptotic upper bound on the rank-adaptive prediction risk under this model selection framework. Then, we adapt the rank-adaptive procedure to propose a data-driven rank-adaptive procedure free of the noise parameter σ. Again, we provide a non asymptotic upper bound on the data-driven rank-adaptive prediction risk. Finally, we study the convex relaxation of the rank-penalized squared Frobenius risk minimization. We provide explicit solutions of this problem and a non asymptotic upper bound on the prediction risk. Numerical results are provided illustrating the theoretical results.

Finally, we consider topic models. We assume we collect a batch of documents and have access to the frequencies of each word of the vocabulary for each document. The columns of this word-document frequency matrix Y are modelled as realizations of multinomial distributions centered on word-document probability vectors. In real world examples, few different topics are covered in corpora of documents. This suggests that the word-document probability matrix Π exhibits a low rank structure. The objective is to factorize this word-document probability matrix Π into a word-topic probability matrix A and a topic-document probability matrix W, that is $\Pi = AW$. In this setting, all these three matrices Π, A and W are left stochastic, that is their entries are non-negative and their columns sum to one. Under specific mild assumptions, the identifiability of both A and W can be established. We also recall the algorithm from [84] for performing this factorization. In this thesis, we assume a temporality in the document collection and model the evolution in time of the topic-document probability matrix W by an autoregressive stationary process, which becomes a time dependent random matrix $\boldsymbol{W}_t$. Specifically, at each time step t, the distribution of topics given a document is a linear combination of the previous distribution and a Dirichlet-distributed noise, which drives the temporal evolution of the topics. Especially we assume that the noise parameters are unknown. Careful attention is devoted to ensuring that this autoregressive model keeps the property that the columns of the topic-document probability matrix sum to one. We first study an oracle case where the full word-document probability matrix $(\boldsymbol{\Pi}_1, \dots, \boldsymbol{\Pi}_T)$ is available. We first provide non asymptotic bounds on the spectrum of the empirical covariance matrix of $(\boldsymbol{W}_1, \dots, \boldsymbol{W}_T)$. Then we adapt the algorithm from [84] to retrieve the word-topic probability matrix A. This allows to recover $(\boldsymbol{W}_1, \dots, \boldsymbol{W}_T)$ by projection. Then we propose estimators of the autoregressive parameters driving the evolution of $\boldsymbol{W}_t$. We provide non asymptotic upper bounds on the estimation risks. Then, we adapt this procedure to the real case where only the full word-document frequency matrix $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_T)$ is available. In the estimation procedure of A, we give more explicit upper bounds than [84] up to log factors. We also provide the dependence on all dimensions of appearing matrices. Finally, we show that the noise due to the multinomial distribution of word-counts and the Dirichlet noise of the stationary distribution of topics given the published documents in time add up in the final estimation rates of the autoregressive parameters. Especially, when the number of words per document grows, that is when the multinomial noise diminishes, we retrieve the oracle rates.

Historically, time series analysis is usually done in an asymptotic framework. The asymptotic analysis of real-valued and vector-valued time series is well understood since [71], [62], [99] and [31] were published. This is still an active field of research both from a theoretical point of view, see [79, 50, 91, 117, 51, 59] and as a tool for studying algorithms, see [142]. Recently, the study of matrix-valued time series and more globally tensor-valued time series has emerged. The studies are still mainly conducted under an asymptotic framework, see [47, 49, 44, 96]. The non-asymptotic analysis of time series is however gaining momentum, see [16, 15, 58, 135]. This thesis is part of this research dynamic and all studied problems are conducted within a non-asymptotic framework. By addressing these challenges and exploring innovative methodologies in each chapter, this thesis contributes to advancing statistical theory in vector-valued and matrix-valued data analysis within high-dimensional settings. The first part of the introduction serves as a comprehensive presentation of the technical tools necessary for understanding the main chapters of this thesis. Then, in the second part, we give the setups and the details of the results.

1.1 A non-asymptotic viewpoint

We begin by providing a rationale for employing a non-asymptotic framework, which is consistently applied throughout the presented research. Subsequently, we delve into a detailed exploration of concentration inequalities, outlining their significance and specifying the classical inequalities that will be employed in our analyses. Additionally, we offer an overview of the problems one may face while working in a high dimensional regime. Finally, we briefly introduce the tools that will be useful to control random matrices and random processes.

1.1.1 From Chebyshev to McDiarmid

The non-asymptotic framework is highly relevant in modern statistical analysis, particularly in scenarios involving high-dimensional data and finite-sample settings. Unlike traditional asymptotic approaches that rely on large sample sizes and convergence to theoretical distributions, the non-asymptotic framework focuses on deriving results that hold for finite sample sizes, providing more practical and immediate insights into statistical properties and performances. Concentration inequalities constitute a cornerstone of our methodological approach, providing rigorous bounds on the deviation of random variables from their expected values.

To avoid unessential technicalities as a first step, we will consider real random variables, *i.e.* $p = 1$. Let's assume the finiteness of $\mathbb{E}[X_1]$ and denote $\bar{X}_n$ the empirical mean of the n random vectors $(X_1, \dots, X_n)$. We are interested in understanding the behaviour of $\bar{X}_n$. To this extent, the strong law of large numbers (SLLN) states that $\bar{X}_n$ converges almost surely towards $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1]$. Once we have defined the asymptotic limit of $\bar{X}_n$, another interesting question is to determine the rate at which $\bar{X}_n$ approaches $\mathbb{E}[\bar{X}_n]$. We assume from now on the finiteness of $\sigma^2 = \mathbb{V}[X_1]$. The Lindeberg–Lévy central limit theorem (LLCLT) then provides the asymptotic convergence rate of $\bar{X}_n$ towards $\mathbb{E}[\bar{X}_n]$.

Lemma 1.1.1 (Lindeberg–Lévy central limit theorem) *Consider $(X_1, \dots, X_n)$ independent and identically distributed random variables with finite second order moment. Let us denote σ^2 their common variance. Then the random variable $\sqrt{n}\left(\bar{X}_n - \mathbb{E}[\bar{X}_n]\right)$ converges in distribution toward $\mathcal{N}(0, \mathbb{V}[X_1])$. Es-*

pecially, considering $U \sim \mathcal{N}(0,1)$*, we get that for all* $\epsilon > 0$*,*

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| > \epsilon\right) \overset{n\to\infty}{\sim} \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right).$$

However, this rate of convergence is only holding true asymptotically. Thus another natural question arises : what can be said about the behaviour of the quantity $|\bar{X}_n - \mathbb{E}[\bar{X}_n]|$ for a finite value of n ? One can already notice that the property of the probability distribution of X_1, denoted $\mathbb{P}_X$, will play a key role in answering this question. Indeed, if X_1 is normally distributed, the sample mean $\bar{X}_n$ is also normally distributed and we get $\bar{X}_n \sim \mathcal{N}\left(\mathbb{E}[\bar{X}_n], \frac{\sigma^2}{n}\right)$. Thus in this context the asymptotic behaviour is satisfied for any sample size, *i.e.* for any $\epsilon > 0$ and for any $n \in \mathbb{N}$, $\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| > \epsilon\right) = \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$ where $U \sim \mathcal{N}(0,1)$. On the other hand if $\mathbb{P}_X$ is not symmetric or exhibits heavy tails (*e.g.*, due to skewness or extreme values) the asymptotic behaviour will appear for larger sample sizes.

The objective is thus to bound from above with high probability the quantity $|\bar{X}_n - \mathbb{E}[\bar{X}_n]|$ for any fixed value of n. The first non asymptotic result that can be used for this purpose is the Chebyshev's inequality.

Lemma 1.1.2 (Chebyshev's Inequality) *Consider* $(X_1, \ldots, X_n)$ *independent and identically distributed random variables with finite second order moment. Denote* $\sigma^2 := \mathbb{V}[X_1]$*,. Then for any* $\epsilon > 0$*,*

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right) < \frac{\sigma^2}{n\epsilon^2}.$$

Notice that in the asymptotic framework, the LLCLT ensures that this probability behaves as $\mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$ where $U \sim \mathcal{N}(0,1)$. Moreover, the tails of the centered reduced normal distribution satisfy for all $\epsilon > 0$ and $n \in \mathbb{N}$,

$$\frac{\sigma^3\sqrt{2}(n\epsilon^2/\sigma^2 - 1)}{n\sqrt{n}\epsilon^3\sqrt{\pi}} \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) \leq \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right) \leq \frac{\sigma\sqrt{2}}{\epsilon\sqrt{n\pi}} \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

Hence the convergence rate of the quantity $\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right)$ towards zero exhibits asymptotically an exponential decay with respect to (*w.r.t*) n while the non-asymptotic rate of decay provided by the Chebyshev's inequality is only linear *w.r.t* n. A natural approach would be to improve the Chebyshev's linear rate of decay by controlling the deviation of $\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right)$ from $\mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$, *i.e.* determine the rate of convergence in the LLCLT. With the additional assumption that X_1 has a finite third order moment, Berry-Esseen central limit theorem (BECLT) provides the answer to this problem.

Lemma 1.1.3 (Berry-Esseen central limit theorem) *Consider* $(X_1, \ldots, X_n)$ *independent and identically distributed random variables with finite third order moment. Denote* $\sigma^2 := \mathbb{V}[X_1]$ *and* $\rho := \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^3]$*. Then, there exists a positive constant* $C > 0$ *such that for any* $\epsilon > 0$ *and for any* $n \in \mathbb{N}^*$ *:*

$$\left|\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right) - \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)\right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}.$$

Thus, the convergence rate in the LLCLT is of order root of n, which can be shown to be optimal, and will therefore dominate the desired exponential decay that arises asymptotically. Indeed, under the previously stated assumptions, the BECLT ensures that there is a positive constant $C > 0$ such that for any $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right) \leq \frac{C\rho}{\sigma^3\sqrt{n}} + \frac{\sigma\sqrt{2}}{\epsilon\sqrt{\pi}} \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

The previously stated result indicates that controlling non-asymptotically the deviation of $\bar{X}_n$ from its expectation by the LLCLT is worse than using directly Chebyshev's inequality. In addition, Chebyshev's inequality is optimal under the stated assumptions. It implies that stronger assumptions are required in order to get the asymptotic behaviour for finite sample sizes. In order to get a better control over the deviation of $\bar{X}_n$ from its expectation, one notices that Chebyshev's inequality is directly obtained from Markov's inequality.

Lemma 1.1.4 (Markov's Inequality) *For a nonnegative random variable Y with finite expectation, it ensures that for all $\epsilon > 0$, $\mathbb{P}(Y \geq \epsilon) \leq \epsilon^{-1}\mathbb{E}[Y]$.*

Thus, for any increasing function Φ, provided that $\Phi(Y)$ is nonnegative and has a finite expectation, Markov's inequality guarantees that for all $\epsilon > 0$,

$$\mathbb{P}(\Phi(Y) \geq \Phi(\epsilon)) \leq \Phi(\epsilon)^{-1}\mathbb{E}[\Phi(Y)].$$

This inequality cannot be improved under these assumptions. Notice that Chebyshev's inequality is derived by considering the square function for Φ and setting $Y := X - \mathbb{E}[X]$. Hence, to derive Chebyshev's inequality, the finiteness of the second order moment of X_1 is needed, as previously stated. It therefore appears that getting a better control of the deviation requires a better control of the law $\mathbb{P}_X$. Thus, if X_1 has a finite higher order moment, a similar reasoning ensures that for any $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right) \leq \min_{p \in \mathbb{N}^*} \epsilon^{-p}\mathbb{E}[\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]|\right)^p].$$

If stronger assumptions are even made, for example the existence of the moment generating function (MGF) of $|\bar{X}_n - \mathbb{E}[\bar{X}_n]|$, *i.e.* the function defined on a real interval $[-\alpha, \alpha]$ with $\alpha > 0$ by $G_X : \lambda \mapsto \mathbb{E}[\exp(\lambda X_1)]$, the Cramer-Chernoff bound (CCB) ensures that for any $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \epsilon\right) \leq \inf_{\lambda > 0} \exp\left(-\lambda\epsilon\right) \mathbb{E}[\lambda\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]|\right)].$$

Hence in order to get Gaussian-like tails for $|\bar{X}_n - \mathbb{E}[\bar{X}_n]|$, *i.e.* an exponential decay, one will need to control the MGF of $|\bar{X}_n - \mathbb{E}[\bar{X}_n]|$. Finally to get a better control over the deviation of $\bar{X}_n$ from its expectation, one needs to control $\mathbb{P}_X$ and avoid any reference to the LLCLT.

The main assumptions that will be made in the core chapters of this thesis is the subGaussianity of the considered random variables. This assumption allows to control the MGF of the variables at hand and thus provides a sharp non-asymptotic rate of convergence for the deviation of $\bar{X}_n$ from its expectation. We start by defining the notion of σ^2-subGaussian random variable. We highlight that it denotes a class of distributions rather than a single specific distribution.

Definition 1.1.1 (σ^2-subGaussian random variable) *A random variable $X \in \mathbb{R}$ is said to be σ^2-subGaussian if, for all $s \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(s(X - \mathbb{E}[X])\right)\right] \leq \exp\left(\frac{s^2\sigma^2}{2}\right).$$

As previously explained, controlling the MGF of a random variable allows to control the tightness of its tails. A subGaussian random variable function then reveals Gaussian-like tails. In addition, it can be shown that if a random variable exhibits Gaussian-like tails, one can control its MGF and thus prove that it has to be subGaussian.

Lemma 1.1.5 (Tails of σ^2-subGaussian r.v., Lemma 1.5 in [115]) *Assume X is a centered random variable such that there exists $\sigma > 0$ satisfying, for all $\epsilon > 0$,*

$$\mathbb{P}\left[X > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \text{ and } \mathbb{P}\left[X < -\epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Then for any $s > 0$, it holds

$$\mathbb{E}\left[\exp\left(sX\right)\right] \leq \exp\left(4s^2\sigma^2\right) \leq \exp\left(\frac{s^2\left(8\sigma^2\right)}{2}\right),$$

that is X is ν^2-subGaussian with $\nu^2 := 8\sigma^2$.

This leads to the following question : what are the characterisations of σ^2-subGaussian variables ? The following lemma provides equivalent characterisations.

Lemma 1.1.6 (Characterization of subGaussian r.v., Proposition 2.5.2 in [130]) *Let X be a centered random variable. Then the following statements are equivalent for finite positive constants $(C_i)_{i=1}^7$:*

$$\begin{aligned}
&\text{for all } \lambda \in \mathbb{R}, \quad \mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(C_1^2\lambda^2\right),\\
&\text{for all } \epsilon \in \mathbb{R}_+, \quad \mathbb{P}\left[|X| \geq \epsilon\right] \leq 2\exp\left(-\epsilon^2/C_2^2\right),\\
&\text{for all } k \in \mathbb{N}^*, \quad \mathbb{E}\left[|X|^k\right]^{1/k} \leq C_3\sqrt{k},\\
&\mathbb{E}\left[\exp\left(X^2/C_4^2\right)\right] \leq 2,\\
&\text{for all } \lambda \in [-\frac{1}{C_5}, \frac{1}{C_5}], \quad \mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(C_5^2\lambda^2\right).
\end{aligned}$$

Thus, Lemma 1.1.6 provides a characterization of subGaussian variables based on the MGF, the tails, the moments, the exponential moment of X^2 and the local MGF of X^2. Note that the σ^2 in definition 1.1.1 provides an upper bound on the variance of X. However, for $\nu \leq \sigma$ a random variable being ν^2-subGaussian will also be σ^2-subGaussian and thus there isn't any notion of optimality in the choice of σ^2. However, this notion of optimality can be helpful in some contexts. Thus, we use the exponential moment of X^2 for this purpose, which leads to a norm on the set of subGaussian random variables.

Definition 1.1.2 (subGaussian norm) *For a subGaussian random variable X, the subGaussian norm of X, denoted $\|X\|_{\Psi_2}$ is defined as follows :*

$$\|X\|_{\Psi_2} := \inf_{s \in \mathbb{R}_+^*} \left(\mathbb{E}\left[\exp\left(X^2/s^2\right)\right] \leq 2\right)$$

Example 1.1.1 *Consider a random variable X and a positive constant C such that $|X| \leq C$ alsmost surely. Then X is subGaussian and satisfies for any $s \in \mathbb{R}_+^*$:*

$$\mathbb{E}\left[\exp\left(X^2/s^2\right)\right] \leq \mathbb{E}\left[\exp\left(C^2/s^2\right)\right].$$

Hence for $s \geq \frac{C}{\sqrt{\log(2)}}$*, there is* $\mathbb{E}\left[\exp\left(X^2/s^2\right)\right] \leq 2$*. This proves that* X *is subGaussian and* $\|X\|_{\Psi_2} = \frac{C}{\sqrt{\log(2)}}$.

Finally, assuming that $(X_1, \dots X_n)$ are independent and subGaussian provides a control over the deviation of $\bar{X}_n$ from its expectation. Notice that the following result is even more general than the firstly considered context as the random variables need not be identically distributed. Only the independence and a control over each probability distribution are enough.

Lemma 1.1.7 (Hoeffding's inequality for σ^2-subGaussian random variables, Proposition 2.5 in [131]) *Suppose that* $(X_i)_{i\in[n]}$ *are independent r.v. and that* X_i *has mean* $\mathbb{E}[X_i]$ *and is* σ_i^2*-subGaussian for all* $i \in [n]$*. Then for all* $\epsilon > 0$*, we have*

$$\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n^2\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

In the case of *i.i.d.* random variables with $\mathbb{P}_X$ being σ^2-subGaussian, one finds that the deviation becomes $\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$. Hence independence and Gaussian-like MGFs ensure that the asymptotic behaviour of $\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right|$ is reached even for finite samples. In addition, it can be noticed that the subGaussianity of the random variables at hand can sometimes be deduced from their definition. The next lemma proves that bounded random variables are indeed subGaussians.

Lemma 1.1.8 (Hoeffding's inequality for bounded random variables) *Suppose that* $(X_i)_{i\in[n]}$ *are independent r.v. and that* X_i *has mean* $\mathbb{E}[X_i]$ *and belongs to some interval* $[a_i, b_i]$ *a.s. for all* $i \in [n]$*. Then for all* $\epsilon > 0$*, we have*

$$\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. If X_i belongs to $[a_i, b_i]$ a.s., then X_i is σ_i^2-subGaussian with $\sigma_i = \frac{b_i - a_i}{2}$. We conclude using Lemma 1.1.7 ■

In the case of *i.i.d.* random variables with $\mathbb{P}_X$ being supported on $[a, b]$, one finds that the deviation becomes $\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$. It can be noticed that in Lemma 1.1.8, there is no assumption on the second order moment of the random variables. However, a more precise bound can be derived if more details are provided on $(X_1, \dots, X_n)$. Indeed, if theses variables have a finite second order moment smaller than half the length of the interval in which they are almost surely, a better control on the deviation of $\bar{X}_n$ can be derived. Hoeffding is in fact valid in the worst case possible under the stated hypotheses cited, *i.e.* for a variable X_i that is fairly distributed between the two ends of the interval $[a_i, b_i]$. The following lemma provides a better control over the deviation of $\bar{X}_n$ for bounded random variables when their second order moment is known.

Lemma 1.1.9 (Bernstein's inequality for bounded random variables, Theorem 2.9 in [29]) *Suppose that the variables* $(X_i)_{i\in[n]}$ *are independent with finite variance and verify for* $M > 0$ *and* $v > 0$*,* $|X_i - \mathbb{E}[X_i]| < M$ *a.s. and* $\sum_{i=1}^n \mathbb{V}[X_i] = v$*. Then for all* $\epsilon > 0$*, we have*

$$\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n^2\epsilon^2/2}{v + nM\epsilon/3}\right)$$

In the case of *i.i.d.* random variables with $\mathbb{P}_X$ being supported on $[a, b]$ with finite variance σ^2, one finds that the deviation becomes $\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right)$. To conclude this section on concentration, we remind that the main tool we presented for controlling the deviation of $\bar{X}_n$ from its expectation is the control of the MGF. The previously stated results focused mainly on Gaussian-like MGFs for the considered independent random variables. Obviously, weaker assumptions can be made which will result as a weaker control on the deviation. This means that the asymptotic behaviour, guaranteed by the LLCLT, will not be reachable for finite sample sizes. To illustrate this fact, we define another class of random variables with a wider MGF than Gaussian ones.

Definition 1.1.3 ((σ^2, α)-subExponential random variable) *A random variable $X \in \mathbb{R}$ is said to be subExponential with parameters (σ^2, α) if $\mathbb{E}[X]$ is finite and its MGF satisfies, for all $s \in \mathbb{R}$ such that $|s| \leq \frac{1}{\alpha}$,*

$$\mathbb{E}\left[\exp\left(s(X - \mathbb{E}[X])\right)\right] \leq \exp\left(\frac{s^2\sigma^2}{2}\right).$$

In this case we say that X is (σ^2, α)-subExponential.

Following the definition, a random variable is subExponential if its MGF is at least Gaussian-like around zero. This will ensure that for small deviations, $\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right|$ will exhibits a Gaussian-like behaviour. However, for larger deviation this will not be the case anymore and we lose the asymptotic regime. This idea is formalised in the following result.

Lemma 1.1.10 (Bernstein's inequality for subExponential random variables, Theorem 2.8.1 in [130]) *Suppose that the variables $(X_i)_{i\in[n]}$ are independent, centered and subExponential with parameters (σ^2, α). Then for all $\epsilon > 0$, we have*

$$\mathbb{P}\left[|X_n| \geq \epsilon\right] \leq 2\exp\left(-cn\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right)\right),$$

where $K := \max_{i\in[\![1,n]\!]}(\|X_i\|_{\psi_1})$ and $c > 0$ is an absolute constant. In addition, $\|\ \|_{\psi_1}$ denotes the subexponential norm.

Hence, for $\epsilon < K$ one notices that the deviation one gets from Lemma 1.1.10 agrees with the asymptotic behaviour provided by the LLCLT. However for $\epsilon > K$ the deviation is wider and one doesn't get the same tight control over $\left|\bar{X}_n\right|$ anymore.

Finally, we have briefly presented a methodology to derive sharp bounds to control the deviation of $\bar{X}_n$ from its expectation given the existence of the MGF. As previously mentioned, if the MGF doesn't exist at least on an open interval around zero, one can leverage the higher finite moment of the random variables at hand. However, at this stage, one last question remains unanswered : is it possible to control the deviation of a quantity other than the empirical mean ? Indeed, the exposed section focused on controlling $\left|\bar{X}_n - \mathbb{E}[X_n]\right|$. Let's consider a real-valued function f. Is it possible to control the deviation of $|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]|$ at least under some regularity conditions for f ? McDiarmid's inequality, which will be used throughout the proofs exposed in this thesis, provides an answer to this question.

Lemma 1.1.11 (McDiarmid's inequality, Theorem B.5 in [64]) *Let $\mathcal{X}$ be some measurable set and f a measurable function taking its arguments in $\mathcal{X}^n$ and with values in $\mathbb{R}$. We assume that f satisfies the bounded difference assumption, meaning there exist constants $\delta_1, \dots, \delta_n$ such that for all $i \in [n]$, for all $\left(x_1, \dots, x_n, x_i^\top\right) \in \mathcal{X}^{n+1}$ such that $x_i \neq x_i^\top$,*

$$\left| f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i^\top, x_{i+1}, \dots, x_n) \right| \leq \delta_i.$$

Then for any $\epsilon > 0$ and any independent random variables $X_1, \dots, X_n$ with values in $\mathcal{X}$, we have

$$\mathbb{P}\left[|f(X_1, \dots, X_n) - \mathbb{E}\left[f(X_1, \dots, X_n)\right]| \geq \epsilon \right] \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^n \delta_i^2} \right).$$

Notice that Lemma 1.1.11 implies Lemma 1.1.8. Indeed consider $(a, b) \in \mathbb{R}^2$ and the real-valued function f defined on $[a, b]^n$ as follows $f : (x_1, \dots, x_n) \mapsto \bar{x}_n$. This function satisfies the bounded difference assumption and if the random variables $(X_1, \dots, X_n)$ are bounded in $[a, b]$ almost surely, McDiarmid's inequality then yields the Hoeffding's bound.

Finally, this section exposes the reasons why assumptions are constantly made on the moment generating functions of the considered random variables in the core chapters of this thesis. These assumptions are necessary in order to be able to control the deviation of an empirical quantity from its probabilistic value in a non-asymptotic way. This allows to have a precise control over the deviation for finite sample sizes. As the next section shows, this will also allow us to control the role played by the dimension in the deviation.

1.1.2 High-dimensional covariance matrix estimation

Let's return to the initial context and recall that we considered real random variables, *i.e.* random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in the measurable space $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ with $p = 1$. The above considerations and developments are based on an asymptotic analysis of the deviation of $\bar{X}_n$ from its expectation. The objective was then to understand how to obtain similar guarantees in a non-asymptotic framework. However, the role played by the dimension p of the random variables was omitted in this first exposition. For a fixed and arbitrarily large value of p, the LLCLT is still valid under similar assumptions. However, this asymptotic result disregards the fact that the dimension p can be of the same order of magnitude as the sample size n. In this context, sending n to infinity while keeping p fixed is not appropriate. To illustrate this point, consider the example of covariance matrix estimation, which will be the focus of chapter 2.

Consider $(X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$ on $\mathbb{R}^p$ with $\mathbb{E}[X_1] = 0$ and covariance matrix $\text{Cov}(X_1) = I_p$. The empirical covariance matrix is defined as $\hat{\Sigma} := \dfrac{1}{n} \sum_{i=1}^n X_i X_i^\top$. For any $(i, j) \in [\![1, p]\!]$, SLLN ensures that $\left[\hat{\Sigma}\right]_{ij}$ converges almost surely towards $\mathbb{1}_{i=j}$ when n goes to infinity. Hence we deduce that $\hat{\Sigma}$ converges almost surely towards I_p when n goes to infinity. It ensures that when p is kept fixed and n goes to infinity, the empirical distribution of the random eigenvalues of $\hat{\Sigma}$ converges almost surely towards the Dirac measure on 1, denoted δ_1. However, in the high dimensional asymptotic regime, when both n and p go to infinity at a constant aspect ratio *i.e.* when p/n converges towards $\beta \in (0, 1]$, the limiting distribution of the empirical spectrum is not δ_1 anymore. In this regime, the empirical distribution of the eigenvalues converges almost surely to the Marchenko-Pastur distribution, see [63]. Hence, in this high dimensional setting, $\hat{\Sigma}$ is not a good estimator of the covariance matrix anymore, even if the sample

size is huge. This exposure reinforces the need not to limit ourselves to an asymptotic study. Indeed, a non asymptotic bound will explicit the dependence on both n and p. The following lemma provides such an example and we underline that the definition of the matrix norm $\| \|_{op}$ is given in the definition 1.1.6.

Definition 1.1.4 (σ^2-subGaussian random vector) *A random vector $Y \in \mathbb{R}^p$ is said to be σ^2-subGaussian if for any vector $u \in \mathbb{R}^p$ such that $\|u\|_2 = 1$, $u^\top Y$ is σ^2-subGaussian.*

Lemma 1.1.12 (Non asymptotic rate of covariance matrix estimation, Theorem 5.7 in [115]) *Consider $\Sigma \in \mathcal{S}_p^{++}$ where $\mathcal{S}_p^{++}$ represents the set of symmetric positive definite matrices of size $p \times p$ and let $Y \in \mathbb{R}^p$ a centered σ^2-subGaussian random vector of parameter 1 such that $\mathbb{E}[YY^\top] = I_d$. Consider $X_1, \dots, X_n$ i.i.d. random vectors with the same distribution as $\Sigma^{1/2}Y$. Then $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_1 X_1^\top] = \Sigma$ and X_1 is $\|\Sigma\|_{op}$-subGaussian. Define $\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$. Then there exists a positive constant C such that for any $\epsilon > 0$,*

$$\mathbb{P}\left[\|\hat{\Sigma} - \Sigma\|_{op} > C\|\Sigma\|_{op} \max\left(\sqrt{\frac{p+\epsilon}{n}}, \frac{p+\epsilon}{n}\right)\right] \leq \exp(-\epsilon).$$

Lemma 1.1.12 indicates that for fixed p, $\hat{\Sigma}$ is a consistent estimator of Σ. However, the bound is not satisfactory when $p > n$. The problem of estimating a covariance matrix in a high dimensional regime is well-studied and we refer the reader to [39], [40], [37] and [38] for more details.

1.1.3 Random matrices with independent entries or rows

In the preceding sections, our focus has been on random vectors (or variables) defined in $\mathbb{R}^p$, with $p \geq 1$. However, chapters 3, 4 and 5 will delve into random matrices, *i.e.* $(X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$ defined on $\mathbb{R}^{p \times q}$. Here, $\mathbb{P}_X$ represents a distribution over the set of all $p \times q$ matrices. Asymptotic results pertaining to the spectrum of such random matrices can be viewed as random matrix analogues of the LLCLT, see [14] for more details. However, this thesis will focus exclusively on non-asymptotic results due to the considerations outlined earlier and the asymptotic results will only be mentioned to have an idea on the sharpest bounds one can expect to derive.

A fundamental aspect to highlight is the concept of convergence within a matrix space. The most straightforward method to define a mode of convergence is by deriving it from a distance, or more restrictively, from a norm. It's crucial to recall that a matrix can be uniquely associated with a linear transformation. Therefore, the key aspect of interest is the expansion or contraction induced by this associated linear transformation on the vectors of a basis. This essential information is fully encapsulated within the spectrum of the matrix, specifically its singular values.

Singular value decomposition (SVD). Consider $M \in \mathbb{R}^{p \times q}$ a matrix of rank r. Then M can be decomposed as $M = \sum_{i=1}^{r} \sigma_i(M) u_i v_i^\top$ where $\sigma_1(M) \geq \dots \geq \sigma_r(M) > 0$ are the singular values of M, $(u_1, \dots, u_r)$ is an orthonormal family of $\mathbb{R}^p$ and $(v_1, \dots, v_r)$ is an orthonormal family of $\mathbb{R}^q$. In addition, the squared singular values are the shared nonzero eigenvalues of $MM^\top$ and $M^\top M$ associated with the eigenvectors $(u_1, \dots, u_r)$ (respectively $(v_1, \dots, v_r)$).

This allows to partially summarize the information of a matrix $M \in \mathbb{R}^{p \times q}$ in a vector $(\sigma_1(M), \dots, \sigma_r(M)) \in \mathbb{R}^r$. Notice that we can omit the rank and define by extension $(\sigma_1(M), \dots, \sigma_m(M)) \in \mathbb{R}^m$, where $m := \min(p, q)$ and extending the definition such that $\sigma_{r+1}(M) = \dots = \sigma_m(M)$. The remaining information pertains to the vectors comprising the eigenvector bases of $MM^\top$ and $M^\top M$. Moreover, the

SVD is valuable for understanding how perturbing a matrix M by another matrix E influences the properties of M. Specifically, if we consider a signal M and a noise E as two matrices in $\mathbb{R}^{p\times q}$ and define $\Delta := M + E$, we are interested in understanding how the singular values of the perturbed signal Δ behave in comparison to those of the pure signal M. Weyl's inequality provides an answer.

Lemma 1.1.13 (Weyl's inequality, Theorem C.6 in [64]) *For two matrices A and B in $\mathbb{R}^{n\times p}$, we have for any $k \leq \min(n,p)$,*

$$|\sigma_k(A) - \sigma_k(B)| \leq \sigma_1(A-B),$$

where $\sigma_k(A)$ (respectively $\sigma_k(B)$) denotes the k^{th} largest singular value of A (respectively B).

Moreover, the SVD leads to a natural way of defining a norm on the matrix space $\mathbb{R}^{p\times q}$. The idea is to look at a matrix $M \in \mathbb{R}^{p\times q}$ as a vector in $\mathbb{R}^m$ and derive a matrix-norm from a vector-norm. Norms defined in this manner are referred to as Schatten norms.

Definition 1.1.5 (k-Schatten norms) *Consider $M \in \mathbb{R}^{p\times q}$ and $m = \min(p,q)$. For $k \in \mathbb{N}$, the k-Schatten norm of M is defined as*

$$\|M\|_k := \left(\sum_{i=1}^{m} \sigma_i(M)^k\right)^{1/k}.$$

Another way to naturally define the norm of a matrix is to consider the largest expansion it causes in any direction. To measure this expansion, we can again consider vector norms.

Definition 1.1.6 ((k,j)-operator norms) *Consider $M \in \mathbb{R}^{p\times q}$. For $(k,j) \in \mathbb{N}^2$, the (k,j) operator norm of M is defined as*

$$\|M\|_{op}^{(k,j)} := \sup_{x\in\mathbb{R}^q} \frac{\|Mx\|_j}{\|x\|_k}.$$

By convention we denote $\|M\|_{op} := \|M\|_{op}^{(2,2)}$ and this quantity is referred to as the operator norm.

Notice that for $k = \infty$, the k-Schatten norm is equal to the operator norm and that Weyl's inequality proves that singular values are 1-Lipschitz *w.r.t.* the operator norm. Finally, it is important to remind that the matrix multiplication is non commutative which will raise new challenges in the study of deviations Once we have established various norms within the matrix space, we can delve into the main topic of convergence. The first natural idea is to derive similar results for $\bar{X}_n$ defined on $\mathbb{R}^{p\times q}$ as those previously derived for $\bar{X}_n$ when the variables $(X_1,\dots,X_n)$ were defined on $\mathbb{R}^p$. The matrix-Bernstein inequality provides such a result when $q = p$ and the matrices X_i are self-adjoint, *i.e.* $X^\top = X$.

Lemma 1.1.14 (Matrix Bernstein inequality, Theorem 1.6.2 in [125]) *Consider $(X_1,\dots,X_n)$ independent centered self-adjoint random $p\times p$ matrices such that there exist positive constants C and v satisfying for all $i \in [\![1,n]\!]$, $\|X_i\|_{op} \leq C$ a.s. and $\left\|\mathbb{E}\left[\sum_{i=1}^n X_i^2\right]\right\|_{op} \leq v$. Then for every $\epsilon > 0$,*

$$\mathbb{P}\left(\left\|\bar{X}_n\right\|_{op} \geq \epsilon\right) \leq 2p\exp\left(-\frac{n^2\epsilon^2}{2v + 2nC\epsilon/3}\right).$$

Lemma 1.1.14 is a matrix generalization of Lemma 1.1.9. It is noteworthy to mention that no assumption is required on how the entries of X_1 are generated. In addition, one may also be interested in understanding the behaviour of a single matrix whose entries are randomly generated. Especially we will now distinguish two types of random matrices : ones with independent real-valued entries, at the core of chapter 3 and ones with independent vector-valued rows/columns, at the core of chapters 4 and 5. The main objective, as previously mentioned, will be to control the spectrum of those random matrices. In order to know the best result we can hope for, we first need to have an idea of the asymptotic regime. Let's first consider a random matrix $M \in \mathbb{R}^{p\times q}$ whose entries are independent centered identically distributed random variables. The limiting behavior of the extreme singular values of M as p and q grow to infinity at a constant aspect ratio $\beta \in (0,1]$ is given by the Bai-Yin's law.

Lemma 1.1.15 (Bai-Yin's law, Theorem 5.31 in [129]) *Consider $M \in \mathbb{R}^{p\times q}$ a random matrix whose entries are centered independent and identically distributed with unit variance, and finite fourth moment. Then as p and q grow to infinity at an aspect ratio $\frac{q}{p} \xrightarrow{p,q\to\infty} \beta \in (0,1]$ there is a.s.*

$$\sigma_m(M) = \sqrt{p} - \sqrt{q} + o(\sqrt{q}) \textit{and} \quad \sigma_1(M) = \sqrt{p} + \sqrt{q} + o(\sqrt{q}).$$

Remind that in the real-valued setting, the bounds provided by the LLCLT were non-asymptotically exact for the deviation of the empirical mean of independent and identically distributed gaussian random variables. The following lemma is a generalization for the spectrum of random matrices with independent gaussian entries.

Lemma 1.1.16 (Spectrum of a Gaussian matrix with independent entries, Corollary 5.35 in [129]) *Consider $M \in \mathbb{R}^{p\times q}$ a random matrix whose entries are independent standard normal random variable. Then for any $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2/2)$ there is :*

$$\sqrt{p} - \sqrt{q} - \epsilon \leq \sigma_m(M) \leq \sigma_1(M) \leq \sqrt{p} + \sqrt{q} + \epsilon.$$

More general results exist for controlling the spectrum of random matrices with independent entries, even for non subGaussian ones and non identically distributed ones. This theory, in line with what has been presented so far, goes beyond the useful framework for a proper understanding of the work presented in this thesis. Interested readers may, however, wish to consult [129] for more details.

Next, our focus is shifted on a relaxed version of the previously exposed result. We now assume that the rows of the random matrix $M \in \mathbb{R}^{p\times q}$ are subGaussian random vectors. This relaxation is important. It is then possible to interpret the random matrix M as a set of p independent random points taken in a space of dimension q. The following lemma proves that the spectrum of M can be control almost as sharply as in Lemma 1.1.16.

Lemma 1.1.17 (Spectrum of a subGaussian matrix with independent rows, Theorem 4.6.1 in [130]) *Consider $M \in \mathbb{R}^{p\times q}$ a random matrix whose rows are independent mean zero, subGaussian isotropic random vectors in $\mathbb{R}^q$. Then for any $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$ there is :*

$$\sqrt{p} - CK^2\sqrt{q} - CK^2\epsilon \leq \sigma_m(M) \leq \sigma_1(M) \leq \sqrt{p} + CK^2\sqrt{q} + CK^2\epsilon,$$

where C is a positive constant and $K := \max_{i\in[\![1,p]\!]} \|[M]_{i.}\|_{\Psi_2}$ with $\|[\ \|_{\Psi_2}$ the subGaussian norm.

Finally, the last extension we will address is the question of controlling the spectrum of a matrix whose columns are non-subGaussian random vectors. Lemma 1.1.18 provides sharp bounds in this context.

Lemma 1.1.18 (Spectrum of random matrix with independent columns) *Let M be a $p \times q$ matrix whose columns $(M_j)_{j\in[n]}$ are independent random vectors in $\mathbb{R}^p$ with the common second moment matrix Σ. Let κ be a number such that $\|M_i\|_2 \leq \sqrt{\kappa}$ almost surely for all $i \in [\![1, q]\!]$. Then for every $\epsilon > 0$, the following inequality holds with probability at least $1 - \exp(-\epsilon^2)$,*

$$\|\frac{1}{q}MM^\top - \Sigma\|_{op} \leq \max\left(\sqrt{\frac{\epsilon^2 + \log(p)}{C}}\sqrt{\frac{\kappa\|\Sigma\|_{op}}{q}}, \frac{\epsilon^2 + \log(p)}{C} \cdot \frac{\kappa}{n}\right),$$

where $C > 0$ is an absolute constant.

Proof. Theorem 5.44 in [129] considers a matrix M of size $p\times q$ whose rows M_i are independent random vectors in $\mathbb{R}^q$ with a common second moment matrix Σ. This matrix is assumed to have a uniform bound which holds almost surely on the L_2 norms of its rows. More specifically there is a number κ such that $\|M_i\|_2 \leq \sqrt{\kappa}$ almost surely for all $i \in [\![1, q]\!]$. Then for every $\delta > 0$, the following inequality holds with probability at least $1 - q\exp(-c\delta^2)$,

$$\|\frac{1}{p}M^\top M - \Sigma\|_{op} \leq \max\left(\delta\sqrt{\frac{\kappa\|\Sigma\|_{op}}{p}}, \delta^2 \cdot \frac{\kappa}{p}\right),$$

where $c > 0$ is an absolute constant. We apply this result to the transposed matrix $M^\top$ and consider $\epsilon^2 := \dfrac{\delta^2 + \log(q)}{c}$. ■

To conclude, this section briefly introduces the notion of random matrices. We are mainly interested in the control that can be obtained on the spectrum, which represents the expansion/contraction that the corresponding linear (random) operator induces on the vectors of a basis. We mainly present the non-asymptotic behaviour of the largest and smallest singular values in the framework of a matrix with independent Gaussian entries, independent subGaussian rows and independent non-subGaussian columns.

1.1.4 Stochastically dependent data

In the previous sections we focused on the study of independent random variables. However, the assumption of independence is sometimes too strong to study some real phenomena. For example, weather data cannot be modelled as a series of independent variables. Indeed the collected data on day t will influence the data that will be collected on day $t + 1$, see [17]. It is therefore necessary, especially when studying phenomena with a time component, to extend the theory presented above to non-independent variables. In a time series context, we consider random processes indexed by T, usually assumed to be a subset of $\mathbb{R}$. It possible to extend this setting and consider a random process indexed on any general abstract set T. Especially, when the set T is a subset of $\mathbb{N}$ which is finite, the random process can be identified with a random vector in $\mathbb{R}^{\mathsf{Card}(T)}$. A comprehensive probabilistic model for a random process involving the random variables $\{X_t\}_{t\in T}$ would ideally describe all the joint probability distributions of the random vectors $(X_1, \ldots, X_t)$, or equivalently all the probabilities $P[X_1 \leq x_1, \ldots, X_t \leq x_t]$ across different time points $t \in T$. A special case of a random process is one in which the joint probability distribution does not change over time. Such a random process is said to be stationary and is at the core of chapter 2.

Definition 1.1.7 (Strictly Stationary Process) *Consider a set $T \subset \mathbb{R}$ and a random process $\{X_t\}_{t\in T}$. Define $n \in \mathbb{N}^*$ and consider $F_X(x_{t_1}, \ldots, x_{t_n})$ the cumulative distribution function of the joint distribution of $\{X_t\}$ at times $t_1, \ldots, t_n$. The process $\{X_t\}$ is strictly stationary if*

$$F_X(x_{t_1+\tau}, \ldots, x_{t_n+\tau}) = F_X(x_{t_1}, \ldots, x_{t_n}) \quad \textit{for all } t_1, \ldots, t_n, t_1+\tau, \ldots, t_n+\tau \in T.$$

Consequently, a strictly stationary process exhibits means and variances that do not change over time. This allows for the definition of a weaker stationarity, named weak stationarity.

Definition 1.1.8 (Weakly Stationary Process) *Consider a set $T \subset \mathbb{R}$ and a random process $\{X_t\}_{t\in T}$. The process $\{X_t\}$ is weakly stationary if*

$$\begin{aligned} \mathbb{E}[X_t] = \mathbb{E}[X_{t+\tau}] \quad &\textit{for all} \quad (t,\tau) \in T^2, \\ \textit{Cov}(X_t, X_s) = \textit{Cov}(X_{t-s}, X_0) \quad &\textit{for all} \quad (t,s) \in T^2, \\ \mathbb{E}[|X_t|^2] < \infty \quad &\textit{for all} \quad t \in T. \end{aligned}$$

Considering a (strictly or weakly) stationary process $(X_t)_{t\in T}$, it can be noticed that its covariance matrix exhibits notable properties. Indeed the covariance between any two observations depends only on their time difference, leading to a simplified and structured representation of the covariance matrix based on a single function of the time lag. This implies that the quantity $\text{Cov}(X_t, X_s)$ for any $(t,s) \in T^2$ only depends on $|t-s|$. Hence, if $T = [\![1, n]\!] \subset \mathbb{N}^*$, the covariance matrix of $(X_t)_{t\in T}$ is a symmetric Toeplitz matrix of size $n \times n$.

Definition 1.1.9 (Symmetric Toeplitz matrix) *A matrix $\Sigma \in \mathbb{R}^{n\times n}$ is symmetric Toeplitz if Σ is symmetric and each descending diagonal from left to right is constant. Thus there exists $(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^n$ such that for any $(i,j) \in [\![1,n]\!]^2$,*

$$[\Sigma]_{ij} = \sigma_{|i-j|}.$$

Another special type of random process, which is briefly mentioned in chapter 2 is the Gaussian process. A random process is said to be Gaussian if any finite collection of random variables follows a joint Gaussian distribution.

Definition 1.1.10 (Gaussian process) *Let $\{X_t\}_{t\in T}$ be a collection of random variables indexed by a set T. The process $\{X_t\}$ is said to be a Gaussian process if for any finite subset $\{t_1, t_2, \ldots, t_n\} \subseteq T$, the random vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ follows a multivariate Gaussian distribution.*

Mathematically, a Gaussian process is completely specified by its mean function $m(t) = \mathbb{E}[X_t]$ and its covariance function $k(t,s) = \text{Cov}(X_t, X_s)$ for all $t, s \in T$. If $m(t) = 0$ for all t (zero-mean Gaussian process), then $k(t,s)$ is called the covariance function.

Chapter 2 considers the p-dimensional observations $X_1, \ldots, X_n$ which are assumed to be independent with Gaussian probability distribution $\mathcal{N}_p(0, \Sigma)$. The objective is to provide a testing procedure to determine if the covariance matrix Σ of the generating stationary process is the identity or not. The considered alternative hypotheses are sub classes of symmetric Toeplitz matrices.

Building upon the foundation of random processes and stationarity, the next step is to introduce fundamental concepts that enable the control of deviations in dependent quantities. These concepts include adapted sequences, martingale differences, and Azuma-Hoeffding inequalities. A random process is said to be adapted if, informally, the information about its value at a given time step t is only accessible for the first time at that same time step t. An adapted process is also referred to as a non-anticipating process.

Definition 1.1.11 (Adapted process) *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $I \subset \mathbb{N}$. For all $i \in I$, let $\mathcal{F}_i$ be a sub-σ-algebra of $\mathcal{A}$ such that for all $(i,j) \in I^2$, if $i \leq j$ then $\mathcal{F}_i \subseteq \mathcal{F}_j$. Let $(X_i)_{i\in I}$ be a random process on Ω with values in the measurable space (E, Ξ). We say that the process $(X_i)_{i\in I}$ is adapted to the filtration $(\mathcal{F}_i)_{i\in I}$ if for all $i \in I$, X_i is $(\mathcal{F}_i, \Xi)$ measurable.*

In addition to adapted processes, the notion of martingale difference sequences (MDS) plays a crucial role in the understanding of dependent quantities. Indeed, in most limit theorems, the assumption of independence can be relaxed with the assumption of MDS which requires weaker restrictions on the dependence structure. Informally, a random process is said to be a MDS if, conditionally on the values taken by the process up to the time step $t-1$, the expected value of X_t is null.

Definition 1.1.12 (Martingale difference) *The process $(X_i)_{i\in\mathbb{N}}$ adapted to the filtration $(\mathcal{F}_i)_{i\in\mathbb{N}}$ is a martingale difference if for all $i \in \mathbb{N}$,*

$$\mathbb{E}\left[|X_i|\right] < \infty, \quad \textit{and } \mathbb{E}\left[X_{i+1}|\mathcal{F}_i\right] = 0.$$

Finally, we present the main tool for controlling the deviation of the empirical mean of dependent quantities.

Lemma 1.1.19 (Azuma-Hoeffding's inequality, Corollary 2.20 in [131]) *Let $(X_i)_{i\in[n]}$ adapted to the filtration $(\mathcal{F}_i)_{i\in[n]}$ be a martingale difference and assume there are constants $\{(a_i, b_i)\}_{i\in[n]}$ such that each X_i belongs to $[a_i, b_i]$ almost surely. Then, for all $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left|\bar{X}_n\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 1.1.19 is a generalization of Lemma 1.1.8 for non independent random variables. To conclude, we mention that it is possible to control random processes in a more general framework. Especially, there is a large body of literature on controlling the supremum of a random process. First, we give some results on Gaussian processes. Indeed, the study of the supremum of a collection of Gaussian random variables is of fundamental importance. In such cases, certain comparison inequalities are helpful in reducing the problem at hand to the same problem for a simpler correlation matrix. Slepian's lemma states that for two Gaussian processes with the same variances, the one with larger intrinsic distances has stochastically larger maximum. Similarly, Sudakov-Fernique inequality ensures that for two Gaussian processes, the one with larger intrinsic distances has larger expected maximum. For subGaussian processes, Dudley's inequality provides an upper bound on the supremum of a random process with subGaussian increments in terms of covering numbers of T, the set used for indexing the random process. Talagrand's comparison inequality ensures that any subGaussian process is bounded by a Gaussian process. For more general processes, the study of the supremum of empirical processes has benefited greatly from the work of Talagrand and the famous result presented in [123], Theorem 1.4. Readers interested in these concepts, which go beyond the scope of this thesis, will benefit from reading the following works : [130] for an introduction to the chaining method and [92] for a more general overview.

1.1.5 Time series analysis

Time series analysis is a fundamental area of study within statistics and data science, focusing on understanding and modelling sequential data points indexed by time. This field plays a crucial role

in various disciplines, including economics, finance, engineering, epidemiology, and climate science, among others. The overarching goal of time series analysis is to extract meaningful insights, identify patterns, and make informed forecasts based on historical observations. We introduce the analysis of vector-valued time series and matrix-valued time series. In the same spirit as the section 1.2.2, we focus the presentation on linear prediction models. In particular, we seek to understand how the sequential dependence of the data impacts the performance of the estimators previously introduced.

Vector-valued time series

A general approach to time series modelling consists into plotting the series and examining the main features of the graph, checking in particular whether there is a trend, a seasonal component, any apparent sharp changes in behaviour or any outlying observations. In a trend-stationary process, the first objective will be to remove the trend in order to model the stationary process. Several methods exist for removing the trend such as a least squares estimation of the trend, a smoothing by means of moving average, or the differencing method. If the process exhibits both a trend and a seasonal component, they can both be removed by the small trend method, the moving average estimation method or the differencing method, see [31] for more details. We then assume to have observations coming from a stationary process. Indeed stationarity simplifies the analysis of time series data. When a process is stationary, its statistical properties (such as mean, variance, and autocovariance) remain constant over time. It also enables the use of time-invariant models, where the parameters of the model do not change over time.

The analysis of univariate time series benefits from an extensive list of references, see [31], [71], [62] and [31] for great introductions. However, situations may arise where the values of interest depend not only on past values but also on other variables. Therefore, it becomes necessary to consider additional variables into the forecasting model to leverage more information. This is not allowed by the standard univariate time series theory and motivates the consideration of vector-valued random processes X_t. The primary objective is to develop effective forecasting methods. We focus on linear models.

Given the d-dimensional vector-valued observations $x_1, \dots, x_p$, realizations of the p-dimensional random vectors $X_1, \dots, X_p$, the objective is to provide a forecast for the period $p+1$. A standard linear model is the following one :

$$X_t = \sum_{i=1}^{p} A_i^* X_{t-i} + \epsilon_t.$$

where for all $i \in [\![1, p]\!]$, $A_i^* \in \mathbb{R}^{d\times d}$ and $\epsilon_t \in \mathbb{R}^d$ is a zero-mean stochastic process with constant variance. However, even if ϵ_t is independent of X_t, it is not independent of X_{t+1} anymore. Thus, a careful investigation is needed to derive well performing estimators. If the dataset contains $T \in \mathbb{N}^*$ observations, the model can be rewritten as follows :

$$Y = A^* Z + E,$$

where

$$
\begin{aligned}
Y &:= \begin{pmatrix} X_p & X_{p+1} & \cdots & X_T \end{pmatrix} \in \mathbb{R}^{d\times(T-p+1)},\\
A^* &= \begin{pmatrix} A_1 & A_2 & \cdots & A_p \end{pmatrix} \in \mathbb{R}^{d\times dp},\\
Z &:= \begin{pmatrix} X_{p-1} & X_p & \cdots & X_{T-1} \\ X_{p-2} & X_{p-1} & \cdots & X_{T-2} \\ \vdots & \vdots & \cdots & \vdots \\ X_1 & X_2 & \cdots & X_{T-p} \end{pmatrix} \in \mathbb{R}^{dp\times(T-p+1)},\\
E &:= \begin{pmatrix} \epsilon_p & \epsilon_{p+1} & \cdots & \epsilon_T \end{pmatrix} \in \mathbb{R}^{d\times(T-p+1)}.
\end{aligned}
$$

In the time series literature, this model is referenced as the vector autoregressive model of order p.

Definition 1.1.13 (Vector autoregressive process of order p : $VAR(p)$) *A discrete-time d-dimensional vector-valued stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is defined as a vector autoregressive process of order p if it can be written, for any $t \in [\![p+1, \infty[\![$, as :*

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + \epsilon_t,$$

where $(A_i)_{i\in[\![1,p]\!]}$ is a set of $d \times d$ matrices corresponding to the parameters of the model, and $\{\epsilon_t\}$ is a serially uncorrelated, zero-mean stochastic process with covariance matrix $\sigma^2 I_d$.

Similar to the univariate autoregressive model, the vector autoregressive model is stationary when the effects of shocks dissipate over time. This condition holds true if all the eigenvalues of the companion-form matrix are less than one in absolute value.

Definition 1.1.14 (Companion-form matrix of $VAR(p)$ model) *The companion-form matrix of the $VAR(p)$ model from definition 1.1.13 is the following matrix :*

$$\Gamma := \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_d & 0_d & \cdots & 0_d & 0_d \\ 0_d & I_d & \cdots & 0_d & 0_d \\ \vdots & \vdots & & \vdots & \vdots \\ 0_d & 0_d & \cdots & I_d & 0_d \end{pmatrix}.$$

The usual main objectives are to estimate the transition matrices $(A_i)_{i\in[\![1,p]\!]}$ and the order of the model p. Notice that the structure of the matrices $(A_i)_{i\in[\![1,p]\!]}$ provides insight into the temporal relationships amongst the time series. In addition, the number of parameters to be estimated in a $VAR(p)$ model with fixed given p is pd^2. In the low-dimensional regime, that is $T > pd^2$, the estimation is carried out by reformulating the problem as a multivariate regression as in (1.1), see [99]. This setting has been extensively studied in the literature. In addition to previous references, see [73], [113] and [126] for an overview. In the high-dimensional regime, that is $T < pd^2$, the VAR model is ill-posed : it suffers from the over-parametrization issue. Hence the estimation is carried out by assuming sparse structures on the transition matrices and adding a regularisation into the optimisation problem. However, Lasso type estimators cannot be used without considering the temporal dependence, as shown in [120]. Hence a careful control is needed on how the dependency affects the rate of estimation.

A popular approach is the low-rank VAR model, introduced in [128]. The rank-penalized least squares estimator, introduced in (1.2) is then leveraged by [1] in this low-rank VAR context. More recently, [135] proposed a method for estimating the transition matrix using the constrained Yule-Walker equations and demonstrated its optimality under the β-mixing dependency condition. Another approach is to assume entry-wise sparsity on the transition matrices, leading to L_1 regularised least squares estimation. Authors in [139] provided guarantees for this estimator that hold even when there is temporal dependence in data. Previously, [16] examined Gaussian Vector Autoregressive models with finite lag and introduced a measure of stability based on the spectral density. The spectral density is defined as the Fourier transform of the autocovariance function of the time series. A subset selection method is proposed for vector autoregressive processes in [78]. Moreover, [15] examined a combination of low-rank and entry-wise sparsity structure on these transition matrices. Finally, these procedures assume a universal lag order applying to all components which constrains the relationship between the components. Providing an adaptive estimation with different lag structures is at the core of current research. This problem is especially tackled in [105].

Another approach for modelling high-dimensional multivariate time series is to use factor models. If we are interested in the linear dynamic structure of the d dimensional vector-valued process $\{X_t\}$, we assume the existence of a static part *i.e.* the serially uncorrelated, zero-mean stochastic process $\{\epsilon_t\}$ and a dynamic component with an unknown low-dimensional structure, denoted B^*Z_t where $B^* \in \mathbb{R}^{d\times r}$ is fixed and $\{Z_t\}$ is a latent r dimensional vector-valued process with $r \leq d$. This leads to the following factor model :

$$X_t = B^*Z_t + \epsilon_t.$$

In this setting, $\{Z_t\}$ is unobserved and thus called the factor process. In addition, the serial dependency in the process $\{X_t\}$ is only driven by the dynamic low-dimensional part B^*Z_t. The objective in this context is to derive an estimator of the fixed parameter B^*, of the factor process $\{Z_t\}$ altogether with an estimator of the dimension r of the factor process. This type of model is well studied in the literature, see *e.g.* [91], [61] and [11]. In addition, matrix factorisation techniques are recently being studied to model vector-valued time series, see [2] and [3].

Finally, we mention the change-point theory in VAR models. Change points refer to sudden or abrupt shifts in time series data, which can signify transitions between different states. Detecting these change points is valuable for modelling and predicting time series. Especially, the previously mentioned models are based on the stationarity assumption of the processes at hand. However, this assumption is broken when the data exhibits structural breaks. Detecting such breaks efficiently heavily depends on understanding the underlying mechanism of the temporal evolution of the data. In a low-rank piecewise stationary VAR model, [56] developed a test of presence of a change-point in the transition matrix with minimax guarantees. For piecewise stationary VAR models, [12] presents a R-package that implements two classes of algorithms to detect multiple change points and [134] proposed a dynamic programming algorithm consistently localizing change points even as the dimensionality, the sparsity of the coefficient matrices, the temporal spacing between two consecutive change points, and the magnitude of the difference of two consecutive coefficient matrices are allowed to vary with the sample size. For global reviews on the topic, see [145], [109] and [13].

Matrix-valued time series

As detailed above, multivariate time series analysis represents a foundational area within the discipline of time series analysis. This multivariate framework not only unveils the temporal dynamics of

the time series but also delves into the relationships among a group of time series, leveraging the available information more comprehensively. While it has been traditional to treat multiple observations as a vector, the relationships among the time series often exhibit additional structure, leading to the concept of matrix-valued time series, introduced by [132]. For example, consider meteorological data where at each time step t, one collects several weather information such as air humidity, wind speed, rainfall level and temperature in five different cities. This leads to a collection of matrices $(X_t)_t$ with 4 rows and 5 columns. In this context, the matrix structure of the data is extremely important. The variables within the same column (representing the meteorological parameter) often exhibit stronger inter-relationships, as do the variables within the same row (related to the same location). Therefore, it is essential to analyze the entire group of variables while fully preserving and leveraging its matrix structure. However, while matrix-valued data are well studied under an independence assumption, see *e.g.* [70], [87], [144] and [125], the impact of the dependence structure is still not well understood. To present the problem, we adopt the same framework as the one previously introduced.

Given the $d \times k$ matrix-valued observations $x_1, \dots, x_p$, realizations of the random $d \times k$ matrices $X_1, \dots, X_p$, the objective remains to provide a forecast for the period $p+1$. A standard linear model is the following one :

$$X_t = \sum_{i=1}^{p} A_i^* X_{t-i} B_i^* + \epsilon_t.$$

where for all $i \in [\![1, p]\!]$, $A_i^* \in \mathbb{R}^{d\times d}$, $B_i^* \in \mathbb{R}^{k\times k}$ and $\epsilon_t \in \mathbb{R}^{d\times k}$ is a serially uncorrelated, zero-mean stochastic process with constant variance. Note that an identifiability problem occurs in this model as for any $\alpha \in \mathbb{R}$, the pairs (A_i^*, B_i^*) and $(\alpha A_i^*, \alpha^{-1} B_i^*)$ lead to the same model. A common practice would be to require for all $i \in [\![1, p]\!]$, $\|A_i^*\|_F = 1$. However this requirement wouldn't be sufficient as taking $\alpha = -1$ would still lead to the same model and satisfy this requirement. The additional requirement that for all $i \in [\![1, p]\!]$, $\mathsf{Tr}(B_i^*)$ solves the issue. This leads to the matrix autoregressive model of order p.

Definition 1.1.15 (Matrix autoregressive process of order p : $MAR(p)$) *A discrete-time $d\times k$ matrix-valued stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is defined as a matrix autoregressive process of order p if it can be written, for any $t \in [\![p+1, \infty[\![$, as :*

$$X_t = \sum_{i=1}^{p} A_i^* X_{t-i} B_i^* + \epsilon_t,$$

where $(A_i^*)_{i\in[\![1,p]\!]}$ is a set of $d \times d$ matrices satisfying $\|A_i^*\|_F = 1$, $(B_i^*)_{i\in[\![1,p]\!]}$ is a set of $k \times k$ matrices satisfying $\mathsf{Tr}(B_i^*) > 0$ and $\{\epsilon_t\}$ is a serially uncorrelated, zero-mean matrix-valued stochastic process with covariance matrix $\sigma^2 I_{d\times k}$.

««««< HEAD One notices that the $MAR(p)$ model can be vectorized. Consider $\{X_t\}$ following the $MAR(p)$ from definition 1.1.15, then it satisfies for all $t \in [\![p+1, \infty[\![$ ======= One notices that the $MAR(p)$ model can be vectorized. Consider $\{X_t\}$ following the $MAR(p)$ from Definition 1.1.15, then it satisfies for all $t \in [\![p+1, \infty[\![$ »»»»> 0ec5dd4 (Intro)

$$\mathsf{vec}(X_t) = \sum_{i=1}^{p} (B_i^* \otimes A_i^*)\mathsf{vec}(X_{t-i}) + \mathsf{vec}(\epsilon_{t-i}),$$

where $\otimes$ denotes the matrix Kronecker product and vec the vectorization of a matrix by stacking its columns. Hence one can define the matrices $\Phi_i := B_i^* \otimes A_i^*$ and assume that $\mathsf{vec}(X_t)$ follows a $VAR(p)$

model. However, the matrices Φ_i present a special structure that is not leveraged in the standard $VAR(p)$ estimation procedures. Thus the initial matrix structure of the problem would be lost.

Matrix-variate time series models have garnered increasing attention within the research community, evidenced by recent publications on this emerging topic. In [47], the $MAR(1)$ model was studied under an asymptotic framework, focusing on probabilistic properties and establishing conditions on A_1^* and B_1^* for model stationarity. Estimators were defined, and their asymptotic properties were rigorously demonstrated. Building on this, [141] introduced an estimation procedure based on alternating least squares tailored for low-rank assumptions on matrices A_1^* and B_1^*, providing further insights into the derived estimators' asymptotic behaviors. These foundational works were extended to tensor scenarios in [96], maintaining autoregression of order 1 through an alternating least squares approach and continuing to offer asymptotic guarantees. They also propose to determine the autoregressive order with an information criterion based procedure. In the context where $T > p$, [77] introduces a comprehensive examination of a generalized rank-R autoregressive model of order p. Their approach involves vectorizing the problem, and their estimation procedure hinges on constrained maximum likelihood, assuming a Gaussian distribution for the vectorized noise matrix. Sparsity in the coefficients is explored in [75] by introducing spatial neighborhoods. The factor model approach is also gaining traction to study matrix-variate time series, see [136], [72] and [45].

1.2 Problems and contributions

In this section we present the statistical problems studied in the core chapters of the thesis. We first detail the hypothesis testing problem, which is central to the understanding of Chapter 2. We then explore the regression problem and especially the multivariate linear regression for which Chapter 3 provides an extension. Then we present the topic model problem, for which a dynamic extension is studied in Chapters 4 and 5.

1.2.1 Hypothesis Testing : deciding where lives a covariance matrix

In all fields, from scientific experimentation to everyday life, we are required to make decisions about risky activities based on the results of experiments or observations of phenomena in an uncertain context. The decision problem consists of deciding, on the basis of observations, between a hypothesis known as the null hypothesis, denoted H_0, and another hypothesis known as the alternative hypothesis, denoted H_1. A hypothesis test is therefore a decision-making procedure used to determine whether or not the null hypothesis can be rejected in favour of the alternative hypothesis given the observed data. We assume that the observations are realizations of the random variables $(X_1, \dots, X_n)$ taking values in $(E, \mathcal{E})$.

Definition 1.2.1 (Test procedure) *A test Δ_n is a measurable function of the observations taking its values in $\{0, 1\}$:*

$$\Delta_n : E^n \to \{0, 1\}.$$

Δ_n then separates the set of possible outcomes of some random event in two contiguous sets, H_0 is rejected whenever $\Delta_n = 1$ and not rejected whenever $\Delta_n = 0$.

We consider in Chapter 2 the observation of n *i.i.d* random vectors $(X_1, \dots, X_n)$ defined on $\mathbb{R}^p$ with a common covariance matrix $\Sigma \in \mathcal{S}_p^{++}$, where $\mathcal{S}_p^{++}$ represents the set of symmetric positive definite

matrices of size $p \times p$. The considered testing problem is :

$$H_0 : \Sigma = \{I_p\}, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}_p,$$

where $\mathcal{F}_p \subset \mathcal{S}_p^{++}$ is a set of sparse Toeplitz matrices. We consider two different alternative hypotheses, either there exists a number s of covariance elements that are significantly positive (the one-sided alternative $\mathcal{F}_p = \mathcal{F}_+(s, S, \sigma)$) or significantly different from zero *i.e.* (the two-sided alternative $\mathcal{F}_p = \mathcal{F}(s, S, \sigma)$). The alternative classes are presented in Definition 2.2.1.

In a decision problem, two types of error are possible. A type I error occurs when we decide that H_1 is true, *i.e.* observing $\Delta_n = 1$, when H_0 is actually true. A type II error occurs when we fail to reject H_0, *i.e.* observing $\Delta_n = 0$, when H_1 is true. The consequences of these two errors can be of varying degrees of importance. Every decision has thus a probability of being right and a probability of being wrong. The type I error probability, in words the worst "chance" of falsely rejecting the null hypothesis, is denoted α and is called the significance level of the test. The type II error probability, in words the worst "chance" of failing to reject the null hypothesis, is denoted $1 - \beta$. Thus β is the probability of correctly rejecting the null hypothesis and is called the power of the test.

Definition 1.2.2 (Type I and type II errors) *Consider the testing procedure Δ_n for the testing problem $H_0 : \Sigma = I_p$, vs. $H_1 : \Sigma \in \mathcal{F}_p$. Then the type I error probability of Δ_n is defined as :*

$$\alpha := \mathbb{P}_{I_p}(\Delta_n = 1).$$

Similarly, the type II error probability of Δ_n is defined as :

$$1 - \beta := \sup_{\Sigma \in \mathcal{F}_p} \mathbb{P}_\Sigma(\Delta_n = 0).$$

To define a test procedure, the ideal would obviously be to find one that minimises both risks of error at the same time. Unfortunately, one can show that they vary in opposite directions, *i.e.* any procedure that decreases α will generally increase $1 - \beta$ and vice versa. Thus there are essentially two ways to define an optimal testing procedure. The first one is the Neyman-Pearson's optimal testing procedure. In this setting, we will consider that one of the two errors is more important than the other, and try to avoid this error. Usually we choose H_0 and H_1 so that the error we are trying to avoid is the type I error. Notice that the ideal test would then almost surely never wrongly reject H_0. However, in usual cases, the only test having $\alpha = 0$ is the trivial test $\Delta_n = 0$. Thus we need to let the other error to occur. For example, in the case of a trial, we generally do everything we can to avoid convicting an innocent person, even if it means taking the risk of acquitting a guilty person. Mathematically, we fix a value for the level $\alpha \in [0, 1]$. The more serious the consequence of the type I error, the smaller α will be. However, for the same decision problem, several tests with a type I error probability smaller than α may exist. In this case, the best of these tests is the one that minimises the probability of the type II error, *i.e.* the one that maximises the power β among the tests with a level being at most α.

Definition 1.2.3 (Neyman-Pearson's optimal testing procedure) *Let's Δ^α denote the set of all testing procedures with level at most α. Then the Neyman-Pearson optimal test, denoted Δ_{NP}, is a test of level α which solves the following :*

$$\text{for all } \Sigma \in \mathcal{F}_p, \quad \mathbb{P}_\Sigma[\Delta_{NP} = 0] = \inf_{\Delta \in \Delta^\alpha} \mathbb{P}_\Sigma[\Delta = 0].$$

If it exists, Δ_{NP} is called a uniformly most powerful test.

Because the problem Δ_{NP} needs to solve does not always have a solution, the notion of optimality defined by the Neyman-Pearson's optimal testing procedure is not universal. Hence, there is a need for a more general approach to finding an optimal testing procedure. As described previously, one can't find a test both minimizing the level α and maximizing the power β as α and $1-\beta$ evolve in opposite directions. However, it is possible to minimize the sum of the type I and type II error probabilities. Hence an equal role is given to H_0 and H_1. This criterion is described as the minimax approach.

Definition 1.2.4 (Maximal testing risk) *Let's consider a testing procedure Δ and define $R(\Delta)$ its maximal testing risk :*

$$R(\Delta, \mathcal{F}_p) := \mathbb{P}_{I_p}(\Delta = 1) + \sup_{\Sigma \in \mathcal{F}_p} \mathbb{P}_{\Sigma}(\Delta = 0).$$

Then a test is said to be minimax optimal if it minimises the maximal testing risk among all testing procedures. Its maximal testing risk is then called the minimax testing risk.

Definition 1.2.5 (Minimax testing risk) *The minimax testing risk is defined as*

$$R^*(\mathcal{F}_p) := \inf_{\Delta} R(\Delta, \mathcal{F}_p).$$

If it exists, the testing procedure achieving the minimax testing risk, denoted Δ_*, is called a minimax test.

Another important point to mention is that the null hypothesis class is a singleton, namely the identity matrix. Hence the objective of the procedure is to determine whether or not it is possible to reject with high probability the hypothesis that Σ is the identity. In addition, we have chosen the alternative hypothesis classes to be a subset of sparse Toeplitz matrices, $\mathcal{F}_p = \mathcal{F}_+(s, S, \sigma)$ or $\mathcal{F}_p = \mathcal{F}(s, S, \sigma)$. Essentially, one can wonder why such a testing problem doesn't take the more general following shape :

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{S}_p^{++} \backslash \{I_p\}.$$

In this scenario, one notices that for any standard choice of distance on $\mathcal{S}_p^{++}$, *e.g.* the distance derived from the Frobenius norm, denoted as $\| \ \|_F$, we have

$$\inf_{\Sigma \in \mathcal{S}_p^{++} \backslash \{I_p\}} \| I_p - \Sigma \|_F = 0.$$

Hence it is not possible to separate the null hypothesis from the alternative one. This leads to the minimax testing risk being equal to one and thus the random guessing test becomes optimal. Hence, in this goodness of fit testing problem, it is mandatory that the alternative hypothesis class is well separated from the null hypothesis singleton. Thus for a fixed $\epsilon > 0$, we need to define $\mathcal{F}_p^{(\epsilon)}$ such that

$$\inf_{\Sigma \in \mathcal{F}_p^{(\epsilon)}} \| I_p - \Sigma \|_F \geq \epsilon.$$

From the definition of our alternative classes we see that both $\mathcal{F}_+(s, S, \sigma)$ and $\mathcal{F}(s, S, \sigma)$ are well separated from the singleton $\{I_p\}$. Finally, the optimal choice of the separation radius ϵ is discussed in the literature and can be defined as the minimax separation radius. This goes beyond the scope of this thesis. However, interested readers may wish to consult [95] and [82] for more details on minimax testing procedures.

Chapter 2 : Covariance matrix testing and support recovery. We consider $(X_i)_{i=1,\dots,n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0,\Sigma)$ where Σ has a Toeplitz structure. We then denote $\sigma_{|i-j|}$ the covariance $\mathrm{Cov}(X^i, X^j)$ for $i,j \in \{1,\dots,p\}$. First, we test whether the covariance matrix Σ is the identity matrix I_p against the one-sided alternative $\mathcal{F}_+(s,S,\sigma)$ or the two-sided alternative $\mathcal{F}(s,S,\sigma)$, see Definition 2.2.1. From an asymptotic point of view, s can tend to infinity as p tends to infinity, thus a nonparametric model is allowed, that is the number of parameters can increase. Such models have only been considered in nonparametric estimation of the spectral density of stationary time series, see [89]. First we define φ_A the linear functional of the covariance matrix Σ associated to the matrix A belonging to $\mathcal{S}_p$ as $\varphi_A(\Sigma) := \mathrm{Tr}(A\Sigma)$. The sample covariance matrix is denoted Σ_n. Thus, the covariance element σ_j, $j \geq 1$, can be written as

$$\sigma_j = \mathbb{E}[X^T A_j X] = \mathrm{Tr}(A_j\Sigma) = \varphi_{A_j}(\Sigma), \quad \text{with } [A_j]_{k\ell} = \frac{1}{2(p-j)}\mathbb{1}(|k-\ell| = j)$$

- a matrix that has 0 elements except on jth upper and lower diagonals. Similarly, the empirical estimator of σ_j can be defined as $\varphi_{A_j}(\Sigma_n)$.

In the moderately sparse case, the sum of all S values will allow to test, whereas in the highly sparse case a search over subsets of size s will be necessary. This is called a scan procedure and it is computationally fast for vectors. Note that, if the sparsity s is unknown a second search over different possible values of s will produce an aggregated procedure, free of s. In the moderately sparse case with the alternative hypothesis being $\mathcal{F}_+(s,S,\sigma)$, we consider for some threshold $t^{MS+}_{n,p}$ the test statistic Δ^{MS+}_n defined in (2.5). When the alternative hypothesis is $\mathcal{F}(s,S,\sigma)$, we consider for some threshold $t^{MS}_{n,p}$ the test statistic Δ^{MS+}_n defined in (2.6). Upper bounds on their maximal testing risks are derived respectively in Theorem 2.3.1 and Theorem 2.3.2. In the highly sparse case, when the alternative hypothesis is $\mathcal{F}_+(s,S,\sigma)$, we consider for some threshold $t^{MS+}_{n,p}$ the test statistic Δ^{HS+}_n defined in (2.7). When the alternative hypothesis is $\mathcal{F}(s,S,\sigma)$, we consider for some threshold $t^{MS}_{n,p}$ the test statistic Δ^{HS+}_n defined in (2.8). The tests Δ^{HS+}_n and Δ^{HS}_n successively try all possible sets $\mathcal{C}$ of s diagonals among the first S diagonal values. If any of these tests decides to reject H_0, then Δ^{HS+}_n also rejects H_0. Upper bounds on their maximal testing risks are derived respectively in Theorem 2.3.3 and Theorem 2.3.4.

To bound from above the maximal testing risks of the stated procedures, we give a new variant of concentration inequality for quadratic forms of large Gaussian vectors and these bounds are specified for covariance matrices that are Toeplitz with few non-null diagonals in Theorem 2.2.2. These bounds are specified for covariance matrices that are Toeplitz with few non-null diagonals in Corollary 2.2.4.

Theorem 2.2.2 The random variable $\varphi_A(\Sigma_n - \Sigma)$ is centered and sub-exponential with parameters $\left(\nu^2 = \frac{2||A\Sigma||_F^2}{n(1-K)}, b = \frac{2||A\Sigma||_\infty}{nK}\right)$, for some arbitrary K in $]0,1[$. Therefore, for any $u > 0$:

$$\mathbb{P}[\varphi_A(\Sigma_n - \Sigma) \geq \max\left\{\sqrt{u}\frac{||A\Sigma||_F}{\sqrt{n(1-K)}}, u\frac{||A\Sigma||_\infty}{nK}\right\}] \leq \exp\left(-\frac{u}{4}\right).$$

Previous concentration inequalities were given for such functionals. The closest to our case is the chi-square type concentration inequality in [121] for standardized Gaussian vectors and generalized to sub-Gaussian vectors. Let us also mention [65] who gave a Bernstein inequality for the empirical covariance element of a stationary centered Gaussian process and generalized it to locally stationary Gaussian processes.

We also propose a method to identify diagonal elements σ_j, $j = 1, ..., S$, with non-null entries in Σ, pinpointing where information may be lost in the modelling process. The objective is to properly select non-null correlation coefficients. It can be defined a lag-selection problem as estimation of η, a vector

with entries $\eta_j = \mathbf{1}(|\varphi_{A_j}(\Sigma)| > 0)$. The aim is to find a selector $\widehat{\eta}$ with $\widehat{\eta_j} = \mathbb{1}(|\varphi_{A_j}(\Sigma_n)| > \tau_n)$ that is consistent in the sense that the risk $R^{LS}(\widehat{\eta}, \mathcal{F}) = \sum_{j=1}^{S} \mathbb{E}_\Sigma[|\widehat{\eta_j} - \eta_j|]$ stays bounded. We provide in Theorem 2.4.1 an explicit value of τ_n such that the risk $R^{LS}(\widehat{\eta}, \mathcal{F})$ remains bounded by a quantity decreasing in S.

1.2.2 Regression framework

Regression analysis is a fundamental statistical method used to explore and quantify the relationship between one or more independent variables (the predictors) and a dependent variable (the target). The goal of regression analysis is to develop a predictive model that can estimate the value of the target based on the values of the predictors. This problem is at the core of chapter 3.

We observe a dataset consisting of $T \subset \mathbb{N}^*$ responses Y_t and T corresponding features X_t. The objective is to develop a model capable of predicting the response Y_{T+1} based on a new feature X_{T+1}. We write our model as follows :

$$\text{for all } t \in [T], \quad Y_t = f^*(X_t) + \epsilon_t,$$

where ϵ_t encompasses measurement errors and factors that cause Y to depend on more than just the considered X. The true function f^* is unknown, leading us to seek an appropriate f that accurately predicts Y values at new points $X = x$. A well-performing function f aids in identifying which components of X are significant for explaining Y and which are not. During data collection, there may be instances where numerous features share the same value, such as $X_i = X_j = x$ with $i \neq j$. Despite this, we might observe $Y_i \neq Y_j$, indicating that ϵ_i and ϵ_j represent irreducible errors in our model. Even with an optimal function f, predicting Y_t using f at each $X_t = x$ can still result in errors because $f(x)$ represents only one value among a distribution of potential Y_t values. One approach is to consider that the function f^* evaluated on x outputs the average of the observed values Y_t corresponding to $X_t = x$. This leads to model the regression function f^* as $f^*(x) = \mathbb{E}[Y|X = x]$. The regression function f^* is the optimal predictor of Y with respect to the mean-squared error :

$$f^* \in \operatorname{argmin}_g \mathbb{E}\left[(Y - g(X))^2 | X = x\right].$$

Furthermore, for any estimate $\hat{f}$ of f^*, we have

$$\mathbb{E}[\left(Y - \hat{f}(X)\right)^2 | X = x] = \left(f^*(x) - \hat{f}(x)\right)^2 + \mathbb{V}(\epsilon).$$

This shows that there is an irreducible error we can't shrink, namely $\mathbb{V}(\epsilon)$, even if we know the true function f^*. We are especially interested in linear models, that is when f^* is a linear function. We refer to this problem as the linear regression problem.

Vector-valued target

In the conventional regression framework, the target variables Y_t are scalar. However, in various applications, the objective is not to predict a scalar variable but rather a vector $Y_{T+1} \in \mathbb{R}^m$. We still consider that the predictors are vector-valued, namely for $t \in [\![1, T]\!]$, $X_t \in \mathbb{R}^p$. As a consequence, the regression function $f^*(x) = \mathbb{E}[Y|X = x]$ takes arguments in $\mathbb{R}^p$ and outputs values in $\mathbb{R}^m$. Without additional assumption, f^* can be estimated independently for each coordinate, leading to independent

linear regressions with real-valued targets. Indeed, the linearity assumption on f allows to rewrite the model as follows :

$$Y = XB^* + E, \tag{1.1}$$

where $Y \in \mathbb{R}^{T\times m}$ is the target matrix, $X \in \mathbb{R}^{T\times p}$ is the predictor matrix and $B^* \in \mathbb{R}^{p\times m}$ is the parameter and $E \in \mathbb{R}^{T\times m}$ is the noise matrix, usually assumed to have *i.i.d.* σ^2-subGaussian entries. One notices that for any $j \in [\![1, m]\!]$, the j^{th} column of Y, denoted $[Y]_{.j}$ only depends on the j^{th} column $[B^*]_{.j}$ of B^* and for any $i \in [\![1, T]\!]$, the i^{th} row of Y, denoted $[Y]_{i.}$, only depends on the i^{th} row $[X]_{i.}$ of X. Hence we can view this problem as p independent linear regression problems with real-valued targets :

$$\text{for all } j \in [\![1, p]\!], \quad [Y]_{.j} = X[B^*]_{.j} + [E]_{.j}.$$

This problem is an instance of multi-task learning, which is heavily studied in the literature [107, 101, 5, 119, 55, 9, 143]. Especially, an estimator of XB^* can be derived by solving p ordinary least squares problems. Let us denote $X\hat{B}$ the corresponding estimator. If E has independent σ^2-subGaussian entries, we derive from the standard OLS analysis, see [115], the existence of a positive constant C such that :

$$\frac{1}{T}\mathbb{E}\left[\left\|X\hat{B} - XB^*\right\|_F^2\right] \leq C\sigma^2\frac{pm}{T}.$$

This result proves that in the high dimensional setting, that is when $T < pm$, the mean squared prediction error of $\hat{B}$ doesn't go to zero. Hence it is natural to ask if another estimator of B^* can be derived solving this problem. Unfortunately, Corollary 4.13 in [115] proves that the least squares estimator achieves the minimax rate of estimation in the univariate Gaussian sequence model. This implies that the least squares estimator is optimal among all estimators without any prior knowledge on the structure of B^*. Since this bound is optimal, it might seem like there's no hope to solve this high-dimensional statistical problem.

Fortunately, it is often noted that high dimensional data exhibit inherent low complexity. When the low-dimensional structures are well-defined, the analysis reverts to more conventional low-dimensional statistics. However, high-dimensional data present challenges due to the unknown underlying low dimensional structures. Therefore, a fundamental task is to identify or approximate these structures. In the multivariate regression setting, there often exist shared structures across coordinates that can be exploited to improve the prediction bounds. For example, one can assume that the columns of B^* share the same sparsity pattern with only s non null entries. If each task is performed individually, this leads to the group-lasso estimator $\hat{B}^{GL}$ studied in [98]. In this setting, there exists a positive constant $C > 0$ such that the mean squared prediction error of $\hat{B}^{GL}$ becomes :

$$\frac{1}{T}\mathbb{E}\left[\left\|X\hat{B}^{GL} - XB^*\right\|_F^2\right] \leq C\sigma^2\frac{sm\log(p)}{T}.$$

We remind that the extra log factor appears because of the unknown support of the non null entries of B^*. Hence in the high dimensional regime under this sparsity structure assumption, the mean squared prediction error is converging to zero as long as $T > sm\log(p)$. Moreover, we underline that this sparsity structure assumption mimics the standard univariate one, solved with the Lasso procedure and its variant, see [124, 22, 114, 33, 19]. Fortunately, more complex structures can be captured in the multivariate regression setting. For example, if the columns of Y are correlated, one can assume a low-rank structure on B^*. This leads to the low-rank multivariate regression.

A possible solution to this problem is to consider an estimator $\hat{B}_\lambda$ of B^* that can be defined as the solution of a rank penalized version of the ordinary least squares problem. Hence for any $\lambda > 0$ we consider :

$$\hat{B}_\lambda \in \operatorname{argmin}_B \|Y - XB\|_F^2 + \lambda r_B, \tag{1.2}$$

where r_B denotes the rank of B. A first question of interest is the selection of the hyperparameter $\lambda > 0$. This problem falls into the category of model selection and we refer the reader to [64, 100] for comprehensive introductions. The first step to compute this estimator is to define the restricted rank estimators, that is $\hat{B}^{(k)}$ which minimizes $\|Y - XB\|_F^2$ among matrices B of rank no larger than k.

Lemma 1.2.1 (Lemma 8.1 in [64]) *Consider $P := X\left(X^\top X\right)^+ X^\top$ the orthogonal projector onto the range of X where $\left(X^\top X\right)^+$ denotes the Moore-Penrose pseudo inverse of $X^\top X$. Denote $\sum_{i=1}^{\operatorname{rank}(PY)} \sigma_i u_i v_i^\top$ the SVD of PY. Then $X\hat{B}^{(k)}$ can be defined as $\sum_{i=1}^{k} \sigma_i(PY) u_i v_i^\top$.*

When the rank of B^* is unknown, the previous estimator can be computed for any value of $r \in \mathbb{N}^*$, leading to $\hat{B}^{(k)}$. The quality of this estimator is given in the following lemma.

Lemma 1.2.2 (Non asymptotic bound on the squared prediction error, Theorem 5 in [32]) *There is a positive constant C such that for any $k \in \mathbb{N}^*$,*

$$\left\| X\hat{B}^{(k)} - XB^* \right\|_F^2 \leq C \left[\sum_{i=r+1}^{\operatorname{rank}(XB^*)} \sigma_i(XB^*)^2 + k\|PE\|_{op}^2 \right].$$

Note that this bound, which exhibit a bias-variance trade-off, holds almost surely but depends on the largest singular valse of the projection of the noise matrix E onto the range of X. One can derive an upper bound not depending on E by controlling the spectrum of the random matrix PE and then provide an upper bound holding true with high probability. The bounds thus derived will be more or less tight depending on the assumptions one makes on the distribution of the noise matrix E. The following lemma provides an example.

Lemma 1.2.3 (Mean squared error in the low-rank multivariate regression, Corollary 6 in [32]) *Assume that the noise matrix E has independent centered gaussian entries with variance σ^2. Then there is a positive constant C such that for any $r \in \mathbb{N}^*$,*

$$\mathbb{E}\left[\left\| X\hat{B}^{(k)} - XB^* \right\|_F^2 \right] \leq C \left[\sum_{i=r+1}^{\operatorname{rank}(XB^*)} \sigma_i(XB^*)^2 + \sigma^2 k(m + r_X) \right],$$

where r_X denotes the rank of X.

Lemma 1.2.3 shows that the mean squared error is bounded by an approximation error and a stochastic term. The approximation error is decreasing in k and vanishes for $k > \operatorname{rank}(XB^*)$. Moreover the mean squared error satisfies for $k > \operatorname{rank}(XB^*)$:

$$\frac{1}{T}\mathbb{E}\left[\left\| X\hat{B} - XB^* \right\|_F^2 \right] \leq C\sigma^2 \frac{k(m + r_X)}{T}.$$

One can then notice that $\mathrm{rank}(B^*) \geq \mathrm{rank}(XB^*)$ and that in a high-dimensional setting with very low rank, $\mathrm{rank}(XB^*)(m + r_X) \ll pm$. However, the value of $\mathrm{rank}(XB^*)$ is unknown and thus the previously stated oracle bound cannot be achieved. A data adaptive procedure is proposed in [32] both in the case of known σ^2 and unknown σ^2, the parameter of the noise. Similar performances are achieved as in the oracle case.

Hence, if the columns of the observed matrix Y are correlated and if we then assume that B^* has a low-rank structure, an estimator $\hat{B}_r$ of B^* can be derived with non-asymptotic guarantees. However, if the rows of Y are correlated, the previously exposed model cannot capture it. This can happen when the observed predictors and targets exhibit serial dependency. This problem is the core of chapter 3. To conclude, generalizing those results for higher order tensors is a matter of considerable interest within the research community. We refer the reader to [97] and references therein for a comprehensive introduction.

Chapter 3 : Two-sided matrix regression. In this chapter, we study a multivariate regression problem where both the columns and the rows of the target quantity Y are assumed to be correlated. We observe the target matrix $Y \in \mathbb{R}^{n\times p}$ and a design matrix $X \in \mathbb{R}^{m\times q}$ related via the two-sided matrix regression (2MR) model. This model involves two parameter matrices $A^* \in \mathbb{R}^{n\times m}$ and $B^* \in \mathbb{R}^{q\times p}$ and is expressed as

$$Y = A^*XB^* + E.$$

The noise matrix E is assumed to have independent centered $\sigma-$subGaussian entries. The objective is to derive predictors $\hat{A}$ and $\hat{B}$ such that $\hat{A}X\hat{B}$ stays close to the signal A^*XB^*, under low-rank assumptions on A^* and B^*.

While this model does not involve time-dependency, the non-asymptotic results obtained here can enhance our understanding of matrix-valued autoregressive time series : $Y_t = A^*X_tB^* + E_t$ (see [47]). The 2MR model also encompasses known models such as matrix regression and matrix factorization. For instance, if $n = m$ and A^* is the identity matrix, the 2MR model reduces to the one-sided matrix regression model $Y = XB^* + E$ (see [108], [32], [104]). Similarly, if $m = q$ and the design matrix X is the identity matrix with rank m smaller than both n and p, the 2MR model becomes a factorization model of the signal $M^* = A^*B^*$ observed with noise.

Another representation of the 2MR model is in the form of a *vector regression model*. By stacking the columns of matrices Y, X and E into $\mathrm{vec}(Y)$, $\mathrm{vec}(X)$ and $\mathrm{vec}(E)$, respectively, we obtain

$$\mathrm{vec}(Y)^\top = \mathrm{vec}(X)^\top \cdot (A^*)^\top \otimes B^* + \mathrm{vec}(E)^\top,$$

where $\otimes$ denotes the tensor product of two matrices. Under this formulation, we predict a row vector of size np using a row vector of size mq (with the feature matrix having rank 1) via a parameter of size $(mq) \times (np)$. This approach is problematic unless the structure of A^* and B^* is trivial. It fails to account for the matrix structure of the features and the matrices A^* and B^*, leading to suboptimal results.

The objective is to build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the squared Frobenius prediction risk under maximal rank constraint, see (3.3). Theorem 3.2.1 provides, for an equivalent problem (3.5), explicit predictors $\hat{A}_{0r}$ and $\hat{B}_{0r}$ with a non-asymptotic upper bound on the prediction risk. We notice especially that this bound can be decomposed as the sum of a bias term, which is the cause of the choice of the rank r of the predictors, potentially lower than the rank of the matrices A^* and B^* and a stochastic term. The analysis of this stochastic term mainly involves random matrix theory, see [129]. These predictors lead to derive $\hat{A}_r$ and $\hat{B}_r$ solution of the initial optimization problem (3.3). This result is stated in Corollary 3.2.2.

However, in the optimization problem (3.3), the question of how to select r arises. We propose a rank-adaptive procedure to answer it. We first select the rank $\hat{r}$ by solving a rank-penalized version of the squared Frobenius minimization problem, (3.8). Then we consider the corresponding predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$. The prediction risk of these predictors is studied in Theorem 3.2.3. The rank selection procedure (3.8) is also proven to be consistent in Proposition 3.2.6. However both these results are stated under the condition that the subGaussian parameter σ^2 of the noise matrix entries is known.

Finally, we propose a data-driven rank-adaptive procedure, allowing to select $\bar{r}$ and derive predictors $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$. These predictors exhibit non asymptotic provable guarantees without requiring the true value σ being known. To do so we modify the penalized minimization problem (3.8) by replacing the rank r with $r\widehat{\sigma}_r^2$, see (3.9), where

$$\widehat{\sigma}_r^2 = \frac{1}{np}\|Y - \hat{A}_r X \hat{B}_r\|_F^2.$$

The performance of this prediction procedure is detailed in Theorem 3.2.7.

Finally, similar to the standard linear regression scenario where the BIC estimator is replaced by its convex relaxed version, the Lasso estimator, we compare the prediction performance achieved using a rank penalty against that obtained using a nuclear norm penalty, which serves as the convex relaxation of the rank penalty. Specifically, we consider the nuclear norm penalized version of the squared Frobenius prediction risk minimization, see (3.10). We provide solutions $\bar{A}$ and $\bar{B}$ to this problem in Theorem 3.3.1 and derive a non asymptotic upper bound on the corresponding prediction risk $\|A^*XB^* - \bar{A}X\bar{B}\|_F^2$.

We conclude by noting that the two-sided matrix regression model suffers from identifiability drawbacks. Indeed many couples of matrices (A, B) solve the equation $M = AXB$ for a given matrix M. We can only hope to identify matrices A and B under very restrictive conditions where $X^\top X$ has full rank and either the matrix A or the matrix B is assumed to have known singular values, *e.g.* like a projector with singular values 1 or 0. Few other setups are known to be identifiable in the literature of factorisation of matrices, *e.g.* non-negative matrix factorisation (NMF), see [54], NMF for topic models [84], [25], [86] or covariance matrix factorization [57].

1.2.3 Topic Modeling

This section is devoted to the presentation of the topic modeling framework, which is at the core of Chapters 4 and 5. Consider a corpus comprising n textual documents written in a language characterised by a dictionary of size p. To analyze and leverage the information conveyed in these n documents, the primary goal is to derive a vector representation for this document set. This mathematical expression will enable the application of analytical tools to extract and scrutinise information more effectively. Given the varying lengths of documents, a straightforward count of each word's occurrence would not be pertinent. Consequently, for each document the focus is shifted to the frequency of appearance for individual words. Each document can thereby be represented as a point within the simplex in $\mathbb{R}^p$. This implies that the whole corpus is depicted as a set of n points within the simplex. Importantly, the order of the documents bears no significance in this context. Additionally, we assume that those n points are not linearly independent but span a subspace of $\mathbb{R}^p$ with dimension $K \ll \min(n, p)$. Interpreted as the number of topics discussed in the corpus, K plays a crucial role in capturing the underlying structure. The principal aim is to find an embedding of these n points within the lower-dimensional space $\mathbb{R}^K$. Consequently, the task is to identify a mapping from $\mathbb{R}^p$ to $\mathbb{R}^K$ such that the initial n points in $\mathbb{R}^p$ can be effectively embedded in $\mathbb{R}^K$ through this mapping.

In a more formal context, each document $j \in [n]$ is modeled as a collection of N_j words drawn from a dictionary of size p. Each document follows a discrete distribution π_j^* on the simplex of $\mathbb{R}^p$. For each document $j \in [n]$, the p-dimensional vector Y_j of word frequencies is observed and assumed to follow a multinomial distribution centered on π_j^* :

$$N_j \boldsymbol{Y}_j \sim \text{Multinomial}_p(N_j, \pi_j^*). \tag{1.3}$$

However, in real world examples, only few different topics are discussed in huge corpora of documents. This leads to assuming that the word-document probability matrix $\Pi^* = (\pi_1^*, \ldots, \pi_n^*) \in \mathbb{R}^{p \times n}$ is of rank $K \ll \min(n, p)$, the number of topics, and can be factorized as :

$$\Pi^* = A^* W^*, \tag{1.4}$$

where $A^* \in \mathbb{R}^{p \times K}$ is the word-topic probability matrix and $W^* \in \mathbb{R}^{K \times n}$ is the topic-document probability matrix.

This framework assumes that the probability of occurrence of word $i \in [p]$ in a document discussing topic $k \in [K]$ is independent of the document itself. Specifically, the probability vector π_j^* of document j, referred to as the word-document probability vector, is a convex combination of K word-topic probability vectors with weights corresponding to the allocation of K topics. From a probabilistic standpoint, this can be expressed with the total probability formula, as :

$$\mathbb{P}(\text{word } i | \text{document } j) = \sum_{k=1}^{K} \mathbb{P}(\text{word } i | \text{topic } k) \mathbb{P}(\text{topic } k | \text{document } j),$$

The primary objective within the traditional topic model framework is to recover A^* and/or W^* based on the observations $\boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n$ with or without a known fixed number of topics K. The estimation of matrices A^* and W^* serves distinct purposes. Indeed, the estimation of matrix A^* discerns the distribution of words in the dictionary given some topic, while the estimation of W^* reveals the distribution of topics given some document.

It is noteworthy that without noise, i.e., the matrix Π^* being observed, the recovery of A^* and W^* becomes an instance of non-negative matrix factorization. The non-negative matrix factorization (NMF) problem has been extensively studied, with algorithms attracting attention due to their ability to generate factors with non-negative constraints, enhancing interpretability. Commonly, NMF is formulated as the minimization of a regularized cost function [94, 93, 112], presenting non-convex optimization challenges, especially in scenarios where numerous words are absent in a single document ($N \ll p$). The main limitation of NMF is that solving the exact NMF problem, i.e., assuming a known rank K of $\Pi^* \in \mathbb{R}^{p \times n}$ and retrieving matrices $A^* \in \mathbb{R}^{p \times K}$ and $W^* \in \mathbb{R}^{K \times n}$ such that $A^* W^* = \Pi^*$, without any additional assumption, is NP-hard, see [127]. This result implies the necessity of additional assumptions to ensure the existence of fast-running algorithms capable of estimating A^* and/or W^*. Moreover, NMF algorithms face an identifiability issue. It is conceivable to find different non-negative matrices $(A_1^*, W_1^*) \in \mathbb{R}^{p \times K} \times \mathbb{R}^{K \times n}$ and $(A_2^*, W_2^*) \in \mathbb{R}^{p \times K} \times \mathbb{R}^{K \times n}$ such that $A_1^* W_1^* = A_2^* W_2^*$. Additional assumptions are required to ensure the uniqueness of the representation. The first such assumption is the *separability assumption* and was initially introduced by [54]. It ensures the uniqueness of NMF. This assumption was later incorporated into the topic model framework by [8], with the interpretation that, for each topic, there exist certain words that exclusively occur in that specific topic. These words are referred to as anchor words. The *anchor word* assumption has subsequently been adopted in most literature on topic models.

Assumption 1 (Anchor word assumption) *For each topic $k \in [K]$, there exists at least one word j such that $[A^*]_{jk} > 0$ and $[A^*]_{jl} = 0$ for $l \in [K]\backslash\{k\}$.*

Model (1.4) assumes that both the matrix of word-topic and the matrix of topic-document are static. In addition it assumes that the documents are exchangeable within the collection. Indeed the model remains the same under a permutation of the columns of the observed matrix $\boldsymbol{Y}$.

Recent works address the algorithmic aspects and give inference results in the problem of estimating the matrix A^* in a static framework under the *anchor words* assumption. For example authors in [84] propose an estimator $\hat{A}$ achieving minimax rates for dense A^*, *i.e.* not sparse, with a known, fixed K. The procedure of [84] performs an SVD on a normalized version of the matrix Y followed by an exhaustive search over a p-dimensional simplex. For unknown K and dense A^*, authors in [24] consider $\hat{A}_K$, provably achieving the minimax optimal rates in this setting. The procedure of [24] starts by recovering the anchor words and then derive an estimator from a scaled version of $YY^\top$. Sparse A^* with unknown K is tackled by [25], proposing a minimax optimal estimation procedure $\hat{A}_{sparse}$ of A^*. The procedure of [25] mainly focuses on the estimation of the portion of A^* corresponding to non-anchor words. To adapt to the sparsity of A^*, their algorithm also requires the solution of a quadratic program for each non-anchor row. Recently, several papers have also studied the problem of estimating a static W^* under various assumptions. When A^* is known, and W^* is assumed to be sparse, [23] suggests a Maximum-Likelihood Estimator (MLE) for W^*. Their analysis proved that the MLE is both minimax optimal and adaptive to the unknown sparsity in a large class of sparse topic distributions. When A^* is unknown, [23] estimates W^* by optimizing the likelihood function corresponding to a plug in estimator $\hat{A}$ of A^*. Hence the estimation error of W^* in their procedure depends on how well $\hat{A}$ estimates A^*. When both A^* and W^* are unknown with a sparsity assumption on the columns of W^* with K allowed to be large, [140] proposes computationally efficient procedures for estimating both matrices. In addition, it is possible to directly estimate W^* by assuming additional structure. Hence [86] assumes another version of the *anchor word* assumption, named *anchor document*. This assumption means that for each topic, there is a document only discussing this topic. Their procedure, called Successive Projection Overlapping Clustering (SPOC) is inspired by the Successive Projection Algorithm (SPA). The idea is to start with the singular value decomposition (SVD) of the matrix Y, and launch an iterative procedure that, at each step, chooses the maximum norm row of the matrix composed of singular vectors. Then it projects on the linear subspace orthogonal to the selected row.

Chapter 4 : Dynamic Expected Topic Model In this chapter, we assume that batches of n documents are collected in T steps over time. The aim is to consider the temporal aspect in the collection of documents and to reflect the dynamic evolution of the topics discussed in the corpora. We assume that the topic-document probability matrix W^* follows a simplex-valued autoregressive model of order one. Hence the matrix $\boldsymbol{W}^{1:T} := (\boldsymbol{W}^1, \dots, \boldsymbol{W}^T)$ is now considered random. Specifically, at each time step t, the distribution of topics given a document is a linear combination of the previous distribution and a Dirichlet-distributed noise, which drives the temporal evolution of the topics. More specifically we consider that for all $t \in [T-1]$:

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t$$

where $c^* \in (0, 1)$, and each $\boldsymbol{\Delta}^t$ is a noise matrix of size $K \times n$ such that the columns are independently and identically drawn from a Dirichlet $\mathcal{D}(\theta^*)$ distribution having parameter $\theta^* \in \mathbb{R}_+^K$. We denote α the L_1 norm of θ^* and $\tilde{\theta}^*$ its L_1-normalization. The objective of this chapter is to estimate the parameters of this autoregressive model, *i.e.* c^*, $\tilde{\theta}^*$ and α, under the assumption that the word-document probability

matrix $\mathbf{\Pi}^{1:T} := (\mathbf{\Pi}_1, \dots, \mathbf{\Pi}_T)$ is available. We call this framework the oracle case. We begin by studying the spectral properties of the empirical covariance matrix $\Sigma_{\boldsymbol{W}}^{1:T} := \frac{1}{nT} \left(\boldsymbol{W}^{1:T}\right) \left(\boldsymbol{W}^{1:T}\right)^\top$. Specifically in Theorem 4.3.3 we provide a control on its smallest eigenvalue and show that it is bounded from above and below by quantities depending on c^*, α and $\tilde{\theta}^*$ with high probability. In Proposition 4.3.1 we control its largest eigenvalue by bounding it from above and below almost surely with quantities depending exclusively on K. These results legitimise a strong assumption we are making on the spectrum of this matrix. Following the work in [84], we present an SVD-based algorithmic procedure that recovers exactly the word-topic probability matrix A^*. Projecting the word-document probability matrix $\mathbf{\Pi}^{1:T}$ on A^* allows to recover exactly the topic-document probability matrix $\boldsymbol{W}^{1:T}$. Then, we estimate the parameters $\tilde{\theta}^*$, c^* and α with the estimators defined respectively in (4.8), (4.9) and (4.11). Non asymptotic bounds on their estimation error are derived respectively in Theorem 4.4.1, Theorem 4.4.2 and Theorem 4.4.3. In particular, we prove that there exist absolute constants $C_1, C_2 > 0$ such that :

$$\mathbb{P}\left[\max\{\|\hat{\theta} - \tilde{\theta}^*\|_2,\, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\} \leq C_1 \cdot \sqrt{\frac{\log(nT)}{nT}}\right] \geq 1 - \frac{C_2}{nT},$$

Note that the dimension of the vector θ^*, which is the number K of topics.

Chapter 5 : Dynamic Topic Model In this chapter, we consider the same setting as in Chapter 4 without the word-document probability matrix $\mathbf{\Pi}^{1:T}$ being available anymore. We assume to only have access to the word-document frequency matrix $\boldsymbol{Y}^{1:T}$. Then, we first define the empirical versions of the quantities involved in the previously exposed procedure recovering A^*. This empirical adapted procedure leads to an estimator $\hat{A}$ of A^*. We provide a careful study of this estimation procedure. More precisely, we give explicit upper bounds up to log factors and their dependence on all dimensions of appearing matrices. Then we project the word-document frequency matrix $\boldsymbol{Y}^{1:T}$ onto the estimated word-topic matrix $\hat{A}$. This leads to an estimated topic-document $\hat{W}^{1:T}$. The estimators of the autoregressive parameters, introduced in Chapter 4, are adapted to this setting. Non asymptotic bounds on their estimation errors are derived respectively in Theorem 5.4.1, Theorem 5.4.2 and Theorem 5.4.3. In particular, we prove that there exist absolute constants $C_1,\, C_2 > 0$ and $a,\, b > 0$ such that :

$$\mathbb{P}\left[\max\{\|\hat{\theta} - \tilde{\theta}^*\|_2,\, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\} \leq C_1 \cdot K^a p^b \left(\sqrt{\frac{\log(nT)}{nT}} + \sqrt{\frac{\log(nT)}{N}}\right)\right] \geq 1 - \frac{C_2}{nT}.$$

This shows the additive contributions to the convergence rates of the Dirichlet noise driving the probability of topics given documents and the multinomial model of word-counts. Moreover, for very long documents, that is when $N \gg nT$, the convergence rates are only driven by the Dirichlet noise up to multiplicative terms in the number of topics K and the size of the vocabulary p.

1.3 List of publications

The core chapters of this thesis are based on the following manuscripts :

— Chapter 2, [21] : "Fast nonasymptotic testing and support recovery for large sparse Toeplitz covariance matrices" (2022) , Nayel Bettache, Cristina Butucea and Marianne Sorba.
Journal of Multivariate Analysis.

— Chapter 3, [20] : "Two-sided matrix regression" (2023), Nayel Bettache and Cristina Butucea. arXiv :2303.04694, *Electronic Journal of Statistics, tentatively accepted*,

— Chapter 4 and Chapter 5 : "Dynamic Topic Model" (2024), Nayel Bettache, Cristina Butucea and Tracy Ke,
under preparation.

Chapitre 2

Covariance matrix testing and support recovery

2.1 Introduction

Covariance matrices of high-dimensional vectors appear in machine learning, signal processing and statistical procedures. In these fields, e.g., in the test-phase of an algorithm or in the validation step of a statistical model, the quality of the residuals (the difference between the observed and the predicted values) is a good indicator of the good performance of the procedure. More precisely, the closer the residuals are to a white noise distribution, the less information was lost by the predictor or the model at hand. It is therefore natural to look for very weak, sparse information in the covariance matrix of such residuals.

Goodness-of-fit tests are designed to assess whether the underlying (unknown) covariance matrix of high-dimensional vectors is the identity (which defines the null hypothesis), or it is far from it with respect to some distance (the alternative hypothesis). The separation radius is a measure of how far the covariance matrix needs to be from the identity matrix in order to be able to distinguish it given the observations. Another important information is to recover the support of the covariance matrix, i.e., the set where the non-null values can be found. As in high-dimensional regression, this support is used to reduce dimension of the problem, produce unbiased estimators of the non-null entries and so on. A selector is a vector with coordinates taking value 1 when the covariance value is non-null, respectively 0 when it is null. The quality of a selector is appreciated with the Hamming loss, which counts the number of miss-classified coordinates. Our main interests are both testing the covariance matrix and recovering the support of significant covariance elements under the alternative hypothesis of weak sparse covariance values.

The p-dimensional observations $X_1, \dots, X_n$ are considered independent with Gaussian probability distribution $\mathcal{N}_p(0, \Sigma)$ where $\Sigma = [\sigma_{ij}]_{1 \leq i,j,p}$ belongs to the set $\mathcal{S}_p^{++}$ of positive definite symmetric matrices. Let us denote by X a generic vector with the same Gaussian $\mathcal{N}_p(0, \Sigma)$ distribution.

More particularly, when the vector X is issued from a stationary process, its covariance matrix Σ has a Toeplitz structure, that is its diagonal elements are all constant and denoted by

$$\sigma_{i,j} = \mathrm{Cov}(X^i, X^j) =: \sigma_{|i-j|}, \quad i, j \in \{1, \dots, p\}.$$

As mentioned in [46], stationary time series are used as approximations of geometrically ergodic time series (whose transition probabilities converge exponentially fast to the stationary distribution). The

information on the Toeplitz matrix is fully contained in the vector $(\sigma_0, \sigma_1, \ldots, \sigma_{p-1})$ of its diagonal values. More generally, any covariance matrix can be similarly studied by looking at the energy of each diagonal of the covariance matrix, that is its Euclidean norm $\sigma_k = \|(\sigma_{1,k+1}, \ldots, \sigma_{p-k,p})\|_2$. Here, our efforts are devoted to quantifying the benefits of the Toeplitz structure in terms of rates for testing and for support recovery. Indeed, the Toeplitz structure helps improving the rates for testing and lag selection when the dimension p grows, and we do not have here a curse but a blessing of dimensionality. All methods are evaluated for all possible values of p less than or greater than n, without restriction.

In this paper is given a new variant of concentration inequality for quadratic forms of large Gaussian vectors and these bounds are specified for covariance matrices that are Toeplitz with few non-null diagonals. We show non-asymptotic separation rates for testing large sparse Toeplitz covariance matrices which are remarkably fast due to the structure of the matrix. The aim is to test here whether the covariance matrix is the identity matrix I_p or there exists a number s of covariance elements among $\sigma_1, \ldots, \sigma_{p-1}$ that are significantly positive (one-sided alternative), respectively significantly different from zero (two-sided alternative). The test procedure combines a sum and a scan procedure in order to detect small (relatively) numerous non-null entries and very few but sufficiently large entries, respectively. This is analogous to but more general than the detection of sparse Gaussian means [53, 80, 81] where observations have the same variance, whereas our model is heteroscedastic.

Moreover, we propose a selector of the diagonals with non-null entries - a lag selector, which is constructed by universal thresholding of some linear estimators. Fast non asymptotic bounds are provided for the expected value of its loss.

Experimental results show the excellent behaviour of these procedures with small values of n (non-asymptotic character of our results) and large values of p. Indeed, by exploiting the Toeplitz structure, the matrix size p does not act as a nuisance parameter anymore, but diminishes the convergence rates. All test procedures and the lag-selector are computationally trivial to implement. Note that the scan procedure is performed on a vector as well and it is therefore computationally fast, in contrast with the scan procedure of matrices, see e.g. [6, 34].

High-dimensional statistics is the major research topic nowadays as attest many recent international events and numerous collections of papers such as [66, 4, 111]. The study of the covariance operator is very often at the core of functional data analysis. Our manuscript contributes in that sense and it makes a first step towards dynamic modelling of time series in the sense that the dimension p may grow when the sample size n increases and, moreover, the sparsity parameter s may evolve with p and n. This may happen within the framework of stationary time series when the sequence of auto-correlations is sparse but infinite : depending on p and n the noise level in the model is more or less important and therefore, s can be viewed as the number of sufficiently significant correlations (above the corresponding noise leve) that obviously increases with the accuracy (that is when p and n increase).

Previously, Cai and Ma [43] considered the same goodness-of-fit test with alternative characterized by covariance values that belong to an $\mathbb{L}_2$ ball of fixed radius. Tests for sparse covariance matrices were given by Arias-Castro, Bubeck and Lugosi [7, 6]. They considered alternative covariance matrices having at most s significant values and also the structured alternative of a clique of size s producing a small submatrix of significant values. Our testing rates are faster, but they are difficult to compare as the Toeplitz structure does not allow for the block or the clique sparsity structure in their paper. Butucea and Zgheib [35, 36] considered the test problem with alternatives that generalize the $\mathbb{L}_2$-ball in [43] to dense ellipsoids for both Toeplitz and not necessarily Toeplitz covariance matrices, respectively. More precisely, it was assumed that σ_k decreased slowly as a polynomial (Sobolev ellipsoids) or faster, as an

exponential of k. The test procedure involved an optimal banding parameter - specific for testing and different from the optimal parameter for estimation of the matrix. It was thus noticed that the minimax rates for goodness-of-fit testing of large covariance matrices are faster for Toeplitz matrices than for non Toeplitz ones, and that they are faster for testing than for estimation of the covariance matrix. In this paper, an alternative class is considered where at most s significant values appear sparsely.

Cai and Liu [41] and Cai, Liu and Xia [42] considered the problem of support recovery in the sense that the estimated set $\hat{\mathcal{C}}_n$ is different from the true set $\mathcal{C}$ with probability tending to 0. To the best of our knowledge, no quantitative rates were given for support recovery in the covariance matrix setup. In the context of Toeplitz covariance matrices, we call this problem lag-selection.

Our bounds for testing and lag selection are non-asymptotic, thus n can be equal to 1 when one cannot observe repeated measurements. However, an important remark is that the rates are faster when the significant covariance values have lags in the recent past : $k \leq S$, for some $S < p$. Indeed, the rates depend on $p - S$. From an asymptotic point of view, s can tend to infinity as p tends to infinity, thus a nonparametric model is allowed (in the sense that the number of parameters increases). Such models have only been considered in nonparametric estimation of the spectral density of stationary time series, see Kreiss, Paparoditis and Politis [89] who uses thresholded empirical covariance coefficients.

2.2 Linear functionals of the covariance matrix

We define φ_A the linear functional of the covariance matrix Σ associated to the matrix A belonging to $\mathcal{S}_p$ (the set of symmetric $p \times p$ matrices) as $\varphi_A(\Sigma) = \mathrm{Tr}(A\Sigma)$.

Recall that $\mathrm{Tr}(A^2)$ is also denoted by $||A||_F^2$, the squared Frobenius norm, for any A in $\mathcal{S}_p$. The largest eigenvalue of the matrix A is denoted by $\|A\|_\infty$.

Recall that a centered real-valued random variable Z is sub-exponential with positive parameters (ν^2, b) if

$$\mathbb{E}[\exp(tZ)] \leq \exp\left(\frac{\nu^2 t^2}{2}\right), \quad |t| \leq \frac{1}{b}. \tag{2.1}$$

The sample covariance matrix is denoted

$$\Sigma_n = \frac{1}{n}\sum_{k=1}^{n} X_k X_k^T.$$

The next theorem states that for $X_1, \ldots, X_n$ independent multivariate Gaussian $\mathcal{N}_p(0, \Sigma)$ vectors, the random variable $Z = \varphi_A(\Sigma_n - \Sigma)$, for A in $\mathcal{S}_p$, is sub-exponential with explicit values for the parameters (ν^2, b). We recall the Bernstein inequality that holds for sub-exponential random variables [131].

Proposition 2.2.1 *If Z is a sub-exponential random variables with parameters (ν^2, b), then*

$$\mathbb{P}[Z \geq t] \leq \begin{cases} \exp\left(-\frac{t^2}{2\nu^2}\right), & \text{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ \exp\left(-\frac{t}{2b}\right), & \text{if } t > \frac{\nu^2}{b}. \end{cases}$$

Equivalently, for $t_u = \max(\nu\sqrt{u}, bu)$, Z satisfies :

$$\mathbb{P}[Z \geq t_u] \leq \exp\left(-\frac{u}{2}\right), \quad u > 0.$$

Thus, a concentration inequality for the plug-in estimator $\varphi_A(\Sigma_n)$ of $\varphi_A(\Sigma)$ follows immediately.

Theorem 2.2.2 *The random variable $\varphi_A(\Sigma_n - \Sigma)$ (respectively $\varphi_A(\Sigma - \Sigma_n)$) is centered and sub-exponential with parameters $\left(\nu^2 = \frac{2||A\Sigma||_F^2}{n(1-K)}, b = \frac{2||A\Sigma||_\infty}{nK}\right)$, for some arbitrary K in $]0,1[$. Therefore :*

$$\mathbb{P}[\varphi_A(\Sigma_n - \Sigma) \geq t_u] \leq \exp\left(-\frac{u}{4}\right), \quad u > 0, \tag{2.2}$$

with $t_u = \max\left\{\sqrt{u}\frac{||A\Sigma||_F}{\sqrt{n(1-K)}}, u\frac{||A\Sigma||_\infty}{nK}\right\}$

Previous concentration inequalities were given for such functionals. The closest to our case is the chi-square type concentration inequality in Spokoiny and Zhilova [121] for standardized Gaussian vectors and generalized to sub-Gaussian vectors. They generalized Hsu, Kakade and Zhang [74] who assumed finite exponential moments of any order for the vector X. Let us also mention Giurcanu and Spokoiny [65] who gave a Bernstein inequality for the empirical covariance element of a stationary centered Gaussian process and generalized it to locally stationary Gaussian processes.

Let us also mention the Hanson-Wright inequality which is stated for more general sub-Gaussian vectors but having independent components i.e. a diagonal covariance matrix (see Rudelson and Vershynin [118] and its improvement under Bernstein condition on moments by Bellec [18]).

The concentration inequality (2.2) is the main tool in the applications considered hereafter to study stationary time series. In this context, $X_1, \ldots, X_n$ are assumed to be repeated, independent observations of length p of an underlying stationary process $X = \{X^1, \ldots, X^p\}$. Note that our results are non-asymptotic, thus n can be equal to 1. Without loss of generality, the process is assumed to be centered. The covariance matrix of a stationary process is a Toeplitz covariance matrix. Let's denote $\sigma_j = \text{Cov}(X^i, X^{i+j})$ for arbitrary integer number i. Let us denote by $\mathcal{T}_p$ the set of $p \times p$ Toeplitz matrices and by $|\mathcal{A}|$ the cardinal of a set $\mathcal{A}$.

Definition 2.2.1 *$\mathcal{F}_+(s, S, \sigma)$ is defined, for $\sigma > 0$ real number and $s \leq S$ integer numbers between 1 and $p-1$, as the set of sparse Toeplitz covariance matrices Σ such that there are s significantly positive covariance elements with lags no larger than S :*

$$\mathcal{F}_+(s,S,\sigma) = \left\{\Sigma \in \mathcal{S}_p^{++} \cap \mathcal{T}_p \text{ and there exists } \mathcal{C} \subseteq \{1,\ldots,S\},\ |\mathcal{C}| = s,\ \forall j \in \{1,p-1\},\ \begin{array}{ll} \sigma_j \geq \sigma > 0, & j \in \mathcal{C}, \\ \sigma_j = 0, & j \notin \mathcal{C} \end{array}\right\}.$$

Similarly, the two-sided set $\mathcal{F}(s, S, \sigma)$ is defined :

$$\mathcal{F}(s,S,\sigma) = \left\{\Sigma \in \mathcal{S}_p^{++} \cap \mathcal{T}_p \text{ and there exists } \mathcal{C} \subseteq \{1,\ldots,S\},\ |\mathcal{C}| = s,\ \forall j \in \{1,p-1\},\ \begin{array}{ll} |\sigma_j| \geq \sigma > 0, & j \in \mathcal{C}, \\ |\sigma_j| = 0, & j \notin \mathcal{C} \end{array}\right\}.$$

Let us apply Theorem 2.2.2 to several choices of the matrices A. First, the covariance element σ_j, $j \geq 1$, can be written as $\sigma_j = \mathbb{E}[X^T A_j X] = \text{Tr}(A_j\Sigma)$, with $[A_j]_{k\ell} = \frac{1}{2(p-j)} I(|k-\ell| = j)$ - a matrix that has 0 elements except on jth upper and lower diagonals. Note that the notation A_j is used instead of $A_{\{j\}}$. The empirical estimator of σ_j is

$$\widehat{\sigma_j} = \frac{1}{n}\sum_{k=1}^{n} X_k^T A_j X_k = \text{Tr}(A_j\Sigma_n).$$

Remark 2.2.1 *It is useful to note that our results can be generalized to time series that are "nearly" stationary, by considering*

$$\widetilde{\sigma_j} = \mathrm{Tr}(A_j\Sigma_n) = \frac{1}{2(p-j)} \sum_{i,k=1,|i-k|=j}^{p} \sigma_{i,k}.$$

In this case, slightly different sets of sparse covariance matrices are considered : $\widetilde{\mathcal{F}}_+(s,S,\sigma)$ *and* $\widetilde{\mathcal{F}}(s,S,\sigma)$*, not necessarily Toeplitz matrices with* s *diagonal average values* $\widetilde{\sigma_j}$ *of the first* S *being significant. By taking into consideration that all studied methods in the sequel for testing and lag selection are exclusively based on the concentration of the mean empirical correlations around their expected values* $\widetilde{\sigma_j}$*, the following results remain valid provided that* $||A\Sigma||_F$ *and* $||A\Sigma||_\infty$ *are controlled.*

Let $W \subseteq \{1,\dots,S\}$ be a set of w values between 1 and S. $\sum_{j\in W} A_j$ is denoted by A_W and $\sum_{j\in W}\sigma_j = \sum_{j\in W}\mathrm{Tr}(A_j\Sigma)$ can then be written $\mathrm{Tr}(A_W\Sigma)$. This allows to estimate $\sum_{j\in W}\sigma_j$ by a plug-in estimator, $\mathrm{Tr}(A_W\Sigma_n)$.

Next Proposition gives properties of the matrix A_W.

Proposition 2.2.3 *Let* $W \subseteq \{1,\dots,S\}$ *contain* w *elements and* $A_W = \sum_{j\in W} A_j$*. Then :*

1. $||A_W||_\infty \le \frac{w}{p-S}$, $||A_W||_F^2 \le \frac{w}{2(p-S)}$
2. *For any covariance matrix* Σ *belonging to* $\mathcal{F}(s,S,\sigma)$*,*

$$||A_W\Sigma||_\infty \le \sigma_0\frac{w(2s+1)}{p-S}, ||A_W\Sigma||_F^2 \le \sigma_0^2\cdot\begin{cases} \frac{\mathcal{K}(2s+1)}{(p-S)}, & w=1,\\ \frac{w(2s+1)^2}{2(p-S)}, & w>1,\end{cases} \text{ with } \mathcal{K} = \begin{cases} 1, & W\subseteq\{1,\dots,\frac{p}{2}-1\},\\ \frac{p}{2}, & W\subseteq\{\frac{p}{2},\dots,p-1\}.\end{cases}$$

The next Corollary specifies the concentration inequality in Theorem 2.2.2 using the bounds in Proposition 2.2.3.

Corollary 2.2.4 *Let* $X_1,\dots,X_n$ *be i.i.d,* $\mathcal{N}_p(0_p,\Sigma)$*,* Σ *belonging to* $\mathcal{F}_+(s,S,\sigma)$ *or* $\mathcal{F}(s,S,\sigma)$ *and* $W \subseteq \{1,\dots,S\}$ *with* $S < \frac{p}{2}$ *having* w *elements. Then, for some arbitrary* K *in* $]0,1[$*,*

$$\mathbb{P}_{I_p}[\varphi_{A_W}(\Sigma_n - I_p) \ge \sigma_0\cdot t] \le \exp\left(-\frac{u}{4}\right), \quad u>0, \tag{2.3}$$

where

$$t = \max\left\{\sqrt{\frac{u}{2(1-K)}}\sqrt{\frac{w}{n(p-S)}}, \frac{u}{K}\frac{w}{n(p-S)}\right\}.$$

Moreover, for any Σ *in* $\mathcal{F}(s,S,\sigma)$*,*

$$\mathbb{P}_{\Sigma}[\varphi_{A_W}(\Sigma_n - \Sigma) \ge \sigma_0\cdot\tilde{t}] \le \exp\left(-\frac{u}{4}\right), \qquad u>0, \tag{2.4}$$

where for $w=1$*,*

$$\tilde{t} = \max\left\{\sqrt{\frac{u}{(1-K)}}\sqrt{\frac{2s+1}{n(p-S)}}, \frac{u}{K}\frac{2s+1}{n(p-S)}\right\}$$

and for $w>1$*,* $\tilde{t} = (2s+1)t$*.*

Similar inequalities hold for $|\varphi_{A_W}(\Sigma_n - I_p)|$ and $|\varphi_{A_W}(\Sigma_n - \Sigma)|$ with the exponential term being multiplied by a factor two respectively in (2.3) and (2.4).

If $W = \{1,\dots,S\}$, it is enough to replace w by S in the previous results. However, if $W=\{j\}$ for some $j \le S$, the previous results are still true with w replaced by 1.

From now on, we assume that $S < \frac{p}{2}$ such that $\mathcal{K}=1$ in the previous proposition. Indeed, in the context of time series, it is natural to look for significant correlations in the recent past.

2.3 Non-parametric testing for stationary time series

From now on is assumed for simplicity that $\sigma_0 = 1$, thus dealing with correlation matrices only. The one-sided test problem is

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}_+(s, S, \sigma).$$

The following two-sided test problem will also be discussed as a generalization :

$$H_0 : \Sigma = I_p \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}(s, S, \sigma).$$

Recall that a test procedure Δ_n is a binary valued random variable $\Delta_n : (\mathbb{R}^p)^{\otimes n} \to \{0, 1\}$. It separates the set of possible outcomes of some random event in two contiguous sets, H_0 is rejected whenever $\Delta_n = 1$ and not rejected whenever $\Delta_n = 0$. The maximal testing risk is defined as

$$R(\Delta_n, \mathcal{F}_+) = \mathbb{P}_{I_p}(\Delta_n = 1) + \sup_{\Sigma \in \mathcal{F}_+} \mathbb{P}_\Sigma(\Delta_n = 0),$$

that is the sum of the type I and the maximal type II error probabilities over the set in the alternative hypothesis. A separation rate is the least possible value for $\sigma > 0$ such that the maximal testing risk stays below some prescribed value.

We proceed by considering successively two measures of the separation between I_p and Σ under the alternative hypothesis H_1. The sets $W = \{1, \dots, S\}$, $W = \mathcal{C}$, and an arbitrary subset of $\{1, \dots, S\}$ with s elements are successively chosen. For testing over $\mathcal{F}_+(s, S, \sigma)$, consider

$$\mathrm{Tr}(A_{1:S}), \qquad \max_{\mathcal{C} \subseteq \{1,\dots,S\}, \#\mathcal{C} = s} \mathrm{Tr}(A_{\mathcal{C}}\Sigma).$$

Correspondingly, over $\mathcal{F}(s, S, \sigma)$ are considered

$$\sum_{j=1}^{S} |\sigma_j| = \sum_{j=1}^{S} |\mathrm{Tr}(A_j\Sigma)|, \qquad \max_{\mathcal{C} \subseteq \{1,\dots,S\}, \#\mathcal{C} = s} \sum_{j \in \mathcal{C}}^{S} |\mathrm{Tr}(A_j\Sigma)|.$$

By analogy to the vector case, moderately sparse and highly sparse covariance structures are distinguished. In the first case, the sum of all S values will allow to test, whereas in the latter a search over subsets of size s will be necessary. This is called a scan procedure and it is computationally fast for vectors. Note that, if the sparsity s is unknown a second search over different possible values of s will produce an aggregated procedure, free of s.

2.3.1 Moderately sparse covariance structure

When the alternative hypothesis is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS+}$ the test statistic

$$\Delta_n^{MS+} = \mathbb{1}\left(\varphi_{A_{1:S}}(\Sigma_n - I_p) \geq t_{n,p}^{MS+}\right). \tag{2.5}$$

Theorem 2.3.1 *The test Δ_n^{MS+} defined in (2.5), with*

$$t_{n,p}^{MS+} = \max\left\{\sqrt{\frac{u\cdot S}{n(p-S)}}, \frac{2u\cdot S}{n(p-S)}\right\}, \quad u>0,$$

is such that if $\sigma \geq \frac{2(s+1)}{s} t_{n,p}^{MS+}$,

$$R(\Delta_n^{MS+}, \mathcal{F}_+) \leq 2\exp\left(-\frac{u}{4}\right).$$

When the alternative set of hypothesis is $\mathcal{F}(s,S,\sigma)$, we consider for some threshold $t_{n,p}^{MS}$ the test statistic

$$\Delta_n^{MS} = \mathbb{1}\left(\sum_{i=1}^{S} |\varphi_{A_i}(\Sigma_n - I_p)| \geq t_{n,p}^{MS}\right). \tag{2.6}$$

Theorem 2.3.2 *The test Δ_n^{MS} defined in (2.6), with*

$$t_{n,p}^{MS} = S\max\left\{\sqrt{\frac{4u\log(S)}{n(p-S)}}, \frac{8u\log(S)}{n(p-S)}\right\}, \quad u>1,$$

is such that if $\sigma \geq t_{n,p}^{MS} + \max\left\{\sqrt{\frac{4(u-1)(2s+1)\log(S)}{n(p-S)}}, \frac{8(u-1)(2s+1)\log(S)}{n(p-S)}\right\}$,

$$R(\Delta_n^{MS}, \mathcal{F}) \leq 4\exp\left(-(u-1)\log(S)\right).$$

2.3.2 Highly sparse covariance structure

Let us consider now for some threshold $t_{n,p}^{HS+}$ the test statistic

$$\Delta_n^{HS+} = \max_{\mathcal{C}\subseteq\{1,\dots,S\}, \#\mathcal{C}=s} \mathbb{1}\left(\varphi_{\mathcal{A}_\mathcal{C}}(\Sigma_n - I_p) \geq t_{n,p}^{HS+}\right). \tag{2.7}$$

The test Δ_n^{HS+} successively tries all possible sets $\mathcal{C}$ of s diagonals among the first S diagonal values. If any of these tests decides to reject H_0, then Δ_n^{HS+} also rejects H_0, otherwise Δ_n^{HS+} accepts the null hypothesis H_0.

Theorem 2.3.3 *The test Δ_n^{HS+} defined in (2.7), with*

$$t_{n,p}^{HS+} = \max\left\{\sqrt{\frac{4u\cdot s\log\binom{S}{s}}{n(p-S)}}, \frac{8u\cdot s\log\binom{S}{s}}{n(p-S)}\right\}, \quad u>1,$$

is such that if $\sigma \geq \frac{1}{s}\left(t_{n,p}^{HS+} + (2s+1)\max\left\{\sqrt{\frac{u\cdot s}{n(p-S)}}, \frac{2u\cdot s}{n(p-S)}\right\}\right)$,

$$R(\Delta_n^{HS+}, \mathcal{F}^+) \leq \exp\left(-(u-1)\log\binom{S}{s}\right) + \exp\left(-\frac{u}{4}\right).$$

When the alternative set of hypotheses is $\mathcal{F}(s,S,\sigma)$, we consider for some threshold $t_{n,p}^{HS}>0$ the test statistic

$$\Delta_n^{HS}=\max_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\mathbb{1}\left(\sum_{j\in\mathcal{C}}|\varphi_{A_j}(\Sigma_n-I_p)|\geq t_{n,p}^{HS}\right). \tag{2.8}$$

Theorem 2.3.4 *The test Δ_n^{HS} defined in (2.8), with*

$$t_{n,p}^{HS}=s\max\left\{\sqrt{\frac{4u\log\left(s\binom{S}{s}\right)}{n(p-S)}},\frac{8u\log\left(s\binom{S}{s}\right)}{n(p-S)}\right\},\quad u>1,$$

is such that if $\sigma\geq t_{n,p}^{HS}+\max\left\{\sqrt{\frac{4(u-1)\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}},\frac{8(u-1)\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}\right\}$,

$$R(\Delta_n^{HS},\mathcal{F})\leq 4\exp\left[-(u-1)\log\left(s\binom{S}{s}\right)\right].$$

Remark 2.3.1 When the separation is measured by $\max_{\mathcal{C}}\sum_{j\in\mathcal{C}}\sigma_j$, its estimator is known as the scan statistic. Note that the computations are not very involved. Indeed, after computing $\xi_1=\varphi_{A_1}(\Sigma_n-I_p),\dots,\xi_S=\varphi_{A_S}(\Sigma_n-I_p)$, these values are sorted in decreasing order : $\xi_{(1)}\geq\xi_{(2)}\geq\cdots\geq\xi_{(S)}$, and then

$$\max_{\mathcal{C}\subseteq\{1,\cdots,S\},\#\mathcal{C}=s}\sum_{j\in\mathcal{C}}\varphi_{A_j}(\Sigma_n-I_p)=\xi_{(1)}+\cdots+\xi_{(s)}$$

Similar calculations hold for $\max_{\mathcal{C}}\sum_{j\in\mathcal{C}}|\sigma_j|$ and $|\xi|_{(1)}\geq|\xi|_{(2)}\geq\cdots\geq|\xi|_{(S)}$. The Toeplitz structure is thus exploited which reduces the matrix structure to a vector and makes the scan statistic computationally efficient.

Remark 2.3.2 *Note that the previous tests must be aggregated over a set of possible values for s in order to be free of the sparsity s : $\tilde{\Delta}_n^{HS}=\max_s\Delta_n^{HS}$ will reject wherever at least one test rejects.*

Remark 2.3.3 *If $S\asymp\log(p)$, giving $p-S\asymp p$, the series has short memory. Then $t_{np}^{MS+}\asymp\sqrt{\log(p)/(np)}$ which gives a test rate smaller than $\sqrt{\log(p)/(np)}$, and with Stirling's approximation, $t_{np}^{HS+}\asymp s\sqrt{\log\left(\frac{\log(p)}{s}\right)/(np)}$ giving the following bound for the testing rate $\sqrt{\frac{\log(\log(p)/s)}{np}}+\sqrt{\frac{s}{np}}$. Thus Δ_n^{HS+} detects smaller values of σ than Δ_n^{MS+} when $s\leq\log(p)$, hence our choice to name the procedures MS and HS respectively.*

Remark 2.3.4 *If the stationary time series has longer memory, for example $S=p/2-1$, this gives $p-S=p/2+1$ and $\frac{S}{p-S}\asymp 1$. In this case, $t_{np}^{MS+}\asymp 1/\sqrt{n}$ and $\sigma\geq 1/\sqrt{n}$, while $t_{np}^{HS+}\asymp s\sqrt{\frac{\log(p/s)}{np}}+\sqrt{\frac{s}{np}}$. Again, if $s/p\to 0$, the test Δ_n^{HS+} detects smaller values of σ then Δ_n^{MS+}. However, if $s=S\asymp\frac{p}{2}$, it is sufficient to use only Δ_n^{MS+}.*

TABLE 2.1 – The four test statistics to test $H_0 : \Sigma = I_p$ vs. $H_1 : \Sigma \in \mathcal{F}_+(s, S, \sigma)$ (MS^+ and HS^+) or $H_1 : \Sigma \in \mathcal{F}(s, S, \sigma)$ (MS and HS). Are presented the threshold values $t_{n,p}$ and the lower bounds σ_0 on σ, conditions under which a small upper bound R_{max} on the maximal testing risk is guaranteed. The threshold $t_{n,p}$ is the smallest value of the test statistic for which the null hypothesis is rejected. The σ_0 value is the smallest non null entry of Σ under the alternative hypothesis that can be tested and R_{max} is the upper bound of the sum of the type I and the maximal type II error probabilities.

Test	Expression	Threshold
Δ_n^{MS+}	$\mathbb{1}\left(\varphi_{A_{1:S}}(\Sigma_n - I_p) \geq t_{n,p}^{MS+}\right)$	$t_{n,p}^{MS+} = \max\left\{C_1\sqrt{\frac{S}{n(p-S)}}, C_2\frac{S}{n(p-S)}\right\}$
Δ_n^{HS+}	$\max_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s} \mathbb{1}\left(\varphi_{\mathcal{A}_\mathcal{C}}(\Sigma_n - I_p) \geq t_{n,p}^{HS+}\right)$	$t_{n,p}^{HS+} = \max\left\{C_1\sqrt{\frac{s\log\binom{S}{s}}{n(p-S)}}, C_2\frac{s\log\binom{S}{s}}{n(p-S)}\right\}$
Δ_n^{MS}	$\mathbb{1}\left(\sum_{i=1}^S \lvert\varphi_{A_i}(\Sigma_n - I_p)\rvert \geq t_{n,p}^{MS}\right)$	$t_{n,p}^{MS} = C\max\left\{C_1\sqrt{\frac{\log(S)}{n(p-S)}}, C_2\frac{\log(S)}{n(p-S)}\right\}$
Δ_n^{HS}	$\max_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s} \mathbb{1}\left(\sum_{j\in\mathcal{C}} \lvert\varphi_{A_j}(\Sigma_n - I_p)\rvert \geq t_{n,p}^{HS}\right)$	$t_{n,p}^{HS} = s\max\left\{C_1\sqrt{\frac{\log\left(s\binom{S}{s}\right)}{n(p-S)}}, C_2\frac{\log\left(s\binom{S}{s}\right)}{n(p-S)}\right\}$

Test	σ_0	R_{max}
Δ_n^{MS+}	$\frac{2(s+1)}{s}t_{n,p}$	$2\exp\left(-\frac{u}{4}\right)$
Δ_n^{HS+}	$\frac{t}{s} + \frac{2s+1}{s}\max\left\{C_1\sqrt{\frac{s}{n(p-S)}}, C_2\frac{s}{n(p-S)}\right\}$	$\exp\left(-(u-1)\log\binom{S}{s}\right) + \exp\left(-\frac{u}{4}\right)$
Δ_n^{MS}	$t + \max\left\{C_1^*\sqrt{\frac{(2s+1)\log(S)}{n(p-S)}}, C_2^*\frac{(2s+1)\log(S)}{n(p-S)}\right\}$	$4\exp\left(-(u-1)\log(S)\right)$
Δ_n^{HS}	$t + \max\left\{C_1^*\sqrt{\frac{\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}}, C_2^*\frac{\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}\right\}$	$4\exp\left[-(u-1)\log\left(s\binom{S}{s}\right)\right]$

Table 2.1 summarizes our results where C_1, C_2, C_1^* and C_2^* denote constants depending only on u.

A detailed numerical study is included in the Supplementary material, containing an example of a sparse $MA(\lfloor p/4 \rfloor)$ series with increasing p. The graphs of the power function are also provided, $\mathbb{E}_\Sigma(\Delta_n = 1)$, for different values of Σ, for the tests Δ_n^{MS} and Δ_n^{HS}. The plots represent the power of the tests by the measure of separation, namely $\sum_{j=1}^S \sigma_j$ for the one sided tests, and $\sum_{j=1}^S |\sigma_j|$ for the two-sided tests. To generate the plots, 5000 samples were generated under the alternative hypothesis and the mean value of the power of the tests is then plotted. The α value will always be 0.1. The plots show very steep power functions, that indicate a narrow band where the decision is hard to make. The power goes from small values near $\alpha = 10\%$ to high values close to 1 in a fast increasing way. There are little differences in the behaviour of moderately and highly sparse tests.

An improvement is noted as p grows (the tests detect matrices closer to the identity), in agreement with theoretical rates that first indicated that p is not a nuisance parameter here. The plots also show that for p smaller than, equal to or bigger than n, the tests behave similarly as the measure of separation increase. However, it can be noticed that the performances are better in high dimension. This is in agreement with our theoretical rates and indicates that p is not a nuisance parameter. The test procedures are not only robust but also more efficient in high dimension. It can also be noticed that the two-sided tests benefit more from the high-dimension than their one-sided versions. The impact, with fixed value of the separation measure, of the number of non null entries in the covariance matrix as well as the impact of their location on the test performances are also studied. The simulations show that the tests are sensitive neither to the number of non null entries nor to their location. Finally the comparison between the moderately and the highly sparse procedures is also provided. When the sparsity of the covariance matrix is known, the simulations show that the highly sparse procedure has a

better detection power than its moderately counter part. It can also be noticed that this outperformance is emphasized when the value of the non null entries increase. When the sparsity level is unknown, an aggregation of the highly sparse procedure with different s values can be compared to the moderately sparse procedure. The simulations prove in this context again that the highly sparse procedure is more efficient.

2.4 Lag-selection for stationary time-series

The objective here is to properly select non-null correlation coefficients. It can be defined a (two-sided) lag-selection problem as estimation of η, a vector with entries $\eta_j = \mathbf{1}(|\varphi_{A_j}(\Sigma)| > 0)$. The aim is to find a selector $\widehat{\eta}$ with $\widehat{\eta}_j = \mathbb{1}(|\varphi_{A_j}(\Sigma_n)| > \tau_n)$ that is consistent in the sense that the risk

$$R^{LS}(\widehat{\eta}, \mathcal{F}) = \sum_{j=1}^{S} \mathbb{E}_{\Sigma}[|\widehat{\eta}_j - \eta_j|]$$

stays bounded (is small). The Hamming loss counts the number of miss-classified elements.

Theorem 2.4.1 *If Σ belongs to $\mathcal{F}(s, S, \sigma)$, with $\sigma \geq 2\tau_n$, the selector $\widehat{\eta}$ with*

$$\tau_n = \max\left\{ \left(\sqrt{\log(s)} + \sqrt{\log(S-s)}\right) \sqrt{u\frac{2s+1}{n(p-S)}}, 2u\log(s(S-s))\frac{2s+1}{n(p-S)} \right\}, \quad u > 1,$$

is such that

$$R_{LS}(\widehat{\eta}, \mathcal{F}) \leq 2\exp\left(-(u-1)\frac{\log(s)}{4}\right) + 2\exp\left(-(u-1)\frac{\log(S-s)}{4}\right).$$

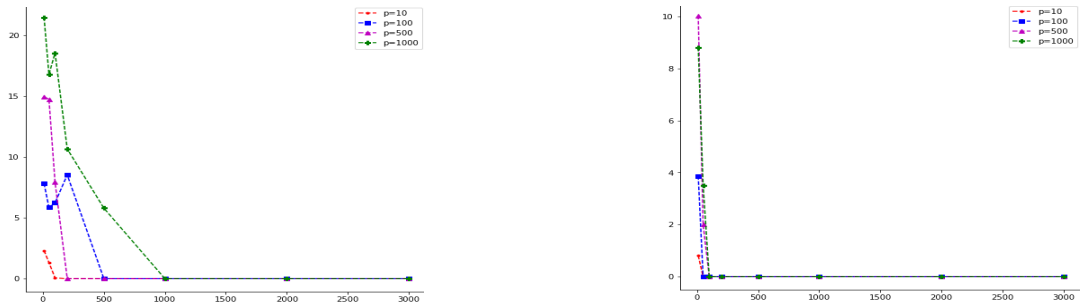
Remark 2.4.1 If the only class considered is $\mathcal{F}^+$, with $\sigma > 2\tau_n$, a one-sided selection is defined by $\eta_j^+ = \mathbf{1}(\varphi_{A_j}(\Sigma) > 0)$ and $\widehat{\eta_j}^+ = \mathbf{1}(\varphi_{A_j}(\Sigma_n) > \tau_n)$ can be considered. Then

$$R_{LS}(\hat{\eta}^+, \mathcal{F}) \leq \exp\left(-(u-1)\frac{\log(s)}{4}\right) + \exp\left(-(u-1)\frac{\log(S-s)}{4}\right).$$

Take for example $S = \frac{p}{2} - 1$, and assume that $s/p = p^{-\beta}$ for some β in (0,1). This implies that $\log(S-s) \sim (1-\beta)\log(p)$ and the asymptotic value of τ_n as p tends to infinity is

$$\tau_n \sim (1 + \sqrt{1-\beta})\sqrt{2u\frac{\log(p)}{np^{\beta}}}, \quad u > 1.$$

Fig. 2.1 shows the good behaviour of our lag selector under $\Sigma \in \mathcal{F}(s, S, \sigma)$ hypothesis. The Hamming loss between η and $\widehat{\eta}$, averaged over 1000 repetitions, is plotted as a function of n, for numerous values of p with $S = \sqrt{p}$. In red is plotted the Hamming loss between η and $\widehat{\eta}$ for $p = 10$, in blue for $p = 100$, in magenta for $p = 500$ and in green for $p = 1000$. The fast decrease to 0 of the Hamming loss can be noted for both $s = S - 1$ in Fig. 2.1 (a) and $s = (S-1)/2$ in Fig. 2.1 (b), despite the small values of $\sigma \asymp \tau_n$ to detect. It can also be noticed that the higher is the value of p, the higher the Hamming loss tends to be. This can be explained by the increase of non null values induced by the increase of p. Mechanically, the bigger is the number of non null values, the higher the Hamming loss is susceptible to be.



(a) $s = S - 1$ (b) $s = \frac{S-1}{2}$

FIGURE 2.1 – Hamming-loss of the lag selector as a function of n, for numerous values of p with $S = \sqrt{p}$. The Hamming-loss is plotted for $s = S - 1$ in (a) and for $s = (S - 1)/2$ in (b).

2.5 Proofs

Proof of Theorem 2.2.2. The following lemma is useful to prove this theorem. A more general statement involving an arbitrary constant K in (0,1) is proved. It is sufficient to take $K = 1/2$ to deduce the theorem.

Lemma 2.5.1 *Let $\Sigma \in \mathcal{S}_p^{++}$ and $\Sigma^{1/2}$ be its square root. Let $A \in S_p$ and $M = \Sigma^{1/2} A \Sigma^{1/2}$. Then, for an arbitrary $K \in]0,1[$, the matrix $I_p - tM$ is invertible and*

$$det\left((I_p - tM)\right)^{-1} \leq \exp\left(t\mathrm{Tr}(A\Sigma) + \frac{t^2||A\Sigma||_F^2}{2(1-K)}\right), \quad |t| < \frac{K}{||A\Sigma||_\infty}.$$

Proof of Lemma 2.5.1. Let $\lambda_1, \ldots, \lambda_p$ be the real eigenvalues of the symmetric matrix M associated to the eigenvectors $x_1, \ldots, x_p$. Then for an arbitrary $K \in]0,1[$, for all $|t| < \frac{K}{||A\Sigma||_\infty}$, $1 - t\lambda_1, \ldots, 1 - t\lambda_p$ are the strictly positive eigenvalues of the matrix $I_p - tM$ associated to the eigenvectors $x_1, .., x_p$. Then

$$\det\left(I_p - tM\right)^{-1} = \exp\left\{-\sum_{k=1}^{p}\log(1 - t\lambda_k)\right\} = \exp\left\{\sum_{k=1}^{p}\sum_{i=1}^{\infty}\frac{1}{i}\left(t\lambda_k\right)^i\right\} = \exp\left\{t\mathrm{Tr}(A\Sigma) + \sum_{k=1}^{p}t^2\lambda_k^2\left(\sum_{i=0}^{\infty}\frac{t^i}{i+2}\lambda_k^i\right)\right.$$

$$\det\left(I_p - tM\right)^{-1} \leq \exp\left\{t\mathrm{Tr}(A\Sigma) + \sum_{k=1}^{p}\frac{t^2\lambda_k^2}{2}\left(\sum_{i=0}^{\infty}t^i\lambda_k^i\right)\right\} = \exp\left(t\mathrm{Tr}(A\Sigma) + \frac{t^2}{2}\sum_{k=1}^{p}\frac{\lambda_k^2}{1 - t\lambda_k}\right).$$

By using the fact that $||A\Sigma||_F^2 = ||M||_F^2 = \sum_{k=1}^{p}\lambda_k^2$ and that $||A\Sigma||_\infty = ||M||_\infty = \max_k|\lambda_k|$, it comes :

$$\det\left(I_p - tM\right)^{-1} \leq \exp\left(t\mathrm{Tr}(A\Sigma) + \frac{t^2||A\Sigma||_F^2}{2(1-K)}\right)$$

which ends the proof. ■

Let us note that if $X \sim \mathcal{N}(0_p, \Sigma)$, then $Y = \Sigma^{-1/2}X \sim \mathcal{N}(0_p, I_p)$. For all $|t| < \frac{nK}{2||A\Sigma||_\infty}$, there is :

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\varphi_A\left(\Sigma_n - \Sigma\right)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{t}{n}\left(X^T A X\right)\right)\right]^n \exp(-t\mathrm{Tr}(A\Sigma)) \\ &= \mathbb{E}\left[\exp\left(\frac{t}{n}\left(Y^T \Sigma^{T1/2} A \Sigma^{1/2} Y\right)\right)\right]^n \exp(-t\mathrm{Tr}(A\Sigma)) \\ &= \mathbb{E}\left[\exp\left(\frac{t}{n}\left(Y^T M Y\right)\right)\right]^n \exp(-t\mathrm{Tr}(A\Sigma)) =: T.\end{aligned}$$

Now, the probability density of Y is used to calculate explicitly

$$\begin{aligned}T &:= \exp(-t\mathrm{Tr}(A\Sigma)) \left(\left(\frac{1}{2\pi}\right)^{p/2} \int \dots \int \exp\left(\frac{t}{n} Y^T M Y - \frac{1}{2} Y^T Y\right) dy_1 \dots dy_p\right)^n \\ &= \exp(-t\mathrm{Tr}(A\Sigma)) \left(\left(\frac{1}{2\pi}\right)^{p/2} \int \dots \int \exp\left(-\frac{1}{2} Y^T (I_p - \frac{2t}{n}) M) Y\right) dy_1 \dots dy_p\right)^n \\ &= \exp(-t\mathrm{Tr}(A\Sigma)) \left(\det\left(I_p - t\frac{2}{n} M\right)\right)^{-n/2}.\end{aligned}$$

By applying Lemma 2.5.1 with $\nu^2 = \frac{2||A\Sigma||_F^2}{n(1-K)}$:

$$\mathbb{E}\left[\exp\left(t\varphi_A\left(\Sigma_n - \Sigma\right)\right)\right] \leq \exp(-t\mathrm{Tr}(A\Sigma)) \exp\left(t\mathrm{Tr}(A\Sigma) + \frac{t^2||A\Sigma||_F^2}{n(1-K)}\right) = \exp\left(\frac{t^2||A\Sigma||_F^2}{n(1-K)}\right) = \exp\left(\frac{\nu^2 t^2}{2}\right).$$

■

Proof of Proposition 2.2.3. First, to bound the operator norm of the matrix A_W, the Gershgorin's circle theorem is used. Let $M = (m_{i,j})_{1 \leq i,j \leq p}$ be a $p \times p$ matrix. Then, all eigenvalues of the matrix M lie within at least one of the Gershgorin discs $D(m_{ii}, \sum_{j \neq i} |m_{ij}|)$.

Gershgorin's circle theorem applied to the matrix A_W gives :

$$||A_W||_\infty = \max_k |\lambda_k| \in D\left(0, 2\sum_{j \in W} \frac{1}{2(p-j)}\right) \Rightarrow ||A_W||_\infty \leq \frac{w}{p-S}.$$

To bound the squared Frobenius norm, sum all the squared elements of A_W, which gives :

$$||A_W||_F^2 = 2\sum_{j \in W} \frac{p-j}{4(p-j)^2} = \sum_{j \in W} \frac{1}{2(p-j)} \leq \frac{w}{2(p-S)}.$$

Then to bound the operator norm of the matrix $A_W\Sigma$ for some Σ in $\mathcal{F}(s, S, \sigma)$, use the Cauchy-Schwarz inequality together with Gershgorin's circle theorem :

$$||A_W\Sigma||_\infty \leq ||A_W||_\infty ||\Sigma||_\infty \leq \sigma_0 \frac{(2s+1)w}{p-S}.$$

To bound the squared Frobenius norm of the matrix $A_W\Sigma$ the following lemma will be used.

Lemma 2.5.2 *Let M and N be two $p \times p$ symmetric matrices. Then $||MN||_F^2 = \mathrm{Tr}(M^2N^2)$ and*

$$||MN||_F^2 \leq \max_{1\leq k\leq p} |\lambda_k|^2 ||N||_F^2 = ||M||_\infty^2 ||N||_F^2.$$

Proof of Lemma 2.5.2. First, $||MN||_F^2 = \mathrm{Tr}(MNN^TM^T) = \mathrm{Tr}(M^2N^2)$, with M^2 and N^2 symmetric and positive semi-definite matrices ($M^2 \geq 0, \quad N^2 \geq 0$). Recall that, if $A \leq B$ (in the sense that $B - A \geq 0$), then $\mathrm{Tr}(AC) \leq \mathrm{Tr}(BC)$, for any $C \geq 0$. Here, $M^2 \leq \lambda_{max}(M^2)I_p \leq \lambda_{max}^2(M)I_p$ and this gives $\mathrm{Tr}(M^2N^2) \leq \lambda_{max}^2(M)\mathrm{Tr}(N^2)$. ■

If $w > 1$, using Lemma 2.5.2 on $M = \Sigma$ and $N = A_W$, it follows

$$||A_W\Sigma||_F^2 \leq ||A_W||_F^2 ||\Sigma||_\infty^2 \leq \sigma_0^2 \frac{w(2s+1)^2}{2(p-S)}.$$

If $w = 1$ and $W = \{j\}$, using Lemma 2.5.2 on $M^2 = \Sigma$ and $N^2 = \Sigma^{1/2}A_j^2\Sigma^{1/2}$, then

$$||A_j\Sigma||_F^2 = \mathrm{Tr}(A_j^2\Sigma^2) \leq ||A_j\Sigma^{1/2}||_F^2 ||\Sigma^{1/2}||_\infty^2 \leq \sigma_0(2s+1)||A_j\Sigma^{1/2}||_F^2.$$

It suffices to prove that $||A_j\Sigma^{1/2}||_F^2 = \mathrm{Tr}(A_j^2\Sigma) \leq \sigma_0 \frac{\mathcal{K}}{(p-S)}$ so that the proof can be finished, namely that $||A_j\Sigma||_F^2 \leq \sigma_0^2 \frac{\mathcal{K}(2s+1)}{p-S}$. Let $B_j = A_j^2 = (b_{k,l}^j)_{1\leq k,l\leq p}$. For every $1 \leq k, l \leq p$,

$$b_{k,l}^j = \sum_{i=1}^p a_{k,i}^j a_{i,l}^j = \sum_{i=1}^p a_{|k-i|}^j a_{|l-i|}^j = \sum_{i=1}^p \frac{\delta_{|k-i|=j}\delta_{|l-i|=j}}{4(p-j)^2} :$$

if $k = l$, $b_{k,k}^j = \begin{cases} \frac{1}{2(p-j)^2}, & j < \frac{p}{2}, \quad j < k \leq p-j, \\ 0, & j \geq \frac{p}{2}, \quad p-j \leq k < j, \\ \frac{1}{4(p-j)^2}, & \text{otherwise.} \end{cases}$

if $k \neq l$, for $\delta_{|k-i|=j}\delta_{|l-i|=j}$ to be non-null requires :

$$\begin{cases} k-i=j, & l-i=-j, \\ l-i=j, & k-i=-j, \end{cases} \Leftrightarrow \begin{cases} k-l=2j, & i=\frac{k+l}{2}, \\ l-k=2j, & i=\frac{k+l}{2}, \end{cases} \Leftrightarrow \begin{cases} |k-l|=2j, \\ i=\frac{k+l}{2}. \end{cases}$$

Therefore, $b_{k,l}^j = \begin{cases} \frac{1}{4(p-j)^2}, & j < \frac{p}{2}, \quad |k-l| = 2j, \\ 0, & \text{otherwise.} \end{cases}$

Summing up the results gives us

$$||A_j\Sigma^{1/2}||_F^2 = \mathrm{Tr}(A_j^2\Sigma) = \sum_{m=1}^p \left(\sum_{i=1}^p b_{m,i}\sigma_{i,m}\right) \leq \sigma_0 \sum_{m=1}^p \left(\sum_{i=1}^p b_{m,i}\right) = \sigma_0 \sum_{m=1}^p b_{m,m} + \sigma_0 \sum_{m\neq i} b_{m,i}$$

$$\leq \sigma_0 \begin{cases} \frac{2(p-j)+2(p-2j)}{4(p-j)^2}, & j < \frac{p}{2}, \\ \frac{2(p-j)}{4(p-j)^2}, & \text{otherwise.} \end{cases} \leq \sigma_0 \begin{cases} \frac{1}{(p-j)}, & j < \frac{p}{2} \\ \frac{1}{2(p-j)}, & \text{otherwise.} \end{cases}$$

This means that

$$||A_j\Sigma||_F^2 \leq \sigma_0(2s+1)||A_j\Sigma^{1/2}||_F^2 \leq \sigma_0^2 \frac{\mathcal{K}(2s+1)}{(p-S)}, \quad \mathcal{K} = \begin{cases} 1, & W \subseteq \{1, \ldots, \frac{p}{2}-1\}, \\ \frac{p}{2}, & W \subseteq \{\frac{p}{2}, \ldots, p\}. \end{cases}$$

■

Proof of Theorem 2.3.1. It is known from Corollary 2.2.4 that the type I error probability is such that

$$\mathbb{P}_{I_p}\left[\varphi_{A_{1:S}}(\Sigma_n - I_p) \geq t_{n,p}^{MS+}\right] \leq \exp\left(-\frac{u}{4}\right)$$

and that, for any Σ in $\mathcal{F}_+(s, S, \sigma)$, there is

$$\mathbb{P}_\Sigma[\varphi_{A_{1:S}}(\Sigma_n - \Sigma) \geq (1+2s)t_{n,p}^{MS+}] \leq \exp\left(-\frac{u}{4}\right), \quad u > 0.$$

The type II error probability can be bounded under the assumption that $\sigma \geq \frac{2(s+1)}{s} t_{n,p}^{MS+}$:

$$\begin{aligned}
\mathbb{P}_\Sigma\left[\varphi_{A_{1:S}}(\Sigma_n - I_p) \leq t_{n,p}^{MS+}\right] &= \mathbb{P}_\Sigma\left[\varphi_{A_{1:S}}(\Sigma_n - \Sigma) \leq t_{n,p}^{MS+} - \varphi_{A_{1:S}}(\Sigma)\right] \\
&= \mathbb{P}_\Sigma\left[\varphi_{A_{1:S}}(\Sigma - \Sigma_n) \geq \varphi_{A_{1:S}}(\Sigma) - t_{n,p}^{MS+}\right] \leq \mathbb{P}_\Sigma\left[\varphi_{A_{1:S}}(\Sigma - \Sigma_n) \geq s\sigma - t_{n,p}^{MS+}\right] \\
&\leq \mathbb{P}_\Sigma\left[\varphi_{A_{1:S}}(\Sigma - \Sigma_n) \geq (2s+1)t_{n,p}^{MS+}\right] \leq \exp\left(-\frac{u}{4}\right), \quad u > 0.
\end{aligned}$$

Finally :

$$R(\Delta_n^{MS+}, \mathcal{F}^+) = \mathbb{P}_{I_p}(\varphi_{A_{1:S}}(\Sigma_n - I_p) \geq t_{n,p}^{MS+}) + \sup_{\Sigma \in \mathcal{F}^+} \mathbb{P}_\Sigma(\varphi_{A_{1:S}}(\Sigma_n - I_p) \leq t_{n,p}^{MS+}) \leq 2\exp\left(-\frac{u}{4}\right).$$

■

Proof of Theorem 2.3.2. Similarly to the proof of Theorem 2.3.1, Corollary 2.2.4 is used to bound the type I error probability

$$\begin{aligned}
\mathbb{P}_{I_p}\left[\sum_{i=1}^{S} |\varphi_{A_i}(\Sigma_n - I_p)| \geq t_{n,p}^{MS}\right] &\leq \mathbb{P}_{I_p}\left[\bigcup_{i=1}^{S}\left\{|\varphi_{A_i}(\Sigma_n - I_p)| \geq \frac{t_{n,p}^{MS}}{S}\right\}\right] \leq \sum_{i=1}^{S} \mathbb{P}_{I_p}\left[|\varphi_{A_i}(\Sigma_n - I_p)| \geq \frac{t_{n,p}^{MS}}{S}\right] \\
&= \sum_{i=1}^{S} \mathbb{P}_{I_p}\left[|\varphi_{A_i}(\Sigma_n - I_p)| \geq \max\left\{\sqrt{\frac{u}{2(1-K)}}\sqrt{\frac{4\log(S)}{n(p-S)}}, \frac{u}{K}\frac{4\log(S)}{n(p-S)}\right\}\right] \\
&\leq \sum_{i=1}^{S} 2\exp\left(-u\log S\right) = 2\exp\left(-(u-1)\log S\right).
\end{aligned}$$

To bound the type II error probability, a condition on σ is used :

$$\begin{aligned}
&\mathbb{P}_\Sigma\left[\sum_{i=1}^{S} |\varphi_{A_i}(\Sigma_n - \mathrm{Id})| \leq t_{n,p}^{MS}\right] \leq \mathbb{P}_\Sigma\left[\bigcap_{i=1}^{S}\left\{|\varphi_{A_i}(\Sigma_n - \mathrm{Id})| \leq t_{n,p}^{MS}\right\}\right] \leq \sup_{1\leq i\leq S} \mathbb{P}_\Sigma\left[|\varphi_{A_i}(\Sigma_n - I_p)| \leq t_{n,p}^{MS}\right] \\
&\leq \sup_{1\leq i\leq S} \mathbb{P}_\Sigma\left[|\varphi_{A_i}(\Sigma_n - \Sigma)| \geq |\varphi_{A_i}(\Sigma - I_p)| - t_{n,p}^{MS}\right] \leq \sup_{1\leq i\leq S} \mathbb{P}_\Sigma\left[|\varphi_{A_i}(\Sigma_n - \Sigma)| \geq \sigma - t_{n,p}^{MS}\right] \\
&\leq \sup_{1\leq i\leq S} \mathbb{P}_\Sigma\left[|\varphi_{A_i}(\Sigma_n - \Sigma)| \geq \max\left\{\sqrt{\frac{u-1}{2(1-K)}}\sqrt{\frac{4\log S(2s+1)}{n(p-S)}}, \frac{(u-1)}{K}\frac{4\log S(2s+1)}{n(p-c)}\right\}\right] \\
&\leq 2\exp\left(-(u-1)\log S\right).
\end{aligned}$$

This finally gives :

$$R(\Delta_n^{MS}, \mathcal{F}) \leq 4\exp\left(-(u-1)\log S\right).$$

■

Proof of Theorem 2.3.3. The type I error probability is bounded by

$$\begin{aligned}\mathbb{P}_{\mathrm{I_p}}[\Delta_n^{HS+}=1] &\leq \sum_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s} \mathbb{P}_{I_p}\left[\varphi_{A_\mathcal{C}}(\Sigma_n - I_p) \geq t_{n,p}^{HS+}\right] \\ &\leq \sum_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s} \exp\left(-u\log\binom{S}{s}\right) = \exp\left(-(u-1)\log\binom{S}{s}\right)\end{aligned}$$

while the type II error probability is, for an arbitrary set $\mathcal{C}\subseteq\{1,\dots,S\}$ containing s values, bounded by

$$\begin{aligned}\mathbb{P}_\Sigma[\Delta_n^{HS+}=0] &= \sup_{\Sigma\in\mathcal{F}^+(s,S,p,\sigma)} \mathbb{P}_\Sigma\left[\bigcap_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\{|\varphi_{A_\mathcal{C}}(\Sigma_n - I_p)| \leq t_{n,p}^{HS+}\}\right] \\ &\leq \sup_{\Sigma\in\mathcal{F}^+(s,S,p,\sigma)} \mathbb{P}_\Sigma\left[\varphi_{A_\mathcal{C}}(\Sigma_n-\Sigma)+\varphi_{A_\mathcal{C}}(\Sigma-I_p) \leq t_{n,p}^{HS+}\right] \\ &= \sup_{\Sigma\in\mathcal{F}^+(s,S,p,\sigma)} \mathbb{P}_\Sigma\left[\varphi_{A_\mathcal{C}}(\Sigma-\Sigma_n) \geq \varphi_{A_\mathcal{C}}(\Sigma) - t_{n,p}^{HS+}\right] \\ &\leq \sup_{\Sigma\in\mathcal{F}^+(s,S,p,\sigma)} \mathbb{P}_\Sigma\left[\varphi_{A_\mathcal{C}}(\Sigma-\Sigma_n) \geq s\sigma - t_{n,p}^{HS+}.\right]\end{aligned}$$

Under the condition

$$s\sigma - t_{n,p}^{HS+} \geq (2s+1)\max\left\{\sqrt{\frac{u}{2(1-K)}}\sqrt{\frac{s}{n(p-S)}}, \frac{u}{K}\frac{s}{n(p-S)}\right\}$$

and Corollary 2.2.4, it comes :

$$\mathbb{P}_\Sigma[\Delta_n^{HS+}=0] \leq \sup_{\Sigma\in\mathcal{F}^+(s,S,p,\sigma)} \mathbb{P}_\Sigma\left[\varphi_{A_\mathcal{C}}(\Sigma-\Sigma_n)\geq\tilde{t}\right] \leq \exp\left(-\frac{u}{4}\right).$$

■

Proof of Theorem 2.3.4. The proof is similar to the proof of Theorem 2.3.2. The type I probability error is bounded by

$$\begin{aligned}\mathbb{P}_{I_p}[\Delta_n^{HS}=1] &\leq \sum_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s} \mathbb{P}_{I_p}\left[\sum_{i\in\mathcal{C}}|\varphi_{A_i}(\Sigma_n-I_p)| \geq t_{n,p}^{HS}\right] \leq \sum_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\sum_{i\in\mathcal{C}} \mathbb{P}_{I_p}\left[|\varphi_{A_i}(\Sigma_n-I_p)| \geq \frac{t_{n,p}^{HS}}{s}\right] \\ &\leq \sum_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\sum_{i\in\mathcal{C}} 2\exp\left[-u\log\left(s\binom{S}{s}\right)\right] = 2\exp\left[-(u-1)\log\left(s\binom{S}{s}\right)\right]\end{aligned}$$

The type II probability is bounded by

$$\begin{aligned}\mathbb{P}_\Sigma[\Delta_n^{HS}=0] &= \mathbb{P}_\Sigma\left[\max_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\sum_{i\in\mathcal{C}}|\varphi_{A_i}(\Sigma_n-\mathrm{Id})| \leq t_{n,p}^{HS}\right] \leq \mathbb{P}_\Sigma\left[\bigcap_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\bigcap_{i\in\mathcal{C}}\left\{|\varphi_{A_i}(\Sigma_n-\mathrm{Id})| \leq t_{n,p}^{HS}\right\}\right] \\ &\leq \sup_{\mathcal{C}\subseteq\{1,\dots,S\},\#\mathcal{C}=s}\sup_{i\in\mathcal{C}} \mathbb{P}_\Sigma\left[|\varphi_{A_i}(\Sigma_n-\Sigma)| \geq \sigma - t_{n,p}^{HS}\right] \leq 2\exp\left[-(u-1)\log\left(s\binom{S}{s}\right)\right].\end{aligned}$$

■

Proof of Theorem 2.4.1. Using Theorem 2.2.2 and Proposition 2.2.3, we have :

$$
\begin{aligned}
R^{LS}(\hat{\eta}, \mathcal{F}^+) = \sum_{j=1}^{S} \mathbb{E}_\Sigma[|\hat{\eta_j} - \eta_j|] &= \sum_{j \in \mathcal{C}} \mathbb{E}_\Sigma[|\hat{\eta_j} - \eta_j|] + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{E}_\Sigma[|\hat{\eta_j} - \eta_j|] = \sum_{j \in \mathcal{C}} \mathbb{E}_\Sigma[|\hat{\eta_j} - 1|] + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{E}_\Sigma[|\hat{\eta_j}|] \\
&= \sum_{j \in \mathcal{C}} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n)| < \tau_n] + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n)| > \tau_n] \\
&\leq \sum_{j \in \mathcal{C}} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \varphi_{A_j}(\Sigma) - \tau_n] + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \tau_n] \\
&\leq \sum_{j \in \mathcal{C}} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \sigma - \tau_n] + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{P}_\Sigma[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \tau_n] \\
&\leq \sum_{j \in \mathcal{C}} \mathbb{P}_\Sigma \left[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \max \left\{ \sqrt{2u\log(s)} \frac{||A_j \Sigma||_F}{\sqrt{n}}, 2u\log(s) \frac{||A_j \Sigma||_\infty}{n} \right\} \right] \\
&\quad + \sum_{j \notin \mathcal{C}, j \leq S} \mathbb{P}_\Sigma \left[|\varphi_{A_j}(\Sigma_n - \Sigma)| > \max \left\{ \sqrt{2u\log(S-s)} \frac{||A_j \Sigma||_F}{\sqrt{n}}, 2u\log(S-s) \frac{||A_j \Sigma||_\infty}{n} \right\} \right] \\
&\leq \sum_{j \in \mathcal{C}} 2\exp\left(-\frac{u\log(s)}{4} \right) + \sum_{j \notin \mathcal{C}, j \leq S} 2\exp\left(-\frac{u\log(S-s)}{4} \right) \\
&\leq 2\exp\left(-(u-1)\frac{\log(s)}{4} \right) + 2\exp\left(-(u-1)\frac{\log(S-s)}{4} \right).
\end{aligned}
$$

■

2.6 Supplementary material

Numerical results related to the presented procedures are available in the supplementary material. They show the good behaviour of the procedures, especially in high dimension, which is in agreement with the theoretical guarantees of the procedures given here. It can be noticed that the higher the value of p, the more efficient the tests are. Indeed, the dimension of the vectors is not a nuisance parameter in this setup. The tests are also robust both to the number of non null entries and to their locations in the covariance matrix. The moderately and highly sparse tests present the same behaviours but an in-depth study shows that the highly sparse procedure behaves better compared to the moderately sparse one in the case of sparser covariance matrices. Moreover the highly sparse test procedure which requires the number of non null entries as an input has a better detection power when this value is known. When the number of non null entries is unknown, a grid-search aggregated procedure is implemented. In such cases the highly sparse test procedure presents similar performances to those of the moderately sparse test procedure.

In order to illustrate a setup where our procedures are of particular interest, we build a moving average stationary process having non-zero coefficients only for even lags and up to $p/4$. Thus the covariance matrix belongs to the considered set of sparse covariance matrices and the entries depend on the parameter ϕ of our MA process. The moderately sparse procedure is applied to this process and the results show how the power of the test procedure increases when the parameter ϕ and the dimension p increase for fixed sample sizes n of 50 and 500, respectively. We conclude our numerical results for synthetic data with a comparison of the presented test procedures to previously existing

ones. The results show that on the set of considered covariance matrices the presented procedures have a better detection power for smaller values of the covariance values than the previously existing ones.

The last section of the supplementary material is focused on a real data set, namely meteorological data available at `http://berkeleyearth.org/data/`. An in-depth study of the data is provided to show that the test procedures detect the significant values in the covariance matrices of the processes from which are issued the data.

2.6.1 Power curves of the test procedures

Several examples are included to illustrate the numerical behavior of our test procedures. First are presented the powers of the Δ_n^{MS} and Δ_n^{HS} tests. Then is highlighted why the plots will be drawn with a logarithmic scale. The power of the following four test procedures are estimated : Δ_n^{MS+}, Δ_n^{MS}, Δ_n^{HS+}, Δ_n^{HS} to test the null hypothesis $\Sigma = I$.

The numbers of non-null entries s and the non-null entries support $\mathcal{C} \subset \{1, \ldots, S\}$ are chosen to be

$$s = (S-1)/2, \quad S = \sqrt{p}.$$

The location of the non zero entries is randomly chosen. The common value of non-null entries are defined as growing fractions of σ. The threshold of the test procedure is defined as $t = t_{n,p,\alpha}$ the empirical $(1-\alpha)$-quantile of the test statistic under the null hypothesis. In order to determine its value empirically, 5000 repeated samples were generated under the null hypothesis. The plots represent the power of the tests by the measure of separation, namely $\sum_{j=1}^{S} \sigma_j$ for the one sided tests, and $\sum_{j=1}^{S} |\sigma_j|$ for the two-sided tests.

To generate the plots, we sample 5000 times under the alternative hypothesis and plot the mean value of the power of the tests. The α value will always be 0.1.

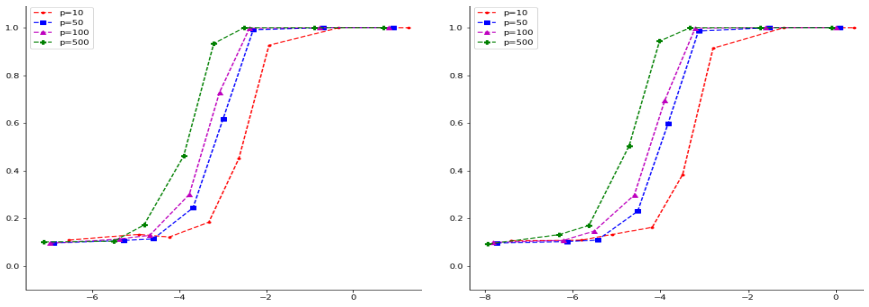
Fig.2.2 and Fig.2.3 show the power for different values of p and n as function of respectively $\sum_{j=1}^{S} |\sigma_j|$ and $\sum_{j \in \mathcal{C}} |\sigma_j|$ - in a logarithmic scale that allow to better read this graphics. The plots show very steep power functions, that indicate a narrow band where the decision is hard to make. The power goes from small values near $\alpha = 10\%$ to high values close to 1 in a fast increasing way. There are little differences in the behaviour of moderately and highly sparse tests.

Fig.2.4 shows that the logarithmic scale should be preferred as it helps to better understand the behaviour of the test procedure when the measure of separation increases. The power of the Δ_n^{MS+} test procedure is now represented as a function of the measure of separation for numerous values of n and p. The best power function goes the fastest from low values above $\alpha = 0.1$ to high values close to 1. The change happens around the theoretical value of the separation rate.

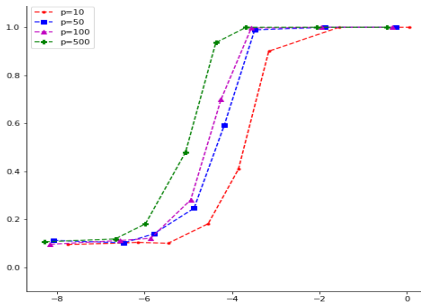
Fig.2.5 shows that for p smaller than, equal to or bigger than n, the Δ_n^{MS+} test presents similar behaviour as the measure of separation increases. However, it can be noticed that the performances are better in high dimension, that is the power curves are shifted to the left. This is in agreement with our theoretical rates and indicates that p is not a nuisance parameter. The Δ_n^{MS+} test is not only robust but also more efficient in high dimension.

Let us consider the two-sided Δ_n^{MS} test and plot its estimated power curve.

Fig.2.6 shows that the Δ_n^{MS} test shows a similar behaviour as the Δ_n^{MS+} test. However, the two-sided test efficiency benefits more from the high-dimension p than the one-sided version, in the sense that the curves shift more to the left, towards the small values of the measure of separation when p is large. Let us consider the Δ_n^{HS+} test.



(a) $n = 100$ (b) $n = 500$



(c) $n = 1000$

FIGURE 2.2 – Power of the Δ_n^{MS} test by the sum of the $S = \sqrt{p}$ entries of the covariance matrix in absolute value. The power is plotted as function of $\sum_{j=1}^{S} |\sigma_j|$ for different values of n in (a), (b), (c) and different values of p in red, blue, magenta, green.

Fig.2.7 shows that the Δ_n^{HS+} test behaves similarly to the Δ_n^{MS+} and Δ_n^{MS} tests. Finally, the two-sided HS test is considered.

Fig.2.8 shows that the Δ_n^{HS} tests also behaves as the previous ones. It can be noticed that the higher the value of p, the better the tests behave. The high dimension improves the efficiency of the tests. It can also be underlined that the power of the tests increase rapidly around -3 on the logarithmic scale of the measure of separation.

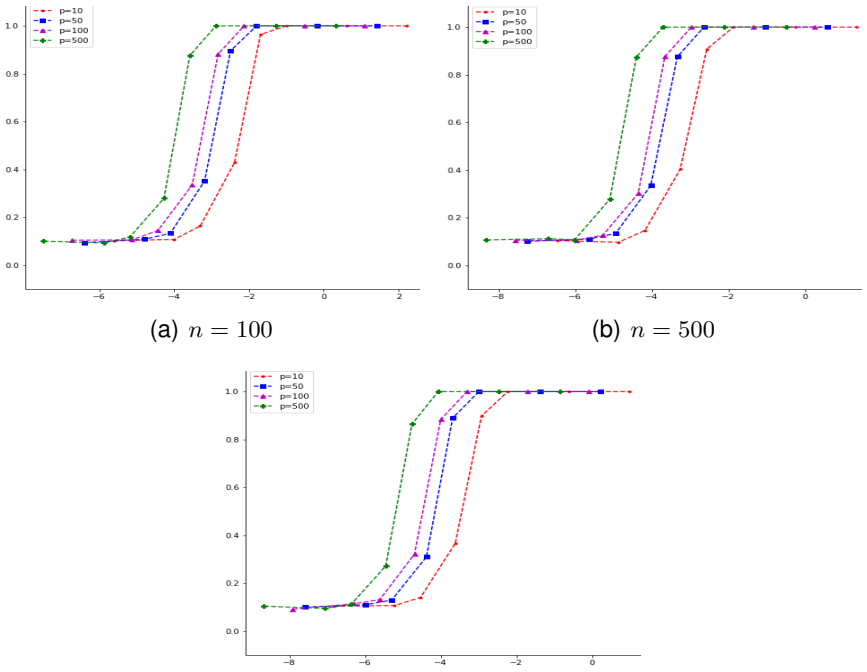
2.6.2 Effect of non null entries

In the previous Section are plotted numerical simulations of the four tests presented in the paper. However we want to understand in more details the impact of the different choices that can be made in this procedures namely : the impact of the number of non null entries s, the impact of the location of non-null entries (close to the main diagonal or far from it).

In this sub-section the focus is put on the Δ_n^{MS+} test as its behaviour can be extrapolated to the other three tests. The underlying covariance matrix belongs to the class $\mathcal{F}_+(s, S, \sigma)$, for some $s \in \{1, \dots, S\}$.

First, is studied the impact of the number of non null entries. For all the previous graphs s was fixed and set to $(S-1)/2$. The objective is to observe how the value of s impacts the behaviour of the test. For this purpose are plotted side by side the Δ_n^{MS+} test with $s = S - 1$ and $s = (S-1)/2$ for $n = 100$ and different values of p ($10, 20$ and 50).

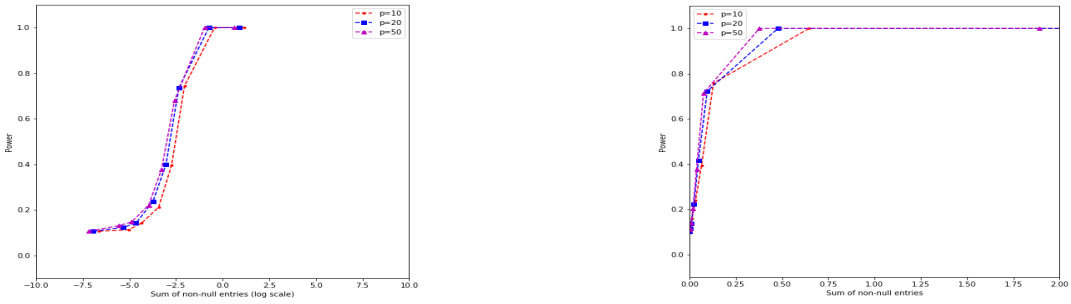
Fig. 2.9 shows that the number of non null entries has no major impact on the power of the test



(a) $n = 100$ (b) $n = 500$

(c) $n = 1000$

FIGURE 2.3 – Power of the Δ_n^{HS} test by the sum of the $s = (S-1)/2$ entries of the covariance matrix in absolute value, with $S = \sqrt{p}$. The power is plotted as function of $\sum_{j \in \mathcal{C}} |\sigma_j|$ for different values of n in (a), (b), (c) and different values of p in red, blue, magenta, green.

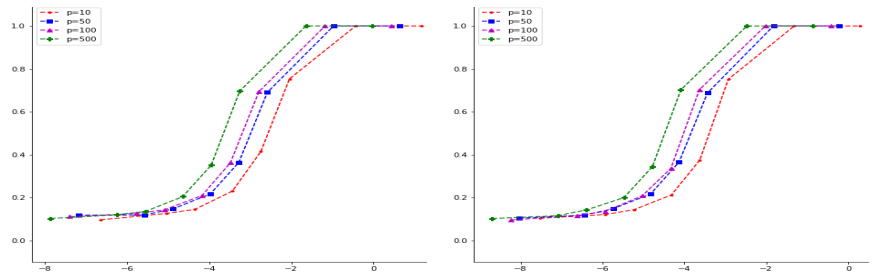


(a) Logarithmic scale (b) Identity scale

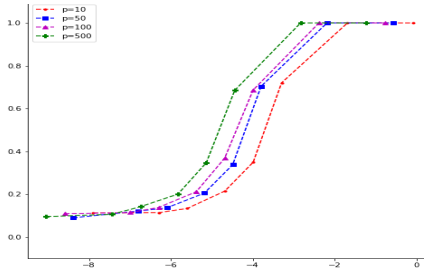
FIGURE 2.4 – Power of the Δ_n^{MS+} test by the sum of the entries of the covariance matrix in absolute value, on a logarithmic scale in (a) and identity scale in (b).

procedure Δ_n^{MS+}.

Second, the impact of the randomness in the location of the non null entries is measured. In all previous graphs the non null entries were randomly located. The objective is to observe how the location of the non null entries impacts the behaviour of the test. To this end is plotted the power function of Δ_n^{MS+} test with $s = (S-1)/2$ for $n = 100$ and different values of p. The non null entries are : (a) randomly located, (b) located next to the main diagonal. The plot (c) shows simultaneously the power functions of Δ_n^{MS+} test for $p = 10$ and $n = 100$, but with non null entries randomly chosen i.e

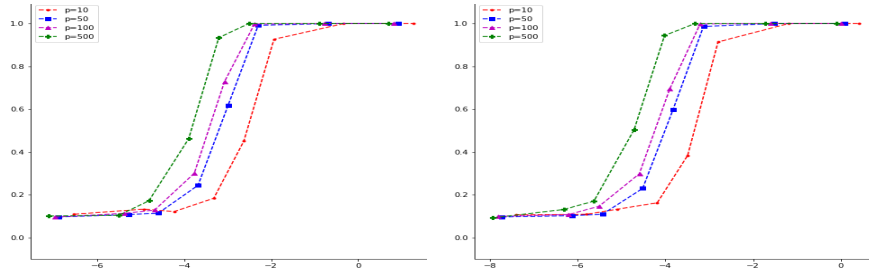


(a) $n = 100$ (b) $n = 500$

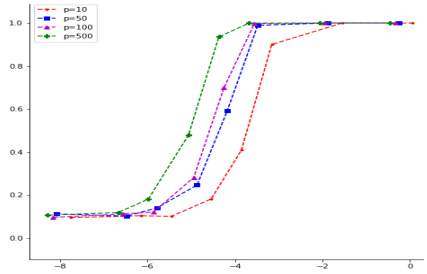


(c) $n = 1000$

FIGURE 2.5 – Power of the Δ_n^{MS+} test by the sum of the entries of the covariance matrix in absolute value, for different values of p in red, blue, magenta, green and different values of n in (a), (b), (c).

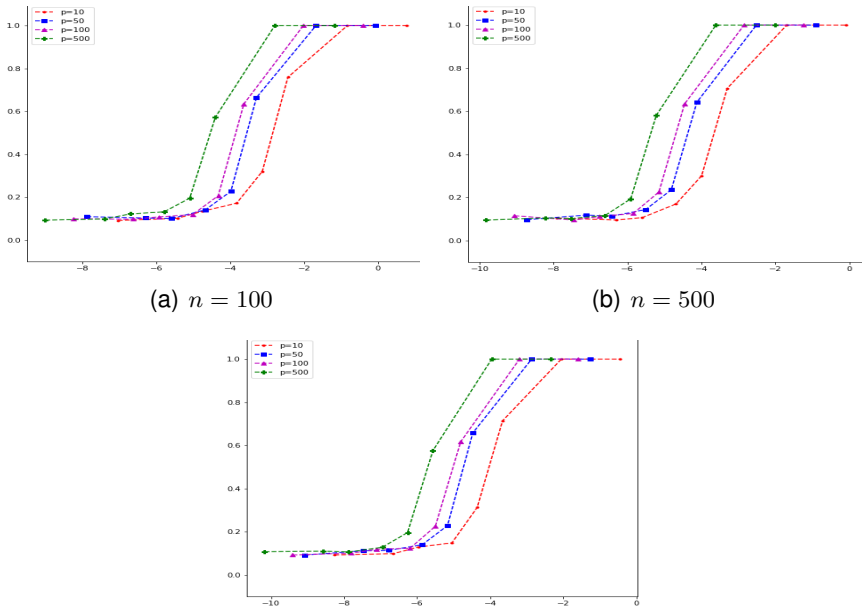


(a) $n = 100$ (b) $n = 500$



(c) $n = 1000$

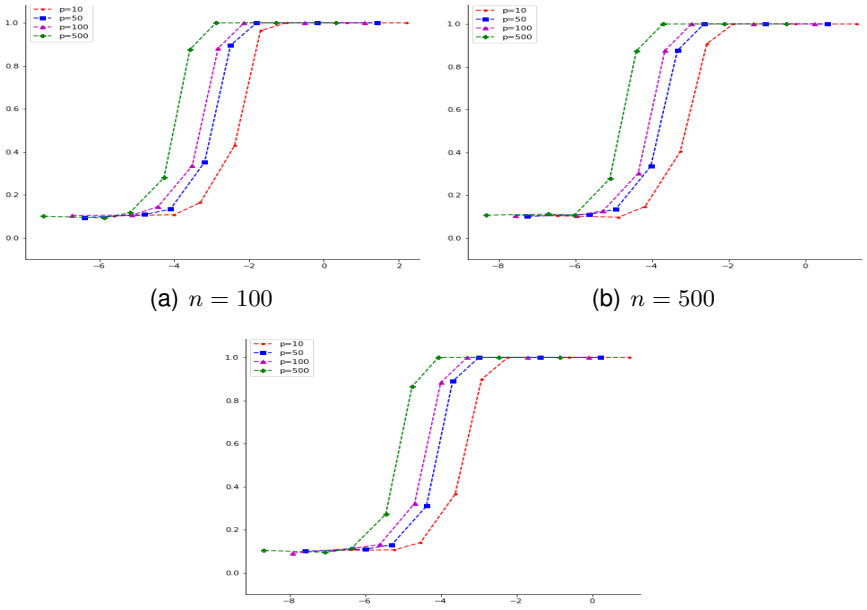
FIGURE 2.6 – Power of the Δ_n^{MS} test by the sum of the entries of the covariance matrix in absolute value, for different values of p in red, blue, magenta, green and different values of n in (a), (b), (c).



(a) $n = 100$ (b) $n = 500$

(c) $n = 1000$

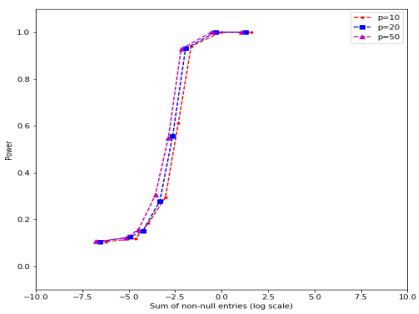
FIGURE 2.7 – Power of the Δ_n^{HS+} test by the sum of the entries of the covariance matrix in absolute value, for different values of p in red, blue, magenta, green and different values of n in (a), (b), (c).



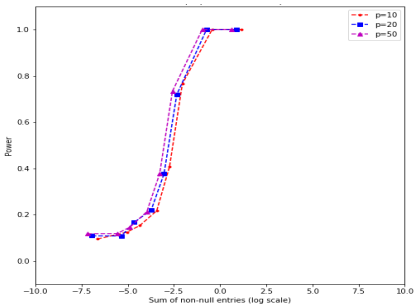
(a) $n = 100$ (b) $n = 500$

(c) $n = 1000$

FIGURE 2.8 – Power of the Δ_n^{HS} test by the sum of the entries of the covariance matrix in absolute value, for different values of p in red, blue, magenta, green and different values of n in (a), (b), (c).



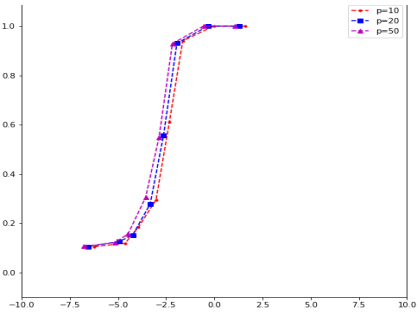
(a) $s = S - 1$



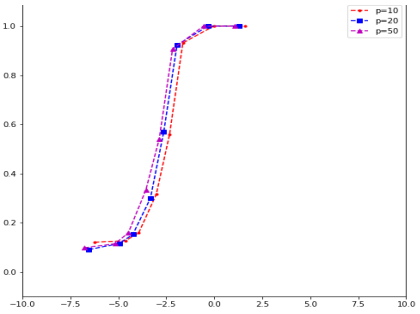
(b) $s = \dfrac{S-1}{2}$

FIGURE 2.9 – Impact of the number of non null entries in the covariance matrix entries on the power of Δ_n^{MS+}.

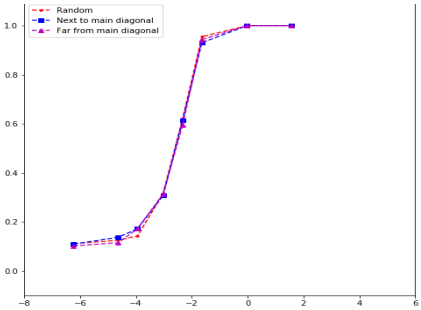
$\mathcal{C} \subset \{1, \ldots, S\}$ with $|\mathcal{C}| = s$ (red), fixed next to the main diagonal i.e $\mathcal{C} = \{1, \ldots, s\}$ (blue) and fixed on the last values of the support i.e $\mathcal{C} = \{S - s, \ldots, S\}$ (magenta).



(a) Randomly chosen



(b) Next to the main diagonal



(c) On the same graph

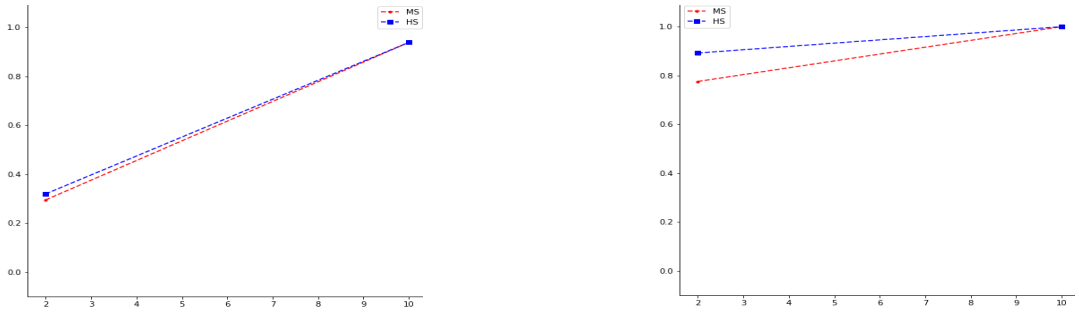
FIGURE 2.10 – Impact of the position of the non null entries in the covariance matrix on the power of Δ_n^{MS+}

Fig. 2.10 shows that the location of the non null entries has no impact on the Δ_n^{MS+} test performances. In conclusion, the tests are sensitive neither to the number of non null entries nor to their location.

2.6.3 Comparison between Δ_n^{MS} and Δ_n^{HS}

The four test procedures Δ_n^{MS+}, Δ_n^{MS}, Δ_n^{HS+} and Δ_n^{HS} present very similar behaviour of their power curves. However, for high sparsity levels of the covariance matrix Δ_n^{HS+} and Δ_n^{HS} were designed to be more efficient than respectively Δ_n^{MS+} and Δ_n^{MS}. The objective is to observe the difference in their behaviours under such high sparsity levels assumption. In this sub-section our study is illustrated on the two-sided Δ_n^{MS} and Δ_n^{HS} tests only, as they are analogous to their one-sided versions.

In order to observe the difference in the impact of sparsity on these two tests the power curves by the number of non null entries s are plotted. The parameters are set as follows $n = 100$, $p = 100$ and $S = \sqrt{p} = 10$. The plot is repeated for the non null entries common value to be $\sigma = t_{n,p,\alpha}/100 \approx 0.01473$ and $\sigma = t_{n,p,\alpha}/50 \approx 0.02945$. As the Δ_n^{HS} test requires a value for s the true value is given in Fig.2.11.



(a) $\sigma = \frac{t_{n,p,\alpha}}{100}$ (b) $\sigma = \frac{t_{n,p,\alpha}}{50}$

FIGURE 2.11 – Power of Δ_n^{MS} in red and Δ_n^{HS} in blue by the number s, known by the procedures, of non null entries. The powers are plotted for different values of the separation rate σ in (a) and (b).

Fig.2.11 shows that indeed the Δ_n^{HS} test procedure with known sparsity s has better detection power than Δ_n^{MS} for higher sparsity, as it was expected. It can also be noticed that larger significant values of the non-null correlations improve even more the power Δ_n^{HS} over Δ_n^{MS}.

Now a new Δ_n^{HS} procedure that is free of knowledge of s is built by aggregating several procedures $\Delta_n^{HS}(s)$ for different values of s. This new procedure is then compared to Δ_n^{MS}. Consider a grid of plausible values of s from 1 to S, build all $\Delta_n^{HS}(s)$ and decide according to

$$\Delta_n^{HS} = \max_s \Delta_n^{HS},$$

that is reject whenever at least one of the tests rejected and accept otherwise.

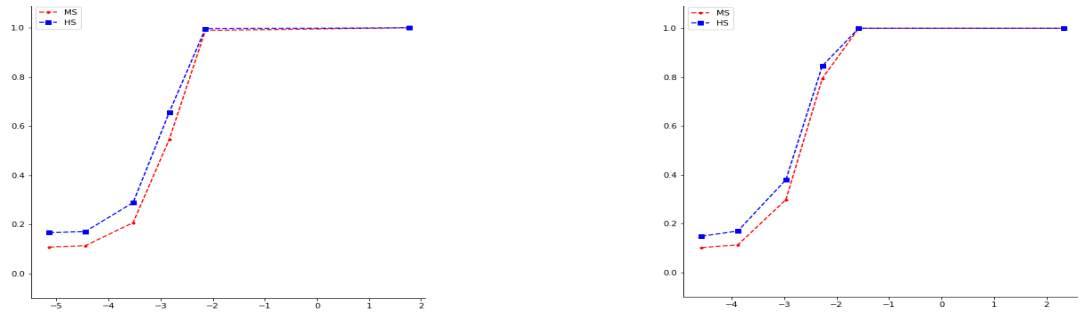
Let us confront the aggregated high-sparsity test and the moderate-sparsity test procedures. The two test procedures have been run in the same setup $n = 100$, $p = 100$ and $S = \sqrt{p} = 10$. The true values of s are being set to $s = 4$ and $s = 7$, respectively. The power curves of the two procedures are plotted by the measure of separation on a log-scale. The latter is rising because of growing values of σ.

In both cases, the grid of plausible sparsity levels has been fixed to two values : 2 and 10, which means that

$$\Delta_n^{HS} = \max\{\Delta_n^{HS}(2), \Delta_n^{HS}(10)\}$$

even though the true underlying sparsity value is not on the grid. This does not seem to be a drawback.

In Fig.2.12 it appears that even with unknown value of s the Δ_n^{HS} test procedure performs better than Δ_n^{MS}. It can be noticed that the curves show larger differences for lower values of the measure of separation.



(a) $s = 4$ (b) $s = 7$

FIGURE 2.12 – Power of Δ_n^{MS} in red and Δ_n^{HS} in blue by the sum of non null entries of the covariance matrix in absolute value. The powers are plotted for different values of s, unknown by the procedures, in (a) and (b).

In conclusion, the theoretical improvements of highly-sparse over moderately sparse procedures show up in the very extreme cases where the underlying signal is very close to white noise either because of very weak correlations or of very few non-null values.

2.6.4 A moderately sparse high-dimensional MA series

Let us construct a stationary process belonging to our set of sparse covariance matrices. Consider the stationary process X_t defined by the following moving average (MA) model :

$$X_t = \sum_{i=0}^{\lfloor \frac{p}{4} \rfloor} \phi^i \epsilon_{t-2i}$$

with $\{\epsilon_t\}_{t\in\mathbb{N}}$ a Gaussian white noise and $|\phi| < 1$. The auto-covariance function of this series is

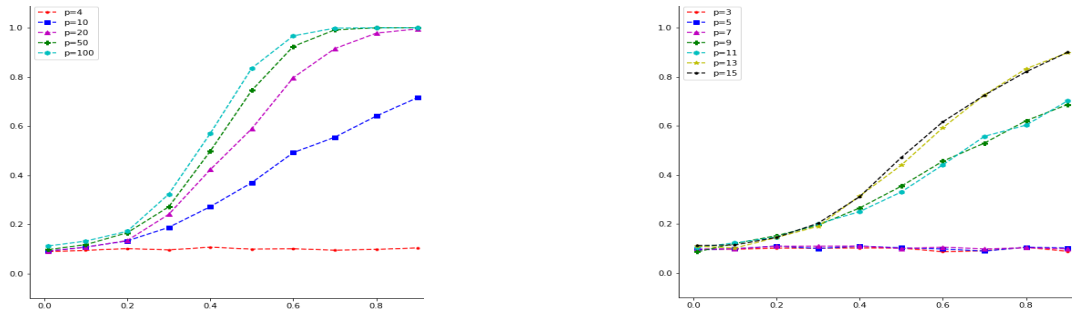
$$\text{Cov}(X_{t+h}, X_t) = \begin{cases} 0, & \text{if } h \text{ odd, or } h \geq \frac{p}{4}, \\ \phi^{-\frac{h}{2}} \left(\frac{\phi^h - \phi^{2(\lfloor \frac{p}{4}+1 \rfloor)}}{1-\phi^2} \right), & \text{otherwise.} \end{cases}$$

In this example, the p-dimensional Gaussian vector $X = (X_t, ..., X_{t+p})$ has a covariance matrix belonging to the class $\mathcal{F}(s, S, \sigma)$ with $s \geq \frac{p}{4} - 1$ tending to infinity with p, $S \leq \frac{p}{2}$ and

$$\sigma = \phi^{-\frac{1}{2}\lfloor \frac{p}{4} \rfloor} \left(\frac{\phi^{\lfloor \frac{p}{4} \rfloor} - \phi^{2\left(\lfloor \frac{p}{4}+1 \rfloor\right)}}{1-\phi^2} \right).$$

The power of the Δ_n^{MS} test is plotted on the y-axis and the value of $\phi < 1$ on the x-axis.

Fig.2.13 shows the power of the Δ_n^{MS} test for this example for various values of p. It can be seen that the Δ_n^{MS} test performs better when the value of p increases showing again that higher the dimension better information on the underlying model. It can be pointed out that for $p < 8$ the $MA(\lfloor p/4 \rfloor)$ is a white noise. It explains why the power of the Δ_n^{MS} test stays constantly low when $p < 8$.



(a) $n = 500$ (b) $n = 50$

FIGURE 2.13 – Power of Δ_n^{MS} test for the $MA(\lfloor p/4 \rfloor)$ for $n = 500$ in (a), $n = 50$ in (b) and for $p = 4$ in red, $p = 10$ in blue, $p = 20$ in magenta, $p = 50$ in green, $p = 100$ in cyan. The horizontal axis represents the value of ϕ and the vertical axis represents the power of Δ_n^{MS}.

2.6.5 Comparison to other test procedures

Our two-sided test procedures are compared with the ones presented in [122] and [60] that are implemented here. In order to calculate these test statistics, is first denoted by α_k the k-th moment of the spectral distribution of Σ, $\alpha_k = \frac{1}{p}\mathrm{Tr}(\Sigma^k)$ and by $\widehat{\beta}_k$ the k-th moment of the spectral distribution of Σ_n, $\widehat{\beta}_k = \frac{1}{p}\mathrm{Tr}(\Sigma_n^k)$.

The authors propose to estimate the α_i, $i \in \{1, 2, 3, 4\}$ using

$$\hat{\alpha}_1 = \hat{\beta}_1, \qquad \hat{\alpha}_2 = \gamma_n^{(2)} \cdot \left(\hat{\beta}_2 - \frac{p}{n}\hat{\beta}_1^2\right), \qquad \hat{\alpha}_3 = \gamma_n^{(3)} \cdot \left(\hat{\beta}_3 - 3\frac{p}{n}\hat{\beta}_2\hat{\beta}_1 + 2(\frac{p}{n})^2\hat{\beta}_1^3\right),$$

$$\hat{\alpha}_4 = \gamma_n^{(4)} \cdot \left(\hat{\beta}_4 - 4\frac{p}{n}\hat{\beta}_3\hat{\beta}_1 - \frac{2n^2+3n-6}{n^2+n+2}\frac{p}{n}\hat{\beta}_2^2 + \frac{10n^2+12n}{n^2+n+2}(\frac{p}{n})^2\hat{\beta}_2\hat{\beta}_1^2 - \frac{5n^2+6n}{n^2+n+2}(\frac{p}{n})^3\hat{\beta}_1^4\right),$$

$$\gamma_n^{(2)} = \frac{n^2}{(n-1)(n+2)}, \gamma_n^{(3)} = \frac{n^4}{(n-1)(n-2)(n+2)(n+4)},$$

$$\gamma_n^{(4)} = \frac{n^5(n^2+n+2)}{(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)}.$$

Using these estimators [122] proposed the test statistic T_{sri} and [60] proposed two test statistics T_{f1} and T_{f2} defined as follows :

$$T_{sri} = \frac{n}{2}(\hat{\alpha}_2 - 2\hat{\alpha}_1 + 1), \qquad T_{f1} = \frac{n}{c\sqrt{8}}\left(\hat{\alpha}_4 - 4\hat{\alpha}_3 + 6\hat{\alpha}_1 + 1\right), \qquad T_{f2} = \frac{n}{\sqrt{8(c^2+12c+8)}}\left(\hat{\alpha}_4 - 2\hat{\alpha}_2 + 1\right),$$

where $c = \lim \frac{p}{n}$, as n and p tend to infinity, is supposed finite and positive.

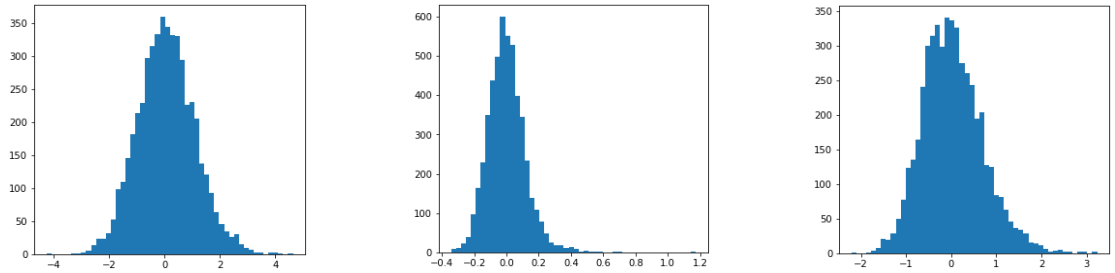
Additional assumptions are needed.

Assumption 1 : There exists $(w_{i,j})_{i,j\geq 1}$ random variables with $\mathbb{E}[w_{11}] = 0$, $\mathbb{E}[w_{11}^2] = 1$ and $\mathbb{E}[w_{11}^4] < \infty$ and for all p, n, $W = (w_{i,j})_{1\leq i\leq p, 1\leq j\leq n}$ such that the observed vector X_j can be represented as $X_j = \Sigma_p^{-1/2} W_{.j}$.

Assumption 2 : The spectral distribution of Σ_p weakly converges to a probability distribution when $p \to \infty$ and the sequence of spectral norms $(||\Sigma_p||)$ is uniformly bounded.

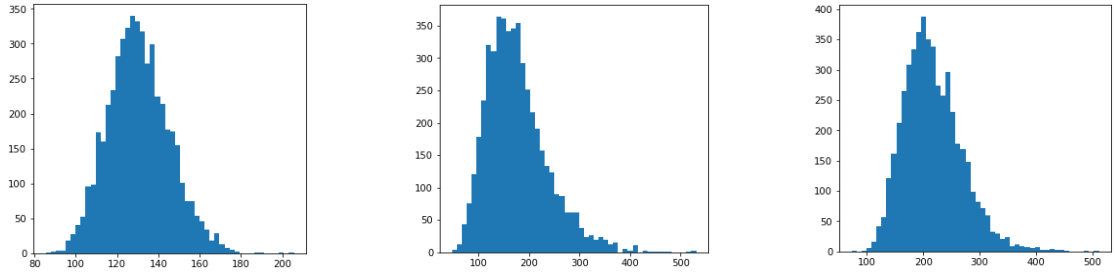
Under $H_0 : \Sigma = I_p$ and assumptions (1) and (2) there are $T_{sri} \to \mathcal{N}(0,1)$, $T_{f1} \to \mathcal{N}(0,1)$ and $T_{f2} \to \mathcal{N}(0,1)$.

Is then plotted the histogram of the defined test statistics under $H_0 : \Sigma = I_p$ in Fig.2.14 and $H_1 : \Sigma \in \mathcal{F}(s, S, \sigma)$, with $n = 200$, $p = 20$, S = $\sqrt{p}$, $s = S-1$ and $\sigma = t_{n,p}^{HS} + \max\left\{\sqrt{\frac{4(u-1)\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}}, \frac{8(u-1)\log\left(s(2s+1)\binom{S}{s}\right)}{n(p-S)}\right\}$, in Fig.2.15.



(a) T_{sri} (b) T_{f1} (c) T_{f2}

FIGURE 2.14 – Histograms of T_{sri} in (a), T_{f1} in (b) and T_{f2} in (c) under $H_0 : \Sigma = I_p$ for $n = 200$, $p = 20$, S = $\sqrt{p}$, $s = S - 1$. On the horizontal axis are represented the values taken by the statistic and on the vertical axis the number of times each value has been taken.



(a) T_{sri} (b) T_{f1} (c) T_{f2}

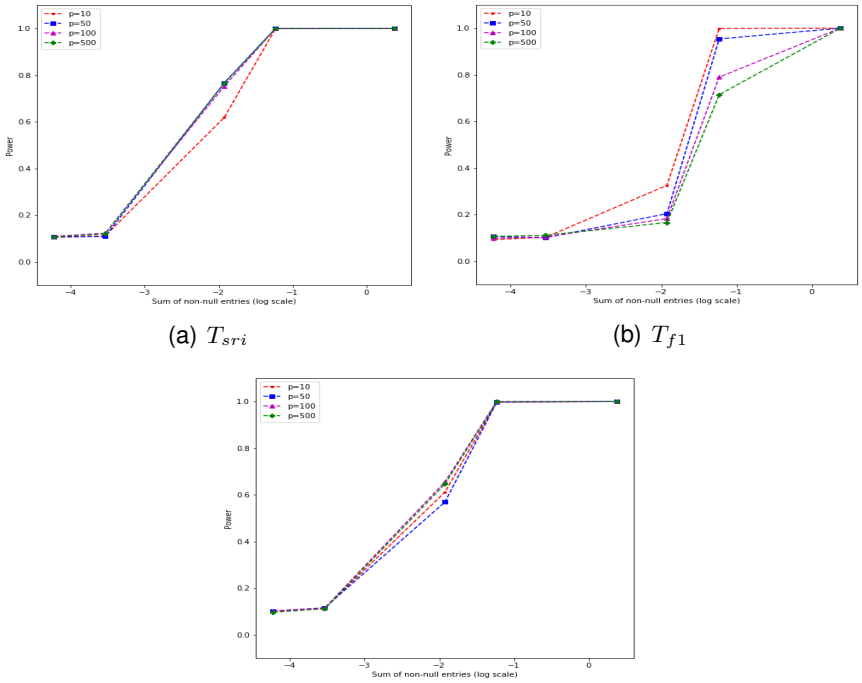
FIGURE 2.15 – Histograms of T_{sri} in (a), T_{f1} in (b) and T_{f2} in (c) under $H_1 : \Sigma \in \mathcal{F}(s, S, \sigma)$ for $n = 200$, $p = 20$, S = $\sqrt{p}$, $s = S - 1$. On the horizontal axis are represented the values taken by the statistic and on the vertical axis the number of times each value has been taken.

In Fig.2.16 are reproduced the powers of the tests associated to T_{sri}, T_{f1} and T_{f2} showing their powers for different values of p, with $n = 200$, as function of $\sum_{j\in\mathcal{C}} |\sigma_j|$ on a logarithmic scale. To plot those powers the same steps are followed as for the powers of Δ_n^{MS} and Δ_n^{HS} tests. The 0.1-quantile of the three tests statistics under the null hypothesis are defined by Monte-Carlo simulation with 5000 samples. Then the value of the non null entries in the alternative hypothesis are gradually increased and it is checked if the test statistics are higher than the defined 0.1-quantile bound. The x-axis represent the sum of non null entries in a logarithmic scale. Under the alternative hypothesis, the covariance matrix is Toeplitz with constant entries on the $s = \lfloor\sqrt{p}\rfloor$ first diagonals equal to σ as defined above.

In Fig.2.17 is plotted the graph for Δ_n^{MS} and Δ_n^{HS} with the exact same parameters in order to provide a fair comparison.

In Fig.2.18 are simultaneously plotted the 5 tests in low dimension as well as in high dimension, respectively.

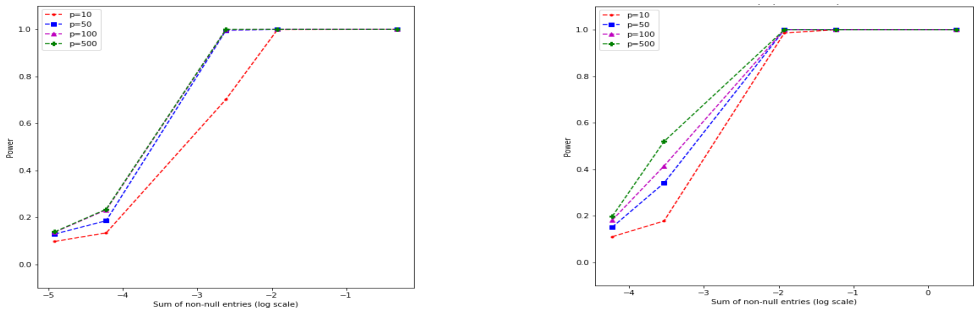
When comparing Fig.2.16 and Fig.2.17 it can be seen that Δ_n^{MS} and Δ_n^{HS} powers are better than those of T_{sri}, T_{f1} and T_{f2} ones. The Δ_n^{MS} and Δ_n^{HS} tests are more sensitive to the non null entries in



(a) T_{sri} (b) T_{f1}

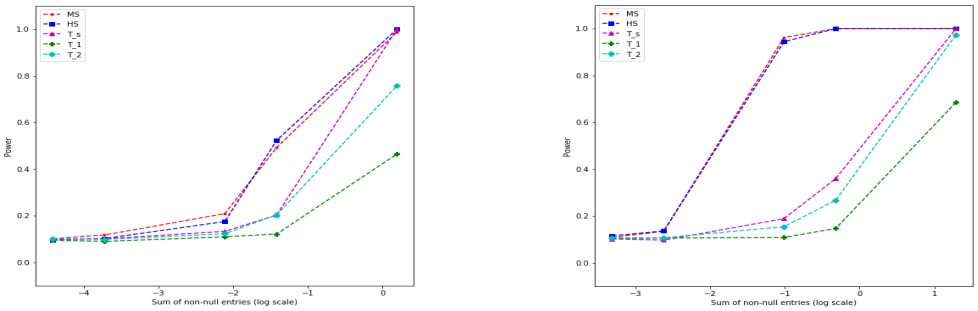
(c) T_{f2}

FIGURE 2.16 – Power of the tests associated with T_{sri}, T_{f1} and T_{f2} for $p = 10$ in red, $p = 50$ in blue, $p = 100$ in magenta, $p = 500$ in green and $n = 200$. The powers are plotted as function of $\sum_{j\in\mathcal{C}} |\sigma_j|$ on a logarithmic scale. The horizontal axis represents the sum of entries in absolute value and the vertical axis represents the value of the power.



(a) Δ_n^{MS} (b) Δ_n^{HS}

FIGURE 2.17 – Power of Δ_n^{MS} and Δ_n^{HS} for $p = 10$ in red, $p = 50$ in blue, $p = 100$ in magenta, $p = 500$ in green and $n = 200$. The powers are plotted as function of $\sum_{j\in\mathcal{C}} |\sigma_j|$ on a logarithmic scale. The horizontal axis represents the sum of entries in absolute value and the vertical axis represents the value of the power.



(a) $n = 20$ and $p = 10$ (b) $n = 20$ and $p = 50$

FIGURE 2.18 – Power of Δ_n^{MS}, Δ_n^{HS}, T_{sri}, T_{f1} and T_{f2} for $n = 20$ and $p = 10$ in (a), $p = 50$ in (b). The powers are plotted as function of $\sum_{j\in\mathcal{C}} |\sigma_j|$ on a logarithmic scale. The horizontal axis represents the sum of entries in absolute value and the vertical axis represents the value of the power.

the sparse Toeplitz covariance matrix. This is confirmed by the Fig.2.18.

Are now compared our two-sided test procedures to the test procedure associated with the statistic $V_{n,k}$ presented in [110] that takes advantage of the sparsity assumption. The authors proposed a generalisation of the $V_n = p^{-1}\text{Tr}\left((\Sigma_n - I_n)^2\right)$ test statistic previously proposed by [83] and [102]. The generalization is $V_{n,k}$ that bands the empirical covariance matrix to its first k diagonals and adds the necessary corrections.

The following assumptions are needed.

Assumption 1 : $\Sigma \in \left\{\Sigma : \max_j \sum_{|i-j|>k_0} |\sigma_{i,j}| \leq Ck_0^{-\alpha}\ \forall k_0 \geq 0,\ 0 < \epsilon_0 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq 1/\epsilon_0\right\}$

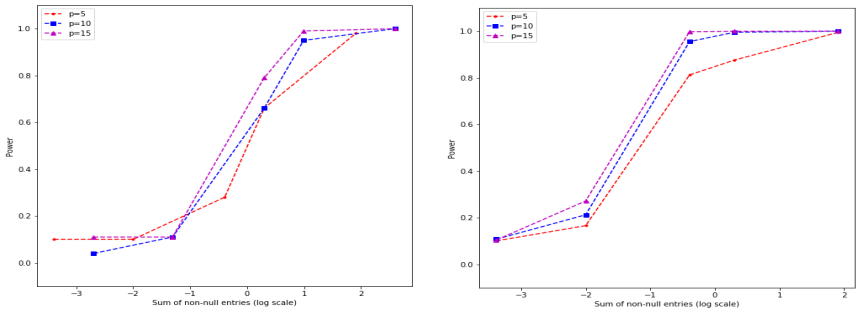
for some constants ϵ_0, C and α which are unrelated to p.

Assumption 2 : Data $X_1, \ldots, X_n$ are independent and identically distributed p-dimensional random vectors such that $X_i = \Gamma Z_i$ where $\Gamma \in \mathbb{R}^{p\times m}$ is a constant loading matrix such that $p \leq m$ and $\Gamma\Gamma^T = \Sigma$ and Z_i are independent and identically p-dimensional random vectors with zero mean and identity covariance.

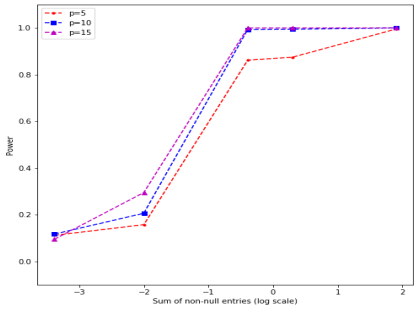
Then $\mathbb{E}(V_{n,k}) = p^{-1}\text{Tr}\left((B_k(\Sigma) - I_p)^2\right)$ and $\mathbb{V}(V_{n,k}) = p^{-2}\sigma_{V_{n,k}}(1 + o(1))$ with $\sigma_{V_{n,k}}$ defined in [110]. Finally under Assumptions 1 and 2, it is proven that $p \cdot \sigma^{-1}_{V_{n,k},\mathcal{H}_0} \cdot V_{n,k} \longrightarrow \mathcal{N}(0,1)$.

In order to limit the computation cost of this simulation the study only focuses on $V_{n,k}$ and it is chosen to not estimate $\sigma_{V_{n,k}}$. We choose $n = 10$ and define the 0.1-quantile of $V_{n,k}$ under the null hypothesis with only 50 samples. Then the non null entries of Σ are gradually increased under the alternative hypothesis for different values of p. The power of the test procedure based on $V_{n,k}$, named $T_{V_{n,k}}$, is then plotted here under. To provide a fair comparison the powers of Δ_n^{MS} and Δ_n^{HS} tests are also plotted under the same environment. The x-axis is the sum of non null entries in a logarithmic scale. To provide a fair comparison again let's choose $k = S$ meaning only non null entries inside the lag support are being looked for.

Fig.2.19 shows that the $V_{n,k}$ procedure is performing well to detect non null entries inside the lag support. It can also be observed that the dimension is improving the performance of the $V_{n,k}$ procedure. However our two-sided test procedures are more sensitive as they detect smaller non-null entries than the $V_{n,k}$ procedure.



(a) $T_{V_{n,k}}$ (b) Δ_n^{MS}



(c) Δ_n^{HS}

FIGURE 2.19 – Power of $T_{V_{n,k}}$, Δ_n^{MS} and Δ_n^{HS} for different values of p and n as function of $\sum_{j\in\mathcal{C}} |\sigma_j|$ on a logarithmic scale.

2.6.6 Application to real data

This section proves that the procedures previously presented can be successfully applied on real data. The test procedures are applied on meteorological data available at `http://berkeleyearth.org/data/`and since they reject the null hypothesis, the lag-selection procedure is also applied. The considered dataset gives the monthly average temperature available in 100 cities since February 1847. Only the four cities with the smallest number of missing values are kept, namely Mexico, New-York, Santo-Domingo, Toronto. The monthly data are then averaged by year in order to avoid seasonality.

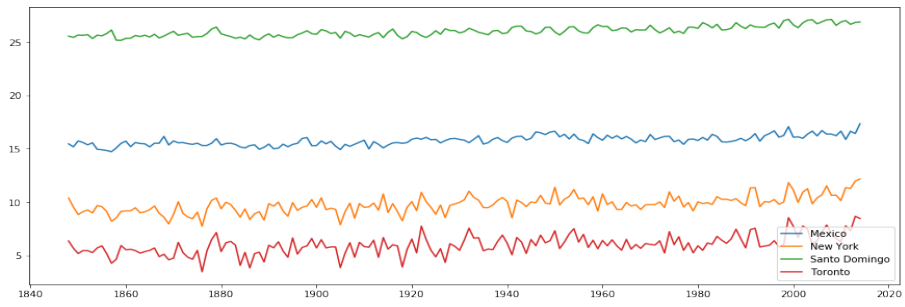


FIGURE 2.20 – Yearly average temperature over time of Mexico, New-York, Santo Domingo and Toronto since 1847.

An augmented Dickey-Fuller Test is performed to verify whether the time series are stationary or not. The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary. The alternative hypothesis is that the time series is stationary. The test is interpreted using its p-value. If the p-value is below the 5% threshold suggests that the null hypothesis has to be rejected and then the time series is assumed to be stationary. If the p-value is above the 5% threshold then the null hypothesis cannot be rejected and the time series is assumed to not be stationary.

TABLE 2.2 – P-value of the augmented Dickey-Fuller Test on yearly average temperatures. The series is not stationary if the p-value is above the 5% threshold.

City	Mexico	New-York	Santo-Domingo	Toronto
p-value	0.583765	0.933036	0.776110	0.952561
Conclusion	Non Stationary	Non Stationary	Non Stationary	Non Stationary

Table 2.2 shows that the time series are not stationary. The first difference method is used to make the time series stationary.

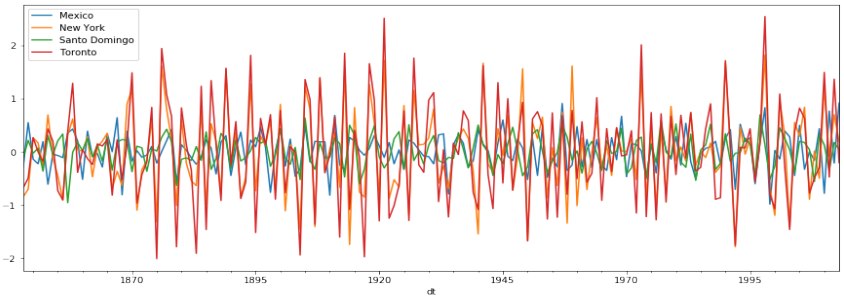


FIGURE 2.21 – Time series after first difference method applied.

Fig. 2.21 shows that after applying the first difference method time series look stationary. The Augmented Dickey-Fuller Test is applied again to verify whether the time series are now indeed stationary or still not.

TABLE 2.3 – P-value of the augmented Dickey-Fuller Test on first difference yearly averaged temperatures. The series is not stationary if the p-value is above the 5% threshold.

City	Mexico	New-York	Santo-Domingo	Toronto
p-value	2.680904e-11	2.529196e-09	1.163820e-15	8.470041e-10
Conclusion	Stationary	Stationary	Stationary	Stationary

Table 2.3 confirms that the time series are stationary after applying the first difference method. We now want to check if the time series are normally distributed.

Fig. 2.22 shows histograms that do not contradict normality of the distributions. To ensure the normality of the time series the Shapiro-Wilk test is performed as well as the D'Agostino's K-squared test. Both tests are interpreted using their p-values. A p-value below the 5% threshold suggests the null hypothesis has to be rejected and that the data can be assumed not to be drawn from a gaussian distribution.

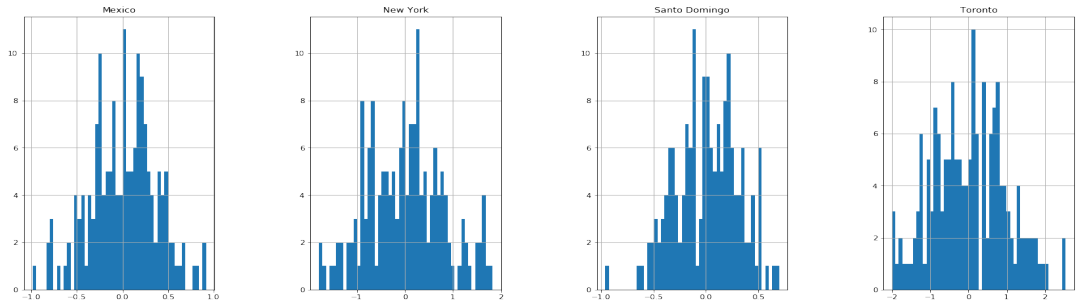
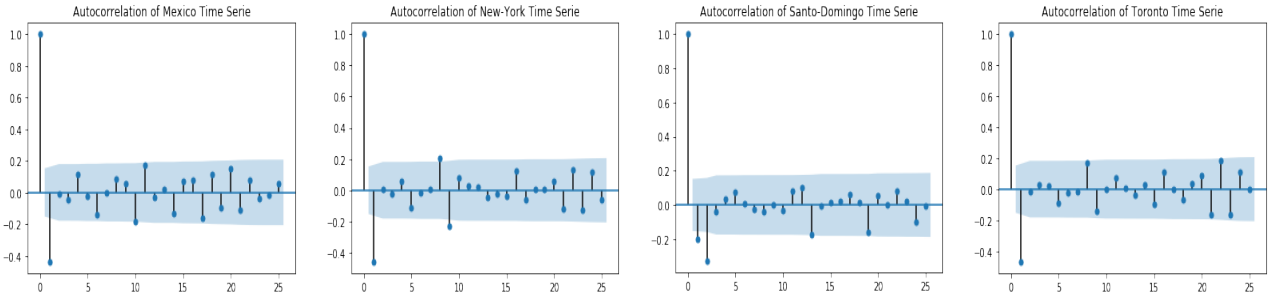


FIGURE 2.22 – Histograms of the temperatures after first difference applied.

TABLE 2.4 – Shapiro-Wilk and D'Agostino's K-squared tests on first difference yearly averaged temperatures. We reject the normal distribution hypothesis when the p-value is below the 5% threshold.

City	Mexico	New-York	Santo-Domingo	Toronto
Shapiro-Wilk p-value	0.773	0.273	0.458	0.409
D'Agostino's K-squared p-value	0.748	0.508	0.493	0.288
Conclusion	Normal	Normal	Normal	Normal

Table 2.4 confirms the normality of the time series. Before applying our procedure the autocorrelations of the four time series are plotted. This will give some additional informations on the structure of the time series.



(a) Mexico (b) New-York (c) Santo-Domingo (d) Toronto

FIGURE 2.23 – Autocorrelation plots of the first difference yearly averaged temperature time series.

The p-dimensional vectors $X_t = (x_{t-i})_{1\leq i\leq p}$ are now created. The value of p is set to be 10, namely $p = 10$. To ensure lack of significant correlations between the vectors $(X_t)_t$ we separate them by the largest non null autocorrelation. The largest non null autocorrelation is considered to be 1 for Mexico, 9 for New-York, 2 for Santo-Domingo and 1 for Toronto. As an example, the Mexico time series will be $X_1^{Mexico} = (x_1^{Mexico}, x_2^{Mexico}, \ldots, x_{10}^{Mexico})$ and $X_2^{Mexico} = (x_{12}^{Mexico}, x_{13}^{Mexico}, \ldots, x_{21}^{Mexico})$.

Our procedures can no be applied. The Δ_n^{MS} and Δ_n^{HS} tests can be performed to verify if the p-dimensional vectors $X_t = (x_{t-i})_{1\leq i\leq p}$ are issued from a $\mathcal{N}_p(0, I_p)$ (null hypothesis) or $\mathcal{N}_p(0, \Sigma)$ for some $\Sigma \in \mathcal{F}(s, S, \sigma)$.

Table 2.5 shows that according to the Δ_n^{MS} and Δ_n^{HS} tests, the vectors X_t are not issued from

TABLE 2.5 – Δ_n^{MS} and Δ_n^{HS} tests on first difference yearly averaged temperatures to test $H_0 : X_t \sim \mathcal{N}_p(0, I_p)$ vs $H_1 : X_t \sim \mathcal{N}_p(0, \Sigma)$. In the table are presented the accepted hypothesis for each test executed on each city.

City	Mexico	New-York	Santo-Domingo	Toronto
Δ_n^{MS} Test	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$
Δ_n^{HS} Test	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$	$\mathcal{N}_p(0, \Sigma)$

$\mathcal{N}_p(0, I_p)$ but from a $\mathcal{N}_p(0, \Sigma)$ with $\Sigma \in \mathcal{F}(s, S, \sigma)$. This demonstrates that our procedure detect the significant correlations in the true underlying covariance matrix Σ.

The support of the non null entries is then recovered by using the lag-selection procedure exposed in Section 4. τ_n is set to be the 0.75-quantile of $|\varphi_{A_j}(\Sigma_n)|$ where Σ_n is the empirical covariance matrix of vectors generated from a $\mathcal{N}_p(0, I_p)$. The results are reported in Table 2.6.

TABLE 2.6 – Support of non null entries recovered by the lag selection procedure.

City	Mexico	New-York	Santo-Domingo	Toronto
Support	{1}	{1, 6, 7, 9}	{2}	{1}

Those results are consistent with the autocorrelations plotted in Fig.2.23. It can be seen that the Mexico time series presents only a non null autocorrelation at lag 1. For the New-York time series the lag 1 is non null as well as the lags 8 and 9. The procedure selects lags 1, 6, 7 and 9 and is not as efficient for this time series as for the others, but that can be explained by the very small number of vectors ($n = 8$) that are available within this series. For the Santo-Domingo time series the lags 1 and 2 are non null with the second being larger than the first. The procedure only selects the second one. Finally for the Toronto time series it can be seen that only the first lag is significantly non null and it is the only one selected by the procedure. The non null lags in the autocorrelation plots are thus consistent with the ones selected by our procedure.

Chapitre 3

Two-sided Matrix Regression

3.1 Introduction

Supervised learning is often performed on large data bases. Matrix regression assumes that the data Y can be well explained by a set of features given by the columns of the matrix X and linear combinations of these columns. It is often the case in real-life that the rows of Y can be explained by linear combinations of the rows of X.

For example, economic data store economic indicators as column features and countries as rows. Such a matrix is usually explained by a smaller matrix roughly containing a smaller number of countries (representatives of groups of geographically or economically close countries) and a few economic features or some factors produced out of all these indicators. We would like to predict a larger number of indicators for a larger number of countries, *i.e.* Y a $n \times p$ matrix, using the features X a $m \times q$ matrix.
Recommendation systems want to predict the opinion of n clients concerning p items. We can use publicly available data on a number m of different groups of clients and their affinity to a number q of large categories of items in order to predict by evaluating the client's correlation to the prescribed groups in the population and the item's weight in its category. We may include a multiple-label situation where the items belonging to a main category are also related to other categories.
Other examples can be given for meteorological data, medical or pharmaceutical data and so on.

Model. We observe the matrix $Y \in \mathbb{R}^{n\times p}$ and a design matrix $X \in \mathbb{R}^{m\times q}$ related via the **two-sided matrix regression (2MR)** model involving two parameter matrices $A^* \in \mathbb{R}^{n\times m}$ and $B^* \in \mathbb{R}^{q\times p}$:

$$Y = A^* X B^* + E, \tag{3.1}$$

where the noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

The 2MR model encompasses known models like, *e.g.* matrix regression and matrix factorisation. Indeed, if $n = m$ and A^* is the identity, the matrix model (3.1) becomes the (one-sided) *matrix regression* (MR) model $Y = XB^* + E$, see [108], [32], [104].
Assume now that $m = q$ and that the design matrix X is the identity matrix of rank m smaller than both n and p. Our model becomes a *factorisation model* of the signal $M^* = A^*B^*$ observed with noise. The idea is to recover a low-rank structure generating the observed data. In [85] the authors have considered structured factorisation of the signal under assumptions that the rows of A^* and the columns of B^* have a common sparsity parameter and X, which they do not observe, has a much smaller dimension than Y.

The 2MR model (3.1) is strongly related to other models, but we argue that it cannot be reduced to these other models of a different nature. Indeed, note that the entry Y_{ij} of the matrix Y can be written

$$Y_{ij} = \mathrm{Tr}(X \cdot B^*_{\cdot,j} A^*_{i,\cdot}) + E_{ij},$$

for any i in $[n]$, where $[n] = \{1, \dots, n\}$, and for any j in $[p]$. Thus every entry Y_{ij} brings information through the same design matrix X on the rank 1 matrix $B^*_{\cdot,j} A^*_{i,\cdot}$. This is unlike *the trace-regression model* or the more general *matrix completion* studied by [116], [88], where a different design matrix brings information on the parameter matrix B^*A^*.
Another way of writing model (3.1) is in the form of *vector regression model*, by stacking the columns of matrices Y, X and E into $vec(Y)$, $vec(X)$ and $vec(E)$, respectively, to get

$$vec(Y)^\top = vec(X)^\top \cdot A^\top \otimes B + vec(E)^\top, \tag{3.2}$$

where $\otimes$ denotes the tensor product of two matrices. Under this relation, we predict a row vector of size np using a row vector of size mq (the matrix of features has rank 1) via a parameter of size $(mq) \times (np)$ which cannot go well unless the structure of A^* and B^* is trivial. This approach cannot take into account the matrix structure of the features, of the matrices A^*, B^*, and it gives poor results on that account.

This model has been introduced in time series by [48] as the *auto-regressive matrix-valued model* of order 1, MAR(1), $Y_t = A^* Y_{t-1} B^* + E_t$, observed at times t in $[T]$. In this case A^* and B^* are squared matrices with spectral radii strictly less than 1 in order to ensure stability of the time series (X_t is thus stationary and causal). The authors propose three estimation methods : first, they use the vector form analogous to (3.2), stack the T lines of $vec(Y_t)^\top$ and they use the nearest Kronecker product (NKP) problem to give estimators of A^* and B^* out of the global least squares estimator of ${A^*}^\top \otimes B^*$; then, their next method minimizes the least squares over A and B

$$\min_{A,B} \frac{1}{T} \sum_{t=1}^{T} \|Y_t - A Y_{t-1} B\|_F^2,$$

by a sequential procedure minimizing over A for fixed given B, then over B for fixed A, and iterating ; finally, they give an MLE procedure over A and B under a particular structure of the covariance matrix of E and proceed also sequentially. Theoretical results state the asymptotic normality as T tends to infinity, for fixed dimensions. However, the first procedure is cumbersome as the estimated matrix is very large, while the other two procedures are based on non-convex minimization without theoretical guarantees as to the limit points of the algorithm.
Least squares and MLE estimators with AIC and BIC penalties have been numerically studied by [76] of a more general time series model

$$Y_t = \sum_{\ell=1}^{L} A_\ell Y_{t-\ell} B_\ell + E_t, \quad t = 1, \dots, T,$$

which is treated as $Y_t = A^* X_t B^* + E_t$, where X_t is the block diagonal matrix containing the $L-$past observed matrices $Y_{t-1}, \dots, Y_{t-L}$ and $A^* = (A_1, \dots, A_L)$ and $B^* = (B_1^\top, \dots, B_L^\top)^\top$ are the concatenated matrices in the previous equation.
Thus, our paper is motivated by the need to deal with high-dimensional data and finite (non-asymptotic) time (say $T = 1$) in order to provide theoretical guarantees for prediction.

Contributions. We show in Section 3.2 that by using the SVD of matrices $Y = U_Y \Sigma_Y V_Y^T$ and $X = U_X \Sigma_X V_X^T$, the least squares procedure can be reduced to fitting predictors of the form $A_0 \Sigma_X B_0$

to the diagonal matrix Σ_Y with explicit relations between A_0, B_0 and A, B. There is a natural choice of predictors of A_0 and of B_0 under diagonal form. We study these predictors for given ranks r and that we transform back into the original space of Y without loss of prediction rate. Then we give a data-dependent rank selector and show that the predictors associated to it attain optimal bounds. We give sufficient conditions so that the rank selector is consistent. Finally, we slightly modify the procedure to be free of the parameter σ of the noise and show new upper bounds in this case. In Section 3.3, we study the nuclear norm penalized least squares and show it attains the optimal bounds too. All proofs are in a dedicated section in the Appendix. Finally, we illustrate in Section 3.4 via numerical simulations the excellent prediction results of these fast running, explicit predictors.

Notations. For any matrix M of size $n \times m$ and rank r_M, we denote its singular value decomposition (SVD) by $M = U_M \Sigma_M V_M^\top$, where U_M belongs to $\mathcal{O}_n$ - the set of orthogonal matrices of size $n \times n$, V_M belongs to $\mathcal{O}_m$ and $\Sigma_M = Diag_{n,m}(\sigma_k(M),\, 1 \le k \le r_M)$. Note that $\sigma_1(M), \dots, \sigma_{r_M}(M)$ are the positive singular values of M listed in decreasing order, and the $n \times m$ diagonal matrix $Diag_{n,m}(\sigma_k(M),\, 1 \le k \le r_M)$ has diagonal entries in the list and 0 elsewhere. Furthermore, denote $\|M\|_F^2 = \sum_{k=1}^{n\wedge m} \sigma_k(M)^2$ its Frobenius norm, $\|M\|_{(2,q)}^2 = \sum_{k=1}^{q} \sigma_k(M)^2$ its Ky-Fan $(2,q)$ norm, $\|M\|_{op} = \sigma_1(M)$ its operator norm, $\|M\|_* = \sum_{k=1}^{n\wedge m} \sigma_k(M)$ its nuclear norm, $M^\dagger$ its Moore-Penrose inverse, r_M its rank and M^T its transpose. For any matrices M_1 and M_2 in $\mathbb{R}^{n\times m}$, $\langle M_1, M_2\rangle_F$ denotes the canonical scalar product, *i.e.* $\langle M_1, M_2\rangle_F = \mathrm{Tr}(M_1^T M_2)$. For any $r \in [r_M]$, we denote $[M]_r$ the best $\operatorname{rank} r$ approximation of M for the Frobenius norm. In the model (3.1), let us denote by r^* the rank of A^*XB^*.

3.2 Rank penalized learning

In this section we propose rank adaptive predictors and provide theoretical guarantees for their error. First we give explicit predictors under the assumption that the ranks of the parameter matrices are known, then a selection procedure will allow to provide a data-dependent rank selector and the associated rank-adaptive predictor. Even though we follow classical results for rank penalized (one-sided) matrix regression, e.g. [32], [63] and [26], we give details for the fixed rank two-sided matrix regression which is novel to the best of our knowledge. Surprisingly, explicit predictors can be proposed despite the identifiability issues of this model. Only after this, we proceed to rank selection and rank-adaptive learning.

3.2.1 Prediction for given ranks

Let r belong to $[n \wedge p \wedge r_X]$. Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex minimization problem

$$\min_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \le r}} \|Y - AXB\|_F^2. \tag{3.3}$$

Notice that the rank constraints on A and B use the same value r. Indeed the objective is to build a predictor for the signal A^*XB^* which satisfies $\operatorname{rank}(A^*XB^*) \le \min(r_{A^*}, r_X, r_{B^*})$. In the steps of the proof of our results, we see that the upper bound of the risk depends on the ranks of A^* and of B^* only through their least value and no information can be recovered on the largest rank of the two. Hence it

makes sense to look for A and B sharing the same rank as a dimension reduction technique without any impact on the final results.

The model (3.1) can be rewritten using the SVD of the observed matrix Y and of the design matrix X as

$$\Sigma_Y = A_0^* \cdot \Sigma_X \cdot B_0^* + E_0, \tag{3.4}$$

where $A_0^* = U_Y^T A^* U_X$, $B_0^* = V_X^T B^* V_Y$ and $E_0 := U_Y^T \cdot E \cdot V_Y$. In the particular case where E has independent entries with distribution $\mathcal{N}(0, \sigma^2)$ than so does E_0, see Lemma 3.5.1. Now, Σ_Y and Σ_X are diagonal matrices, not necessarily squared, not necessarily full rank. Given the invariance of the Frobenius norm by left or right multiplication with orthogonal matrices, we get that for any matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{q\times p}$ we have

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2,$$

where $A_0 = U_Y^T A U_X$ and $B_0 = V_X^\top B V_Y$ are obtained via analogous transformations to those relating the true underlying parameters.

Obviously, matrices A and A_0 have the same rank, and the same holds for B and B_0. Therefore, solving (3.3) is equivalent to solving for $\hat{A}_{0r}$ and $\hat{B}_{0r}$ solutions of

$$\min_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2. \tag{3.5}$$

Theorem 3.2.1 *Let us define for* $r \in [n \wedge p \wedge r_X]$

$$\hat{A}_{0r} = Diag_{n,m}(\sigma_k(Y),\, 1 \le k \le r \wedge r_Y) \quad \textit{and} \quad \hat{B}_{0r} = Diag_{q,p}(\sigma_k(X)^{-1},\, 1 \le k \le r). \tag{3.6}$$

Then, $(\hat{A}_{0r}, \hat{B}_{0r})$ *belong to the set of solutions of problem* (3.5) *and the predictor* $\hat{A}_{0r}\Sigma_X \hat{B}_{0r}$ *satisfies for an absolute constant* $C > 0$ *and for any* $t > 0$*, the oracle inequality*

$$\begin{aligned}\|A_0^* \Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X \hat{B}_{0r}\|_F^2 \le 9 \inf_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|A_0^* \Sigma_X B_0^* - A_0 \Sigma_X B_0\|_F^2 \\ + 24 C \sigma^2 (1+t)^2 \cdot r(n+p),\end{aligned}$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

Next, from the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ of (3.5) we deduce explicit solutions of (3.3).

Corollary 3.2.2 *Let us define for* $r \in [n \wedge p \wedge r_X]$

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^T \quad \textit{and} \quad \hat{B}_r = V_X \hat{B}_{0r} V_Y^T, \tag{3.7}$$

with $\hat{A}_{0r}$ *and* $\hat{B}_{0r}$ *defined in* (3.6)*. Then* $(\hat{A}_r, \hat{B}_r)$ *are solution to the problem* (3.3) *and the predictor* $\hat{A}_r X \hat{B}_r$ *satisfies for an absolute constant* $C > 0$ *and for any* $t > 0$*, the oracle inequality*

$$\|A^* X B^* - \hat{A}_r X \hat{B}_r\|_F^2 \le 9 \inf_{\substack{A, B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \le r}} \|A^* X B^* - AXB\|_F^2 + 24 C \sigma^2 (1+t)^2 \cdot r(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

The proofs of Theorem 3.2.1 and of Corollary 3.2.2 can be found in Section 3.5. In the proofs we explicit the bias in terms of the unknown matrix parameters :

$$\inf_{\substack{A,B:\\ \operatorname{rank} A\wedge \operatorname{rank} B\leq r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*}.$$

Note that our choice for the couple of predictors $(\hat{A}_{0r}, \hat{B}_{0r})$ is not unique and we can easily derive families of solutions to the problem (3.5). Each family of solutions can be turned into a solution to the problem (3.3). Indeed, consider $(\alpha\hat{A}_{0r}, \frac{1}{\alpha}\hat{B}_{0r})$ with arbitrary $\alpha > 0$. Alternatively, let λ_i for all $i \leq m \wedge q$ be arbitrary positive numbers, then

$$(\hat{A}_{0r} Diag_{m,m}(\lambda_1, \ldots, \lambda_{m\wedge q}), Diag_{q,q}(\lambda_1^{-1}, \ldots, \lambda_{m\wedge q}^{-1})\hat{B}_{0r})$$

give the same prediction. Let us see that the same transformations applied to the parameter matrices A_0^* and B_0^* also lead to the same signal matrix $A_0^*\Sigma_X B_0^*$. Indeed, the model is non-identifiable and so, without further strong assumptions, we can only hope to learn the global signal, and not the parameters of the model.

Alternative predictors. Let us define a second couple of predictors $(\tilde{A}, \tilde{B}_r)$ producing exactly the same prediction as $(\hat{A}_r, \hat{B}_r)$ with the same theoretical properties, but having the advantage that $\tilde{A}$ is full rank and does not depend on r. Define

$$\tilde{A}_0 = I_{n,m} \quad \text{and} \quad \tilde{B}_{0r} = Diag_{q,p}\left(\frac{\sigma_k(Y)}{\sigma_k(X)}, 1 \leq k \leq r \wedge r_Y\right)$$

where $I_{n,m}$ denotes the identity matrix of dimension $n \times m$, whereas $\tilde{B}_{0r}$ has rank $r \wedge r_Y$. Using the analogous transformations we obtain

$$\tilde{A} = U_Y I_{n,m} U_X^T \quad \text{and} \quad \tilde{B}_r = V_X \tilde{B}_{0r} V_Y^T.$$

It is easy to see that Theorem 3.2.1 is valid for $\tilde{A}_0$ and $\tilde{B}_{0r}$, and that Corollary 3.2.2 is valid for $\tilde{A}$ and $\tilde{B}_r$.

3.2.2 Rank-adaptive prediction

In this section, we propose rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ which are selected from the family $\{(\hat{A}_r, \hat{B}_r) : r \in [n \wedge p \wedge r_X]\}$ by a model selection procedure analogous to that of [32]. Let us first define, for a generic matrix M and any $\lambda > 0$, the $\lambda-$rank of M as

$$r_M(\lambda) = 1 \vee \sum_{k=1}^{\operatorname{rank} M} \mathbf{1}_{\sigma_k(M)^2 \geq \lambda}.$$

For given $\lambda > 0$, let

$$\hat{r} := \arg\min_{r\in[n\wedge p\wedge r_X]} \left\{ \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r \right\}. \tag{3.8}$$

Consider the predictors introduced in (3.7) for the data-driven rank $\hat{r}$ as defined in (3.8). The next Theorem extends the oracle inequality to the rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ associated to the estimated rank $\hat{r}$ and to some $\lambda > 0$ large enough.

Theorem 3.2.3 *The rank-adaptive predictors* $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ *associated to* $\hat{r}$ *in* (3.8) *and to* λ *such that, for some absolute constant* $C > 0$ *and for any* $t > 0$, $\lambda \geq 4C(1+t)^2\sigma^2(n+p)$, *satisfy the oracle inequality*

$$\|A^*XB^* - \hat{A}_{\hat{r}}X\hat{B}_{\hat{r}}\|_F^2 \leq \min_{r\in[n\wedge p\wedge r_X]} \left\{ 9\sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*} + 6\lambda r \right\},$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Note that the minimum on the right-hand side of the previous display is always smaller than the value at $r = r^*$, giving under the assumptions of Theorem 3.2.3 that

$$\|A^*XB^* - \hat{A}_{\hat{r}}X\hat{B}_{\hat{r}}\|_F^2 \leq 6r^*\lambda,$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.
The bounds of order $r^*(n+p)$ attained by our procedure are analogous to those for the low-rank matrix regression models in [116] and [63]. Indeed, the 2MR model is more difficult than the MR model, (*i.e.* one of the matrices is known) and we will suppose known the matrix with larger rank in order to achieve the correct lower bounds. Thus the lower bounds for prediction in the low-rank MR model will be valid for our model.

3.2.3 Consistent rank selection

We study the consistency of the rank selector $\hat{r}$ in (3.8) and see when it recovers the true rank r^* with high probability. First, we show that, for properly chosen λ, the data-driven rank $\hat{r}$ is actually the unique solution and coincides with the $\lambda-$rank of Y, $\hat{r} = r_Y(\lambda)$.

Proposition 3.2.4 *If* $\lambda > \sigma_{r_Y}(Y)^2$, *there is a unique solution* $\hat{r}$ *to the optimisation problem in* (3.8) *and it is actually the* $\lambda-$*rank of* Y, *i.e.* $\hat{r} = r_Y(\lambda)$.

Next, we prove that $\hat{r}$ recovers with high probability the $\lambda-$rank of A^*XB^*.

Proposition 3.2.5 *Let* $\lambda > 0$ *and denote by* $r^*(\lambda)$ *the* $\lambda-$*rank of* A^*XB^* . *If for some constant* c *in (0,1),* $\sigma_{r^*(\lambda)}(A^*XB^*)^2 > (1+c)^2\lambda$ *and* $\sigma_{r^*(\lambda)+1}(A^*XB^*)^2 < (1-c)^2\lambda$, *then*

$$\mathbb{P}(\hat{r} = r^*(\lambda)) \geq \mathbb{P}(\|E\|_{op}^2 \leq c^2\lambda).$$

In particular, if $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ *for some absolute constant* $C > 0$ *and for any* $t > 0$, *then* $\hat{r} = r^*(\lambda)$ *with probability larger than* $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Finally, remember that the fact that $r^*(\lambda)$ coincides with the true underlying rank r^* is equivalent to having $\sigma_{r^*}(A^*XB^*)^2 \geq \lambda > 0$. The rank selector will then coincide with r^* if λ also satisfies $\sigma_1(E)^2 \leq c^2\lambda$, for some absolute constant $c > 0$. It is therefore necessary that a signal-to-noise ratio, given here by $\sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ be significant in order to have the true underlying rank selected by $\hat{r}$. By combining this with the previous Propositions we get the following.

Proposition 3.2.6 *Let* $\lambda > 0$. *If for some constant* c *in (0,1),* $\sigma_{r^*}(A^*XB^*)^2 > (1+c)^2\lambda$, *then*

$$\mathbb{P}(\hat{r} = r^*) \geq \mathbb{P}(\|E\|_{op}^2 \leq c^2\lambda).$$

In particular, if $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ *for some absolute constant* $C > 0$ *and for any* $t > 0$, *then* $\hat{r} = r^*$ *with probability larger than* $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

3.2.4 Data-driven rank-adaptive prediction

The rank selector $\hat{r}$ in (3.8) is used for building consistent predictors as detailed in Theorem 3.2.3 provided that the condition $\lambda \geq 4C(1+t)^2\sigma^2(n+p)$ is satisfied. However the noise parameter σ is not known in general settings. Thus a data dependent rank selector is needed for building consistent predictors in those cases. Motivated by the previous case where σ^2 was supposed known, we proceed as follows. First, we change the penalty to $\lambda \cdot r\widehat{\sigma}_r^2$ with

$$\widehat{\sigma}_r^2 = \frac{1}{np}\|Y - \hat{A}_r X \hat{B}_r\|_F^2.$$

Note that in the particular case of Gaussian noise $\widehat{\sigma}_r^2$ estimates the variance σ^2 of the noise. Next, given a largest possible value for the true rank $r_{max} \leq n \wedge p \wedge r_X$, we define the data-driven rank selector

$$\bar{r} := \arg\min_{r \in [r_{max}]} \left\{ \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda \cdot r\widehat{\sigma}_r^2 \right\}. \tag{3.9}$$

Finally, we use the predictors $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$. The next theorem extends the upper bounds of Theorem 3.2.3 to these data-driven rank-adaptive predictors.

Theorem 3.2.7 *The data-driven rank-adaptive predictors $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$ associated to $\bar{r}$ in* (3.9) *with $r_{max} \leq n \wedge p \wedge r_X$, and to $\lambda = (1+\varepsilon)np/(r_{max} \vee r_Y)$ for some $\varepsilon > 0$, satisfy for some absolute constant $C > 0$ and for any $t > 0$ the oracle inequality*

$$\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq \min_{r \in [r_{max}]} \left\{ 9\|A^*XB^* - \hat{A}_r X \hat{B}_r\|_F^2 + 6(1+\varepsilon) \cdot r\sigma_{r+1}(A^*XB^*)^2 \right\} + 12C(2+\varepsilon)(1+t)^2 \cdot \sigma^2 r_{max}(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Apply the Corollary 3.2.2, to get under the assumptions of Theorem 3.2.7 that

$$\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq \min_{r \in [r_{max}]} \left\{ 9^2 \inf_{\substack{A,B: \\ r_A \wedge r_B \leq r}} \|A^*XB^* - A_r X B_r\|_F^2 + 6(1+\varepsilon) \cdot r\sigma_{r+1}(A^*XB^*)^2 \right\} + 12(20+\varepsilon)C(1+t)^2 \cdot \sigma^2 r_{max}(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Note that the minimum on the right-hand side of the previous display is always smaller than its value at $r = r^*$ if r_{max} is larger than r^*, giving under the assumptions of Theorem 3.2.7 that

$$\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq 12(20+\varepsilon)C(1+t)^2 \cdot \sigma^2 r_{max}(n+p).$$

In order to compare to the previous results, note that the upper bound derived from Theorem 3.2.3 for the value $r = r^*$ and the least value $\lambda = 4C(1+t)^2\sigma^2(n+p)$ gives the very similar bound

$$\|A^*XB^* - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 \leq 24C(1+t)^2 \cdot \sigma^2 r^*(n+p).$$

From a computational point of view, it is preferable to change $\widehat{\sigma}_r^2$ in some cases. For example, we use in our numerical simulations

$$\widehat{\sigma}_r^2 = \frac{1}{np - (m \wedge q)r_X}\|Y - \hat{A}_r X \hat{B}_r\|_F^2$$

when $n \geq m$, $p \geq q$ and thus $np > (m \wedge q)r_X$. It is straightforward to prove the analogue of Theorem 3.2.7 by considering $\lambda = (1+\varepsilon)(np - (m \wedge q)r_X)/(r_{max} \vee r_Y)$.

3.3 Nuclear norm penalized learning

Nuclear norm penalized least squares is known to exhibit good properties, see [10] or [103]. Hence it may show advantages over rank-penalized methods. Let us define the nuclear norm penalized (NNP) optimisation problem

$$\min_{A,B} \|Y - AXB\|_F^2 + 2\lambda \cdot \|AXB\|_*, \tag{3.10}$$

for some $\lambda > 0$. The objective of the optimization problem is non-jointly convex in A and B. Note that in matrix regression (when A^* is the identity matrix) the nuclear norm of XB has been used , see [88], or other adaptive forms depending on the feature matrix X, [90]. However, we exhibit explicit predictors belonging to the set of solutions of this problem and show an oracle inequality they satisfy.

Theorem 3.3.1 *The predictors* $(\bar{A}, \bar{B})$ *defined by*

$$\bar{A} = U_Y I_{n,m} U_X^\top \quad \textit{and} \quad \bar{B} = V_X \cdot Diag_{q,p}\left(\frac{(\sigma_k(Y) - \lambda)_+}{\sigma_k(X)}, 1 \le k \le r_Y \wedge r_X\right) V_Y^\top \tag{3.11}$$

are solutions to the problem in (3.10)*. Moreover, if* λ *is such that, for some absolute constant* $C > 0$ *and for any* $t > 0$*,* $\lambda \ge 2C(1+t)^2\sigma^2(n+p)$*, they satisfy the oracle inequality*

$$\|A^*XB^* - \bar{A}X\bar{B}\|_F^2 \le 9 \min_{r \in [n \wedge p \wedge r_X]} \left\{ \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*} + 16\lambda r \right\},$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$*.*

The proof can be found in Section 3.5.

Remark. Another approach could be to consider the model under the vectorized form (3.2) and solve the problem

$$\min_{A,B} \|vec(Y)^\top - vec(X)^\top \cdot A^\top \otimes B\|_2^2 + 2\lambda \|A^\top \otimes B\|_*,$$

for some $\lambda > 0$. Recall that $A^\top \otimes B$ denotes the tensor product of matrices $A^\top$ and B and that we can write $\|A^\top \otimes B\|_* = \sum_{k,j \ge 1} \sigma_k(A)\sigma_j(B)$. However, the features are 1-dimensional and we loose the structured information contained in the original matrix X. This approach could make more sense in the case of repeated observation (Y_t, X_t) for t in $[T]$, by stacking the rows $vec(Y_t)^\top$ and $vec(X_t^\top)$ into matrices $\mathbb{Y}$ and $\mathbb{X}$, respectively, and do a classical matrix regression. Even so, the usual assumptions on the feature matrix $\mathbb{X}$ in order to achieve good prediction are not reasonable in this context as they are not much related to the original matrix data sets X_t, t in $[T]$.

Remark (Sufficient conditions for identifiability) We have indicated at several times that many couples of matrices (A, B) solve the equation $M = AXB$ for a given matrix M. Given the SVD of the matrix M, we may reduce the dimensionality of the problem by choosing the solution (A, B) given by $A = U_M A_0 U_X^\top$ and $B = V_X B_0 V_M^\top$, with A_0 and B_0 diagonal matrices such that

$$\sigma_k(A)\sigma_k(X)\sigma_k(B) = \sigma_k(M), \quad \text{for all } k \le r_X \wedge r_A \wedge r_B.$$

Thus, even under diagonal forms we can only identify the product of respective singular values of A and B. We can only hope to identify matrices A and B under very restrictive conditions where $X^\top X$ has full rank and either the matrix A or the matrix B is assumed to have known singular values, *e.g.* like a projector with singular values 1 or 0. Few other setups are known to be identifiable in the literature of factorisation of matrices, *e.g.* non-negative matrix factorisation (NMF), see [54], NMF for topic models [84], [25], [86] or covariance matrix factorization [57].

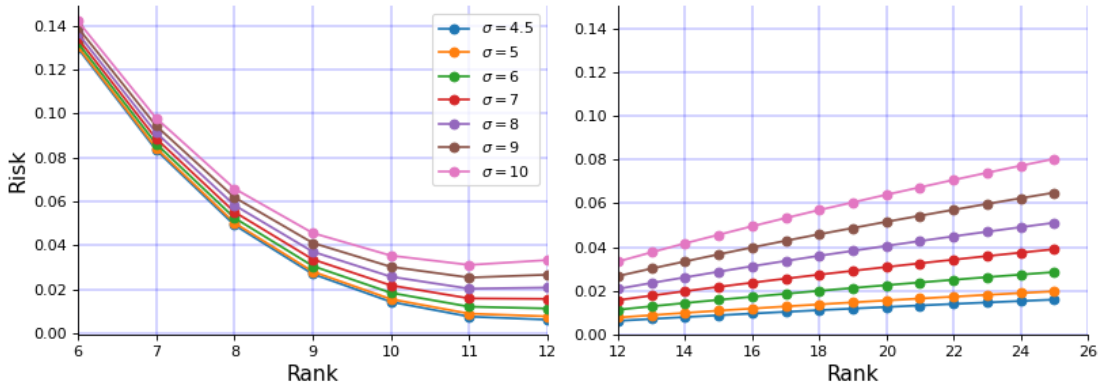


FIGURE 3.1 – Evolution of the risk $\frac{\|\hat{A}_r X \hat{B}_r - A^* X B^*\|_F^2}{\|A^* X B^*\|_F^2}$ in function of r for different values of σ

3.4 Numerical Results

Let us set the dimensions of the observed matrix Y to be $n = 100$ and $p = 300$, the dimensions of the design matrix X to be $m = 50$ and $q = 60$. We randomly generate three matrices : A^*, B^*, and X, with independent random gaussian entries with mean 0 and variance 1. These matrices are then projected onto the best low-rank matrix approximation, with the matrix A^* having a rank $r_A^* = 16$, the matrix B^* having a rank $r_B^* = 12$, and the matrix X having a rank $r_X = 25$. The signal matrix is defined as A^*XB^* and shows a rank of 12 in all experiments. We also define various settings for the variance σ^2 of the Gaussian noise E so that the signal-to-noise ratio $SNR := \sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ varies approximately in the range $[0.5, 2]$.

Figure 3.1 illustrates the prediction performances of the predictor $\hat{A}_r X \hat{B}_r$, defined in (3.7), for different values of r. For $\sigma < 8$ giving the SNR approximately above the value 1, the prediction risk decreases when the rank increases while remaining bounded from above by 12 and then increases with the rank when the rank is above 12. For $\sigma \geq 8$ giving the SNR below the value 1, the prediction risk decreases when the rank increases while remaining bounded from above by 11 and then increases with the rank when the rank is above 11. It highlights that the best predictor is achieved when $r = r^* = 12$ for small noise variance levels (*i.e.* $\sigma < 8$) and when $r = 11$ for strong noise variance levels (*i.e.* $\sigma \geq 8$). This shows that there is a strong overfitting phenomenon in the case of strong noise and that it is therefore better to slightly underestimate the rank in these situations.

Figure 3.2 represents the predicted $\hat{r}$, defined in (3.8), for various values of λ. Independently of the noise variance level, for small values of λ the estimated $\hat{r}$ is maximal and there is $\hat{r} = r_X = 25$. This illustrates the previously exposed overfitting phenomenon, that is the higher the rank r, the lower the error $\|Y - \hat{A}_r X \hat{B}_r\|_F^2$. As λ increases the penalty on the rank r becomes more important in the minimization procedure and $\hat{r}$ decreases. However, for moderate values of λ (*i.e.* approximately $\log(\lambda) \leq 5$) the smaller the noise variance level σ, the faster $\hat{r}$ decreases. Ultimately, for large values of λ (*i.e.* approximately $\log(\lambda) > 5$) the rate of decay of $\hat{r}$ as a function of λ no longer depends on σ.

The numerical value of λ is an important issue. We exhibit explicit (fast to calculate) procedures for the choice of this tuning parameter. In the case of *known noise variance*, the rule of thumb suggested

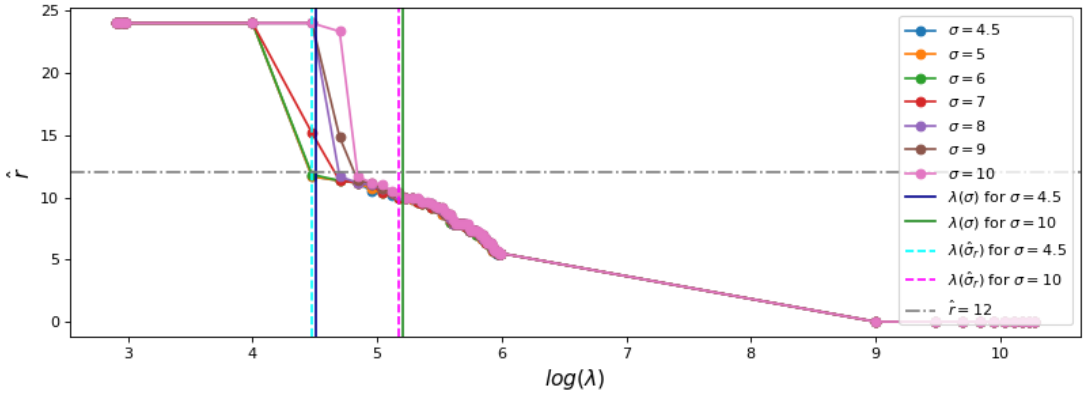


FIGURE 3.2 – Evolution of the estimated $\hat{r}$ as a function of $\log(\lambda)$ for different values of σ

by [27] indicates to choose

$$\lambda(\sigma) = 2C(n+p)\sigma^2(1+t)^2$$

in Theorem 3.2.3 with $t = 0$, and $C = 2$. The two solid vertical lines represent $\lambda(4.5)$ (blue) and $\lambda(10)$ (green). With these choices of the tuning parameter we get successful estimators of the underlying rank of the signal $\hat{r} \approx 12 = r^*$. We underline that in the small noise regime the rank is slightly overestimated and in the strong noise regime it is slightly underestimated. This behaviour perfectly matches the results drawn from Figure 3.1 showing that overestimating the rank in small noise regime does not impact the performances and slightly underestimating it in strong noise regime improves the performances.

However, in real world applications the noise has *unknown variance*. This raises the question of how to choose a data-driven λ in this case, without deteriorating the prediction. This situation is more challenging as it first requires an estimator of σ^2 before using the previously exposed rule of thumb. We choose the initial value of r equal to $r_X \wedge n \wedge p$ and propose the r-dependent estimator $\widehat{\sigma}_r^2 := \dfrac{\|Y - \hat{A}_r X \hat{B}_r\|_F^2}{np - (m \wedge q) r_X}$. It allows to compute the previously defined $\lambda(\widehat{\sigma}_r)$ and using this data-driven tuning parameter we produce the rank estimator $\bar{r}$. This procedure takes r as an argument and returns $\lambda(\widehat{\sigma}_r)$ and $\bar{r}$. However, when r is substantially larger than r^*, $\hat{A}_r X \hat{B}_r$ is overfitting Y and performing this procedure once will not lead to a satisfying output $\bar{r}$. Hence we iterate while $\bar{r} < r$. We note $\lambda(\widehat{\sigma}_{\bar{r}})$ and $\bar{r}$ the final outputs of the procedure. The two dashed vertical lines represent $\lambda(\widehat{\sigma}_{\bar{r}})$ when $\sigma = 4.5$ (cyan) and $\sigma = 10$ (magenta). The proposed procedure exhibits great numerical properties.

Finally, numerical simulations generated in the same context, with different values for the true underlying ranks, show similar excellent prediction bounds, combined with correct rank selection. Together with the current case where $\min(r_A^*, r_X, r_B^*) = r_B^*$, we have explored successfully the cases $\min(r_A^*, r_X, r_B^*) = r_A^*$, $\min(r_A^*, r_X, r_B^*) = r_X$ and $\min(r_A^*, r_X, r_B^*) = r_A^* = r_X = r_B^*$.

3.5 Proofs

Basic facts For any matrix $M \in \mathbb{R}^{n\times m}$, $\|M\|_*^2 \leq r_M \|M\|_F^2$. In addition, for any matrices M_1 and M_2 in $\mathbb{R}^{n\times m}$, the following inequalities hold $\langle M_1, M_2\rangle_F \leq \|M_1\|_* \|M_2\|_{op}$ and $\|M_1 + M_2\|_F \leq \|M_1\|_F + \|M_2\|_F$. Furthermore, if we set $a = \operatorname{rank} M_1 \wedge \operatorname{rank} M_2$ then $\langle M_1, M_2\rangle_F \leq \|M_1\|_{(2,a)} \|M_2\|_{(2,a)}$.

Lemma 3.5.1 *Let E be a $n \times p$ random matrix whose entries are independent and having Gaussian distribution $\mathcal{N}(0, \sigma^2)$. If U and V belong to $\mathcal{O}_n$ and $\mathcal{O}_p$ respectively, then $E_0 := U^\top EV$ has independent entries with Gaussian distribution $\mathcal{N}(0, \sigma^2)$.*

Proof of Lemma 3.5.1. Note that we can vectorize the matrix E_0 and get that

$$vec(E_0) = (V^\top \otimes U^\top) \cdot vec(E),$$

where $vec(E)$ is a Gaussian vector of dimension np, centered, with variance $\sigma^2 I_{np}$. Moreover, the tensor product $V^\top \otimes U^\top$ belongs to $\mathcal{O}_{np}$, thus $vec(E_0)$ is still a Gaussian vector with distribution $\mathcal{N}_{np}(0, \sigma^2 I_{np})$. ■

Recall that, for an arbitrary matrix M, we denote $U_M \Sigma_M V_M^\top$ its SVD.

Lemma 3.5.2 *If M^* is a $n \times p$ matrix of rank r^*, than for any $r \leq n \wedge p$, we have*

$$\inf_{M:\text{rank } M \leq r} \|M - M^*\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(M^*)^2 \cdot \mathbf{1}_{r<r^*},$$

and the infimum is attained by the projection $[M^*]_r$ of M^* on the space of $n \times p$ matrices with rank r given by the matrix

$$[M^*]_r = U_{M^*} \cdot Diag_{n,p}(\sigma_1(M^*), ..., \sigma_{r \wedge r^*}(M^*)) \cdot V_{M^*}^\top.$$

3.5.1 Proof of Theorem 3.2.1

Let $r \in [n \wedge p \wedge r_X]$ and $(\hat{A}_{0r}, \hat{B}_{0r})$ defined in (3.6). Let us denote here $M_0^* = A_0^* \Sigma_X B_0^*$ and $\hat{M}_0 = \hat{A}_{0r} \Sigma_X \hat{B}_{0r}$. By construction, $\hat{M}_0$ is the projection $[\Sigma_Y]_r$ of Σ_Y onto the set of matrices with rank less than or equal to r, in the sense of Lemma 3.5.2. Therefore,

$$\|\Sigma_Y - \hat{M}_0\|_F^2 \leq \|\Sigma_Y - [M_0^*]_r\|_F^2$$

We recall that in our model $\Sigma_Y = M_0^* + E_0$ which leads to

$$\|M_0^* - \hat{M}_0 + E_0\|_F^2 \leq \|M_0^* - [M_0^*]_r + E_0\|_F^2.$$

We expand the squares and arrange terms to get

$$\|M_0^* - \hat{M}_0\|_F^2 \leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\langle \hat{M}_0 - [M_0^*]_r, E_0 \rangle_F.$$

Now, since $\text{rank}(\hat{M}_0) = r$ and $\text{rank}([M_0^*]_r) \leq r$, we get that $\text{rank}(\hat{M}_0 - [M_0^*]_r) \leq 2r$. This inequality gives

$$\begin{aligned}
\|M_0^* - \hat{M}_0\|_F^2 &\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \|\hat{M}_0 - [M_0^*]_r\|_{(2,2r)} \\
&\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \|\hat{M}_0 - [M_0^*]_r\|_F \\
&\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \left(\|\hat{M}_0 - M_0^*\|_F + \|M_0^* - [M_0^*]_r\|_F \right).
\end{aligned}$$

We apply the inequality $2xy \leq \alpha x^2 + \alpha^{-1} y^2$ with $x, y \geq 0$ and $\alpha > 0$. We obtain, for real numbers $\alpha > 1$ and $\beta > 0$,

$$(1 - \alpha^{-1}) \cdot \|M_0^* - \hat{M}_0\|_F^2 \leq (1 + \beta^{-1}) \cdot \|M_0^* - [M_0^*]_r\|_F^2 + (\alpha + \beta) \cdot \|E_0\|_{(2,2r)}^2.$$

Let us use that $\|E_0\|^2_{(2,2r)} \leq 2r \cdot \|E_0\|^2_{op}$ and Lemma 3.5.2 to further get

$$\|M_0^* - \hat{M}_0\|_F^2 \leq \frac{1+\beta^{-1}}{1-\alpha^{-1}} \cdot \inf_{M:\operatorname{rank} M \leq r} \|M_0^* - M\|_F^2 + \frac{\alpha+\beta}{1-\alpha^{-1}} \cdot 2r\|E_0\|_{op}^2. \tag{3.12}$$

Noticing that for any matrices A_0, B_0 having rank less than or equal to r, $\operatorname{rank}(A_0\Sigma_X B_0) \leq r_{A_0} \wedge r_X \wedge r_{B_0} \leq r$, we deduce that

$$\inf_{M:\operatorname{rank} M \leq r} \|M_0^* - M\|_F^2 \leq \inf_{\substack{A_0,B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \leq r}} \|M_0^* - A_0\Sigma_X B_0\|_F^2.$$

Indeed, the second inf is taken over a possibly smaller family of matrices. We actually show that equality holds in the previous display. Indeed, by Lemma 3.5.2 we have that $\inf_{M:\operatorname{rank} M\leq r} \|M_0^* - M\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(M_0^*)^2 \cdot \mathbf{1}_{r<r^*}$, where $r^* = \operatorname{rank}(M_0^*)$. Recall that $M_0^* = A_0^*\Sigma_X B_0^*$ is a product of diagonal matrices, giving that $r^* = \min(r_X, r_{A_0^*}, r_{B_0^*})$ and $\sigma_k(M_0^*) = \sigma_k(A_0^*)\sigma_k(X)\sigma_k(B_0^*) \cdot \mathbf{1}_{k\leq r^*}$. Thus, the particular choice

$$A_{0r} := Diag_{n,m}(\sigma_1(A_0^*), \ldots, \sigma_{r\wedge r_{A_0^*}}(A_0^*)) \text{ and } B_{0r} := Diag_{q,p}(\sigma_1(B_0^*), \ldots, \sigma_{r\wedge r_{B_0^*}}(B_0^*))$$

solves exactly the problem giving $M_0^* = A_{0r}\Sigma_X B_{0r}$. Finally,

$$\inf_{M:\operatorname{rank} M \leq r} \|M_0^* - M\|_F^2 = \inf_{\substack{A_0,B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \leq r}} \|M_0^* - A_0\Sigma_X B_0\|_F^2. \tag{3.13}$$

Plugging this into (3.12) and considering the particular choice $\alpha = 3/2$ and $\beta = 1/2$ give the theorem :

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X\hat{B}_{0r}\|_F^2 \leq 9 \inf_{\substack{A_0,B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \leq r}} \left(\|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2\right) + 12r\|E_0\|_{op}^2.$$

The last step is the high-probability bound on $\|E_0\|_{op}$. Recall that $E_0 = U_Y^\top E V_Y$ with U_Y in $\mathcal{O}_n$ and V_Y in $\mathcal{O}_p$ and therefore E_0 and E have the same singular values. Therefore $\|E\|_{op} = \|E_0\|_{op}$. The noise matrix E has independent, centered, $\sigma-$sub-Gaussian entries and its spectral norm verifies (see [130]) for some absolute constant $C > 0$

$$\mathbb{P}\left(\|E\|_{op}^2 \leq 2C\sigma^2 \cdot (1+t)^2(n+p)\right) \geq 1 - 2e^{-t^2(\sqrt{n}+\sqrt{p})^2}, \quad \text{for any } t > 0. \tag{3.14}$$

Moreover, $\mathbb{E}\left[\|E\|_{op}\right] \leq \sqrt{C}\sigma(\sqrt{n}+\sqrt{p})$.

3.5.2 Proof of Corollary 3.2.2

Recall the notation $M_0^* = A_0^*\Sigma_X B_0^*$ and $\hat{M}_0 = \hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ with $\hat{A}_{0r}$ and $\hat{B}_{0r}$ given by (3.6) and let us denote $M^* = A^*XB^*$ and $\hat{M} = \hat{A}_r X \hat{B}_r$ with $\hat{A}_r$ and $\hat{B}_r$ given by (3.7). Notice that the Frobenius norm and the rank are invariant under left or right multiplication by orthogonal matrices. Therefore, we follow the lines of the proof of Theorem 3.2.1 and see that $\|Y - \hat{M}\|_F^2 = \|\Sigma_Y - \hat{M}_0\|_F^2$ and $\operatorname{rank} M^* = \operatorname{rank} M_0^* = r^*$. Also, $\hat{M}$ is the projection $[Y]_r$ of Y on the space of matrices with rank less than or equal to r. Finally, the equality (3.13) can be pushed forward

$$\inf_{M:\operatorname{rank} M \leq r} \|M_0^* - M\|_F^2 = \inf_{\substack{A_0,B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \leq r}} \|M_0^* - A_0\Sigma_X B_0\|_F^2 = \inf_{\substack{A,B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \|M^* - AXB\|_F^2.$$

Indeed, we have one-to-one transformations of A_0, B_0 into A, B, respectively, and equality of the Frobenius norms. This finishes the proof.

3.5.3 Proof of Theorem 3.2.3

By definition of $\hat{r} = \hat{r}(\lambda)$, we have that, for all $r \in [n \wedge p \wedge r_X]$,

$$\|Y - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 + \lambda \hat{r} \leq \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r.$$

Since $\hat{A}_r X \hat{B}_r$ is the projection $[Y]_r$ of Y on the space of matrices M with $\operatorname{rank} M \leq r$, we get that for all matrices A and B such that $\operatorname{rank} A \wedge \operatorname{rank} B \leq r$

$$\|Y - \hat{A}_r X \hat{B}_r\|_F^2 \leq \|Y - AXB\|_F^2.$$

Indeed, $\operatorname{rank}(AXB) \leq r$ and Pythagora's theorem gives the former inequality. We deduce that

$$\|Y - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 + \lambda \hat{r} \leq \|Y - AXB\|_F^2 + \lambda r.$$

Next, replace $Y = A^* X B^* + E$, expand the squares and rearrange terms to get

$$\begin{aligned}\|A^* X B^* - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 \leq \|A^* X B^* - AXB\|_F^2 + \lambda(r - \hat{r}) \\ + 2\langle E, \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}} - AXB\rangle.\end{aligned}$$

Let us denote by $\hat{M}(\hat{r}) = \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}$, $M(r) = AXB$ and see that $\operatorname{rank}(\hat{M}(\hat{r}) - M(r)) \leq \hat{r} + r$. We have

$$\begin{aligned}\langle E, \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}} - AXB\rangle &\leq \|E\|_{op} \cdot \|\hat{M}(\hat{r}) - M(r)\|_* \\ &\leq \|E\|_{op} \cdot \sqrt{\hat{r} + r} \|\hat{M}(\hat{r}) - M(r)\|_F \\ &\leq \|E\|_{op} \cdot \sqrt{\hat{r} + r} (\|M^* - \hat{M}(\hat{r})\|_F + \|M^* - M(r)\|_F).\end{aligned}$$

Then, using twice the inequality $2xy \leq \alpha x^2 + \alpha^{-1} y^2$ with $x, y \geq 0$ and $\alpha > 0$, we obtain for arbitrary real numbers $\alpha > 1$, $\beta > 0$:

$$\begin{aligned}(1 - \alpha^{-1})\|M^* - \hat{M}(\hat{r})\|_F^2 \leq (1 + \beta^{-1})\|M^* - M(r)\|_F^2 \\ + (\alpha + \beta)\|E\|_{op}^2 (r + \hat{r}) + \lambda(r - \hat{r}).\end{aligned}$$

Consequently, if $(\alpha + \beta)\|E\|_{op}^2 \leq \lambda$:

$$(1 - \alpha^{-1})\|M^* - \hat{M}(\hat{r})\|_F^2 \leq (1 + \beta^{-1})\|M^* - M(r)\|_F^2 + 2\lambda r,$$

for all r in $[n \wedge p \wedge r_X]$ and all $M(r) = AXB$ with $\operatorname{rank} A \wedge \operatorname{rank} B \leq r$. We get the result by replacing again $\alpha = 3/2$ and $\beta = 1/2$. Then we use that

$$\min_{\substack{A,B \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \|A^* X B^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^* X B^*)^2$$

and the high-probability bounds in (3.14).

3.5.4 Proofs of results in Section 3.2.3

Proof of Proposition 3.2.4. For any r in $[n \wedge p \wedge r_X]$, we have that $\hat{A}_r X \hat{B}_r = [Y]_r$ is the projection of Y on the space of matrices having rank smaller than or equal to r. Now, write

$$\begin{aligned} F(r) :&= \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r \\ &= \sum_{k=r+1}^{r_Y} \sigma_k(Y)^2 \cdot \mathbf{1}_{r<r_Y} + \lambda r \\ &= \sum_{k=r+1}^{r_Y} (\sigma_k(Y)^2 - \lambda) \cdot \mathbf{1}_{r<r_Y} + \lambda r_Y. \end{aligned}$$

It is easy to see that F as a function of r has a unique minimum at $r_Y(\lambda)$ if $\lambda > \sigma_{r_Y}(Y)^2$, but is minimal and constant for $r = r_Y, \ldots, (n \wedge p \wedge r_X)$ whenever $\lambda \leq \sigma_{r_Y}(Y)^2$. ■

Proof of Proposition 3.2.5. By definition of $\hat{r}$, we have $k > \hat{r}$ if and only if $\lambda > \sigma_k(Y)^2$ and $k < \hat{r}$ if and only if $\lambda \leq \sigma_{k+1}(Y)^2$. In our model $Y = A^* X B^* + E$, the Weyl inequality gives $|\sigma_k(A^* X B^*) - \sigma_k(Y)| \leq \sigma_1(E)$ for all k. The events on $\hat{r}$ can be written in terms of $\sigma_1(E) = \|E\|_{op}$ as follows. We have

$$\begin{aligned} \{k > \hat{r}\} \quad &\text{implies} \quad \lambda > (\sigma_k(A^* X B^*) - \sigma_1(E))^2, \\ \{k < \hat{r}\} \quad &\text{implies} \quad \lambda \leq (\sigma_{k+1}(A^* X B^*) + \sigma_1(E))^2. \end{aligned}$$

Thus $\{\hat{r} \neq k\}$ implies either $\sigma_1(E) > \sigma_k(A^* X B^*) - \sqrt{\lambda}$ or $\sigma_1(E) \geq \sqrt{\lambda} - \sigma_{k+1}(A^* X B^*)$. Let us take $k = r^*(\lambda)$. Then the assumption that $\sigma_{r^*(\lambda)}(A^* X B^*) > (1+c)\sqrt{\lambda}$ gives that $\sigma_1(E) > c\sqrt{\lambda}$ and the assumption that $\sigma_{r^*(\lambda)+1}(A^* X B^*) < (1-c)\sqrt{\lambda}$ gives also that $\sigma_1(E) > c\sqrt{\lambda}$. Thus,

$$\mathbb{P}\left(\hat{r} \neq r^*(\lambda)\right) \leq \mathbb{P}\left(\sigma_1(E) > c\sqrt{\lambda}\right).$$

The proof is finished using the inequality (3.14). ■

3.5.5 Proof of Theorem 3.2.7

The optimization problem (3.9) can be written, after replacing $\widehat{\sigma}_r^2$, as follows

$$\bar{r} \in \arg\min_{r \in [r_{max}]} \|Y - \hat{A}_r X \hat{B}_r\|_F^2 \left(1 + \frac{\lambda r}{np}\right).$$

We denote by $\bar{M} = \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}$, $\hat{M}_r = \hat{A}_r X \hat{B}_r$ and $M^* = A^* X B^*$. With this notation it follows that, for $r \leq r_{max}$,

$$\|Y - \bar{M}\|_F^2 \left(1 + \frac{\lambda \bar{r}}{np}\right) \leq \|Y - \hat{M}_r\|_F^2 \left(1 + \frac{\lambda r}{np}\right).$$

Developing the squares and using the equality $Y = M^* + E$, we get

$$\|M^* - \bar{M}\|_F^2 \leq \|M^* - \hat{M}_r\|_F^2 + 2\langle E, \bar{M} - \hat{M}_r\rangle_F + \frac{\lambda r}{np}\|Y - \hat{M}_r\|_F^2 - \frac{\lambda \bar{r}}{np}\|Y - \bar{M}\|_F^2.$$

We now use the upper bound $\langle E, \bar{M} - \hat{M}_r\rangle_F \leq \|E\|_{op}\|\bar{M} - \hat{M}_r\|_*$ and the definition of $\bar{M}$ and $\hat{M}_r$ to derive

$$\|M^* - \bar{M}\|_F^2 \leq \|M^* - \hat{M}_r\|_F^2 + 2\|E\|_{op}\|\bar{M} - \hat{M}_r\|_* + \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2.$$

Let us note that we use $\sigma_k(Y) = 0$ in case $k > r_Y$. We recall that $\|\bar{M} - \hat{M}_r\|_* \leq \sqrt{r+\bar{r}} \cdot \|\bar{M} - \hat{M}_r\|_F$ and further obtain

$$\|M^* - \bar{M}\|_F^2 \leq \|M^* - \hat{M}_r\|_F^2 + 2\|E\|_{op}\sqrt{r+\bar{r}}\left(\|M^* - \bar{M}\|_F + \|M^* - \hat{M}_r\|_F\right) + \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2.$$

Using twice the inequality $2ab \leq \alpha a^2 + \alpha^{-1}b^2$ for $a, b > 0$, with $\alpha > 1$ first and with $\beta > 0$ second, we get

$$(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \leq (1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + (\alpha+\beta)\|E\|_{op}^2(r+\bar{r}) + \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2. \tag{3.15}$$

We now distinguish the two cases : $r \leq \bar{r}$ and $r > \bar{r}$. In the first case, namely $r \leq \bar{r}$, we bound from above as follows :

$$\begin{aligned}
\frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2 &= \frac{\lambda}{np}\left(r\sum_{k=r+1}^{\bar{r}}\sigma_k(Y)^2 + (r-\bar{r})\sum_{k>\bar{r}}\sigma_k(Y)^2\right) \\
&\leq \frac{\lambda}{np}r(\bar{r}-r)\sigma_{r+1}(Y)^2 \\
&\leq \frac{2\lambda r}{np}(\bar{r}-r)(\sigma_{r+1}(M^*)^2 + \|E\|_{op}^2) \\
&\leq \frac{2\lambda r}{np}r_{max}\sigma_{r+1}(M^*)^2 + \frac{2\lambda r_{max}}{np}(\bar{r}-r)\|E\|_{op}^2,
\end{aligned}$$

where we used Weyl inequality $\sigma_{r+1}(Y) \leq \sigma_{r+1}(M^*) + \|E\|_{op}$ leading to $\sigma_{r+1}(Y)^2 \leq 2\|E\|_{op}^2 + 2\sigma_{r+1}(M^*)^2$. We plug this into (3.15) to get

$$\begin{aligned}
(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \leq (1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 &+ \frac{2\lambda r_{max}}{np}r\sigma_{r+1}(M^*)^2 \\
&+ r\|E\|_{op}^2(\alpha+\beta-\frac{2\lambda r_{max}}{np}) \\
&+ \bar{r}\|E\|_{op}^2(\alpha+\beta+\frac{2\lambda r_{max}}{np}),
\end{aligned}$$

for all $r \leq \bar{r}$ belonging to $[r_{max}]$. Thus, for λ such that $\frac{2\lambda\cdot(r_{max}\vee r_Y)}{np} = (1+\varepsilon)(\alpha+\beta)$ for some $\varepsilon > 0$ we get

$$(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \leq \min_{r\in[\bar{r}]}\left\{(1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + (1+\varepsilon)(\alpha+\beta)r\sigma_{r+1}(M^*)^2\right\} + (2+\varepsilon)(\alpha+\beta)r_{max}\|E\|_{op}^2.$$

We now focus on the second case, namely $r > \bar{r}$. We observe that in this case,

$$\begin{aligned}\frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2 &= \frac{\lambda}{np}\left((r-\bar{r})\sum_{k>r}\sigma_k(Y)^2 - \bar{r}\sum_{k=\bar{r}+1}^{r}\sigma_k(Y)^2\right)\\ &\leq \frac{\lambda(r-\bar{r})}{np}(r_Y - r)\sigma_{r+1}(Y)^2\\ &\leq \frac{2\lambda r}{np} r_Y \cdot \sigma_{r+1}(M^*)^2 + \frac{2\lambda(r-\bar{r})}{np}\cdot(r_Y \vee r_{max})\|E\|_{op}^2,\end{aligned}$$

by a similar reasoning in the previous case. We plug this into (3.15) to get

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \leq (1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 &+ \frac{2\lambda\cdot r_{max}\vee r_Y}{np} r\sigma_{r+1}(M^*)^2\\ &+ r\|E\|_{op}^2(\alpha+\beta+\frac{2\lambda\cdot r_{max}\vee r_Y}{np})\\ &+ \bar{r}\|E\|_{op}^2(\alpha+\beta-\frac{2\lambda\cdot r_{max}\vee r_Y}{np}).\end{aligned}$$

With the same choice of λ such that $\frac{2\lambda\cdot r_{max}\vee r_Y}{np} = (1+\varepsilon)(\alpha+\beta)$ for some $\varepsilon > 0$ we get also in this case that

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \leq \min_{\bar{r}<r\leq r_{max}}&\left\{(1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + (1+\varepsilon)(\alpha+\beta)r\sigma_{r+1}(M^*)^2\right\}\\ &+ (2+\varepsilon)(\alpha+\beta)r_{max}\|E\|_{op}^2.\end{aligned}$$

Taking $\alpha = 3/2$ and $\beta = 1/2$ and combining both cases leads to the following result

$$\|M^* - \bar{M}\|_F^2 \leq \min_{r\in[r_{max}]}\left\{9\|M^* - \hat{M}_r\|_F^2 + 6(1+\epsilon)\cdot r\sigma_{r+1}(M^*)^2\right\} + 6(2+\epsilon)\cdot r_{max}\|E\|_{op}^2,$$

where we choose λ such that $\lambda\cdot r_{max}\vee r_Y = (1+\varepsilon)np$ for some $\varepsilon > 0$. We conclude by using the inequality (3.14).

3.5.6 Proof of Theorem 3.3.1

We proceed by solving the problem in two steps for solving the optimization problem (3.10) which can be equivalently written as

$$\min_{\substack{A,B\\ M=AXB}}\ \min_{M}\|Y-M\|_F^2 + 2\lambda\cdot\|M\|_*,$$

for $\lambda > 0$. The solution to the problem in M is explicit and it is known to be obtained from Y by soft-thresholding of its eigenvalues : $\bar{M} = U_Y Diag_{n,p}((\sigma_k(Y)-\lambda)_+)V_Y^\top$, where we used the SVD of Y : $U_Y\Sigma_Y V_Y^\top$. Next, we project $\bar{M}$ on the space of matrices AXB for A and B in Frobenius norm. It is easy to check that our choice of $\bar{A}, \bar{B}$ are exact solutions, that is $\bar{M} = \bar{A}X\bar{B}$.

Similarly to the proof of Theorem 3.2.3, by applying the definition of $\bar{M}$, expanding the squares and rearranging terms we get for all M :

$$\begin{aligned}\|\bar{M} - M^*\|_F^2 &\leq \|M^* - M\|_F^2 + 2\langle E, \bar{M} - M\rangle + 2\lambda(\|M\|_* - \|\bar{M}\|_*)\\ &\leq \|M^* - M\|_F^2 + 2\sqrt{\lambda}(\|\bar{M} - M\|_* + \|M\|_* - \|\bar{M}\|_*),\end{aligned}$$

under the event that $\|E\|_{op}^2 \leq \lambda$. We use the decomposability of the nuclear norm of matrices as in [32], to find $\bar{M}_1$ and $\bar{M}_2$ such that $\bar{M} = \bar{M}_1 + \bar{M}_2$, $\|\bar{M}\|_* = \|\bar{M}_1\|_* + \|\bar{M}_2\|_*$ and $\|\bar{M} - M\|_* = \|\bar{M}_1 - M\|_* + \|\bar{M}_2\|_*$. Moreover, $\operatorname{rank}(\bar{M}_1) \leq 2\operatorname{rank}(M)$. This implies

$$\begin{aligned}\|\bar{M} - M^*\|_F^2 &\leq \|M^* - M\|_F^2 + 4\sqrt{\lambda}\|\bar{M}_1 - M\|_* \\ &\leq \|M^* - M\|_F^2 + 4\sqrt{\lambda}\sqrt{3\operatorname{rank}(M)} \cdot \|\bar{M}_1 - M\|_F \\ &\leq \|M^* - M\|_F^2 + 4\sqrt{\lambda}\sqrt{3\operatorname{rank}(M)} \cdot (\|\bar{M} - M^*\|_F + \|M - M^*\|_F).\end{aligned}$$

We obtain for arbitrary real numbers $\alpha > 1$ and $\beta > 0$, for all M,

$$(1 - \alpha^{-1})\|\bar{M} - M^*\|_F^2 \leq (1 + \beta^{-1})\|M^* - M\|_F^2 + 4(\alpha + \beta)\lambda \cdot 6\operatorname{rank}(M).$$

For the particular values $\alpha = 3/2$ and $\beta = 1/2$, we get

$$\begin{aligned}\|\bar{M} - M^*\|_F^2 &\leq \min_M \left\{9\|M^* - M\|_F^2 + 144\lambda \cdot \operatorname{rank}(M)\right\} \\ &\leq 9 \min_{r \in [n \wedge p \wedge r_X]} \left\{ \min_{M:\operatorname{rank} M = r} \|M^* - M\|_F^2 + 16\lambda \cdot r \right\}.\end{aligned}$$

Recall that $\min_{M:\operatorname{rank} M = r} \|M^* - M\|_F^2 = \sum_{K=r+1}^{r^*} \sigma_K(M^*)^2 \cdot \mathbf{1}_{r<r^*}$ to get the final result.

3.6 Auxiliary results

Algorithm 1 Data-driven procedure for selecting $\bar{r}$ and λ

Input : data X, Y

Require : $np \geq (m \wedge q) r_X > 0$

Define : $\widehat{\sigma}_r^2 := \dfrac{\|Y - \hat{A}_r X \hat{B}_r\|_F^2}{np - (m \wedge q) r_X}$

Define : $\lambda(\sigma) := 4(n+p)\sigma^2$

Define : $\hat{r}_\lambda := \operatorname{argmin}_{r \in [n \wedge p \wedge r_X]} \left(\|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda \cdot r \right)$

Initialize : $r \leftarrow r_X \wedge n \wedge p$, $\bar{r} \leftarrow \hat{r}_{\lambda(\widehat{\sigma}_r^2)}$

  while $\bar{r} < r$ **do**

    $r \leftarrow \bar{r}$

    $\bar{r} \leftarrow \hat{r}_{\lambda(\widehat{\sigma}_r^2)}$

  end while

  Output : $\bar{r}, \lambda(\widehat{\sigma}_{\bar{r}}^2)$

Chapitre 4

Dynamic Expected Topic Models

4.1 Introduction

Topic modeling is a widely used statistical technique that has gained significant attention in the field of natural language processing (NLP) and text mining. It provides a valuable framework for uncovering latent thematic structures within large collections of textual data. The fundamental goal of topic modeling is to automatically discover underlying topics or themes that are present in a corpus of documents, without any prior knowledge or manual annotation. Topic models have found applications in various domains, including biology, collaborative filtering, population genetics, social networks and image analysis. These models provide researchers with a means to effectively organize, condense, and scrutinize textual data, facilitating a deeper understanding of the inherent semantic organization within documents. For instance, by employing topic modeling techniques, one can uncover thematic patterns in large corpora. This allows to discern topics present in a collection of documents and explore how they relate to each other, thereby extracting valuable insights about the underlying content and structure.

Notations : In addition to the notation introduced at the beginning of this manuscript, we add some specific notations for the following two chapters. For any matrix M of size $n \times m$ and rank r_M, we denote $M^\top$ its transpose and for all $i \in [n]$ and for all $j \in [m]$, $[M]_{ij}$ denotes the entry of M in the i^{th} row and j^{th} column, $[M]_{i.}$ its i^{th} row and $[M]_{.j}$ its j^{th} column. We denote $M = U_M \Sigma_M V_M^\top$ its singular value decomposition (SVD), where U_M belongs to $\mathbb{R}^{n \times r_M}$ and satisfies $U_M^\top U_M = I_{r_M}$, V_M belongs to to $\mathbb{R}^{m \times r_M}$ and satisfies $V_M^\top V_M = I_{r_M}$. Σ_M is a diagonal matrix containing the non null singular values $\sigma_1(M), \ldots, \sigma_{r_M}(M)$ of M listed in decreasing order and null entries elsewhere. We denote $\sigma_{\min}(M) := \sigma_{r_M}(M)$ the smallest singular value of M. For $k \leq \min(n, m)$ we define $U_M^{(k)} \Sigma_M^{(k)} \left(V_M^{(k)}\right)^\top$ the k-SVD of M where $U_M^{(k)}$ belongs to $\mathbb{R}^{n \times k}$ and satisfies $\left(U_M^{(k)}\right)^\top U_M^{(k)} = I_k$, $V_M^{(k)}$ belongs to to $\mathbb{R}^{m \times k}$ and satisfies $\left(V_M^{(k)}\right)^\top V_M^{(k)} = I_k$ and

$$\Sigma_M^{(k)} := \text{diag}(\sigma_i(M),\, 1 \leq i \leq k) \in \mathcal{D}_k(\mathbb{R}_+^*) \quad \text{if} \quad k \leq r_M,$$
$$\Sigma_M^{(k)} := \text{diag}(\sigma_1(M), \ldots, \sigma_{r_M}(M), 0, \ldots, 0) \in \mathcal{D}_k(\mathbb{R}_+^*) \quad \text{if} \quad k > r_M.$$

For any diagonalizable matrix $Q \in \mathbb{R}^{n \times n}$, we denote $\lambda_1(Q), \ldots, \lambda_n(Q)$ the eigenvalues of Q listed in decreasing order. We denote $\lambda_{\min}(Q)$ the smallest non zero eigenvalue of Q. For any set E and any integer p we denote $\mathcal{D}_p(E)$ the set of diagonal matrices of size p with entries in E. For any matrix M,

we denote M_+ the matrix obtained by setting all negative entries in M to 0. We denote by $\Phi_{row}(M)$ the matrix obtained by normalizing each row of M to have a unit $\mathbb{L}_1$-norm and, analogously, $\Phi_{col}(M)$ is the matrix obtained by normalizing each column of M to have a unit $\mathbb{L}_1$-norm. Random quantities are written in bold, except for the estimators which are marked with a hat.

4.2 Dynamic topic model framework

In this study, we assume that n textual documents are observed successively in time, and that the topics distribution given a document follows a stationary time series whereas the distribution of words given a topic remains the same. This is akin to the regular intervals at which daily newspapers publish. Rather than treating this collection of documents independently of their publication date, our objective is to develop a model capable of capturing the temporal evolution inherent in the successive corpora.

In our study, we make the assumption that the number of topics discussed remains constant over time. Additionally, we assert that the word-topic probability matrix A^*, remains static over time. This assumption is grounded in the interpretation of the columns of matrix A^* as the distribution of each topic across the vocabulary. It is asserting that the same words distribution is systematically used to discuss a given topic.

More precisely, the collection process unfolds in T steps, where at each time step $t \in [T]$, a fixed number n of documents is collected. In this context, the j^{th} document at step t comprises N_j^t words. As we exclusively focus on the frequencies of each word, for simplicity and without loss of generality, we presume uniformity in the word count, i.e., $N_j^t = N$ for all j and t. The overall number of documents collected throughout the entire procedure is nT. The model of interest becomes, for all t in $[T]$ and $j \in [n]$:

$$N\boldsymbol{Y}_j^t | \boldsymbol{W}_j^t \sim \text{Multinomial}_p(N, \boldsymbol{\Pi}_j^t), \tag{4.1}$$

where for all columns j, the vectors $(\boldsymbol{Y}_j^t)_t$ are assumed to be conditionally independent given $(\boldsymbol{W}_j^t)_t$. We also assume that for all time step t, the vectors $(\boldsymbol{W}_j^t)_j$ are independent. We still assume that the word-document probability matrix $\boldsymbol{\Pi}^t = \left(\pi_1^t, \ldots, \pi_n^t\right)$, can be factorized as follows :

$$\boldsymbol{\Pi}^t = A^* \boldsymbol{W}^t, \quad t \in [T]. \tag{4.2}$$

We remind that the topic-document probability matrix at the step t, namely $\boldsymbol{W}^t$ which belongs to $\mathbb{R}^{K \times n}$ is now a random matrix following a simplex-valued autoregressive model and that the anchor word assumption on the word-topic probability matrix A^*, which belongs to $\mathbb{R}^{p \times K}$, is still made :

Assumption 2 (Anchor word assumption) *For each topic $k \in [K]$, there exists at least one word j such that $[A^*]_{jk} > 0$ and $[A^*]_{jl} = 0$ for $l \in [K] \backslash \{k\}$.*

Let us denote the concatenated matrices by

$$\boldsymbol{W}^{1:T} = \left(\boldsymbol{W}^1, \ldots, \boldsymbol{W}^T\right) \text{ and by } \boldsymbol{\Pi}^{1:T} = \left(\boldsymbol{\Pi}^1, \ldots, \boldsymbol{\Pi}^T\right),$$

which belong respectively to $\mathbb{R}^{K \times (nT)}$ and to $\mathbb{R}^{p \times (nT)}$, respectively. The model (4.2) can be re-written as

$$\boldsymbol{\Pi}^{1:T} = A^* \boldsymbol{W}^{1:T}.$$

Let us consider an autoregressive model of order 1 for the matrices $(\boldsymbol{W}^t)_{t\in[T]}$. However, at each time step t, it is crucial to emphasize that each column $\boldsymbol{W}_j^t$ is structured as a probability vector, meaning it consists of non-negative entries that sum up to one. Moreover, an insightful observation underlies our modeling approach : a topic that enjoys high popularity at time t is anticipated to sustain its prevalence at time $t+1$. Given these considerations, we define the autoregressive model with the following constraints for all $t \in [T-1]$:

$$\boldsymbol{W}^{t+1} = (1-c^*)\cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t \tag{4.3}$$

where $c^* \in (0,1)$, and each $\boldsymbol{\Delta}^t$ is a noise matrix of size $K \times n$ such that the columns are independently and identically drawn from a Dirichlet $\mathcal{D}(\theta^*)$ distribution having parameter $\theta^* \in \mathbb{R}_+^K$. The primary focus of this study is to estimate the parameters associated with this dynamic evolution and to establish non-asymptotic rates of convergence.

Model (4.3) encapsulates the notion that at time $t+1$, $\boldsymbol{W}_j^{t+1}$ serves as a barycenter between $\boldsymbol{W}_j^t$ and a noise vector $\boldsymbol{\Delta}_j^t$ drawn from $\mathcal{D}(\theta^*)$. As a consequence of this formulation, $\boldsymbol{W}_j^{t+1}$ is a probability vector for all $j \in [n]$.

We then assume that the value c^* is included in a closed subset of $(0,1)$. Therefore, c^* is mixing the contribution of the present value $\boldsymbol{W}^t$ in trade-off with that of the noise in order to get the future value $\boldsymbol{W}^{t+1}$ in (4.3). It is thus natural to exclude that c^* gets too close to either 1 (no influence of the current value, only noise) or 0 (no time evolution, static model).

Assumption 3 *There exist two real values $\underline{c}$ and $\overline{c}$ in $(0,1)$ such that the parameter c^* satisfies :*

$$\underline{c} \leq c^* \leq \overline{c}.$$

Recall that a $K-$dimensional vector distributed according to the Dirichlet distribution $\mathcal{D}(\theta^*)$ lives on the simplex of dimension K and has expected value and variance given by

$$\tilde{\theta}^* \text{ and } \Sigma := \Sigma(\theta^*) = \frac{1}{\alpha+1}\left(\text{diag}(\tilde{\theta}^*) - \tilde{\theta}^* \cdot (\tilde{\theta}^*)^\top\right), \tag{4.4}$$

respectively, where we denote by $\alpha := \|\theta^*\|_1 > 0$ and by $\tilde{\theta}^* := \theta^*/\alpha$ which belongs to the simplex $S_{(K-1)}$. We denote by $\text{diag}(\tilde{\theta}^*)$ the $K \times K$ diagonal matrix with values $\tilde{\theta}^*(k)$ on its diagonal.

The third assumption is focused on giving a lower bound to the variance of the Dirichlet distribution from which the noise matrices $\left(\boldsymbol{\Delta}^t\right)_{t\in[T]}$ are drawn. Let us note that for any parameter θ^*, the Trace of $\Sigma(\theta^*)$ can be expressed and bounded from above by one as follows :

$$\text{Tr}(\Sigma(\theta^*)) = \frac{1-\|\tilde{\theta}^*\|_2^2}{\alpha+1} \leq 1.$$

Assumption 4 *There exist real values $0 < \underline{\theta} < 1$ and $0 < m < 1$ such that the parameter θ^* satisfies :*

$$\min_{k\in[K]} \tilde{\theta}^*(k) \geq \underline{\theta} \quad \textit{and} \quad m \leq \textit{Tr}(\Sigma(\theta^*)) \leq 1.$$

This assumption prevents $\text{Tr}(\Sigma(\theta^*))$ to be too close to 0. Implicitly, this gives on the one hand that $\alpha = \|\theta^*\|_1$ cannot tend to infinity and stays bounded by some constant $A(m) < \infty$ and on the other hand that $\|\tilde{\theta}^*\|_2$ does not get too close to 1. The latter can happen only when $\tilde{\theta}^*$ gets close to a corner of the simplex, where the euclidean and the $\mathbb{L}_1$ norms are both equal to 1.

Finally, Assumption 5 states that we start our study when the stationary regime is already reached and thus avoid any transitional regime. We also assume that the initial vectors $\left(\boldsymbol{W}_j^1\right)_j$ are random with a continuous distribution and that their first and second moments are compatible with the stationary regime.

Assumption 5 (Stationary regime) *We assume that the initial vectors $\boldsymbol{W}_j^1$, $j = 1, \ldots, n$ are independent and identically distributed following the continuous stationary distribution.*

Hence for all $j \in [n]$, $\boldsymbol{W}_j^1$ is almost surely in the simplex S_{K-1} and

$$\mathbb{E}\left[\boldsymbol{W}_j^1\right] = \tilde{\theta}^* \text{ and } \mathbb{V}\left[\boldsymbol{W}_j^1\right] = \frac{c^*}{2 - c^*} \cdot \Sigma.$$

Combining equations (4.2) and (4.3) leads to the following dynamic expected topic model (DETM).

Definition 4.2.1 (Dynamic Expected Topic Model) *We refer to the Dynamic Expected Topic Model (DETM) described by the following equation :*

$$\boldsymbol{\Pi}_j^{t+1} = (1 - c^*)\boldsymbol{\Pi}_j^t + c^* A^* \cdot \boldsymbol{\Delta}_j^t, \tag{4.5}$$

where we observe $\boldsymbol{\Pi}_j^t$ for $t \in [T]$, $j \in [n]$, satisfying $\boldsymbol{\Pi}^t = A^* \boldsymbol{W}^t$ with $\boldsymbol{W}^t$ given by the AR(1) model (4.3)*. We assume that Assumptions 2, 3, 4, and 5 are met.*

Notice that (4.5) is a simplex-valued autoregressive model of order one and can be further developed as

$$\boldsymbol{\Pi}^t = (1 - c^*)^{t-1} A^* \boldsymbol{W}^1 + c^* \sum_{s=1}^{t-1} (1 - c^*)^{t-1-s} A^* \boldsymbol{\Delta}^s.$$

Our first objective will be to estimate the parameters c^*, $\tilde{\theta}^*$ and α in the DETM. However, the DETM is an oracle case where the word-document probability vectors $\boldsymbol{\Pi}_j^t$ are available. The real case where only the word-document frequency vectors $\boldsymbol{Y}_j^t$ are available will be considered in the next chapter.

The primary goal of this work is to grasp the temporal dynamics embedded in textual data. We introduce a model designed to accommodate the evolution and shifting of topics across discrete time periods. Indeed, themes within textual data often exhibit temporal variations. For instance, during election periods, news articles may emphasize different topics compared to periods of economic downturns. By integrating temporal information, our objective is to facilitate the discovery of how topics evolve, emerge, or diminish over time, thereby offering valuable insights into the dynamic nature of textual data. The question of modeling dynamic components in the topic model framework has been first treated by [28]. They introduced Dynamic Topic Model (DTM) as a solution to the limitations of Latent Dirichlet Allocation (LDA) when modeling topics across a series of documents. Numerous papers followed this initial work, mainly using variational approximate inference algorithms [137, 133, 138, 52]. However, these estimation procedures lack statistical guarantees.

In this chapter we assume that we have access to the word-document probability matrix $\boldsymbol{\Pi}^{1:T}$. This is equivalent to assuming that we observe the word-document frequency matrix $Y^{1:T} := \left(Y^1, \ldots, Y^T\right)$ where each document has an infinite number of words, *i.e.* $N = +\infty$ in (4.1). The randomness here is only due to the time series describing the distribution of topics in the document at time t. Hence, our attention is focused on the DETM. The goal of this chapter is to recover the data $\boldsymbol{W}^{1:T}$ following the AR(1) model (4.3), and then to estimate the underlying parameters of this model, by giving non-asymptotic high-probability bounds.

4.3 Recovery of the word-topic matrix

In this subsection, we follow the work by [84] and recall the procedure to recover the static deterministic word-topic matrix A^* given $\mathbf{\Pi}^{1:T}$ under their assumptions. Then, we project $\mathbf{\Pi}^{1:T}$ on the linear space spanned by the columns of A^* and retrieve $\boldsymbol{W}^{1:T}$. Finally, we show that under the AR(1) model (4.3), assumptions on $\boldsymbol{W}$ are valid with high probability.

Definition 4.3.1 *We define $H := diag(h_1, \dots, h_p) \in \mathcal{D}_p(\mathbb{R}_+^*)$, where for $i \in [p]$, $h_i := \|A_{i.}^*\|_1$ sums the frequencies of each word across all topics. Define the topic-topic overlapping matrix $\Sigma_A \in \mathbb{R}^{K\times K}$ as follows*

$$\Sigma_A := (A^*)^\top H^{-1} A^*.$$

The quantities $h_1, \dots, h_p$ reflect the variability in the frequency of occurrence of each word. The matrix Σ_A measures the affinity of topics using the same words. The authors in [84] require that the frequencies of the words considered in the vocabulary stay bounded from below by some positive constant. This condition aligns with the prevalent pre-processing practice of eliminating exceedingly low-frequency words or aggregating them into a pseudo-word. In addition, we underline that extreme heterogeneity remains allowed.

Assumption 6 (Minimal word frequency) *We assume that for some constant $c_1 \in (0, 1)$,*

$$\min_i h_i := h_{\min} \geq c_1 \frac{K}{p}.$$

Definition 4.3.2 *Define the topic-topic concurrence matrix $\Sigma_{\boldsymbol{W}}^{1:T} \in \mathbb{R}^{K\times K}$ as follows*

$$\Sigma_{\boldsymbol{W}}^{1:T} := \frac{1}{nT} \left(\boldsymbol{W}^{1:T}\right) \left(\boldsymbol{W}^{1:T}\right)^\top.$$

The matrix $\Sigma_{\boldsymbol{W}}^{1:T}$ captures the affinity of topics to be covered together in the same document.

The following assumption coupled with Assumption 2 ensure the identifiability of A^* and $\boldsymbol{W}^{1:T}$. We recall that by design, the topic model assumes that the matrix $\mathbf{\Pi}^{1:T} \in \mathbb{R}^{p\times nT}$ is of rank K and thus can be written as the product of maximal rank matrices $A^* \in \mathbb{R}^{p\times K}$ and $\boldsymbol{W}^{1:T} \in \mathbb{R}^{K\times nT}$. Hence A^* and $\boldsymbol{W}^{1:T}$ are of rank K when $K \leq p \wedge (nT)$ which implies that Σ_A and $\Sigma_{\boldsymbol{W}}^{1:T}$ are also of rank K. The following assumption allows to control the smallest eigenvalue of both matrices. We also consider $\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\mathbf{\Pi}^{1:T} 1_{nT}\right) \in \mathcal{D}_p(\mathbb{R}_+^*)$.

Assumption 7 *We assume θ^* is a vector with positive entries and that for some constants $c_2 > 0$ and $c_3 > 0$,*

$$\lambda_K\left(\Sigma_A\right) \geq c_2 \text{ and } \min_{k,l} [\Sigma_A]_{kl} \geq c_2, \quad \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq c_2, \quad \text{a.s.},$$
$$c_3^{-1} \geq \left|\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)) - \lambda_2(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right))\right| \geq c_3, \quad \text{a.s.}.$$

The matrix A^* is fixed and thus the assumptions on Σ_A are mild. The assumption on the smallest singular value of $\Sigma_{\boldsymbol{W}}^{1:T}$ can be relaxed as it holds true with high probability as shown in Theorem 4.3.3. Finally, we justify the last assumption using Perron-Frobenius theorem, see Lemma 5.6.9. Note that $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)$ is a $K \times K$ symmetric matrix with entries in $[0, 1]$ a.s. because both $\Sigma_{\boldsymbol{W}}^{1:T}$ and

$[A^*]^\top \boldsymbol{M}_*^{-1} A^*$ are. We need to prove that the entries of $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)$ are positive. Let us re-write the matrix $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)$ as $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top H^{-1} A^*\right) + \Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \left(\boldsymbol{M}_*^{-1} - H^{-1}\right) A^*\right)$. Proposition 5.2.8 gives that $\boldsymbol{M}_*^{-1} - H^{-1}$ is a diagonal matrix with almost surely non-negative entries. Moreover assumptions on Σ_A ensure that the entries of $[A^*]^\top H^{-1} A^*$ are bounded from below by c_2. Finally this proves that the $K \times K$ matrix $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)$ is a square matrix with positive entries almost surely. Conditionally on $\boldsymbol{W}^{1:T}$, Perron–Frobenius theorem gives that $\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)$ has a unique positive largest eigenvalue which is also its operator norm. We deduce that conditionally on $\boldsymbol{W}^{1:T}$, the following inequality holds almost surely

$$\left|\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)) - \lambda_2(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right))\right| > 0.$$

Moreover, Proposition 4.3.1 allows to get a milder assumption if one accepts c_3 to depend on K.

Proposition 4.3.1 *The entries of Σ_A are in $[0,1]$ and the entries of $\Sigma_{\boldsymbol{W}}^{1:T}$ are almost surely in $[0,1]$. In addition their spectral norm satisfies :*

$$\frac{1}{\sqrt{K}} \leq \lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}) \leq \sqrt{K} \quad \textit{a.s.}, \quad \frac{1}{\sqrt{K}} \leq \lambda_1\left(\Sigma_A\right) \leq \sqrt{K}.$$

Proof. By definition of $\Sigma_{\boldsymbol{W}}^{1:T}$ and Σ_A we get immediately that their coefficients are positive and bounded from above by one a.s.. In addition Lemma 5.6.8 ensures the bounds on the spectral norm of those two matrices. ■

Under these assumptions, we recover exactly A^* following the steps below. We remind that in the DETM setting, the matrix $\boldsymbol{\Pi}^{1:T}$ is accessible and thus all the random quantities introduced in the procedure are available.

1. *Pre-SVD normalization :* Consider $\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\boldsymbol{\Pi}^{1:T} 1_{nT}\right) \in \mathcal{D}_p(\mathbb{R}_+^*)$. Then derive $\boldsymbol{\Pi}_* := \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}$. This multiplication mimics the pre-SVD normalization to be used in the real case. The matrix $\boldsymbol{M}_*$ addresses word frequency heterogeneity in real corpora in order to boost the signal-to-noise ratio in SVD. In the DETM, pre-SVD normalization is optional and the procedure exhibits the same performance for any choice of $\boldsymbol{M}_*$ among diagonal matrices of dimension p with positive entries.

2. *SVD :* Compute the Singular Value Decomposition of $\boldsymbol{\Pi}_* \in \mathbb{R}^{p \times nT}$ which satisfies $\text{rank}(\boldsymbol{\Pi}_*) = K$ a.s. :

$$\boldsymbol{\Pi}_* := \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

   Let $[\boldsymbol{U}]_{.1}, \ldots, [\boldsymbol{U}]_{.K}$ be the column vectors of $\boldsymbol{U} \in \mathbb{R}^{p \times K}$ and notice that Perron-Frobenius's theorem, Lemma 5.6.9, guarantees that $[\boldsymbol{U}]_{.1}$ does not possess any null entry a.s. . The SVD creates a low dimensional word embedding into $\mathbb{R}^K$ but these vectors do not directly lead to the recovery of A^*.

3. *Post-SVD normalization :* Compute $\boldsymbol{R} \in \mathbb{R}^{p \times (K-1)}$ defined as follows, for $i \in [p]$ and $k \in [K-1]$:

$$[\boldsymbol{R}]_{ik} = \frac{[\boldsymbol{U}]_{i(k+1)}}{[\boldsymbol{U}]_{i1}}.$$

   This post-SVD normalization yields normalized vectors $[\boldsymbol{R}]_{1.}, \ldots, [\boldsymbol{R}]_{p.}$, the row vectors of $\boldsymbol{R}$. Proposition 5.2.14 ensures that there exist $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K \in \mathbb{R}^{(K-1)}$ such that the row vectors of $\boldsymbol{R}$ are

located in $G_{\boldsymbol{\eta}} \subset \mathbb{R}^{(K-1)}$ defined as follows :

$$G_{\boldsymbol{\eta}} := \left\{ x : x = \sum_{k=1}^{K} \alpha_k \boldsymbol{\eta}_k,\ \forall k \in [K],\ \alpha_k \geq 0\ \sum_{k=1}^{K} \alpha_k = 1 \right\}.$$

The vertices $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ of $G_{\boldsymbol{\eta}}$ are determined in the following step.

4. *Vertex Hunting :* The vertices $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ of $G_{\boldsymbol{\eta}}$ are recovered by computing the convex hull of the point cloud $[\boldsymbol{R}]_{1.}, \ldots, [\boldsymbol{R}]_{p.}$. Subsequently we define the matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times K}$ by solving the following system, for all $i \in [p]$,

$$[\boldsymbol{R}]_{i.} = \sum_{k=1}^{K} [\boldsymbol{\Lambda}]_{ik} \boldsymbol{\eta}_k,$$
$$\sum_{k=1}^{K} [\boldsymbol{\Lambda}]_{ik} = 1, \quad [\boldsymbol{\Lambda}]_{ik} \geq 0, \quad k \in [K].$$

5. *Word-topic matrix estimation :* Define the matrix $\boldsymbol{\Gamma} := \boldsymbol{M}_*^{1/2} \mathsf{diag}([\boldsymbol{U}]_{.1}) \boldsymbol{\Lambda}$. Normalize each column of $\boldsymbol{\Gamma}$ by its $\mathbb{L}_1$ norm. The resulting matrix is $\hat{A}$ which is almost surely equal to A^*, as stated in Theorem 4.3.2.

Finally, in the DETM setting, the matrix $\hat{A}$, estimator of A^*, can be represented as

$$\hat{A} = \Phi_{col} \left(\boldsymbol{M}_*^{1/2} \mathsf{diag}([\boldsymbol{U}]_{.1}) \Phi_{row} \left(\boldsymbol{\Lambda}_+ \right) \right). \tag{4.6}$$

Theorem 4.3.2 *In the DETM setting, the matrices $\hat{A}$ and A^* are equal almost surely.*

Proof. See Lemma 2.1, 2.2 and 2.3 in [84]. ■

It is important to highlight that under our assumptions, the matrix $(A^*)^\top (A^*)$ becomes full rank, facilitating the precise reconstruction of the matrix $\boldsymbol{W}^{1:T}$ through regression of $\boldsymbol{\Pi}^{1:T}$ onto A^*. Specifically, $\boldsymbol{W}^{1:T}$ can be recovered as :

$$\boldsymbol{W}^{1:T} = \left[(A^*)^\top (A^*) \right]^{-1} (A^*)^\top \boldsymbol{\Pi}^{1:T}.$$

Let us recall that here, we assume that the matrix $\boldsymbol{W}^{1:T}$ is issued by an AR(1) model and thus $\Sigma_{\boldsymbol{W}}^{1:T}$ is random. We show that its smallest eigenvalue is bounded away from 0 with high probability in Theorem 4.3.3. Then we prove in Proposition 4.3.4 that each topic is well-represented across documents. We demonstrate in Proposition 4.3.5 that the covariance matrix of each $\boldsymbol{W}_j^t$, namely $\Sigma(\theta^*)$, is singular. This explains why we focus on the second order moment matrix and its empirical version, the topic-topic concurrence matrix $\Sigma_{\boldsymbol{W}}^{1:T}$. Then we control the spectral norms of $\Sigma_{\boldsymbol{W}}^{1:T}$ and Σ_A in Proposition 4.3.1.

Theorem 4.3.3 *Consider the DETM under Assumptions 2, 3, 4 and 5.Denote $\tilde{\theta}_{(1)}^* \geq \tilde{\theta}_{(2)}^* \geq \ldots \geq \tilde{\theta}_{(K)}^*$ the components of $\tilde{\theta}^* \in \mathbb{R}^K$ in increasing order and $\gamma := \dfrac{c^*}{(2-c^*)(\alpha+1)}$. Then, for an absolute constant $C > 0$ and for any $\epsilon > 0$, we have, with probability at least $1 - T\exp(-\epsilon)$,*

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \leq \gamma \tilde{\theta}_{(K-1)}^* + \max\left(\sqrt{\frac{\epsilon + \log(K)}{C}} \sqrt{\frac{\gamma \tilde{\theta}_{(1)}^* + (1-\gamma)}{n}}, \frac{\epsilon + \log(K)}{nC} \right),$$

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq \gamma \tilde{\theta}_{(K)}^* - \max\left(\sqrt{\frac{\epsilon + \log(K)}{C}} \sqrt{\frac{\gamma \tilde{\theta}_{(1)}^* + (1-\gamma)}{n}}, \frac{\epsilon + \log(K)}{nC} \right).$$

In particular, under the assumptions of the previous theorem we get for $\epsilon = \log(nT)$ that, with probability at least $1 - \frac{1}{n}$,

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \leq \gamma\tilde{\theta}^*_{(K-1)} + \max\left(\sqrt{\frac{\log(nTK)\left(\gamma\tilde{\theta}_{(1)} + (1-\gamma)\right)}{nC}}, \frac{\log(nTK)}{nC}\right),$$

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq \gamma\tilde{\theta}^*_{(K)} - \max\left(\sqrt{\frac{\log(nTK)\left(\gamma\tilde{\theta}_{(1)} + (1-\gamma)\right)}{nC}}, \frac{\log(nTK)}{nC}\right).$$

Proof. See Proof in Subsection 4.5.1 ■

Under Assumption 3 and 4, ensuring that $\tilde{\theta}^*_{(K)} \geq \underline{\theta} > 0$ and leading to $\alpha < A(m)$, we get that $\gamma\tilde{\theta}^*_{(K)} \geq (A(m)+1)^{-1}\frac{\underline{c}\cdot\underline{\theta}}{(2-\underline{c})}$. Thus, for n large enough, we can find $c_2 > 0$ such that $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq c_2$ with high probability, which is a relaxed version of the a.s. constraint in Assumption 7.

It is important to note that Theorem 4.3.3 guarantees that each topic is well-represented across documents. Indeed, this implies a uniform lower bound on the frequency of each topic as shown in the next proposition. We reiterate the probabilistic interpretation of the matrix $\boldsymbol{W}^{1:T}$: for all $(j,t,k) \in [n]\times[T]\times[K]$, $\boldsymbol{W}_j^t(k)$ is the probability to observe the topic k given the document j at time t, $\mathbb{P}(\text{topic } k|\text{ document } j,\text{ step } t)$.

Proposition 4.3.4 (Topic distribution among documents) *Consider the model* (4.3) *under assumption 5. Then for all* $k \in K$, $\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k) \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})$ *almost surely.*

Proof. See Proof in Subsection 4.5.2 ■

In addition we highlight that the variance $\Sigma(\theta^*) = \mathbb{V}\left[\boldsymbol{W}_j^t\right]$ is singular. Thus we work with the second order moment matrix $\mathbb{E}\left[\boldsymbol{W}_j^t(\boldsymbol{W}_j^t)^\top\right]$.

Proposition 4.3.5 ($\Sigma(\theta^*)$ is singular) *For any* $\theta^* \in \mathbb{R}_+^K$, $\Sigma(\theta^*) := \frac{1}{\alpha+1}\left(\textit{diag}(\tilde{\theta}^*) - \tilde{\theta}^*\cdot(\tilde{\theta}^*)^\top\right) \in \mathbb{R}_+^{K\times K}$ *is singular, where* $\alpha := \|\theta^*\|_1 > 0$ *and* $\tilde{\theta}^* := \theta^*/\alpha$.

Proof. It is sufficient to note that $1_K/\sqrt{K}$, the $K-$dimensional vector with entries equal to $1/\sqrt{K}$, is an eigenvector of $\Sigma(\theta^*)$ associated to the eigenvalue 0. Therefore, the rank of this matrix is at most $K-1$ and the matrix is singular. ■

Following the recovery of $\boldsymbol{W}^{1:T}$, the subsequent section outlines a detailed procedure to estimate the key parameters of interest, c^*, $\tilde{\theta}^*$ and α. More specifically, we leverage the recovery of $\boldsymbol{W}^{1:T}$ using (4.3) to derive estimators for $\hat{c}$, $\hat{\theta}$, and $\hat{\alpha}$. However, it's worth noting that an equivalent procedure can also be applied. By utilizing the availability of $\boldsymbol{\Pi}^{1:T}$ and directly using (4.5), one can estimate the key parameters. Although $\hat{c}$ can be readily derived using this equation, estimating $\tilde{\theta}$ and α still requires recovering A^* and the projection of an estimated $\boldsymbol{\Delta}^{1:T}$ onto the span of A^*. Our approach here involves the projection of $\boldsymbol{\Pi}^{1:T}$ onto the span of A^*, followed by the estimation of all scalar parameters. Therefore, in this context, both approaches yield similar results, and our approach offers theoretical results that are easier to derive.

As mentioned in [86], it may appear possible to apply the results on the recovery of the matrix A^* to the recovery of $\boldsymbol{W}^{1:T}$ by merely transposing equation (4.2) and interchanging the roles of these two

matrices. However, such an inference is not possible due to the inherent disparity between the resulting models. In fact, the independence assumption among the columns of $\mathbf{\Pi}^1$, stated in assumption 5, does not hold after transposition. In addition, the row-wise summation of matrices A^*, $\boldsymbol{W}^{1:T}$ and $\mathbf{\Pi}^{1:T}$ does not yield unity, leading to the need of a distinct statistical treatment. This discrepancy underscores the need for a nuanced analysis, recognizing that the implications and statistical properties of estimating A^* differ substantially from those associated with recovering $\boldsymbol{W}^{1:T}$. This justifies our prioritization of initially recovering A^* and subsequently leveraging A^* to infer $\boldsymbol{W}^{1:T}$. A direct focus on $\boldsymbol{W}^{1:T}$ would require an additional set of assumptions, surpassing the scope of this study. Therefore, our focus remains on describing the methodology for recovering A^* and subsequently utilizing it to infer $\boldsymbol{W}^{1:T}$.

4.4 Estimation of the autoregressive model

In this section, we present non-asymptotic rates of convergence in the case where the matrix $\boldsymbol{W}^{1:T}$ is observed. This scenario arises in the DETM, once A^* is recovered and $\mathbf{\Pi}^{1:T}$ is regressed on A^*. We observe that model (4.3) can be alternatively expressed as a collection of n independent vector autoregressive processes of order 1, denoted VAR(1), as follows, where $(t, j) \in [T-1] \times [n]$,

$$\boldsymbol{W}_j^{t+1} = (1 - c^*)\boldsymbol{W}_j^t + c^*\tilde{\theta}^* + c^* \cdot \left(\boldsymbol{\Delta}_j^t - \tilde{\theta}^*\right). \tag{4.7}$$

Assumption 5 asserts that our analysis commences after the system has entered a stationary regime, bypassing any transitional phase. Additionally, we assume that the initial vectors $\left(\boldsymbol{W}_j^1\right)_j$ are stochastic, and their first and second moments align with the characteristics of the stationary regime.

To estimate the parameters of model (4.3), we adopt the method of moments. We define $\hat{\theta}$ as the empirical mean of the observed $\left(w_j^{t+1}\right)_{j,t}$:

$$\hat{\theta} := \frac{1}{n(T-1)} \sum_{j=1}^{n} \sum_{t=1}^{T-1} w_j^t. \tag{4.8}$$

We estimate $1 - c^*$ by the normalized sum of scalar products between the centered consecutive vectors $w_j^{t+1} - \overline{w}^{+1}$ and $w_j^t - \overline{w}$:

$$\widehat{(1-c)} := \frac{\sum\limits_{t=1}^{T-1} \sum\limits_{j=1}^{n} \left\langle w_j^{t+1} - \overline{w}^{+1};\ w_j^t - \overline{w} \right\rangle}{\sum\limits_{t=1}^{T-1} \sum\limits_{j=1}^{n} \left\| w_j^t - \overline{w} \right\|_2^2}, \tag{4.9}$$

where $\overline{w}^{+1} := \dfrac{1}{n(T-1)} \sum\limits_{t=1}^{T-1} \sum\limits_{j=1}^{n} w_j^{t+1}$ and $\overline{w} := \dfrac{1}{n(T-1)} \sum\limits_{t=1}^{T-1} \sum\limits_{j=1}^{n} w_j^t = \hat{\theta}$.

Finally, using the variance of the stationary sequence w_j^t and the explicit expression of the matrix Σ, we see that :

$$\mathsf{Tr}(\mathbb{V}(w_j^t)) = \frac{c^*}{2 - c^*} \frac{1 - \|\tilde{\theta}^*\|_2^2}{\alpha + 1}. \tag{4.10}$$

Thus, we plug-in estimators $\hat{\theta}$, $\hat{c}$ and the empirical variance to get

$$\hat{\alpha} = \frac{\hat{c}}{2 - \hat{c}} \frac{1 - \|\hat{\theta}\|_2^2}{\mathcal{V}} - 1, \quad \text{where} \quad \mathcal{V} := \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| w_j^t - \overline{w} \right\|_2^2. \tag{4.11}$$

Next we give the convergence rates of these three estimators. It is worth emphasizing that the convergence rates of these estimators are independent of the dimension K.

Theorem 4.4.1 (Estimation of $\tilde{\theta}^*$) *In the DETM, under the Assumptions 3, 4 and 5, the estimator $\hat{\theta}$ defined in* (4.8) *is such that for $T \geq 2 + \frac{2}{\overline{c}}$ and any $0 < \epsilon < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:*

$$\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2 \leq \frac{\epsilon + 1}{\sqrt{n(T-1)}} \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right), \tag{4.12}$$

with probability larger than $1 - 2\exp\left(-\epsilon^2/4\right)$.

Proof. See Proof in Subsection 4.5.3 ■

Theorem 4.4.2 (Estimation of c^*) *In the DETM, under the Assumptions 3, 4 and 5, the estimator $\widehat{(1-c)}$ defined in* (4.9) *is such that for n and T large enough, for all $0 < \epsilon < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:*

$$|\widehat{(1-c)} - (1-c^*)| \leq \frac{C_1 \cdot \epsilon}{\sqrt{n(T-1)}} + \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right) \frac{C_2\left(\epsilon^2 + 1\right)}{n(T-1)}, \tag{4.13}$$

with probability larger than $1 - 15\exp(-\epsilon^2/4)$ where $C_1 := \dfrac{44}{\underline{c}m}$ and $C_2 := \dfrac{8}{\underline{c}m}$.

Proof. See Proof in Subsection 4.5.4 ■

Theorem 4.4.3 (Estimation of α) *In the DETM, under the Assumptions 3, 4 and 5, the estimator $\hat{\alpha}$ defined in* (4.11) *is such that for n and T large enough, for all $0 < \epsilon < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:*

$$|\hat{\alpha} - \alpha^*| \leq \frac{C_3 \cdot \epsilon}{\sqrt{n(T-1)}} + \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right)\left[\frac{C_4\left(\epsilon + 1\right)}{\sqrt{n(T-1)}} + \frac{C_5\left(\epsilon^2 + 1\right)}{n(T-1)}\right], \tag{4.14}$$

with probability larger than $1 - 17\exp\left(-\epsilon^2/4\right)$ where $C_3 := \dfrac{176\left(1+\overline{c}\right)}{\underline{c}2^2m^2}$, $C_4 := \frac{8\overline{c}}{\underline{c}m(2-\overline{c})} + \frac{16(A(m)+1)}{\underline{c}m}$, and $C_5 := \dfrac{32\left(1+\overline{c}\right)}{\underline{c}^2m^2}$, and $A(m)$ is defined after the Assumption 4.

Proof. See Proof in Subsection 4.5.5 ■

It's worth mentioning that an alternative model could involve assigning distinct parameters θ_j^* to the n columns of the noise matrices $\mathbf{\Delta}^t$. In this scenario, we forfeit the benefit of multiple vectors sharing a common parameter. Nevertheless, our results remain valid for estimating the n parameters θ_j^* when $n = 1$. In particular we may have n different estimators $\hat{\theta}_j$ showing a convergence rate of order $\mathcal{O}\left(1/\sqrt{T-1}\right)$. Such a model is useful for capturing the distinct ways in which newspapers address current affairs, exhibiting unique preferences and avoidances. By considering different θ_j^*, we enable newspapers to have distinct stationary distributions, reflecting differences in their treatment of information. This flexibility allows us to capture variations in information dissemination among different newspapers. Another possible extension is to consider a matrix distribution on the noise, which would make it possible to lift the hypothesis of independence between newspapers and to consider that journalists influence each other in the processing of information. This model goes beyond the scope of this paper and is left for future works.

4.5 Proofs

4.5.1 Proof of Theorem 4.3.3

Proof of Theorem 4.3.3. First, we recall that

$$\Sigma_{\boldsymbol{W}}^{1:T} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top\right),$$

where $\boldsymbol{W}^t := [\boldsymbol{W}_1^t, \dots, \boldsymbol{W}_n^t]$ is a matrix of size $K \times n$. However, the matrices $[\boldsymbol{W}^t]_{t\in[T]}$ are not independent. By Lemma 5.6.2 we have that

$$\lambda_K\left(\Sigma_{\boldsymbol{W}}^{1:T}\right) \geq \frac{1}{T}\sum_{t=1}^{T}\lambda_K\left(\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top\right).$$

Next, we see that, for each $t \in [T]$, $\boldsymbol{W}^t$ has independent columns $[\boldsymbol{W}_j^t]_{j\in[n]}$ which are random probability vectors. Therefore, $\left\|\boldsymbol{W}_j^t\right\|_2 \leq \left\|\boldsymbol{W}_j^t\right\|_1 \leq 1$. Let us denote by

$$\Omega := \mathbb{E}\left[\boldsymbol{W}_j^t\left(\boldsymbol{W}_j^t\right)^\top\right]$$

the common second order moment matrix of the vectors $\left(\boldsymbol{W}_j^t\right)_{j,t}$. By Proposition 5.1.1 we see that

$$\Omega = \gamma\mathsf{diag}(\tilde{\theta}^*) + (1-\gamma)\,\tilde{\theta}^*\left(\tilde{\theta}^*\right)^\top.$$

By assumption, $\tilde{\theta}^*$ has positive entries and thus $\mathsf{diag}(\tilde{\theta}^*) \in \mathcal{D}_K\left(\mathbb{R}_+^*\right)$ is full rank with positive coefficient : $1-\gamma > 0$. Elementary linear algebra results, see [67], give that :

$$\gamma\tilde{\theta}_{(K)}^* \leq \lambda_K(\Omega) \leq \gamma\tilde{\theta}_{(K-1)}^* \quad \text{and} \quad \gamma\tilde{\theta}_{(1)}^* \leq \lambda_1(\Omega) \leq \gamma\tilde{\theta}_{(1)}^* + (1-\gamma).$$

The Weyl's inequality, see Lemma 1.1.13, provides, for all $t \in [T]$, almost surely :

$$\lambda_K(\Omega) + \left\|\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top - \Omega\right\|_{op} \geq \lambda_K\left(\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top\right) \geq \lambda_K(\Omega) - \left\|\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top - \Omega\right\|_{op}.$$

Finally, we apply Lemma 1.1.18 with $\kappa = 1$ to get that for all $t \in [T]$ and for all $\epsilon > 0$, with probability at least $1 - \exp\left(-\epsilon^2\right)$,

$$\left\|\frac{1}{n}\boldsymbol{W}^t(\boldsymbol{W}^t)^\top - \Omega\right\|_{op} \leq \max\left(\sqrt{\frac{\epsilon^2+\log(K)}{C}}\sqrt{\frac{\lambda_1(\Omega)}{n}}, \frac{\epsilon^2+\log(K)}{nC}\right),$$

where $C > 0$ is an absolute constant. We conclude, using a union bound in $t \in [T]$, the upper bounds on $\lambda_1(\Omega)$ and $\lambda_K(\Omega)$ and the lower bound on $\lambda_K(\Omega)$.

■

4.5.2 Proof of Proposition 4.3.4

Proof of Proposition 4.3.4. Let us consider the model (4.3) under Assumption 5 . Under this setting, for all $k \in [K]$ and for all $(j,t) \in [n] \times [T]$, $\boldsymbol{W}_j^t(k) \leq 1$ almost surely. This implies almost surely that

$$\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k) \geq \frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k)^2 = \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk}.$$

However, $\Sigma_{\boldsymbol{W}}^{1:T}$ is positive definite under Assumption 7. Moreover, diagonal entries of a positive definite matrix cannot be smaller than the smallest eigenvalue. Indeed by definition

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) := \min_{\|x\|_2=1} x^\top \Sigma_{\boldsymbol{W}}^{1:T} x.$$

Fixing $k \in [K]$ and considering $x = e_k$ where $(e_1, \ldots, e_K)$ is the canonical basis of $\mathbb{R}^K$ leads to $\|x\|_2 = 1$ and $\langle \Sigma_{\boldsymbol{W}}^{1:T} x, x\rangle = \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk}$. This proves that almost surely we have

$$\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k) \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}).$$

■

4.5.3 Proof of Theorem 4.4.1

Proof of Theorem 4.4.1. We use repeatedly the equation of our model to get that, for any integer $t \geq 2$ and any $j \geq 1$, we have :

$$\begin{aligned} w_j^t - \tilde{\theta}^* &= (1-c^*)(w_j^{t-1} - \tilde{\theta}^*) + c^*(\Delta_j^{t-1} - \tilde{\theta}^*) \\ &= (1-c^*)^{t-1}\left(w_j^1 - \tilde{\theta}^*\right) + c^* \sum_{s=1}^{t-1}(1-c^*)^{t-1-s}\left(\Delta_j^s - \tilde{\theta}^*\right). \end{aligned} \tag{4.15}$$

We plug this in the estimator to get :

$$\begin{aligned} \hat{\theta} - \tilde{\theta}^* &= \frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}(w_j^t - \tilde{\theta}^*) \\ &= \frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}(1-c^*)^{t-1}\left(w_j^1 - \tilde{\theta}^*\right) \\ &+ \frac{c^*}{n(T-1)}\sum_{j=1}^{n}\sum_{t=2}^{T-1}\sum_{s=1}^{t-1}(1-c^*)^{t-1-s}\left(\Delta_j^s - \tilde{\theta}^*\right) \\ &= \frac{1-(1-c^*)^{T-1}}{c^*(T-1)n}\sum_{j=1}^{n}(w_j^1 - \tilde{\theta}^*) \\ &+ \frac{c^*}{n(T-1)}\sum_{j=1}^{n}\sum_{s=1}^{T-2}\sum_{t=s+1}^{T-1}(1-c^*)^{t-1-s}\left(\Delta_j^s - \tilde{\theta}^*\right). \end{aligned}$$

Finally, we get :

$$\hat{\theta} - \theta^* = \frac{1-(1-c^*)^{T-1}}{c^*(T-1)} \left[\frac{1}{n} \sum_{j=1}^{n} w_j^1 - \tilde{\theta}^* \right] + \sum_{t=1}^{T-2} \frac{1-(1-c^*)^{T-2-t}}{n(T-1)} \sum_{j=1}^{n} \left(\Delta_j^t - \tilde{\theta}^* \right).$$

We apply a vector-Bernstein inequality (see Lemma 5.6.1) successively to each term above.

On the one hand, Assumption 5 ensures that $\left(w_j^1 - \tilde{\theta}^* \right)_{j \in [n]}$ are centered and independent. In addition, $\left(w_j^1 \right)_{j \in [n]}$ are in the simplex as well as $\tilde{\theta}^*$. This implies that for all $j \in [n]$, $\left\| w_j^1 - \tilde{\theta}^* \right\|_2 \leq \left\| w_j^1 - \tilde{\theta}^* \right\|_1 \leq 2$ almost surely. Let us define $V_1 := \sum_{j=1}^{n} \mathbb{E}\left[\left\| w_j^1 - \tilde{\theta}^* \right\|_2^2 \right]$ and note that $V_1 \leq 4n$. More precisely, Assumption 5 gives that :

$$\begin{aligned} V_1 &= \sum_{j=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[\left(w_j^1(k) - \tilde{\theta}^*(k) \right)^2 \right] \\ &= n \cdot \mathsf{Tr}\left(\mathbb{V}(w_j^1) \right) = n \cdot \frac{c^*}{2-c^*} \mathsf{Tr}\left(\Sigma \right) \leq n. \end{aligned}$$

Therefore, by Lemma 5.6.1, we get that for all $\epsilon \in (0, \sqrt{V_1}/2)$ we have :

$$\left\| \frac{1}{n} \sum_{j=1}^{n} w_j^1 - \tilde{\theta}^* \right\|_2 \leq \frac{(\epsilon+1)\sqrt{V_1}}{n} \leq \frac{\epsilon+1}{\sqrt{n}}, \tag{4.16}$$

with probability larger than $1 - \exp\left(-\frac{\epsilon^2}{4} \right)$.

On the other hand, let us denote by $a_t := \left[1 - (1-c^*)^{T-1-t} \right]$, and see that $\left(a_t \cdot \left(\Delta_j^t - \tilde{\theta}^* \right) \right)_{t \in [T-1]}$ are independent vectors, centered, uniformly bounded from above in Euclidean norm and have variance bounded from above by 1 :

$$\begin{aligned} \mathbb{E}\left[\left[1 - (1-c^*)^{T-1-t} \right] \left(\Delta_j^t - \tilde{\theta}^* \right) \right] &= 0, \\ \left\| \left[1 - (1-c^*)^{T-1-t} \right] \left(\Delta_j^t - \tilde{\theta}^* \right) \right\|_2 &\leq 2, \; a.s. \\ \mathbb{E}\left[\left\| \left[1 - (1-c^*)^{T-1-t} \right] \left(\Delta_j^t - \tilde{\theta}^* \right) \right\|_2^2 \right] &= a_t^2 \cdot \mathsf{Tr}(\Sigma) \leq 1. \end{aligned}$$

Note that the last inequality is due to the Assumption 4. Let us denote by

$$V_2 := \sum_{j=1}^{n} \sum_{t=1}^{T-2} \mathbb{E}\left[\left\| \left[1 - (1-c^*)^{T-1-t} \right] \left(\Delta_j^t - \tilde{\theta}^* \right) \right\|_2^2 \right] = n \cdot \mathsf{Tr}(\Sigma) \cdot \sum_{t=1}^{T-2} a_t^2$$

and note that $V_2 \leq n(T-1)$. Lemma 5.6.1 finally proves that for all $\epsilon \in (0, \sqrt{V_2}/2)$,

$$\begin{aligned}\left\| \frac{1}{n(T-1)} \sum_{j=1}^{n} \sum_{t=1}^{T-2} a_t \left(\Delta_j^t - \tilde{\theta}^* \right) \right\|_2 &\leq \frac{(\epsilon+1)\sqrt{V_2}}{n(T-1)} \\ &\leq \frac{(\epsilon+1)}{\sqrt{n(T-1)}}, \\ &\leq \frac{(\epsilon+1)}{\sqrt{n(T-1)}}, \end{aligned} \tag{4.17}$$

with probability larger than $1 - \exp\left(-\frac{\epsilon^2}{4}\right)$.

Let us define $V_* := \min(V_1;\ V_2)$ and for $\epsilon \in (0, \sqrt{V_*}/2)$ we get using a union bound and inequalities (4.16) and (4.17) that :

$$\begin{aligned}\left\| \hat{\theta} - \tilde{\theta}^* \right\|_2 &\leq \left(\frac{1-(1-c^*)^{T-1}}{c^*\sqrt{n}(T-1)} + \frac{1}{\sqrt{n(T-1)}} \right)(\epsilon+1) \\ &\leq \frac{\epsilon+1}{\sqrt{n(T-1)}} \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1 \right),\end{aligned}$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{4}\right)$. We conclude by giving a lower bound on V_* using Assumptions 3, 4 and 5. We first bound from below V_1 and V_2 as follows :

$$\begin{aligned} V_1 &= n \cdot \mathsf{Tr}\left(\mathbb{V}\left(w_1^1\right)\right) = n\frac{c^*}{2-c^*}\mathsf{Tr}(\Sigma) \geq n\frac{\underline{c}}{2-\underline{c}}m, \\ V_2 &= n \cdot \mathsf{Tr}\left(\mathbb{V}\left[\Delta_1^1\right]\right) \sum_{t=1}^{T-1} \left(1-(1-c^*)^{T-1-t}\right)^2 \\ &= n \cdot \mathsf{Tr}(\Sigma) \sum_{t=1}^{T-1} \left(1 - 2(1-c^*)^{T-t-1} + (1-c^*)^{2T-2t-2}\right) \\ &= n \cdot \mathsf{Tr}(\Sigma) \left((T-1) - 2\frac{1-(1-c^*)^{T-1}}{c^*} + \frac{1-(1-c^*)^{2(T-1)}}{1-(1-c^*)^2} \right) \\ &\geq n \cdot \mathsf{Tr}(\Sigma) \left(T - 1 - \frac{2}{c^*} \right) \geq nm \left(T - 1 - \frac{2}{\overline{c}} \right). \end{aligned}$$

We conclude that $V_* \geq nm \min\left(\frac{\underline{c}}{2-\underline{c}},\ T - 1 - \frac{2}{\overline{c}} \right)$. In particular, for $T \geq 2 + \frac{2}{\overline{c}}$,

$$V_* \geq nm\frac{\underline{c}}{2-\underline{c}}.$$

■

4.5.4 Proof of Theorem 4.4.2

Proof of Theorem 4.4.2. We denote $\overline{\Delta} := \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \Delta_j^t$ and see that

$$\widehat{(1-c)} - (1-c^*) = c^* \frac{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^t - \tilde{\theta}^* \right\rangle - n(T-1) \left\langle \overline{\Delta} - \tilde{\theta}^*;\ \overline{w} - \tilde{\theta}^* \right\rangle}{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| w_j^t - \overline{w} \right\|_2^2}.$$

We note that $\overline{w}^{+1} = (1-c^*)\,\overline{w} + c^* \overline{\Delta}$. This implies the following :

$$\left\langle w_j^{t+1} - \overline{w}^{+1};\ w_j^t - \overline{w} \right\rangle = (1-c^*) \left\| w_j^t - \overline{w} \right\|_2^2 + c^* \left\langle \Delta_j^t - \overline{\Delta};\ w_j^t - \overline{w} \right\rangle.$$

Thus,

$$\widehat{(1-c)} = (1-c^*) + c^* \frac{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \overline{\Delta};\ w_j^t - \overline{w} \right\rangle}{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| w_j^t - \overline{w} \right\|_2^2}.$$

By expansion of the numerator above and using the bilinearity of the scalar product, we get :

$$\begin{aligned} &\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \tilde{\theta}^* - \left(\overline{\Delta} - \tilde{\theta}^* \right);\ w_j^t - \tilde{\theta}^* - \left(\overline{w} - \tilde{\theta}^* \right) \right\rangle \\ &= \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^t - \tilde{\theta}^* \right\rangle - n(T-1) \left\langle \overline{\Delta} - \tilde{\theta}^*;\ \overline{w} - \tilde{\theta}^* \right\rangle. \end{aligned} \tag{4.18}$$

We then bound from above with high probability the first term of the right-hand side of the equation (4.18) :

$$\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^t - \tilde{\theta}^* \right\rangle.$$

We use the expansion in (4.15) to get :

$$\begin{aligned} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^t - \tilde{\theta}^* \right\rangle &= \sum_{t=1}^{T-1} \sum_{j=1}^{n} (1-c^*)^{t-1} \left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^1 - \tilde{\theta}^* \right\rangle \\ &\quad + c^* \sum_{t=2}^{T-1} \sum_{j=1}^{n} \sum_{s=1}^{t-1} (1-c^*)^{t-1-s} \left\langle \Delta_j^t - \tilde{\theta}^*;\ \Delta_j^s - \tilde{\theta}^* \right\rangle \\ &= \sum_{t=1}^{T-1} \sum_{j=1}^{n} Z_j^t + \sum_{t=2}^{T-1} \sum_{j=1}^{n} X_j^t, \end{aligned}$$

where we denote by Z_j^t and X_j^t the real-valued random variables defined as follows

$$\begin{aligned} Z_j^t &:= (1-c^*)^{t-1}\left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^1 - \tilde{\theta}^* \right\rangle, \\ X_j^t &:= c^* \sum_{s=1}^{t-1} (1-c^*)^{t-1-s} \left\langle \Delta_j^t - \tilde{\theta}^*;\ \Delta_j^s - \tilde{\theta}^* \right\rangle \\ &= c^*(1-c^*)^{t-1} \left\langle \Delta_j^t - \tilde{\theta}^*;\ \sum_{s=1}^{t-1} (1-c^*)^{-s} \left(\Delta_j^s - \tilde{\theta}^*\right) \right\rangle. \end{aligned}$$

We first notice that the $(Z_j^t)_{j,t}$ are centered *i.e.* $\mathbb{E}\left[Z_j^t\right] = 0$, and uniformly bounded. Indeed, for all (j,t), Cauchy-Schwarz inequality ensures that

$$|Z_j^t| \leq (1-c^*)^{t-1} \left\|\Delta_j^t - \tilde{\theta}^*\right\|_2 \left\|w_j^1 - \tilde{\theta}^*\right\|_2 \leq 4 \quad \text{a.s..}$$

Moreover, they are independent conditionally on $(w_j^1)_j$. Thus Hoeffding's concentration inequality, Lemma 1.1.8, ensures that for any $\epsilon > 0$, conditionally on $(w_j^1)_j$,

$$\left| \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} Z_j^t \right| \leq \frac{2\sqrt{2}}{\sqrt{n(T-1)}}\epsilon, \tag{4.19}$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{4}\right)$. Since this bound is free of the $(w_j^1)_j$ it remains unchanged after integrating with respect to the stationary distribution of these r.v..

Similarly, the real-valued random variables $(X_j^t)_{j,t}$ are uniformly bounded. Indeed for all (j,t), we have almost surely that :

$$\begin{aligned} |X_j^t| &\leq c^*(1-c^*)^{t-1} \sum_{s=1}^{t-1} (1-c^*)^{-s} \left\|\Delta_j^s - \tilde{\theta}^*\right\|_2 \left\|\Delta_j^t - \tilde{\theta}^*\right\|_2 \\ &\leq 4 \cdot \left(1 - (1-c^*)^{t-1}\right) \leq 4 \quad \text{a.s..} \end{aligned}$$

However, for each $j = 1, \ldots, n$, the $(X_j^t)_t$ are dependent random variables that form a martingale difference. Indeed, Δ_j^t is independent of $(\Delta_j^s)_{s<t}$. We denote $\mathcal{F}_j^{t-1} := \sigma\left(\Delta_j^1, \ldots, \Delta_j^{t-1}\right)$ the natural filtration of the random process $\left(\Delta_j^t\right)_t$. This ensures that for all t,

$$\begin{aligned} \mathbb{E}[X_j^t] &= 0, \\ \mathbb{E}[X_j^t | \mathcal{F}_j^{t-1}] &= 0 \quad \text{a.s.}, \\ \mathbb{E}[|X_j^t|] &\leq 4\left(1 - (1-c^*)^{t-1}\right) \leq 4. \end{aligned}$$

Hence, for all $j \in [n]$, the adapted sequence $\left(\left\{X_j^t, \mathcal{F}_j^{t-1}\right\}\right)_{t\in[T-1]}$ is a martingale difference, see Definition 1.1.12. Azuma-Hoeffding's inequality, see Lemma 1.1.19, ensures that for all $j \in [n]$ and for any $\epsilon > 0$:

$$\left| \frac{1}{T-1} \sum_{t=2}^{T-1} X_j^t \right| \leq \frac{T-2}{T-1} \frac{4\epsilon}{\sqrt{T-2}} \leq \frac{4\epsilon}{\sqrt{T-1}},$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{2}\right)$. We also deduce that for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T-1}\sum_{t=2}^{T-1} X_j^t\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2(T-1)}{32}\right)$$
$$\leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \quad \text{with} \quad \sigma^2 = \frac{16}{T-1}.$$

Lemma 1.1.6 shows that the random variables $X_j := 1/(T-1)\sum_{t=1}^{T-1} X_j^t$ for $j = 1, \ldots, n$ are ν^2-subGaussian with $\nu^2 = 8\sigma^2 = \frac{128}{T-1}$. As the $(X_j)_j$ are independent, the more general Hoeffding's inequality for independent sub-Gaussian random variables, Lemma 1.1.7, ensures that for all $\epsilon > 0$,

$$\left|\frac{1}{n}\sum_{j=1}^{n} X_j\right| \leq \frac{8\epsilon}{\sqrt{n(T-1)}}, \tag{4.20}$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{4}\right)$. Indeed Hoeffding's inequality, Lemma 1.1.7 ensures that

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_j X_j\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\nu^2}\right)$$
$$\leq 2\exp\left(-\frac{n(T-1)\epsilon^2}{256}\right),$$

from which (4.20) follows. Putting together (4.19) and (4.20) we get, for any $\epsilon > 0$,

$$\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\langle \Delta_j^t - \tilde{\theta}^*;\ w_j^t - \tilde{\theta}^*\right\rangle\right| \leq \frac{\left(8+2\sqrt{2}\right)\epsilon}{\sqrt{n(T-1)}} \leq \frac{11\epsilon}{\sqrt{n(T-1)}}, \tag{4.21}$$

with probability larger than $1 - 4\exp\left(-\frac{\epsilon^2}{4}\right)$.

We now bound from above with high probability the second term of the right-hand side of the equation (4.18), namely :

$$\left\langle \overline{\Delta} - \tilde{\theta}^*;\ \overline{w} - \tilde{\theta}^*\right\rangle.$$

First, recall for convenience that $\overline{w} = \hat{\theta}$ and Theorem 4.4.1 ensures that for any $0 < \epsilon < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:

$$\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2 \leq \frac{\epsilon+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right),$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{4}\right)$, see (4.12).

In addition, the vectors $\left(\Delta_j^t - \tilde{\theta}^*\right)$ are centered and satisfy for any $j \in [n]$ and any $t \in [T-1]$, $\left\|\Delta_j^t - \tilde{\theta}^*\right\|_2 \leq \left\|\Delta_j^t - \tilde{\theta}^*\right\|_1 \leq 2$ a.s.. Hence we define

$$V_3 := \sum_{j=1}^{n}\sum_{t=1}^{T-1} \mathbb{E}\left[\left\|\Delta_j^t - \tilde{\theta}^*\right\|_2^2\right] = n(T-1)\mathsf{Tr}(\Sigma),$$

which verifies $mn(T-1) \leq V_3 \leq n(T-1)$. Thus, Lemma 5.6.1 gives that for any $\epsilon \in \left(0, \sqrt{m}/2 \cdot \sqrt{n(T-1)}\right)$,

$$\left\|\overline{\Delta} - \tilde{\theta}^*\right\|_2 \leq \frac{(\epsilon+1)}{\sqrt{n(T-1)}}, \tag{4.22}$$

with probability larger than $1 - \exp\left(-\frac{\epsilon^2}{4}\right)$. Indeed vector Bernstein's inequality ensures that for all $\epsilon \in (0, \sqrt{V_3}/2)$,

$$\mathbb{P}\left[\left\|\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left(\Delta_j^t - \tilde{\theta}^*\right)\right\|_2 \geq (\epsilon+1)\sqrt{V_3}\right] \leq \exp\left(-\frac{\epsilon^2}{4}\right),$$

which implies (4.22). By Cauchy-Schwarz,

$$\left\langle \overline{\Delta} - \tilde{\theta}^*;\, \overline{w} - \tilde{\theta}^* \right\rangle \leq \left\|\overline{\Delta} - \tilde{\theta}^*\right\|_2 \left\|\hat{\theta} - \theta^*\right\|_2.$$

The bounds in (4.12) and (4.22) combined with a union bound give, for any $\epsilon \in (0, \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2)$,

$$\left|\left\langle \overline{\Delta} - \tilde{\theta}^*;\, \overline{w} - \tilde{\theta}^* \right\rangle\right| \leq \frac{(\epsilon+1)^2}{n(T-1)}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right), \tag{4.23}$$

with probability larger than $1 - 3\exp\left(-\frac{\epsilon^2}{4}\right)$.

The final step is to bound from below with high probability the empirical variance

$$\mathcal{V} := \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|w_j^t - \overline{w}\right\|_2^2. \tag{4.24}$$

Note that we can write using the stationarity of $(w_j^t)_t$ for all j :

$$\begin{aligned}
\mathcal{V} &= \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|w_j^t - \tilde{\theta}^*\right\|_2^2 - \left\|\tilde{\theta}^* - \overline{w}\right\|_2^2 \\
&= \mathbb{E}\left[\left\|w_1^1 - \tilde{\theta}^*\right\|_2^2\right] \\
&\quad + \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left(\left\|w_j^t - \tilde{\theta}^*\right\|_2^2 - \mathbb{E}\left[\left\|w_j^t - \tilde{\theta}^*\right\|_2^2\right]\right) - \left\|\tilde{\theta}^* - \overline{w}\right\|_2^2.
\end{aligned}$$

Let us define

$$U_1 := \mathbb{E}\left[\left\|w_1^1 - \tilde{\theta}^*\right\|_2^2\right],$$
$$U_2 := \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left(\left\|w_j^t - \tilde{\theta}^*\right\|_2^2 - \mathbb{E}\left[\left\|w_j^t - \tilde{\theta}^*\right\|_2^2\right]\right),$$
$$U_3 := \left\|\tilde{\theta}^* - \overline{w}\right\|_2^2.$$

Recall that $U_1 = \frac{c^*}{2-c^*}\mathrm{Tr}(\Sigma) \geq \frac{c}{2-c}m$ and that we use (4.12) to bound from above U_3 with high probability. Hence $U_1 - U_3$ is bounded from below with high probability. Next notice that

$$\mathcal{V} = U_1 + U_2 - U_3 \geq U_1 - U_3 - |U_2|.$$

The last step is thus to give a high probability bound from above for U_2. Recall that (4.15) is giving that $w_j^t - \tilde{\theta}^* = (1-c^*)^{t-1}(w_j^1 - \tilde{\theta}^*) + c^*\sum_{k=1}^{t-1}(1-c^*)^{t-1-k}E_j^k$, for all $t \geq 2$, where $E_j^k = \Delta_j^k - \tilde{\theta}^*$ is the centered, bounded, noise random variable. Thus we can decompose U_2 in the following terms

$$\begin{aligned}
U_2 &= \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}(1-c^*)^{2(t-1)}\left(\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2 - \mathbb{E}\left[\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2\right]\right)\\
&+ \frac{(c^*)^2}{n(T-1)}\sum_{t=2}^{T-1}\sum_{j=1}^{n}\left(\left\|\sum_{k=1}^{t-1}(1-c^*)^{t-1-k}E_j^k\right\|_2^2 - \mathbb{E}\left[\left\|\sum_{k=1}^{t-1}(1-c^*)^{t-1-k}E_j^k\right\|_2^2\right]\right)\\
&+ 2\frac{c^*}{n(T-1)}\sum_{t=2}^{T-1}\sum_{j=1}^{n}(1-c^*)^{t-1}\langle w_j^1 - \tilde{\theta}^*, \sum_{k=1}^{t-1}(1-c^*)^{t-1-k}E_j^k\rangle\\
&\leq T_1 + T_2 + T_3, \quad \text{say.}
\end{aligned}$$

We bound successively these last three terms in absolute values. First,

$$|T_1| = \frac{1-(1-c^*)^{2(T-1)}}{c^*(2-c^*)(T-1)}\left|\frac{1}{n}\sum_{j=1}^{n}\left(\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2 - \mathbb{E}\left[\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2\right]\right)\right|.$$

Remark that the random variables $\left(\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2\right)_{j\in[n]}$ are almost surely bounded in $[0,4]$ and independent. Applying Hoeffding's inequality for bounded random variables, see Lemma 1.1.8, leads to :

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2 - \mathbb{E}\left[\left\|w_j^1 - \tilde{\theta}^*\right\|_2^2\right]\right)\right| \leq \frac{\sqrt{2}\epsilon}{\sqrt{n}},$$

with probability larger than $1 - 2\exp\left(-\frac{\epsilon^2}{4}\right)$. Hence this leads to

$$|T_1| \leq \frac{1-(1-c^*)^{2(T-1)}}{c^*(2-c^*)(T-1)}\cdot\frac{\sqrt{2}\epsilon}{\sqrt{n}}, \tag{4.25}$$

with probability larger than $1-2\exp\left(-\frac{\epsilon^2}{4}\right)$.

Next, we write $T_2 := \frac{1}{n}\sum_{j=1}^{n}\Big(\Phi(E_j^1,\dots,E_j^{T-1}) - \mathbb{E}\Phi(E_j^1,\dots,E_j^{T-1})\Big)$, where for all $j\in[n]$,

$$\Phi : \mathcal{B}_2(2)^{T-1} \longrightarrow \mathbb{R}_+$$
$$\Phi(E_j^1,\dots,E_j^{T-1}) := \frac{(c^*)^2}{T-1}\sum_{t=2}^{T-1}\left\|\sum_{k=1}^{t-1}(1-c^*)^{t-1-k}E_j^k\right\|_2^2.$$

We show that Φ is a function with bounded differences in each argument. More precisely, for an arbitrary ℓ from 1 to $T-1$ and any x and x' having euclidean norm not larger than 2, we have :

$$\begin{aligned}
&\Phi(E_j^1,\dots,E_j^{\ell-1},x,E_j^{\ell+1},\dots,E_j^{T-1}) - \Phi(E_j^1,\dots,E_j^{\ell-1},x',E_j^{\ell+1},\dots,E_j^{T-1})\\
=\;& \frac{(c^*)^2}{T-1}\sum_{t=\ell+1}^{T-1}\Big((1-c^*)^{2(t-1-\ell)}(\|x\|_2^2-\|x'\|_2^2)\\
&+2(1-c^*)^{t-1-\ell}\langle x-x';\sum_{k=1,k\neq\ell}^{t-1}(1-c^*)^{t-1-k}E_j^k\rangle\Big)\\
\leq\;& \frac{(c^*)^2}{T-1}\sum_{t=\ell+1}^{T-1}\Big(4(1-c^*)^{2(t-1-\ell)}\\
&+2(1-c^*)^{t-1-\ell}\|x-x'\|_2\cdot\sum_{k=1}^{t-1}(1-c^*)^{t-1-k}\|E_j^k\|_2\Big)\\
\leq\;& \frac{(c^*)^2}{T-1}\left(4\frac{1-(1-c^*)^{2(T-\ell-1)}}{c^*(2-c^*)}+16\sum_{t=\ell+1}^{T-1}(1-c^*)^{t-1-\ell}\frac{1-(1-c^*)^{t-1}}{c^*}\right)\\
\leq\;& \frac{4c^*}{T-1}+\frac{16}{T-1}\leq\frac{20}{T-1}.
\end{aligned}$$

Thus, we deduce using McDiarmid's inequality, see Lemma 1.1.11, that for all $\epsilon>0$ and for any j in $[n]$:

$$|\Phi(E_j^1,\dots,E_j^{T-1}) - \mathbb{E}\Phi(E_j^1,\dots,E_j^{T-1})| \leq \frac{10\epsilon}{\sqrt{T-1}},$$

with probability larger than $1-2\exp(-\epsilon^2/2)$. Lemma 1.1.6 thus proves that the random variables $\Big(\Phi(E_j^1,\dots,E_j^{T-1}) - \mathbb{E}\Phi(E_j^1,\dots,E_j^{T-1})\Big)_{j\in[n]}$ are σ^2-subGaussian with $\sigma^2=\frac{800}{T-1}$.

Using the independence with respect to j and Hoeffding inequality for sub-Gaussian random variables, see Lemma 1.1.7, we get :

$$|T_2| \leq \frac{20\epsilon}{\sqrt{n(T-1)}}, \tag{4.26}$$

with probability larger than $1-2\exp(-\epsilon^2/4)$, for all $\epsilon>0$.

Finally, for T_3 we use the Hoeffding inequality conditionnaly on $(w_j^1)_j$. Indeed,

$$\begin{aligned} T_3 &= \frac{2c^*}{n(T-1)} \sum_{j=1}^{n} \sum_{t=2}^{T-1} \sum_{k=1}^{t-1} (1-c^*)^{2(t-1)-k} \langle w_j^1 - \tilde{\theta}^*, E_j^k \rangle \\ &= \frac{2c^*}{n(T-1)} \sum_{j=1}^{n} \sum_{k=1}^{T-1} \sum_{t=k+1}^{T-1} (1-c^*)^{2(t-1)-k} \langle w_j^1 - \tilde{\theta}^*, E_j^k \rangle \\ &= \frac{2}{n(T-1)} \sum_{j=1}^{n} \sum_{k=1}^{T-1} (1-c^*)^k \frac{1-(1-c^*)^{2(T-1-k)}}{2-c^*} \langle w_j^1 - \tilde{\theta}^*, E_j^k \rangle. \end{aligned}$$

Conditionally on $(w_j^1)_j$ the random variables $(U_j^k)_{j,k}$ are independent, centered and bounded :

$$|U_j^k| := \left| (1-c^*)^k \frac{1-(1-c^*)^{2(T-1-k)}}{2-c^*} \langle w_j^1 - \tilde{\theta}^*, E_j^k \rangle \right| \leq 4, \text{ a.s..}$$

By Hoeffding's inequality, see Lemma 1.1.8, we get that

$$\mathbb{P}\left(\frac{1}{n(T-1)} \left| \sum_{j=1}^{n} \sum_{k=1}^{T-1} \left(U_j^k - \mathbb{E}U_j^k \right) \right| \geq \epsilon \right) \leq 2\exp\left(-\frac{2\epsilon^2 n(T-1)}{64} \right).$$

This immediately leads to

$$|T_3| \leq \frac{4\sqrt{2}\epsilon}{\sqrt{n(T-1)}}, \tag{4.27}$$

with probability larger than $1 - 2\exp(-\epsilon^2/4)$. Putting together (4.25), (4.26) and (4.27), we get for all $\epsilon > 0$,

$$\begin{aligned} |U_2| &\leq |T_1| + |T_2| + |T_3| \\ &\leq \frac{1-(1-c^*)^{2(T-1)}}{c^*(2-c^*)(T-1)} \cdot \frac{\sqrt{2}\epsilon}{\sqrt{n}} + \frac{20\epsilon}{\sqrt{n(T-1)}} + \frac{4\sqrt{2}\epsilon}{\sqrt{n(T-1)}} \\ &\leq \frac{1}{\underline{c}(T-1)} \cdot \frac{\sqrt{2}\epsilon}{\sqrt{n}} + \frac{\left(20+4\sqrt{2}\right)\epsilon}{\sqrt{n(T-1)}} \end{aligned} \tag{4.28}$$

with probability larger than $1 - 6\exp(-\epsilon^2/4)$. This leads to, for all $0 < \epsilon < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$

$$\begin{aligned} \mathcal{V} &\geq |U_1| - |U_3| - |U_4| \\ \mathcal{V} &\geq m\frac{\underline{c}}{2-\underline{c}} - \frac{\epsilon+1}{\sqrt{n(T-1)}} \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1 \right) - \frac{\sqrt{2}\epsilon}{\underline{c}\sqrt{n}(T-1)} - \frac{\left(20+4\sqrt{2}\right)\epsilon}{\sqrt{n(T-1)}} \\ &\geq m\frac{\underline{c}}{2-\underline{c}} - \frac{\left(1+\sqrt{2}\right)\epsilon+1}{\underline{c}\sqrt{n}(T-1)} - \frac{\left(21+4\sqrt{2}\right)\epsilon+1}{\sqrt{n(T-1)}} \geq \frac{\underline{c}m}{4}, \end{aligned} \tag{4.29}$$

for n and T large enough, with probability larger than $1 - 6\exp(-\epsilon^2/4)$. Large enough means we need

$$\frac{2\left(1+\sqrt{2}\right)\epsilon+2}{m\underline{c}^2\sqrt{T-1}}(2-\underline{c}) + \frac{2(21+4\sqrt{2})\epsilon+2}{m\underline{c}}(2-\underline{c}) \leq \sqrt{n(T-1)}.$$

And thus, as long as $T \geq 2$, it is sufficient to have

$$\frac{2\left(1+\sqrt{2}\right)\epsilon+2}{m\underline{c}^2}\left(2-\underline{c}\right)+\frac{2(21+4\sqrt{2})\epsilon+2}{m\underline{c}}\left(2-\underline{c}\right)\leq\sqrt{n(T-1)}. \tag{4.30}$$

We conclude using (4.21), (4.23) and (4.29) that, for n and T satisfying (4.30), for all $0<\epsilon<\sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$,

$$\begin{aligned}
|\widehat{1-c}-(1-c^*)| &\leq \frac{1}{\mathcal{V}}\cdot\left[\frac{(\epsilon+1)^2}{n(T-1)}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)+\frac{11\epsilon}{\sqrt{n(T-1)}}\right],\\
&\leq \frac{4}{\underline{c}m\sqrt{n(T-1)}}\cdot\left[11\epsilon+\frac{(\epsilon+1)^2}{\sqrt{n(T-1)}}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)\right],\\
&\leq \frac{4}{\underline{c}m\sqrt{n(T-1)}}\cdot\left[11\epsilon+\frac{2(\epsilon^2+1)}{\sqrt{n(T-1)}}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)\right],
\end{aligned}$$

with probability larger than $1-15\exp(-\epsilon^2/4)$, see (4.13). ■

4.5.5 Proof of Theorem 4.4.3

Proof of Theorem 4.4.3. Using (4.10) and (4.24) we get new expressions for $\alpha^* := \|\tilde{\theta}^*\|_1$ and $\hat{\alpha}$:

$$\begin{aligned}
\alpha^* &= \frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathsf{Tr}(\mathbb{V}(w_j^t))}-1,\\
\hat{\alpha} &= \frac{\hat{c}}{2-\hat{c}}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}}-1, \quad \text{where} \quad \mathcal{V}:=\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|w_j^t-\overline{w}\right\|_2^2.
\end{aligned}$$

We decompose this difference and bound from above as follows :

$$\begin{aligned}
|\hat{\alpha}-\alpha^*| \leq &\left|\frac{\hat{c}}{2-\hat{c}}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}}-\frac{c^*}{2-c^*}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}}\right|\\
&+\left|\frac{c^*}{2-c^*}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}}-\frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathcal{V}}\right|\\
&+\left|\frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathcal{V}}-\frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathsf{Tr}(\mathbb{V}(w_j^t))}\right|.
\end{aligned}$$

Then we bound from above the three following quantities :

$$Q_1 := \left|\frac{\hat{c}}{2-\hat{c}}-\frac{c^*}{2-c^*}\right|, \quad Q_2 := \left|\|\hat{\theta}\|_2^2-\|\tilde{\theta}^*\|_2^2\right| \text{ and } Q_3 := \left|\mathcal{V}-\mathsf{Tr}(\mathbb{V}(w_j^t))\right|.$$

We first bound from above Q_1 :

$$\begin{aligned}
\left|\frac{\hat{c}}{2-\hat{c}}-\frac{c^*}{2-c^*}\right| &= \left|\frac{1}{2-\hat{c}}\left(\hat{c}-c^*\right)+c^*\left(\frac{1}{2-\hat{c}}-\frac{1}{2-c^*}\right)\right| \leq (1+c^*)\cdot|\hat{c}-c^*|\\
&\leq \frac{4\left(1+\overline{c}\right)}{\underline{c}m\sqrt{n(T-1)}}\cdot\left[11\epsilon+\frac{2(\epsilon^2+1)}{\sqrt{n(T-1)}}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)\right],
\end{aligned}$$

with probability larger than $1 - 15\exp(-\epsilon^2/4)$, see Theorem 4.4.2.

We next bound from above Q_2 :

$$\begin{aligned}\left|\|\hat{\theta}\|_2^2 - \|\tilde{\theta}^*\|_2^2\right| &= \left|\langle \hat{\theta} - \tilde{\theta}^*; \hat{\theta} + \tilde{\theta}^*\rangle\right| \leq \left\|\hat{\theta} + \tilde{\theta}^*\right\|_2 \cdot \left\|\hat{\theta} - \tilde{\theta}^*\right\|_2, \\ &\leq 2 \cdot \left[\frac{\epsilon+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right)\right],\end{aligned}$$

with probability larger than $1 - 2\exp\left(-\epsilon^2/4\right)$, see Theorem 4.4.1.

Recalling $\overline{w} := \hat{\theta}$, we then bound from above Q_3

$$\begin{aligned}\left|\mathcal{V} - \mathsf{Tr}(\mathbb{V}(w_j^t))\right| &= \frac{1}{n(T-1)}\left|\sum_{jt}\left\|w_j^t - \overline{w}\right\|_2^2 - \sum_{jt}\left\|w_j^t - \tilde{\theta}^*\right\|_2^2\right|, \\ &= \frac{1}{n(T-1)}\left|2\sum_{jt}\left\langle w_j^t; \tilde{\theta}^* - \overline{w}\right\rangle + n(T-1)\left(\|\overline{w}\|_2^2 - \left\|\tilde{\theta}^*\right\|_2^2\right)\right|, \\ &\leq 2\left\|\overline{w}\right\|_2 \cdot \left\|\tilde{\theta}^* - \overline{w}\right\|_2 + \left|\|\overline{w}\|_2^2 - \left\|\tilde{\theta}^*\right\|_2^2\right|, \\ &\leq 2\|\hat{\theta}\|_2 \cdot \left\|\tilde{\theta}^* - \hat{\theta}\right\|_2 + \|\hat{\theta} - \tilde{\theta}^*\|_2 \cdot \left(\|\hat{\theta}\|_2 + \|\tilde{\theta}^*\|_2\right), \\ &\leq \|\hat{\theta} - \tilde{\theta}^*\|_2 \cdot \left(3\|\hat{\theta}\|_2 + \|\tilde{\theta}^*\|_2\right), \\ &\leq 4\|\hat{\theta} - \tilde{\theta}^*\|_2, \\ &\leq 4 \cdot \left[\frac{\epsilon+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right)\right],\end{aligned}$$

with probability larger than $1-2\exp\left(-\epsilon^2/4\right)$, see Theorem 4.4.1. Notice that the high probability bounds on Q_2 and Q_3 are based on the same event, realised with high probability.

Those three bounds together with (4.29) allow to bound from above the distance between $\hat{\alpha}$ and α^* from above :

$$\begin{aligned}|\hat{\alpha} - \alpha^*| &= \left|\frac{\hat{c}}{2-\hat{c}} \cdot \frac{1 - \left\|\hat{\theta}\right\|_2^2}{\mathcal{V}} - \frac{c^*}{2-c^*} \cdot \frac{1 - \left\|\tilde{\theta}^*\right\|_2^2}{\mathsf{Tr}(\mathbb{V}(w_j^t))}\right|, \\ &\leq Q_1 \cdot \frac{1 - \left\|\hat{\theta}\right\|_2^2}{\mathcal{V}} + \frac{c^* Q_2}{\mathcal{V}\,(2-c^*)} + \frac{1+\alpha}{\mathcal{V}} Q_3 \\ &\leq \frac{16\,(1+\overline{c})}{\underline{c}^2 m^2 \sqrt{n(T-1)}} \cdot \left[11\epsilon + \frac{2(\epsilon^2+1)}{\sqrt{n(T-1)}}\left(1 + \frac{1}{\underline{c}\sqrt{T-1}}\right)\right] \\ &+ \frac{8\overline{c}}{\underline{c} m\,(2-\overline{c})}\left[\frac{\epsilon+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right)\right] \\ &+ \frac{16(1+\alpha)}{\underline{c} m} \cdot \left[\frac{\epsilon+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right)\right],\end{aligned}$$

with probability larger than $1 - 17\exp\left(-\epsilon^2/4\right)$. ■

Chapitre 5

Dynamic topic model

5.1 Introduction

We consider the same framework as the one exposed in Chapter 4. In this new chapter, we shift our attention to the estimation of parameters in the dynamic topic model, presented in definition 5.1.1. In this scenario, we make the assumption that both the matrix $\boldsymbol{W}^{1:T}$ and the matrix $\boldsymbol{\Pi}^{1:T}$ are not directly accessible. We use the same notation as in Chapter 4.

Definition 5.1.1 (Dynamic Topic Model) *We call Dynamic Topic Model (DTM) the model summarized by the following equations, where $t \in [T]$, $j \in [n]$ and $c^* \in (\underline{c}, \overline{c})$ and satisfying assumptions 2, 3, 4 and 5 :*

$$\begin{aligned} N\boldsymbol{Y}_j^t | \boldsymbol{W}_j^t &\sim \textit{Multinomial}_p\left(N, A^* \boldsymbol{W}_j^t\right), \\ \boldsymbol{W}_j^{t+1} &:= (1-c^*)\,\boldsymbol{W}_j^t + c^* \boldsymbol{\Delta}_j^t, \quad t \in [T-1], \\ \boldsymbol{\Delta}_j^t &\overset{\textit{i.i.d}}{\sim} \mathcal{D}(\theta^*). \end{aligned}$$

The definition entails many properties for the matrix process at hand. Indeed, the columns of the matrix $\boldsymbol{W}^1$ are assumed independent and having the stationary distribution by Assumption 5 and the noise vectors are i.i.d. imply that column vectors of $\boldsymbol{W}^t$ are independent and have the stationary distribution at any time $t \in [T]$. Also, $\boldsymbol{Y}_j^1, \ldots, \boldsymbol{Y}_j^T$ are independent given $\boldsymbol{W}_j^1, ..., \boldsymbol{W}_j^T$. This is summarized in the following Proposition.

Proposition 5.1.1 (DTM attributes) *The Dynamic Topic Model satisfies the following :*

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{W}_j^t\right] := \tilde{\theta}^* \quad \textit{and} \quad \mathbb{V}\left[\boldsymbol{W}_j^t\right] &:= \frac{c^*}{2-c^*} \cdot \Sigma(\theta^*), \\ \mathbb{P}_{(\boldsymbol{W}_1^t, ..., \boldsymbol{W}_n^t)} &:= \bigotimes_{j=1}^{n} \mathbb{P}_{\boldsymbol{W}_j^t}, \quad \textit{for all } t \in [T], \\ \mathbb{P}_{(\boldsymbol{Y}_j^1, ..., \boldsymbol{Y}_j^T)|(\boldsymbol{W}_j^1, ..., \boldsymbol{W}_j^T)} &:= \bigotimes_{t=1}^{T} \mathbb{P}_{\boldsymbol{Y}_j^t | \boldsymbol{W}_j^t}, \quad \textit{for all } j \in [n]. \end{aligned}$$

Our only available information is the word-document frequencies $Y^{1:T}$. The conditional distribution of the j^{th} column at time step t in this matrix, given $\boldsymbol{W}^{1:T}$, follows a multinomial distribution with an

expectation of $A^*\boldsymbol{W}_j^t$. For simplicity, it is presumed that all documents share the same word count, denoted as N. We still assume that the previously stated assumptions are holding true. The subsequent proposition outlines the first and second moments, as well as the conditional moments of $Y^{1:T}$ given $\boldsymbol{W}^{1:T}$.

Proposition 5.1.2 *In the model* (4.1) *under the constraints defined in* (4.2) *and* (4.3)*, we have, for all* t *in* $[T-1]$ *and* $j \in [n]$*,*

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}_j^t|\boldsymbol{W}_j^t] &= A^*\boldsymbol{W}_j^t \\
\mathbb{V}[\boldsymbol{Y}_j^t|\boldsymbol{W}_j^t] &= N^{-1}\left(\textit{diag}(A^*\boldsymbol{W}_j^t) - (A^*\boldsymbol{W}_j^t)^\top(A^*\boldsymbol{W}_j^t)\right) \\
\mathbb{E}[\boldsymbol{Y}_j^t] &= A^*\tilde{\theta}^*, \\
\mathbb{V}[\boldsymbol{Y}_j^t] &= N^{-1}\mathbb{E}[\left(\textit{diag}(A^*\boldsymbol{W}_j^t) - (A^*\boldsymbol{W}_j^t)^\top(A^*\boldsymbol{W}_j^t)\right)] + A\mathbb{V}[\boldsymbol{W}_j^t]A^\top.
\end{aligned}$$

5.2 Estimation of the word-topic matrix A^*

The estimation procedure of A^* uses the recovery procedure presented in Chapter 4 applied to the matrix of empirical frequencies $\boldsymbol{Y}^{1:T}$ instead of the true underlying $\boldsymbol{\Pi}^{1:T}$. Hence, in this case, all the previously introduced random quantities will be replaced by there empirical versions. First, $\boldsymbol{M}_* \in \mathbb{R}^{p\times p}$ defined as $\boldsymbol{M}_* := (nT)^{-1}\text{diag}\left(A^*\boldsymbol{W}^{1:T}1_{nT}\right) \in \mathbb{R}^{p\times p}$ is replaced by a data driven $\hat{M}$ as follows

$$\hat{M} := (nT)^{-1}\text{diag}\left(\boldsymbol{Y}^{1:T}1_{nT}\right) \in \mathbb{R}^{p\times p}.$$

Similarly $\boldsymbol{\Pi}_* := \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}$ is replaced by $\hat{\Pi} := \hat{M}^{-1/2}\boldsymbol{Y}^{1:T} \in \mathbb{R}^{p\times nT}$ and $\boldsymbol{R} := [\text{diag}([\boldsymbol{U}]_{.1})]^{-1}[[\boldsymbol{U}]_{.2}, \ldots, [\boldsymbol{U}]_{.K}] \in \mathbb{R}^{p\times(K-1)}$ is replaced by its empirical version $\hat{R}$ where $[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}$ are the first K left singular vectors of $\hat{\Pi}$. We recall that for all $i \in [p]$, the quantity h_i denotes the $\mathbb{L}_1$ norm of the i^{th} row of A^*. In this procedure, we need to control the noise introduced by replacing the population quantities by their sample estimates. We update the procedure with the following steps :

— *Pre-SVD Normalization :* Consider $\hat{M} := (nT)^{-1}\text{diag}\left(\boldsymbol{Y}^{1:T}1_{nT}\right) \in \mathcal{D}_p(\mathbb{R}_+^*)$ and derive $\hat{\Pi} := \hat{M}^{-1/2}\boldsymbol{Y}^{1:T} \in \mathbb{R}^{p\times nT}$.

— *SVD Computation :* $\hat{\Pi}$ is not guaranteed to have rank K, so we compute the K-SVD of $\hat{\Pi} \in \mathbb{R}^{p\times nT}$:

$$\hat{U}\hat{\Sigma}\hat{V}^\top := U_{\hat{\Pi}}^{(K)}\Sigma_{\hat{\Pi}}^{(K)}\left(V_{\hat{\Pi}}^{(K)}\right)^\top.$$

Let $[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}$ be the column vectors of $\hat{U} \in \mathbb{R}^{p\times K}$.

— *Post-SVD Normalization :* Compute $\hat{R} \in \mathbb{R}^{p\times(K-1)}$ defined as follows, for $i \in [p]$ and $k \in [K-1]$:

$$[\hat{R}]_{ik} := \frac{[\hat{U}]_{i(k+1)}}{[\hat{U}]_{i1}}.$$

This post-SVD normalization yields normalized vectors $[\hat{R}]_{1.}, \ldots, [\hat{R}]_{p.}$, the row vectors of $\hat{R}$.

— *Vertex Hunting :* We run the vertex hunting procedure as in the DETM on the estimated $[\hat{R}]_{1.}, \ldots, [\hat{R}]_{p.}$. It outputs estimated vertices $\hat{\eta}_1, \ldots, \hat{\eta}_K \in \mathbb{R}^{K-1}$. Further, we obtain $\hat{\Lambda} \in \mathbb{R}^{p\times K}$ such that for all

$i \in [p]$, $\sum_{k=1}^{K} [\hat{\Lambda}]_{ik} = 1$ and for all $i \in [p]$,

$$[\hat{R}]_{i.} = \sum_{k=1}^{K} [\hat{\Lambda}]_{ik} \hat{\eta}_k.$$

— *Topic Matrix Estimation :* Normalize each column of the matrix $\hat{M}^{1/2}\text{diag}([\hat{U}]_{.1})\Phi_{row}\left(\hat{\Lambda}_+\right)$ to derive an estimator $\hat{A}$ of the word-topic matrix A^*.

Finally, in this setting, the matrix $\hat{A}$ can be represented as

$$\hat{A} = \Phi_{col}\left(\hat{M}^{1/2}\text{diag}([\hat{U}]_{.1})\Phi_{row}\left(\hat{\Lambda}_+\right)\right). \tag{5.1}$$

Our primary objective remains to derive estimators of the autoregressive parameters, namely c^*, $\tilde{\theta}^*$, and α. To accomplish this, we follow the approach outlined in Chapter 4, using a projection of the observed $\boldsymbol{Y}^{1:T}$ onto $\hat{A}$ to derive a data-driven version $\hat{W}$ of $\boldsymbol{W}^{1:T}$. We then adapt the previously introduced estimators to this $\hat{W}$. However, in order to establish non-asymptotic convergence rates for the estimators $\hat{c}$, $\hat{\theta}$, and $\hat{\alpha}$, theoretical guarantees on the deviation of $\hat{A}$ from A^* are necessary. Specifically, we need to analyze how $\hat{M}$ deviates from $\boldsymbol{M}_*$, how $[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}$ deviate from $[\boldsymbol{U}]_{.1}, \ldots, [\boldsymbol{U}]_{.K}$, and finally, how the vertex hunting algorithm behaves with noisy entries. We adapt the theoretical analysis of [84] to our setting with random matrices and further improve their results by providing explicit constants and probability control.

We first consider a vertex hunting procedure that satisfies specific assumptions.

Assumption 8 (Vertex Hunting procedure) *When the vertex hunting algorithm is given the noisy point cloud $[\hat{R}]_{1.}, \ldots, [\hat{R}]_{p.}$, the algorithm outputs $\hat{\eta}_1, \ldots, \hat{\eta}_K$ such that, up to a permutation and for a constant $C_{VH} > 0$,*

$$\max_{k\in[K]} \|\hat{\eta}_k - \boldsymbol{\eta}_k\|_2 \leq C_{VH} \max_{i\in[p]} \left\| \left[\hat{R}\right]_{i.} - [\boldsymbol{R}]_{i.} \right\|_2 \quad \textit{a.s.}.$$

Deviation of $\hat{M}$ from $\boldsymbol{M}_*$

In this subsection we study the deviation of $\hat{M}$ from $\boldsymbol{M}_*$ in the Dynamic Topic Model framework, see Definition 5.1.1.

Proposition 5.2.1 (Estimation error of $\boldsymbol{M}_*$) *For all $i \in [p]$, for any $\epsilon > 0$, with probability at least $1 - 2\exp\left(-\epsilon^2\right)$, we have*

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < 2\epsilon\sqrt{\frac{\min(2, h_i)}{NnT}}.$$

Proof. See Proof in Subsection 5.5.1 ■

Remark 5.2.1 *Proposition 5.2.1 improves the result presented in Lemma E.1 in [84]. Specifically, by setting $\epsilon^2 = 5\log(nT)$, it establishes that for all $i \in [p]$, with probability at least $1 - 2(nT)^{-5}$, we have*

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < 2\sqrt{\frac{5h_i \log(nT)}{NnT}}.$$

Notably, unlike Lemma E.1 in [84], Proposition 5.2.1 does not require any assumption on the asymptotic behavior of $NnTh_{\min}/\log(nT)$, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided.

Corollary 5.2.2 (Estimation error of M_*) *For any $\epsilon > 0$, with probability at least $1 - 2p\exp\left(-\epsilon^2\right)$, we have*

$$\max_{i\in[p]} h_i^{-1/2} \left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < \frac{2\epsilon}{\sqrt{NnT\max(h_i/2, 1)}}.$$

Proof. See Proof in Subsection 5.5.1 ■

Next, we control for $i \in [p]$ and $k \in [K]$, the norm of the scalar products $[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}$ where $\boldsymbol{Z}^{1:T} := \boldsymbol{Y}^{1:T} - A^*\boldsymbol{W}^{1:T}$.

Proposition 5.2.3 (Concentration of cross products) *For all $i \in [p]$ and for all $k \in [K]$, for any $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$, we have*

$$\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| < 2\epsilon\sqrt{\frac{\min(2, h_i)nT}{N}}.$$

Proof. See Proof in Subsection 5.5.2 ■

Remark 5.2.2 *Proposition 5.2.3 improves the first result presented in Lemma E.2 in [84]. Specifically, by setting $\epsilon^2 = 3\log(nT)$, it establishes that for all $i \in [p]$, with probability at least $1 - 2(nT)^{-5}$, we have*

$$\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| < 2\sqrt{\frac{5h_i\log(nT)nT}{N}}.$$

Notably, unlike Lemma E.2 in [84], Proposition 5.2.3 does not require any assumption on the asymptotic behavior of $NnTh_{\min}/\log(nT)$, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided.

Corollary 5.2.4 (Concentration of cross products) *For all $k \in [K]$, for any $\epsilon > 0$, with probability at least $1 - 2p\exp\left(-\epsilon^2\right)$, we have*

$$\max_{i\in[p]} h_i^{-1/2} \left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| < 2\epsilon\sqrt{\frac{nT}{N}}.$$

Proof. See Proof in Subsection 5.5.2 ■

The following corollary gives for all $k \in [K]$, an upper bound on the norm of the vectors $\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2$.

Corollary 5.2.5 *Consider the Dynamic Topic Model, see definition 5.1.1. Then, for all $\epsilon > 0$ with probability at least $1 - 2pK\exp(-\epsilon^2)$, we have, for all $k \in [K]$,*

$$\max_{k\in[K]} \left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2 \leq 2\epsilon\sqrt{\frac{pnT}{c_2N}}.$$

Proof. See Proof in Subsection 5.5.2 ■

Remark 5.2.3 *Corollary 5.2.5 improves the second result presented in Lemma E.2 in [84]. Specifically, by setting $\epsilon^2 = 5\log(nT)$, it establishes that with probability at least $1 - 2pK(nT)^{-5}$, we have*

$$\max_{k\in[K]} \left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2 < 2\sqrt{\frac{5pnT\log(nT)}{c_2N}}.$$

Notably, unlike Lemma E.2 in [84], Proposition 5.2.3 does not require any assumption on the asymptotic behavior of $NnTh_{\min}/\log(nT)$, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided.

We now state a proposition which controls the deviation of the entries of the matrix $[\boldsymbol{Z}^{1:T}]^\top[\boldsymbol{Z}^{1:T}]$ from its expectation.

Proposition 5.2.6 *For the absolute constant $c > 0$ introduced in Lemma 1.1.10, for any $\epsilon > 0$, with probability at least $1 - 4\exp\left(2\log(p) - \min\left(\epsilon^2; \sqrt{cnT}\epsilon\right)\right)$ we have*

$$\max_{(i,m)\in[p]^2} \left| \frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}\right]}{\sqrt{h_i \cdot h_m}} \right| < \frac{576 \cdot e}{\log(2)\sqrt{c}} \cdot \frac{\epsilon\sqrt{nT}}{N \max(h_{\min}/2, 1)}.$$

Proof. See Proof in Subsection 5.5.3 ■

Remark 5.2.4 *Proposition 5.2.6 improves the results presented in Lemmas E.3 and E.4 in [84]. Specifically, by setting $\epsilon^2 = 5\log(nT)$, it establishes that for all $(i, m) \in [p]^2$, with probability at least $1 - 4(nT)^{-5}$ if $c \geq \dfrac{5\log(nT)}{nT}$, we have*

$$\left| \frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}\right]}{\sqrt{h_i \cdot h_m}} \right| \leq \frac{576 \cdot e}{\log(2)\sqrt{c}} \cdot \frac{\sqrt{5\log(nT)nT}}{N}.$$

Notably, unlike Lemmas E.3 and E.4 in [84], Proposition 5.2.6 does not require any assumption on the asymptotic behavior of $\log(nT)$, the probability of the stated event is controlled non-asymptotically. Finally, the upper bound in Proposition 5.2.6 does not contain additional terms in contrast to the ones in [84], which represents an improvement.

Finally we derive deviation bounds for the matrix

$$\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}.$$

Proposition 5.2.7 *For the absolute constant $c > 0$ introduced in Lemma 1.1.10, for any $\epsilon > 0$, with probability at least $1 - 2\exp\left(p\log(9) - \min\left(\epsilon^2, \sqrt{cnT}\epsilon\right)\right)$, we have*

$$\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\|_{op} \leq \frac{576 \cdot e}{c_2 \log(2)} \cdot \frac{\sqrt{nT}\epsilon}{N\sqrt{c}\max(h_{\min}/2, 1)}.$$

Proof. See Proof in Subsection 5.5.4 ■

Remark 5.2.5 *Proposition 5.2.7 improves the result presented in Lemmas E.5 and E.6 in [84]. Specifically, by setting $\epsilon^2 = p\log(9) + 5\log(nT)$, it establishes that with probability at least $1 - 2(nT)^{-5}$ if $c \geq \dfrac{p\log(9) + 5\log(nT)}{nT}$ we have*

$$\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\|_{op} \leq \frac{2304}{c_2} \cdot \frac{\sqrt{5nT\left(p\log(9) + 5\log(nT)\right)}}{N\sqrt{c}}.$$

Notably, unlike Lemmas E.5 and E.6 in [84], Proposition 5.2.1 does not require any assumption on either the asymptotic behavior of $\log(nT + N)$ or the asymptotic behaviour of p. Moreover, the probability of the stated event is controlled non-asymptotically. Finally, the upper bound in Proposition 5.2.6 does not contain additional terms in contrast to the ones in [84], which represents an improvement.

Deviation of the estimated singular space from the true

In this subsection, we consider the Dynamic Topic Model framework, see Definition 5.1.1 and give deviation bounds for the estimated left-singular vectors with respect to the true ones in the factorization procedure. More precisely, the vectors $[\boldsymbol{U}]_{.K}, \dots, [\boldsymbol{U}]_{.K}$ (respectively $[\hat{U}]_{.1}, \dots, [\hat{U}]_{.K}$) are the singular vectors of $\boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T}$ (respectively $\hat{M}^{-1/2} \boldsymbol{Y}^{1:T}$). We define two symmetric matrices of size $p \times p$ as follows,

$$\hat{G} := \hat{M}^{-1/2} \boldsymbol{Y}^{1:T} \left[\boldsymbol{Y}^{1:T}\right]^\top \hat{M}^{-1/2} - \frac{nT}{N} I_p, \quad \boldsymbol{G}_* := \left(1 - \frac{1}{N}\right) \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T} \left[A^* \boldsymbol{W}^{1:T}\right]^\top \boldsymbol{M}_*^{-1/2}. \quad (5.2)$$

We also consider the following $K \times K$ matrices :

$$\hat{\Phi} := \hat{A}^\top \hat{M}^{-1} \hat{A}, \quad \boldsymbol{\Phi}^* := (A^*)^\top \boldsymbol{M}_*^{-1} A^*. \quad (5.3)$$

We then consider the eigenvalues $[\lambda_1(\boldsymbol{G}_*), \dots, \lambda_{\min}(\boldsymbol{G}_*)]$ of $\boldsymbol{G}_*$ and $[\lambda_1(\hat{G}), \dots, \lambda_{\min}(\hat{G})]$ of $\hat{G}$. We notice that Theorem 4.3.3 ensures that $\operatorname{rank} \boldsymbol{G}_* = K$ almost surely and thus $\lambda_{\min}(\boldsymbol{G}_*) > 0$ almost surely. In addition we notice that $\boldsymbol{G}_* = \boldsymbol{\Pi}_* \boldsymbol{\Pi}_*^\top$ and $\hat{G} = \hat{\Pi} \hat{\Pi}^\top$. Hence $[\hat{U}]_{.1}, \dots, [\hat{U}]_{.K}$ are the eigenvectors of $\hat{G}$ and $[\boldsymbol{U}]_{.K}, \dots, [\boldsymbol{U}]_{.K}$ are the eigenvectors of $\boldsymbol{G}_*$.

Proposition 5.2.8 *Consider the Assumption 7. Then the following inequality holds almost surely for all* $i \in [p]$*,*

$$c_2 h_i \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

Proof. See Proof in Subsection 5.5.5 ■

Proposition 5.2.9 *Consider the matrix* $\boldsymbol{G}_*$ *in* (5.2) *and Assumption 7. Then* $\operatorname{rank}(\boldsymbol{G}_*) = K$ *and its eigenvalues satisfy almost surely the following inequalities,*

$$\left(1 - \frac{1}{N}\right) \frac{nTK}{c_2} \geq \lambda_1(\boldsymbol{G}_*) \quad \textit{and} \quad \lambda_K(\boldsymbol{G}_*) \geq \left(1 - \frac{1}{N}\right) nTc_2^2 \quad \textit{and} \quad \lambda_1(\boldsymbol{G}_*) \geq \left(1 - \frac{1}{N}\right) nTc_3 + \lambda_2(\boldsymbol{G}_*).$$

Proof. See Proof in Subsection 5.5.6 ■

Remark 5.2.6 *Proposition 5.2.9 extends the first result presented in Lemma F.2 in [84]. Notably, Proposition 5.2.9 does not require any assumption on either the asymptotic behavior of* $\log^2(nT)$ *or the asymptotic behaviour of* $p \log(nT)$*. Moreover, the constants are explicitly provided.*

Proposition 5.2.10 *Consider the Assumption 7. We denote* $[\boldsymbol{U}]_{.K}, \dots, [\boldsymbol{U}]_{.K}$ *the eigenvectors of* $\boldsymbol{G}_*$ *and* $U = [[\boldsymbol{U}]_{.K}, \dots, [\boldsymbol{U}]_{.K}] \in \mathbb{R}^{p \times K}$*. Then for all* $i \in [p]$*, we have almost surely,*

$$\|U_{i.}\|_2 \leq \frac{\sqrt{Kh_i}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}.$$

Proof. See Proof in Subsection 5.5.7 ■

Remark 5.2.7 *Proposition 5.2.10 improves the result presented in Lemma F.3 in [84]. Notably, Proposition 5.2.10 does not require any assumption on either the asymptotic behavior of* $\log^2(nT)$ *or the asymptotic behaviour of* $p \log(nT)$*. Moreover, the constants are explicitly provided.*

Theorem 5.2.11 *Consider the Assumptions 6 and 7. Then for N, n and T large enough, for all $i \in [p]$ and for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2K\exp(-\epsilon_2^2) - 2pK\exp(-\epsilon_3^2) - 4p\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$, the quantity $h_i^{-1/2}\left\|e_i^\top\left(\hat{G} - \boldsymbol{G}_*\right)\right\|_2$ is bounded from above by*

$$2\sqrt{\frac{nTp}{N}}\left[\epsilon_1\left(2\frac{\sqrt{p}}{Nc_1c_2K} + 4\frac{K}{c_2^2} + 3\frac{\sqrt{K}}{c_2^2\sqrt{c_1}}\right) + 4\frac{\epsilon_3 + \epsilon_2\sqrt{K/c_1}}{c_2} + 4\epsilon_4\frac{288\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\sqrt{p}}{c_2\sqrt{Nc_1K}}\right],$$

where c is an absolute constant appearing in Lemma 1.1.10. In addition, N, n and T large enough means :

$$NnT \geq \epsilon_1^2\max\left(\frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right).$$

Proof. See Proof in Subsection 5.5.8 ■

Remark 5.2.8 *We set $\epsilon_1^2 = \epsilon_4^2 = \log(nTp)$, $\epsilon_2^2 = \log(nTK)$ and $\epsilon_3^2 = \log(nTpK)$ and we assume $c \geq \frac{\log(nTp)}{nT}$. We notice that for N, n and T large enough the sample size conditions of Theorem 5.2.11 is fulfilled :*

$$NnT \geq \log(nTp)\max\left(\frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right).$$

Then, there exists χ, a positive constant only depending on K such that for all $i \in [p]$, with probability at least $1 - \frac{10}{nT}$:

$$h_i^{-1/2}\left\|e_i^\top\left(\hat{G} - \boldsymbol{G}_*\right)\right\|_2 \leq \chi\sqrt{\frac{nTp\log(nTp)}{N}}\left(\sqrt{\frac{p}{N}} + 1\right).$$

This convergence rate matches with the one stated in Lemma F.4 in [84].

Theorem 5.2.12 *Consider the Assumptions 6 and 7. Then for N, n and T large enough, for all $i \in [p]$ and for any $\epsilon_1, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p\exp\left(-\epsilon_1^2\right) - 2pK\exp\left(-\epsilon_3^2\right) - 2\cdot 9^p\exp\left(-\min\left(\epsilon_4^2, \sqrt{cnT}\epsilon_4\right)\right)$, the quantity $\left\|\left(\hat{G} - \boldsymbol{G}_*\right)\right\|_{op}$ is bounded from above by*

$$\frac{4\epsilon_1\sqrt{nTp}}{Nc_2\sqrt{Nc_1K}} + \frac{8\epsilon_3K\sqrt{nTp}}{c_2\sqrt{N}} + \frac{4\epsilon_4\sqrt{nT}}{N}\cdot\frac{288\cdot e}{c_2\log(2)\sqrt{c}} + \frac{24\epsilon_1\sqrt{nTp}K^2}{c_2^2\sqrt{N}}.$$

where c is an absolute constant appearing in Lemma 1.1.10. In addition, N, n and T large enough means :

$$NnT \geq \epsilon_1^2\max\left(\frac{16p^2}{c_2^4c_1K}, \frac{4pK}{c_2^3}\right).$$

Proof. See Proof in Subsection 5.5.9 ■

Remark 5.2.9 *We set $\epsilon_1^2 = \log(nTp)$, $\epsilon_3^2 = \log(nTpK)$ and $\epsilon_4^2 = \log(nT) + p\log(9)$ and we assume $c \geq \frac{\log(nT) + p\log(9)}{nT}$. We notice that for N, n and T large enough the sample size conditions of Theorem 5.2.12 is fulfilled :*

$$NnT \geq \log(nTp)\max\left(\frac{16p^2}{c_2^4c_1K}, \frac{4pK}{c_2^3}\right).$$

Then, there exists χ, a positive constant only depending on K such that, with probability at least $1-\frac{6}{nT}$:

$$\left\|\left(\hat{G}-\boldsymbol{G}_*\right)\right\|_{op} \leq \chi\sqrt{\frac{nTp\log(nT)}{N}}.$$

This convergence rate matches with the one stated in Lemma F.5 in [84].

We now derive a large-deviation bound for singular vectors. We recall that $\hat{U} = \left[[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}\right]$ contains the first K left singular vectors of the noisy quantity $\hat{\Pi}$. Their population counterparts are denoted respectively U and $\mathbf{\Pi}_*$. For any matrix M we denote $[M]_{i.}$ the i^{th} row of M.

Theorem 5.2.13 *Consider Assumptions 6 and 7. Then there exists a matrix $\Omega = diag(\omega, \Omega_{2:K}) \in \mathbb{R}^{K\times K}$ where $\omega \in \{-1,1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ is an orthogonal matrix such that for N, n and T large enough, for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ satisfying*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)},$$

for all $i \in [p]$ and with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)-2\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$ we have

$$\left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2 \leq C_{tot}(p,N)\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\sqrt{\frac{h_i p}{nT(N-2)}},$$

where

$$C_{tot}(p,N) \leq \frac{40K^{3/2}}{c_2^4\min\left(c_3,c_2^2\right)}\left(\frac{K+2+\sqrt{2K/c_1}+2K^2c_2^{-1}+2880(cpN)^{-1/2}+N^{-1}}{c_2}+2+2\sqrt{K/c_1}+\frac{\sqrt{p}}{Nc_1K}+\sqrt{\frac{Kp}{c_1N}}\right).$$

Moreover, N, n and T large enough means :

$$NnT \geq \epsilon_1^2\max\left(\frac{36K}{c_2^3};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right).$$

Proof. See Proof in Subsection 5.5.10 ■

We note that the dependency of $C_{tot}(p,N)$ is of the order of magnitude of $\frac{K^{3/2}}{c_2^4\min\left(c_3,c_2^2\right)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)$.

Remark 5.2.10 *We set $\epsilon_1^2 = \log(nTp)$, $\epsilon_2^2 = \log(nTK)$, $\epsilon_3^2 = \log(nTpK)$ and $\epsilon_4^2 = \log(nT)+\log(2p+9^p) \leq \log(nT)+p$ once $p \geq 2$ and we assume $c \geq \frac{\log(nT)+p}{nT}$. We notice that for N, n and T large enough the sample size conditions of Theorem 5.2.13 are fulfilled :*

$$\begin{aligned} nT &\geq \psi\cdot(\log(nT)+p)\,(\sqrt{p}+1),\\ NnT &\geq \log(nTp)\max\left(\frac{36K}{c_2^3};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right),\end{aligned}$$

for any positive constant ψ only depending on K. Then, there exists χ, a positive constant only depending on K such that for all $i \in [p]$, with probability at least $1 - \frac{8}{nT}$:

$$\left\|\Omega[\hat{U}]_{i.} - [U]_{i.}\right\|_2 \leq \chi\sqrt{\frac{h_i p(\log(nT)+p)}{nT(N-2)}}\left(1+\sqrt{\frac{p}{N}}\right).$$

This convergence rate matches with the one stated in Theorem 3.1 in [84].

Proposition 5.2.14 (Rows of R Lie in a Simplex) *There exist K vectors of $\mathbb{R}^{(K-1)}$ denoted $\eta_1, \ldots, \eta_K$ such that the matrix $R \in \mathbb{R}^{p\times(K-1)}$ defined in the Post-SVD Normalization step in Chapter 4 has its rows embedded in G_η, where*

$$G_\eta := \left\{x \in \mathbb{R}^{K-1} : x = \sum_{k=1}^{K}\alpha_k\eta_k,\ \forall k \in [K],\ \alpha_k \geq 0,\ \sum_{k=1}^{K}\alpha_k = 1\right\}.$$

Furthermore, we denote $N := [\eta_1, \ldots, \eta_K]^\top \in \mathbb{R}^{K\times(K-1)}$.

Proof. See proof in Subsection 5.5.11 ■

Theorem 5.2.15 *Consider the Assumptions 6 and 7. Consider the matrices R and $\hat{R}$ defined in the Post-SVD Normalization step. Then, for N, n and T large enough, for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ satisfying*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3, c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N} + 2K\sqrt{p} + \frac{576e}{\log(2)\sqrt{Nc}} + \frac{4K^2}{c_2}\sqrt{p}\right)},$$

for all $i \in [p]$, with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2K\exp(-\epsilon_2^2) - 2pK\exp(-\epsilon_3^2) - 2\cdot(2p + 9^p)\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$, there exists $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$, an orthogonal matrix, such that

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 \leq \left(2\frac{C_{tot}(p,N)\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{c_1c_2^{9/2}K}\frac{p^{3/2}}{\sqrt{nT(N-2)}}\right)\left(2+\frac{p}{c_2^5c_1K}\right),$$

with $C_{tot}(p,N)$ defined in Theorem 5.2.13. Moreover, N, n and T large enough means :

$$NnT \geq \epsilon_1^2\max\left(\frac{36K}{c_2^3};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right) \quad \textit{and} \quad (N-2)nT \geq C_{tot}(p,N)^2\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)^2\frac{p^3}{c_2^9c_1^2K^2}.$$

Proof. See proof in Subsection 5.5.12 ■

Remark 5.2.11 *We set $\epsilon_1^2 = \log(nTp)$, $\epsilon_2^2 = \log(nTK)$, $\epsilon_3^2 = \log(nTpK)$ and $\epsilon_4^2 = \log(nT) + \log(2p + 9^p) \leq \log(nT) + p$ once $p \geq 2$. We notice that for N, n and T large enough the sample size conditions of Theorem 5.2.15 are fulfilled :*

$$\begin{aligned}
nT &\geq \psi\cdot(\log(nT)+p)\,(\sqrt{p}+1),\\
(N-2)nT &\geq \psi\cdot p^3\,(\log(nT)+p)\left(\frac{p}{N}+1\right),\\
NnT &\geq \psi\cdot\log(nTp)\max\left(\frac{36K}{c_2^3};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right),
\end{aligned}$$

for any positive constant ψ only depending on K. Then, there exists χ, a positive constant only depending on K such that for all $i \in [p]$, with probability at least $1 - \frac{8}{nT}$:

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 \leq \chi p\sqrt{\frac{p\log(nT)+p^2}{nT(N-2)}}\left(1+\sqrt{\frac{p}{N}}\right)(1+p)\,.$$

This convergence rate matches with the one stated in Theorem 3.2 in [84].

Theorem 5.2.16 *Consider the Assumptions 6 and 7. Let $\hat{A}$ be the estimator of A^* defined in* (5.1)*. Let $\mathcal{D}_K$ be the set of matrices $\Omega = diag(\omega, \Omega_{2:K}) \in \mathbb{R}^{K\times K}$ where $\omega \in \{-1,1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ is an orthogonal matrix. Then, for N, n and T large enough, for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ satisfying*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)},$$

for all $i \in [p]$, with probability at least $1 - 2p^2\exp(-\epsilon_1^2) - 2pK\exp(-\epsilon_2^2) - 2p^2K\exp(-\epsilon_3^2) - 2p\cdot(2p + 9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$ and up to a permutation of columns of $\hat{A}$ we have

$$\max_{i\in[p]}\left(\frac{\left\|\,[\hat{A}]_{i.} - [A^*]_{i.}\right\|_1}{h_i}\right) \leq C_A(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{\frac{p}{nT(N-2)}} + \epsilon_1\frac{C_B}{\sqrt{NnT}}.$$

Moreover, N, n and T large enough imply that :

$$\begin{aligned}
NnT &\geq \epsilon_1^2\max\left(\frac{36K}{c_2^3};\ \frac{4K}{c_2^2};\ \frac{4p}{c_1K};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right),\\
\sqrt{NnT} &\geq \frac{4}{c_2^{9/2}c_1}\left[C_{tot}(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{p}\left[1+\left(2+\frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}}+\frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right]+\frac{2\sqrt{K}\epsilon_1}{c_2}\right],\\
(N-2)nT &\geq C_{tot}(p,N)^2\max(\epsilon_i)_{i\in[4]}^2p^3\max\left(\frac{1}{c_2^9c_1^2K^2};\ \frac{4K^2C_{VH}^2}{c_2^2}\left(2+\frac{p}{c_2^5c_1K}\right)^2\right).
\end{aligned}$$

In addition $C_{tot}(p,N)$ is bounded from above in Theorem 5.2.13 and the quantities $C_A(p,N)$ and $C_B(p,N)$ are defined as follows :

$$C_A(p,N) := C_{tot}(p,N)\frac{4\sqrt{2}}{\left(c_2^{9/2}c_1\right)}\left[1+\left(2+\frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}}+\frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right],\quad C_B := \frac{8\sqrt{2K}}{c_2^{11/2}c_1}.$$

Proof. See proof in Subsection 5.5.13 ■

We note that the dependency of $C_A(p,N)$ is of the order of magnitude of $\frac{K^{5/2}p^2}{c_2^{21}c_1^3\min\left(c_3,c_2^2\right)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)$.

Remark 5.2.12 *We set $\epsilon_1^2 = 2\log(nTp)$, $\epsilon_2^2 = \log(nTpK)$, $\epsilon_3^2 = 2\log(nTpK)$ and $\epsilon_4^2 = \log(nT)+\log(2p^2+p9^p) \leq \log(nT)+p$ once $p \geq 34$. We notice that for N, n and T large enough the sample size conditions*

of Theorem 5.2.16 are fulfilled :

$$\begin{aligned} nT &\geq \psi \cdot (\log(nT) + p)\,(\sqrt{p} + 1), \\ (N-2)nT &\geq \psi \cdot p^3\,(\log(nT) + p)\left(\frac{p}{N} + 1\right)(1+p)^2, \\ NnT &\geq \psi \cdot \log(nTp) \max\left(\frac{36K}{c_2^3};\ \frac{4K}{c_2^2};\ \frac{4p}{c_1 K};\ \frac{64p}{c_2^4 c_1 K};\ \frac{16}{c_2^2 h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3 h_{\min}^3}\right), \\ \sqrt{NnT} &\geq \psi \cdot \sqrt{p}\,(\log(nT) + p)\left(\sqrt{\frac{p}{N}} + 1\right)\left(1 + \max_{x \in \mathcal{G}_\eta} \|x\|_2\right), \end{aligned}$$

for any positive constant ψ only depending on K. Then, there exists χ, a positive constant only depending on K such that with probability at least $1 - \dfrac{8}{nT}$:

$$\max_{i \in [p]} \left(\frac{\left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{h_i} \right) \leq \chi \sqrt{\frac{p \log(nT) + p^2}{nT(N-2)}} p \left(1 + \sqrt{\frac{p}{N}}\right)(1+p)(1 + \max_{x \in \mathcal{G}_\eta} \|x\|_2).$$

This convergence rate matches with the one stated in Lemma G.1 in [84].

Theorem 5.2.17 *Consider the Assumptions 6, 7 and 8. Let $\hat{A}$ be the estimator of A^* defined in* (5.1)*. Then, under the same conditions and with the same notations as in Theorem 5.2.16 we have*

$$\sum_{i=1}^{p} \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \leq K C_A(p, N) \max(\epsilon_i)_{i \in [4]} \sqrt{\frac{p}{nT(N-2)}} + \epsilon_1 \frac{K C_B}{\sqrt{NnT}},$$

with probability at least $1 - 2p^2 \exp(-\epsilon_1^2) - 2pK \exp(-\epsilon_2^2) - 2p^2 K \exp(-\epsilon_3^2) - 2p \cdot (2p + 9^p) \exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$.

Proof. See proof in Subsection 5.5.14 ■

Remark 5.2.13 *We set $\epsilon_1^2 = 2\log(nTp)$, $\epsilon_2^2 = \log(nTpK)$, $\epsilon_3^2 = 2\log(nTpK)$ and $\epsilon_4^2 = \log(nT) + \log(2p^2 + p9^p) \leq \log(nT) + p$ once $p \geq 34$. We notice that for N, n and T large enough the sample size conditions of Theorem 5.2.17 are fulfilled, see the remark under Theorem 5.2.16. Then, there exists χ, a positive constant only depending on K such that with probability at least $1 - \dfrac{8}{nT}$:*

$$\sum_{i=1}^{p} \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \leq \chi \sqrt{\frac{p \log(nT) + p^2}{nT(N-2)}} p \left(1 + \sqrt{\frac{p}{N}}\right)(1+p)(1 + \max_{x \in \mathcal{G}_\eta} \|x\|_2).$$

This convergence rate matches with the one stated in Theorem 3.3 in [84].

5.3 Estimation of the topic-document matrix

This section is devoted to giving a proxy random matrix $W^{1:T}$ of the unobserved $\boldsymbol{W}^{1:T}$, once an estimator $\hat{A}$ of A^* is derived, see Section 5.2. First, let us denote

$$\boldsymbol{\Phi}^* := (A^*)^\top \boldsymbol{M}_*^{-1} A^* \quad \text{and} \quad \hat{\Phi} := \hat{A}^\top \hat{M}^{-1} \hat{A}.$$

Notice that for all $j \in [n]$ and for all $t \in [T]$,

$$\begin{aligned}\boldsymbol{W}_j^t &= \left((A^*)^\top \boldsymbol{M}_*^{-1} A^*\right)^{-1} \cdot \left((A^*)^\top \boldsymbol{M}_*^{-1} \boldsymbol{\Pi}_j^t\right), \\ &= (\boldsymbol{\Phi}^*)^{-1} \cdot \left((A^*)^\top \boldsymbol{M}_*^{-1} \boldsymbol{\Pi}_j^t\right).\end{aligned}$$

This motivates to define for all $j \in [n]$ and for all $t \in [T]$,

$$\begin{aligned}\tilde{W}_j^t &= \left(\hat{A}^\top \hat{M}^{-1} \hat{A}\right)^{-1} \cdot \left(\hat{A}^\top \hat{M}^{-1} \boldsymbol{Y}_j^t\right), \\ &= \left(\hat{\Phi}\right)^{-1} \cdot \left(\hat{A}^\top \hat{M}^{-1} \boldsymbol{Y}_j^t\right).\end{aligned}$$

However, each $\boldsymbol{W}_j^t$ lies in the K dimensional simplex non-negative entries and unit $\mathbb{L}_1$ norm. Hence we derive for all $j \in [n]$ and for all $t \in [T]$ the estimator $\hat{W}_j^t$ by setting negative entries of $\tilde{W}_j^t$ to zero and normalizing it to have a unit $\mathbb{L}_1$ norm.

Theorem 5.3.1 *For every $t \in [T]$ and for every $j \in [n]$, for N, n and T large enough, for any $(\epsilon_i)_{i\in[5]} \in (\mathbb{R}_+^*)^5$ satisfying*

$$\max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \leq \sqrt{nT} \frac{c_2^3 \min\left(c_3, c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N} + 2K\sqrt{p} + \frac{576e}{\log(2)\sqrt{Nc}} + \frac{4K^2}{c_2}\sqrt{p}\right)},$$

with probability at least $1 - 2p^2\exp(-\epsilon_1^2) - 2pK\exp(-\epsilon_2^2) - 2Kp^2\exp(-\epsilon_3^2) - 2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right) - 2K\exp(-\epsilon_5^2)$, we have :

$$\left\|\hat{W}_j^t - \boldsymbol{W}_j^t\right\|_1 \leq \frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}},$$

where $\nu_1(p,N) := 32K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p} + C_B\right]\left[\frac{2+c_2}{c_2^2} + \frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^4}\right]$, $\nu_2 := \frac{4\sqrt{2}K^{3/2}}{c_2^2}$ and $\nu_3 := \frac{16K^3}{c_2^5}$. Moreover, N, n and T large enough imply that :

$$\begin{aligned}
NnT &\geq \epsilon_1^2 \max\left(\frac{36K}{c_2^3};\ \frac{4K}{c_2^2};\ \frac{4p}{c_1K};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2h_{\min}^2};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right), \\
\sqrt{NnT} &\geq \frac{4}{c_2^{9/2}c_1}\left[C_{tot}(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{p}\left[1 + \left(2 + \frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}} + \frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right] + \frac{2\sqrt{K}\epsilon_1}{c_2}\right], \\
nT(N-2) &\geq 8K^{3/2}\max(\epsilon_i^2)_{i\in[4]}\left[C_A(p,N)\sqrt{p} + C_B\right]\sqrt{p}, \\
(N-2)nT &\geq C_{tot}(p,N)^2\max(\epsilon_i)_{i\in[4]}^2 p^3 \max\left(\frac{1}{c_2^9c_1^2K^2};\ \frac{4K^2C_{VH}^2}{c_2^2}\left(2 + \frac{p}{c_2^5c_1K}\right)^2\right), \\
\sqrt{nT(N-2)} &\geq \frac{4K^{3/2}\epsilon_1}{c_2^2} + \frac{\sqrt{16K^3\epsilon_1^2/c_2^4 + 32\left[2\sqrt{K}+1\right]\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\left[C_A(p,N)\sqrt{p} + C_B\right]\sqrt{p}}}{c_2}.
\end{aligned}$$

Proof. See proof in Subsection 5.5.15 ■

We note that the dependency of $\nu_1(p,N)$ is of the order of magnitude of $\frac{p^3K^6}{c_2^{21}c_1^3\min\left(c_3,c_2^2\right)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)$ and the sample size condition can be restated, in order of magnitude, as :

$$\sqrt{NnT}\geq\frac{p^{5/2}K^{5/2}}{c_2^{21}c_1^3\min(c_3,c_2^2)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right),\quad nT(N-2)\geq\max(\epsilon_i^2)_{i\in[4]}\frac{K^4p^3}{c_2^{21}c_1^3\min\left(c_3,c_2^2\right)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right).$$

Remark 5.3.1 *We set* $\epsilon_1^2=2\log(nTp)$, $\epsilon_2^2=\log(nTpK)$, $\epsilon_3^2=2\log(nTpK)$, $\epsilon_5^2=\log(nTK)$ *and* $\epsilon_4^2=\log(nT)+\log(2p^2+p9^p)\leq\log(nT)+p$ *once* $p\geq34$. *We notice that for* N, n *and* T *large enough the sample size conditions of Theorem 5.3.1 are fulfilled. Then, with probability at least* $1-\frac{8}{nT}$, *in order of magnitude there is :*

$$\left\|\hat{W}_j^t-\boldsymbol{W}_j^t\right\|_1\leq\frac{p^3K^6(\log(nT)+p)}{c_2^{21}c_1^3\min\left(c_3,c_2^2\right)nTN}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)+\frac{K^{3/2}\sqrt{\log(nTK)}}{c_2^2\sqrt{N}}+\frac{K^3\sqrt{\log(nTp)}}{c_2^5\sqrt{NnT}}.$$

5.4 Estimation of the underlying parameters of the autoregressive model

This subsection is devoted to the estimation of the parameters of the autoregressive model (4.3) once an estimator $\hat{W}^{1:T}$ of the realization $W^{1:T}$ of $\boldsymbol{W}^{1:T}$ is defined, as in the previous subsection. Similarly as in Chapter 4, we define the estimators of $\tilde{\theta}^*$, c^* and α via the method of moments. We define $\hat{\theta}$ as the empirical mean of the estimated $\left(\hat{W}_j^t\right)_{j,t}$:

$$\hat{\theta}:=\frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\hat{W}_j^t. \tag{5.4}$$

We estimate $1-c^*$ by the normalized sum of scalar products between the centered consecutive vectors $\hat{W}_j^{t+1}-\overline{\hat{W}^{+1}}$ and $\hat{W}_j^t-\overline{\hat{W}}$:

$$\widehat{(1-c)}:=\frac{\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\langle\hat{W}_j^{t+1}-\overline{\hat{W}^{+1}};\ \hat{W}_j^t-\overline{\hat{W}}\right\rangle}{\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|\hat{W}_j^t-\overline{\hat{W}}\right\|_2^2}, \tag{5.5}$$

where $\overline{\hat{W}^{+1}}:=\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\hat{W}_j^{t+1}$ and $\overline{\hat{W}}:=\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\hat{W}_j^t=\hat{\theta}$.

Finally, using the variance of the stationary sequence $\boldsymbol{W}_j^t$ and the explicit expression of the matrix Σ, we see that :

$$\mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))=\frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\alpha+1}.$$

Thus, we plug-in estimators $\hat{\theta}$, $\hat{c}$ and the empirical variance to get

$$\hat{\alpha} = \frac{2-\hat{c}}{\hat{c}\cdot(1-\|\hat{\theta}\|_2^2)} \cdot \frac{1}{n(T-1)} \sum_j \sum_t \mathsf{Tr}((\hat{W}_j^t - \overline{\hat{W}})(\hat{W}_j^t - \overline{\hat{W}})^\top). \tag{5.6}$$

Next we give the convergence rates of these three estimators.

Theorem 5.4.1 (Estimation of $\tilde{\theta}^*$) *In the DTM model, under the Assumptions 2, 3, 4 and 5, the estimator $\hat{\theta}$ defined in* (5.4) *is such that for N, n and T large enough, for every $(\epsilon_i)_{i\in[6]} \in (\mathbb{R}_+^*)^6$ satisfying $\epsilon_6 < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$ and*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)},$$

we have

$$\begin{aligned}\left\|\hat{\theta}-\tilde{\theta}^*\right\|_2 \leq & \frac{\epsilon_6+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right)+\epsilon_5\frac{4\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}}+\epsilon_1\frac{16K^3}{c_2^5\sqrt{NnT}} \\ & + \frac{32\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right].\end{aligned}$$

with probability larger than $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)-2\exp\left(-\epsilon_6^2/4\right)$. *Moreover the condition on N, n and T is similar to the one stated in Theorem 5.3.1.*

Proof. Let us denote $\bar{\theta}$ the empirical mean of the unobserved $\boldsymbol{W}_j^t$. Then use the triangle inequality to get :

$$\left\|\hat{\theta}-\tilde{\theta}^*\right\|_2 \leq \left\|\hat{\theta}-\bar{\theta}\right\|_2 + \left\|\bar{\theta}-\tilde{\theta}^*\right\|_2.$$

The quantity $\left\|\bar{\theta}-\tilde{\theta}^*\right\|_2$ is bounded from above in Theorem 4.4.1 and the $\mathbb{L}_1$-$\mathbb{L}_2$ inequality combined with the triangle inequality ensure that the quantity $\left\|\hat{\theta}-\bar{\theta}\right\|_2$ is bounded from above by $\frac{1}{nT}\sum_{j=1}^n\sum_{t=1}^T\left\|\left(\hat{W}_j^t-\boldsymbol{W}_j^t\right)\right\|_1$. We conclude using Theorem 5.3.1. ■

Remark 5.4.1 *We set $\epsilon_1^2 = 2\log(nTp)$, $\epsilon_2^2 = \log(nTpK)$, $\epsilon_3^2 = 2\log(nTpK)$, $\epsilon_5^2 = \log(nTK)$, $\epsilon_4^2 = \log(nT)+\log(2p^2+p9^p) \leq \log(nT)+p$ once $p \geq 34$ and $\epsilon_6^2 = 4\log(nT)$. Then, for N, n and T large enough with probability at least $1-\frac{12}{nT}$, in order of magnitude there is :*

$$\begin{aligned}\left\|\hat{\theta}-\tilde{\theta}^*\right\|_2 \leq & \frac{\sqrt{\log(nT)}}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right)+\frac{K}{c_2^2}\sqrt{\frac{K\log(nTK)}{N}}+\frac{\sqrt{\log(nTp)}K^3}{c_2^5\sqrt{NnT}} \\ & + \frac{[\log(nT)+p]K^6p^3\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)}{c_2^{23}c_1^3\min\left(c_3,c_2^2\right)nT(N-2)}\left[c_2+\frac{K^2}{c_2^2}\right].\end{aligned}$$

This rate of convergence can be compared to the one obtained in Theorem 4.4.1. In the oracle DETM, the convergence rate is of order $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}}\right)$ *with probability at least* $1-\frac{2}{nT}$*. In the realistic DTM, fixing the number of topics* K *and the vocabulary size* p*, the rate of convergence is of order* $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}}+\sqrt{\frac{\log(nT)}{N}}\right)$ *with probability at least* $1-\frac{12}{nT}$*. Hence the probability control is weaker by a constant factor in the DTM and an extra term* $\sqrt{\frac{\log(nT)}{N}}$ *appears due to the multinomial noise. We underline that the other terms come from the estimation error of* A^**.*

Theorem 5.4.2 (Estimation of c^*) *In the DTM, under the Assumptions 2, 3, 4 and 5, the estimator* $\widehat{(1-c)}$ *defined in* (5.5) *is such that for* n *and* T *large enough, for any* $(\epsilon_i)_{i\in[7]} \in (\mathbb{R}_+^*)^7$ *satisfying* $\max(\epsilon_6,\epsilon_7) < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$ *and*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)},$$

we have

$$\begin{aligned}|\widehat{(1-c)}-(1-c^*)| \leq{}& \frac{64(1-c^*)}{\underline{c}m}\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+\frac{\nu_2\epsilon_5}{\sqrt{N}}+\frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)\\ &+\frac{8c^*}{\underline{c}m}\left[\frac{(\epsilon_7+1)^2}{n(T-1)}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)+\frac{11\epsilon_7}{\sqrt{n(T-1)}}\right],\end{aligned}$$

with probability larger than $1-2n(T-1)p^2\exp(-\epsilon_1^2)-2n(T-1)pK\exp(-\epsilon_2^2)-2n(T-1)Kp^2\exp(-\epsilon_3^2)-2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2n(T-1)K\exp(-\epsilon_5^2)-2n(T-1)\exp\left(-\epsilon_6^2/4\right)-13\exp(-\epsilon_7^2/4)$*. Moreover,* N*,* n *and* T *large enough means :*

$$\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+\frac{\nu_2\epsilon_5}{\sqrt{N}}+\frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)\leq 2,$$

where $\nu_1(p,N)$*,* ν_2 *and* ν_3 *are defined in Theorem 5.3.1.*

Proof. See proof in Subsection 5.5.16 ■

Remark 5.4.2 *We set* $\epsilon_1^2=2\log(nTp)$*,* $\epsilon_2^2=2\log(nTpK)$*,* $\epsilon_3^2=2\log(nTpK)$*,* $\epsilon_5^2=2\log(nTK)$*,* $\epsilon_4^2=2\log(nT)+2\log(2p^2+p9^p)\leq 2\log(nT)+2p$ *once* $p\geq 34$*,* $\epsilon_6^2=8\log(nT)$ *and* $\epsilon_7^2=4\log(nT)$*. Then, for* N*,* n *and* T *large enough with probability at least* $1-\frac{25}{nT}$*, in order of magnitude there is :*

$$\begin{aligned}|\widehat{(1-c)}-(1-c^*)| \leq{}& \frac{(1-c^*)}{\underline{c}m}\left(\frac{p^3K^6(\log(nT)+p)}{NnTc_2^{21}c_1^3\min\left(c_3,c_2^2\right)}\left(\sqrt{\frac{Kp}{c_1N}}+K^2\right)+\frac{\sqrt{\log(nTK)}K^{3/2}c_2^2}{\sqrt{N}}+\frac{K^3\sqrt{\log(nTp)}}{c_2^5\sqrt{NnT}}\right)\\ &+\frac{c^*}{\underline{c}m}\left[\frac{\log(nT)}{n(T-1)}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)+\sqrt{\frac{\log(nT)}{n(T-1)}}\right].\end{aligned}$$

This rate of convergence can be compared to the one obtained in Theorem 4.4.2. In the oracle case, the convergence rate is of order $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}}\right)$ *with probability at least* $1-\frac{15}{nT}$*. In the real case, fixing the number of topics* K *and the vocabulary size* p*, the rate of convergence is of order* $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}}+\sqrt{\frac{\log(nT)}{N}}\right)$ *with probability at least* $1-\frac{25}{nT}$*. Again, we see that for a probability control of the same order, an extra term* $\sqrt{\frac{\log(nT)}{N}}$ *appearing the upper bounds due to the multinomial noise.*

Theorem 5.4.3 (Estimation of α) *In the DTM, under the Assumptions 2, 3, 4 and 5, the estimator* $\hat{\alpha}$ *defined in* (5.6) *is such that for* n *and* T *large enough, for any* $(\epsilon_i)_{i\in[7]}\in(\mathbb{R}_+^*)^7$ *satisfying* $\max(\epsilon_6,\epsilon_7)<\sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$ *and*

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\leq\sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)},$$

we have :

$$\begin{aligned}
|\hat{\alpha}-\alpha^*|\leq&\frac{256(1-c^*)^2}{\underline{c}^2m^2}\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+\frac{\nu_2\epsilon_5}{\underline{c}m\sqrt{N}}+\frac{\nu_3\epsilon_1}{\underline{c}m\sqrt{NnT}}\right)\\
&+\frac{32c^*(1-c^*)}{\underline{c}^2m^2}\left[\frac{(\epsilon_7+1)^2}{n(T-1)}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)+\frac{11\epsilon_7}{\sqrt{n(T-1)}}\right]\\
&+\frac{4c^*}{\underline{c}m\left(2-c^*\right)}\left[\frac{2(\epsilon_6+1)}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right)+\epsilon_5\frac{8\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}}+\epsilon_1\frac{32K^3}{c_2^5\sqrt{NnT}}\right]\\
&+\frac{256c^*\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2\underline{c}m\left(2-c^*\right)nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right]\\
&+\frac{1+A(m)}{\underline{c}m}\frac{16(\epsilon_6+1)}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right)+\epsilon_5\frac{32\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}}+\epsilon_1\frac{128K^3}{c_2^5\sqrt{NnT}}\\
&+\frac{1+A(m)}{\underline{c}m}\frac{256\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right],
\end{aligned}$$

with probability larger than $1-2n(T-1)p^2\exp(-\epsilon_1^2)-2n(T-1)pK\exp(-\epsilon_2^2)-2n(T-1)Kp^2\exp(-\epsilon_3^2)-2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2n(T-1)K\exp(-\epsilon_5^2)-2n(T-1)\exp\left(-\epsilon_6^2/4\right)-19\exp(-\epsilon_7^2/4)$.

Proof. See proof in Subsection 5.5.17 ■

Remark 5.4.3 *We set* $\epsilon_1^2=2\log(nTp)$*,* $\epsilon_2^2=2\log(nTpK)$*,* $\epsilon_3^2=2\log(nTpK)$*,* $\epsilon_5^2=2\log(nTK)$*,* $\epsilon_4^2=2\log(nT)+2\log(2p^2+p9^p)\leq2\log(nT)+2p$ *once* $p\geq34$*,* $\epsilon_6^2=8\log(nT)$ *and* $\epsilon_7^2=4\log(nT)$*. Then, for*

N, n and T large enough with probability at least $1 - \frac{31}{nT}$, in order of magnitude there is :

$$
\begin{aligned}
|\hat{\alpha} - \alpha^*| \leq & \left(1 + \frac{c^*}{\underline{c}m} + \frac{(1-\underline{c})^2}{\underline{c}^3 m^3}\right)\left[\sqrt{\log(nTK)}\frac{K^{3/2}}{c_2^2\sqrt{N}} + \sqrt{\log(nTp)}\frac{K^3}{c_2^5\sqrt{NnT}}\right] \\
& + \frac{A(m)+c^*}{\underline{c}m}\frac{\sqrt{\log(nT)}}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}} + 1\right) \\
& + \frac{c^*(1-c^*)}{\underline{c}^2 m^2}\left[\frac{\log(nT)}{n(T-1)}\left(1 + \frac{1}{\underline{c}\sqrt{T-1}}\right) + \sqrt{\frac{\log(nT)}{n(T-1)}}\right] \\
& + \left(\frac{(1-\underline{c})^2}{\underline{c}^2 m^2} + \frac{p(A(m)+c^*)}{c_2^2 \underline{c} m}\left[c_2 + \frac{K^2}{c_2^2}\right]\right)\frac{(\log(nT)+p)K^6 p^3\left(\sqrt{\frac{Kp}{c_1 N}} + K^2\right)}{c_2^{21} c_1^3 \min\left(c_3, c_2^2\right) nT(N-2)}.
\end{aligned}
$$

This rate of convergence can be compared to the one obtained in Theorem 4.4.3. In the oracle DETM, the convergence rate is of order $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}}\right)$ with probability at least $1 - \frac{17}{nT}$. In the realistic DTM, fixing the number of topics K and the vocabulary size p, the rate of convergence is of order $\mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}} + \sqrt{\frac{\log(nT)}{N}}\right)$ with probability at least $1 - \frac{31}{nT}$. Hence the same probability of control up to a constant factor, an extra term $\sqrt{\frac{\log(nT)}{N}}$ appears due to the multinomial noise.

In conclusion, the convergence rates obtained in the dynamic topic model show an additive behavior of the noise contained at different levels in the model. The bounds are driven by the Dirichlet noise driving the probability of topics given documents and by the noise in the multinomial model of word-counts. In particular, for very long documents, that is when $N \gg nT$, the convergence rates are only driven by the Dirichlet noise in the autoregressive model.

5.5 Proofs

5.5.1 Proof of Proposition 5.2.1 and its Corollary

Proof of Proposition 5.2.1. We start by reminding $\boldsymbol{M}_* := (nT)^{-1}\text{diag}\left(A^*\boldsymbol{W}^{1:T}1_{nT}\right)$ and $\hat{M} := (nT)^{-1}\text{diag}\left(\boldsymbol{Y}^{1:T}1_{nT}\right)$. Thus the two following equations hold where $A^*\boldsymbol{W}_j^t(i) := \sum_{k=1}^{K} A_{ik}^*\boldsymbol{W}_j^t(k) \in \mathbb{R}$,

$$
\begin{aligned}
\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| &= (nT)^{-1}\left|\sum_{j=1}^{n}\sum_{t=1}^{T}\left(\boldsymbol{Y}_j^t(i) - A^*\boldsymbol{W}_j^t(i)\right)\right|, \quad i \in [p] \\
\left\|\boldsymbol{M}_* - \hat{M}\right\|_{op} &= \max_{i\in[p]}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right|.
\end{aligned}
$$

Let us fix $i \in [p]$ and consider any $u > 0$. The tail of $\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right|$ can be defined through its conditional distribution given $\boldsymbol{W}^{1:T}$ as follows,

$$\mathbb{P}\left(\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > u\right) = \mathbb{E}_{\boldsymbol{W}}\left[\mathbb{P}\left(\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > u\right)|\boldsymbol{W}^{1:T}\right].$$

In addition, the variables $\left(\boldsymbol{Y}_j^t(i) - A^*\boldsymbol{W}_j^t(i)\right)_{j,t}$ are real-valued and independent conditionally on $\boldsymbol{W}^{1:T}$. From the definition of the multinomial distribution, they can be expressed, conditionally on $\boldsymbol{W}^{1:T}$, for all $(t, j, i) \in [T] \times [n] \times [p]$, as,

$$\boldsymbol{Y}_j^t(i) - A^*\boldsymbol{W}_j^t(i) = \frac{1}{N}\sum_{l=1}^{N}\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right), \tag{5.7}$$

where for all $l \in [N]$ and for all $(t, j) \in [T] \times [n]$, $Q_{jl}^t|\boldsymbol{W}_j^t \sim \text{Multinomial}_p\left(1, A^*\boldsymbol{W}_j^t\right)$ and $\mathbb{P}_{(Q_{j1}^1,\dots,Q_{jN}^1,Q_{j1}^2\dots,Q_{jN}^T)|(\boldsymbol{W}_j^1,\dots,\boldsymbol{W}_j^T)} = \bigotimes_{t=1}^{T}\bigotimes_{l=1}^{N}\mathbb{P}_{Q_{jl}^t|\boldsymbol{W}_j^t}$. Then the following equalities hold for all $(t, j, i, l) \in [T] \times [n] \times [p] \times [N]$,

$$\begin{aligned}
\mathbb{E}\left[Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]|\boldsymbol{W}^{1:T}\right] &= 0 \quad a.s.,\\
\mathbb{P}\left[\left|Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right| > 2|\boldsymbol{W}^{1:T}\right] &= 0 \quad a.s.,\\
\mathbb{V}\left[Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]|\boldsymbol{W}^{1:T}\right] &= \sum_{k=1}^{K} A_{ik}^*\boldsymbol{W}_j^t(k)\left(1 - \sum_{k=1}^{K} A_{ik}^*\boldsymbol{W}_j^t(k)\right)\\
&= A^*\boldsymbol{W}_j^t(i)\left(1 - A^*\boldsymbol{W}_j^t(i)\right) \ a.s.
\end{aligned}$$

Hence applying Hoeffding's inequality, Lemma 1.1.8, conditionally on $\boldsymbol{W}^{1:T}$ to $\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{l=1}^{N}\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)$ gives, for all $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left[\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \epsilon|\boldsymbol{W}^{1:T}\right] &\leq 2\exp\left(-\frac{NnT\epsilon^2}{8}\right) \quad a.s.,\\
\mathbb{P}\left[\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \epsilon\right] &\leq 2\mathbb{E}_{\boldsymbol{W}}\left[\exp\left(-\frac{NnT\epsilon^2}{8}\right)\right].
\end{aligned}$$

The last inequality proves that, for all $i \in [p]$, with probability at least $1 - \exp(-\epsilon^2)$ we have

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < \sqrt{\frac{8}{NnT}}\epsilon. \tag{5.8}$$

For the second part of the proof, we adapt the proof of Hoeffding's lemma as follows to control the moment generating function of $Q_{jl}^t(i)$. It will allow to control the deviation of $[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}$ with the conditional variance of $Q_{jl}^t(i)$. We first consider, for all $(j, t, l, i) \in [n] \times [T] \times [N] \times [p]$ an identical and independent copy of $Q_{jl}^t(i)$, conditionally on $\boldsymbol{W}^{1:T}$, that we name $R_{jl}^t(i)$. We name this step the symmetrisation. We then consider their conditionally centered version, namely $Q_{jlt}(i)^\top = Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)|\boldsymbol{W}^{1:T}]$ and $R_{jlt}(i)^\top = R_{jl}^t(i) - \mathbb{E}[R_{jl}^t(i)|\boldsymbol{W}^{1:T}]$. We first notice that the following equality holds for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] = \mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i) - \lambda\mathbb{E}_R\left[R_{jlt}^\top(i)\right]\right)|\boldsymbol{W}^{1:T}\right],$$

where $\mathbb{E}_Q$ (respectively $\mathbb{E}_R$) denotes the conditional expectation taken *w.r.t* the distribution of $Q_{jlt}(i)^\top$ (respectively $R_{jlt}(i)^\top$). Then applying conditional Jensen's inequality provides, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] \leq \mathbb{E}_{Q,R}\left[\exp\left(\lambda Q_{jlt}^\top(i) - \lambda R_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right].$$

We notice that the random variable $Q_{jlt}^\top(i) - R_{jlt}^\top(i)$ is symmetric and centered conditionally on $\boldsymbol{W}^{1:T}$. Indeed the random variables $Q_{jlt}^\top(i) - R_{jlt}^\top(i)$ and $R_{jlt}^\top(i) - Q_{jlt}^\top(i)$ share the same distribution conditionally on $\boldsymbol{W}^{1:T}$. This proves that for all $k \in \mathbb{N}$, if k is odd we get $\mathbb{E}_{Q,R}\left[\left(Q_{jlt}^\top(i) - R_{jlt}^\top(i)\right)^k|\boldsymbol{W}^{1:T}\right] = 0$ almost surely. Indeed for all $k \in \mathbb{N}$ such that k is odd we get

$$\mathbb{E}_{Q,R}\left[\left(Q_{jlt}^\top(i) - R_{jlt}^\top(i)\right)^k|\boldsymbol{W}^{1:T}\right] = \mathbb{E}_{Q,R}\left[\left(R_{jlt}^\top(i) - Q_{jlt}^\top(i)\right)^k|\boldsymbol{W}^{1:T}\right].$$

We also note that conditionally on $\boldsymbol{W}^{1:T}$, the variable $Q_{jlt}^\top(i) - R_{jlt}^\top(i)$ are bounded almost surely in [-4, 4]. Taylor's theorem ensures that for all $\lambda \in \mathbb{R}$, for all $x \in [-4, 4]$, there exists $\gamma \in [\min(0, x); \max(0, x)]$ such that

$$\exp(\lambda x) = 1 + \lambda x + \frac{\lambda^2 x^2}{2} + \frac{\lambda^3 x^3 \exp(\lambda\gamma)}{6}$$

If x is positive, then x^3 is positive and $\gamma \leq x$. We get $x^3\exp(\lambda\gamma) \leq x^3\exp(\lambda x)$. If x is negative, then x^3 is negative and $\gamma \geq x$. We get $x^3\exp(\lambda\gamma) \leq x^3\exp(\lambda x)$. Finally this leads to

$$\begin{aligned}\exp(\lambda x) &= 1 + \lambda x + \frac{\lambda^2 x^2}{2} + \frac{\lambda^3 x^3 \exp(\lambda x)}{6},\\ &\leq 1 + \lambda x + \frac{\lambda^2 x^2}{2} + \frac{\lambda^3 x^3 \exp(4\lambda)}{6}.\end{aligned}$$

Finally this leads to the following inequality which holds almost surely,

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] \leq 1 &+ \lambda\mathbb{E}_{Q,R}\left[(Q_{jlt}^\top(i) - R_{jlt}^\top(i))|\boldsymbol{W}^{1:T}\right]\\ &+ \frac{\lambda^2}{2}\mathbb{E}_{Q,R}\left[(Q_{jlt}^\top(i) - R_{jlt}^\top(i))^2|\boldsymbol{W}^{1:T}\right]\\ &+ \frac{\lambda^3\exp(4\lambda)}{6}\mathbb{E}_{Q,R}\left[(Q_{jlt}^\top(i) - R_{jlt}^\top(i))^3|\boldsymbol{W}^{1:T}\right].\end{aligned}$$

The conditional symmetry of $Q_{jlt}^\top(i) - R_{jlt}^\top(i)$ around zero ensures that its conditional odd moments are almost surely null and we get,

$$\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] \leq 1 + \frac{\lambda^2}{2}\mathbb{V}_{Q,R}\left[(Q_{jlt}^\top(i) - R_{jlt}^\top(i))|\boldsymbol{W}^{1:T}\right].$$

By independence and identical conditional distributions of $Q_{jlt}^\top(i)$ and $R_{jlt}^\top(i)$ we have

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] &\leq 1 + \lambda^2\mathbb{V}_Q\left[Q_{jlt}^\top(i)|\boldsymbol{W}^{1:T}\right],\\ &\leq 1 + \lambda^2 A^*\boldsymbol{W}_j^t(i)\left(1 - A^*\boldsymbol{W}_j^t(i)\right).\end{aligned}$$

We notice that for all $i \in [p]$ and for all $k \in [K]$ the quantity A_{ik}^* is bounded from above by 1. In addition the random vector $\boldsymbol{W}_j^t$ is non negative and sum to one almost surely. Hence for all $i \in [p]$ the

quantity $A^*\boldsymbol{W}_j^t(i) := \sum_{k=1}^{K} A_{ik}^* \boldsymbol{W}_j^t(k)$ is bounded from above by 1. This leads to, for all $i \in [p]$ and for all $(j,t) \in [n] \times [T]$, $A^*\boldsymbol{W}_j^t(i)\left(1 - A^*\boldsymbol{W}_j^t(i)\right) := \sum_{k=1}^{K} A_{ik}^* \boldsymbol{W}_j^t(k)\left(1 - A_{ik}^*\boldsymbol{W}_j^t(k)\right) \leq \sum_{k=1}^{K} A_{ik}^* := h_i$ almost surely. Hence, using the inequality $\exp(s) \geq 1 + s$ for all $s \in \mathbb{R}$, we finally have, almost surely, and for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda Q_{jlt}^\top(i)\right)|\boldsymbol{W}^{1:T}\right] &\leq 1 + \lambda^2 h_i, \\ &\leq \exp\left(\lambda^2 h_i\right).\end{aligned}$$

This ensures the following bound which holds for all $\lambda > 0$ and for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left[\sum_{j,t,l} Q_{jtl}^\top(i) \geq t|\boldsymbol{W}^{1:T}\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_{j,t,l} Q_{jtl}^\top(i)\right) \geq \exp\left(\lambda t\right)|\boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t)\mathbb{E}\left[\exp\left(\lambda \sum_{j,t,l} Q_{jtl}^\top(i)\right)|\boldsymbol{W}^{1:T}\right], \quad \text{by Markov's inequality} \\ &\leq \exp(-\lambda t)\prod_{j,t,l}\mathbb{E}\left[\exp\left(\lambda Q_{jtl}^\top(i)\right)|\boldsymbol{W}^{1:T}\right], \quad \text{by conditional independence} \\ &\leq \exp(-\lambda t)\prod_{j,t,l}\exp\left(\lambda^2 h_i\right), \\ &\leq \exp(-\lambda t)\exp\left(NnT\lambda^2 h_i\right).\end{aligned}$$

Choosing $\lambda := \dfrac{t}{2NnTh_i}$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to, for all $\epsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{j,t,l} Q_{jtl}^\top(i)\right| \leq t\right] \geq 1 - 2\exp\left(-\frac{t^2}{4NnTh_i}\right).$$

Finally, for all $i \in [p]$ and for all $\epsilon > 0$, it comes

$$\mathbb{P}\left[\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{NnT\epsilon^2}{4h_i}\right).$$

We finally get for all $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < \sqrt{\frac{4h_i}{NnT}}\epsilon. \tag{5.9}$$

We conclude by combining (5.8) and (5.9). ■

Proof of Corollary 5.2.2. We consider Proposition 5.2.1 which leads to the following inequalities,

holding for all $i \in [p]$,

$$\mathbb{P}\left[h_i^{-1/2}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \epsilon\right] \leq 2\exp\left(-\frac{NnTh_i\epsilon^2}{4\min(2,h_i)}\right),$$
$$\mathbb{P}\left[h_i^{-1/2}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \epsilon\right] \leq 2\exp\left(-\frac{NnT\epsilon^2}{4\min(2/h_i,1)}\right),$$
$$\mathbb{P}\left[h_i^{-1/2}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \epsilon\right] \leq 2\exp\left(-\frac{NnT\max(h_i/2,1)\epsilon^2}{4}\right).$$

This provides, for all $\epsilon > 0$,

$$\mathbb{P}\left[h_i^{-1/2}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| > \frac{2\epsilon}{\sqrt{NnT\max(h_i/2,1)}}\right] \leq 2\exp\left(-\epsilon^2\right).$$

Using a union bound leads to the stated result. ■

5.5.2 Proof of Proposition 5.2.3 and its Corollaries

Proof of Proposition 5.2.3. We define, for $(j,l,t) \in [n]\times[N]\times[T]$, the random variables Q_{jl}^t similarly as in (5.7). This leads, for all $(i,k) \in [p]\times[K]$, to

$$\begin{aligned}
[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.} &= \sum_{j=1}^n \sum_{t=1}^T Z_j^t(i)\boldsymbol{W}_j^t(k), \\
&= \frac{1}{NnT}\sum_{j=1}^n\sum_{t=1}^T\sum_{l=1}^N nT\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)\boldsymbol{W}_j^t(k), \\
&= \frac{1}{N}\sum_{j=1}^n\sum_{t=1}^T\sum_{l=1}^N \left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)\boldsymbol{W}_j^t(k).
\end{aligned}$$

Let us remind that for all $(j,t,k) \in [n]\times[T]\times[K]$, $|\boldsymbol{W}_j^t(k)| \leq 1$ almost surely. This implies the following equalities for all $(t,j,i,l) \in [T]\times[n]\times[p]\times[N]$,

$$\begin{aligned}
\mathbb{E}\left[\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)W_j^t(k)|\boldsymbol{W}^{1:T}\right] &= 0 \quad a.s., \\
\mathbb{P}\left[\left|\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)\boldsymbol{W}_j^t(k)\right| > 2|\boldsymbol{W}^{1:T}\right] &= 0 \quad a.s., \\
\mathbb{V}\left[\left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)\boldsymbol{W}_j^t(k)|\boldsymbol{W}^{1:T}\right] &= A^*\boldsymbol{W}_j^t(i)\left(1 - A^*\boldsymbol{W}_j^t(i)\right)\boldsymbol{W}_j^t(k)^2 \quad a.s..
\end{aligned}$$

Applying Hoeffding's inequality, Lemma 1.1.8, conditionally on $\boldsymbol{W}^{1:T}$ to

$$\sum_{j=1}^n\sum_{t=1}^T\sum_{l=1}^N \left(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t(i)]\right)\boldsymbol{W}_j^t(k)$$

gives, for all $\epsilon > 0$, for all $(i,k) \in [p]\times[K]$,

$$\begin{aligned}
\mathbb{P}\left[\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon|\boldsymbol{W}^{1:T}\right] &\leq 2\exp\left(-\frac{N\epsilon^2}{8nT}\right) \quad a.s., \\
\mathbb{P}\left[\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon\right] &\leq 2\mathbb{E}_{\boldsymbol{W}}\left[\exp\left(-\frac{N\epsilon^2}{8nT}\right)\right].
\end{aligned}$$

The last inequality proves that, for all $i \in [p]$, for all $k \in [K]$ and for all $\epsilon > 0$, with probability at least $1 - \exp(-\epsilon^2)$ we have,

$$\left| [\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.} \right| < \sqrt{\frac{8nT}{N}} \epsilon. \tag{5.10}$$

For the second part of the proof, we adapt the proof of Hoeffding's lemma as follows to control the moment generating function of $Q_{jl}^t(i)\boldsymbol{W}_j^t(k)$ for all $i \in [p]$ and for all $k \in [K]$. It will allow to control the deviation of $[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}$ with the conditional variance of $Q_{jl}^t(i)$. We follow the same technical steps as in the proof of Proposition 5.2.1. Namely we consider the symmetrisation step, then we apply conditional Jensen's inequality, we notice that the random variables $\boldsymbol{W}_j^t(k)\left(Q_{jlt}^\top(i) - R_{jlt}^\top(i)\right)$ are symmetric and centered conditionally on $\boldsymbol{W}^{1:T}$ and that the variables $\boldsymbol{W}_j^t(k)\left(Q_{jlt}^\top(i) - R_{jlt}^\top(i)\right)$ are bounded almost surely in [-4, 4]. Then we use Taylor's theorem and notice that the conditional symmetry of $Q_{jlt}^\top(i) - R_{jlt}^\top(i)$ around zero ensures that its conditional odd moments are almost surely null. In addition $\boldsymbol{W}_j^t(k)$ is almost surely bounded from above by one. We finally use the independence and identical conditional distributions of $Q_{jlt}^\top(i)$ and $R_{jlt}^\top(i)$ to get for all $\lambda > 0$:

$$\mathbb{E}_Q\left[\exp\left(\lambda \boldsymbol{W}_j^t(k) Q_{jlt}^\top(i)\right) | \boldsymbol{W}^{1:T}\right] \leq \exp\left(\lambda^2 h_i\right).$$

This then ensures for all $\lambda > 0$ and all $t > 0$:

$$\mathbb{P}\left[\sum_{j,t,l} \boldsymbol{W}_j^t(k) Q_{jtl}^\top(i) \geq t | \boldsymbol{W}^{1:T}\right] \leq \exp(-\lambda t) \exp\left(NnT\lambda^2 h_i\right).$$

Choosing $\lambda := \dfrac{t}{2NnTh_i}$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to, for all $\epsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{j,t,l} \boldsymbol{W}_j^t(k) Q_{jtl}^\top(i)\right| \leq t\right] \geq 1 - 2\exp\left(-\frac{t^2}{4NnTh_i}\right).$$

Finally, for all $i \in [p]$, for all $k \in [K]$ and for all $\epsilon > 0$, it comes

$$\mathbb{P}\left[\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{N\epsilon^2}{4nTh_i}\right).$$

We finally get for all $\epsilon > 0$, for all $i \in [p]$ and for all $k \in [K]$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| < \sqrt{\frac{4nTh_i}{N}} \epsilon. \tag{5.11}$$

Combining (5.10) and (5.11) gives the stated result. ■

Proof of Corollary 5.2.4. We consider Proposition 5.2.3 which leads to the following inequalities, holding for all $i \in [p]$, for all $k \in [K]$ and for all $\epsilon > 0$,

$$\mathbb{P}\left[h_i^{-1/2}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon\right] \leq 2\exp\left(-\frac{Nh_i\epsilon^2}{4nT\min(2, h_i)}\right),$$

$$\mathbb{P}\left[h_i^{-1/2}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon\right] \leq 2\exp\left(-\frac{N\epsilon^2}{4nT\min(2/h_i, 1)}\right),$$

$$\mathbb{P}\left[h_i^{-1/2}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon\right] \leq 2\exp\left(-\frac{N\max(h_i/2, 1)\epsilon^2}{4nT}\right).$$

This leads to, for all $i \in [p]$, for all $k \in [K]$ and for all $\epsilon > 0$,

$$\mathbb{P}\left[h_i^{-1/2}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right| > \epsilon\sqrt{\frac{4nT}{N\max(h_i/2,1)}}\right] \leq 2\exp\left(-\epsilon^2\right).$$

Using a union bound leads to the stated result. ■

Proof of Corollary 5.2.5. We start by recalling that, for all $k \in [K]$, the quantity $\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2$ can be expressed as follows,

$$\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2^2 = \sum_{i=1}^p \frac{1}{[\boldsymbol{M}_*]_{ii}}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right|^2.$$

Moreover for all $i \in [p]$, the following inequalities hold,

$$\begin{aligned}
[\boldsymbol{M}_*]_{ii} := \frac{1}{nT}\sum_{j=1}^n\sum_{t=1}^T\sum_{k=1}^K A_{ik}^*\boldsymbol{W}_j^t(k) &= \sum_{k=1}^K A_{ik}^*\left(\frac{1}{nT}\sum_{j=1}^n\sum_{t=1}^T\sum_{k=1}^K \boldsymbol{W}_j^t(k)\right), \\
&\geq \sum_{k=1}^K A_{ik}^*\left(\frac{1}{nT}\sum_{j=1}^n\sum_{t=1}^T\sum_{k=1}^K \left(\boldsymbol{W}_j^t(k)\right)^2\right), \\
&\geq \sum_{k=1}^K A_{ik}^*\left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk}.
\end{aligned}$$

In addition, as is proved in Proposition 4.3.4 we have for all $k \in [K]$, $\left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})$. This result ensures that for all $i \in [p]$,

$$\begin{aligned}
[\boldsymbol{M}_*]_{ii} &\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\cdot\sum_{k=1}^K A_{ik}^*, \\
&\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\cdot h_i.
\end{aligned}$$

Hence this leads to, for all $k \in [K]$,

$$\begin{aligned}
\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2^2 &\leq \sum_{i=1}^p \frac{1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\cdot h_i}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right|^2, \\
&\leq \frac{p}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\max_{i\in[p]} h_i^{-1}\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}\right|^2.
\end{aligned}$$

Finally, using the result of Corollary 5.2.4 gives, for all $\epsilon > 0$ and for all $k \in [K]$, with probability at least $1 - 2p\exp(-\epsilon^2)$,

$$\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2^2 < \epsilon^2\frac{4pnT}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})N}.$$

■

5.5.3 Proof of Proposition 5.2.6

Proof of Proposition 5.2.6. We start by fixing $(i,m) \in [p]^2$ and we remind that the diagonal matrix $H \in \mathbb{R}^{p\times p}$ is defined as $H := \text{diag}\,(h_1,\dots,h_p)$ where for all $s \in [p],\ h_s := \|A_{s.}\|_1$. We also remind that $A \in \mathbb{R}^{p\times K}$ and $A_{s.} \in \mathbb{R}^K$. In addition we have the following developments,

$$[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.} = \sum_{j=1}^n \sum_{t=1}^T Z_{ij}^t Z_{mj}^t,$$
$$\frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}}{\sqrt{h_i \cdot h_m}} = \sum_{j=1}^n \sum_{t=1}^T \frac{Z_{ij}^t}{\sqrt{h_i}} \frac{Z_{mj}^t}{\sqrt{h_m}}.$$

The Parallelogram law then ensures that

$$\frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}}{\sqrt{h_i \cdot h_m}} = \sum_{j=1}^n \sum_{t=1}^T \left(\frac{Z_{ij}^t}{\sqrt{h_i}} + \frac{Z_{mj}^t}{\sqrt{h_m}} \right)^2 - \left(\frac{Z_{ij}^t}{\sqrt{h_i}} - \sum_{j=1}^n \sum_{t=1}^T \frac{Z_{mj}^t}{\sqrt{h_m}} \right)^2.$$

For all $s \in [p]$ we denote e_s the standard basis vector of $\mathbb{R}^p$ of order s, *i.e.* the vector with all coordinates equal to zero except the s^{th} coordinate which equals one. We then define $\epsilon_{im}^+ := \dfrac{e_i + e_m}{2}$ and $\epsilon_{im}^- := \dfrac{e_i - e_m}{2}$ and we derive the following equality,

$$\frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}}{\sqrt{h_i \cdot h_m}} = \sum_{j=1}^n \sum_{t=1}^T \left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 - \sum_{j=1}^n \sum_{t=1}^T \left(\left[\epsilon_{im}^-\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2.$$

Finally we get

$$\begin{aligned}\frac{[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{Z}^{1:T}]_{m.}\right]}{\sqrt{h_i \cdot h_m}} &= \sum_{j=1}^n \sum_{t=1}^T \left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 - \sum_{j=1}^n \sum_{t=1}^T \mathbb{E}\left[\left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 \right] \\ &\quad - \sum_{j=1}^n \sum_{t=1}^T \left(\left[\epsilon_{im}^-\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 + \sum_{j=1}^n \sum_{t=1}^T \mathbb{E}\left[\left(\left[\epsilon_{im}^-\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 \right].\end{aligned}$$

We start by deriving an upper bound for the first term :

$$A := \sum_{j=1}^n \sum_{t=1}^T \left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 - \sum_{j=1}^n \sum_{t=1}^T \mathbb{E}\left[\left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t \right)^2 \right].$$

We define here again the random variables (Q_{jl}^t) for $(j,l,t) \in [p]\times[N]\times[T]$ as in (5.7). It allows to express, for all $(j,l,t) \in [p]\times[N]\times[T]$, the random variables $\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t$ as follows,

$$\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Z_j^t = \frac{1}{N} \sum_{l=1}^N \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot \left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t | \boldsymbol{W}^{1:T}\right] \right) \quad a.s..$$

We recall that conditionally on $\boldsymbol{W}^{1:T}$,

$$\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot \mathsf{diag}(A\boldsymbol{W}_j^t)\cdot H^{-1/2}\cdot\left[\epsilon_{im}^{+}\right]=\frac{1}{4}\left[\frac{\sqrt{A\boldsymbol{W}_j^t(i)}}{\sqrt{h_i}}+\frac{\sqrt{A\boldsymbol{W}_j^t(m)}}{\sqrt{h_m}}\right]^2\leq 1\quad a.s.,$$

because we have almost surely, $A\boldsymbol{W}_j^t(i)\leq h_i$ and $A\boldsymbol{W}_j^t(m)\leq h_m$. Hence, given the definition of $\left(Q_{jl}^t\right)_{j,l,t}$ the following equalities hold almost surely for $(j,l,t)\in[p]\times[N]\times[T]$,

$$\mathbb{V}\left[\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jl}^t-\mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)|\boldsymbol{W}^{1:T}\right]\leq\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\mathsf{diag}(A\boldsymbol{W}_j^t)\cdot H^{-1/2}\cdot\left[\epsilon_{im}^{+}\right]\leq 1,$$
$$\left|\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jl}^t-\mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)\right|\leq\frac{1}{\sqrt{h_i}}+\frac{1}{\sqrt{h_m}}\leq\frac{2}{\sqrt{h_{\min}}}.$$

Hence applying Hoeffding's inequality, for bounded random variables, Lemma 1.1.8, conditionally on $\boldsymbol{W}^{1:T}$ to $\sum_{l=1}^{N}\frac{1}{N}\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jl}^t-\mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ gives, for all $(j,t)\in[n]\times[T]$, for all $s>0$,

$$\mathbb{P}\left[\left|\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t\right|>s|\boldsymbol{W}^{1:T}\right]\leq 2\exp\left(-\frac{Nh_{\min}s^2}{8}\right).\tag{5.12}$$

On the other hand we adapt the proof of Hoeffding's lemma as follows to control the moment generating function of $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot(Q_{jl}^t(i)-\mathbb{E}[Q_{jl}^t|\boldsymbol{W}^{1:T}])$. It will allow to control the deviation of $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t$ with the conditional variance of Q_{jl}^t. We first consider, for all $(j,t,l,i)\in[n]\times[T]\times[N]\times[p]$ an identical and independent copy of Q_{jl}^t, conditionally on $\boldsymbol{W}^{1:T}$, that we name R_{jl}^t. We then consider their conditionally centered version, namely $Q_{jlt}^{\top}=Q_{jl}^t-\mathbb{E}[Q_{jl}^t|\boldsymbol{W}^{1:T}]$ and $R_{jlt}^{\top}=R_{jl}^t-\mathbb{E}[R_{jl}^t|\boldsymbol{W}^{1:T}]$. We first notice that the following equality holds for all $\lambda\in\mathbb{R}$, $\mathbb{E}_Q\left[\exp\left(\lambda\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Q_{jlt}^{\top}\right)|\boldsymbol{W}^{1:T}\right]$ is equal to $\mathbb{E}_Q\left[\exp\left(\lambda\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-\mathbb{E}_R\left[R_{jlt}^{\top}\right]\right)\right)|\boldsymbol{W}^{1:T}\right]$ where $\mathbb{E}_Q$ (respectively $\mathbb{E}_R$) denotes the conditional expectation taken *w.r.t* the distribution of $Q_{jlt}^{\top}$ (respectively $R_{jlt}^{\top}$). Then applying conditional Jensen's inequality provides, for all $\lambda\in\mathbb{R}$,

$$\mathbb{E}_Q\left[\exp\left(\lambda\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Q_{jlt}^{\top}\right)|\boldsymbol{W}^{1:T}\right]\leq\mathbb{E}_{Q,R}\left[\exp\left(\lambda\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-R_{jlt}^{\top}\right)\right)|\boldsymbol{W}^{1:T}\right].$$

We notice that the random variables $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-R_{jlt}^{\top}\right)$ are symmetric and centered conditionally on $\boldsymbol{W}^{1:T}$. Indeed the random variables $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-R_{jlt}^{\top}\right)$ and $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(R_{jlt}^{\top}-Q_{jlt}^{\top}\right)$ share the same distribution conditionally on $\boldsymbol{W}^{1:T}$. This proves that for all $k\in\mathbb{N}$, if k is odd we get $\mathbb{E}_{Q,R}\left[\left(\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-R_{jlt}^{\top}\right)\right)^k|\boldsymbol{W}^{1:T}\right]=0$ almost surely. We also note that conditionally on $\boldsymbol{W}^{1:T}$, the variables $\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jlt}^{\top}-R_{jlt}^{\top}\right)$ are bounded almost surely in $[-4/\sqrt{h_{\min}},4/\sqrt{h_{\min}}]$. Taylor's theorem ensures that for all $\lambda\in\mathbb{R}$, for all $x\in[-4/\sqrt{h_{\min}},4/\sqrt{h_{\min}}]$, there exists $\gamma\in[\min(0,x);\max(0,x)]$ such that

$$\exp(\lambda x)=1+\lambda x+\frac{\lambda^2x^2}{2}+\frac{\lambda^3x^3\exp(\lambda\gamma)}{6}$$

If x is positive, then x^3 is positive and $\gamma \leq x$. We get $x^3 \exp(\lambda\gamma) \leq x^3 \exp(\lambda x)$. If x is negative, then x^3 is negative and $\gamma \geq x$. We get $x^3 \exp(\lambda\gamma) \leq x^3 \exp(\lambda x)$. Finally this leads to

$$\begin{aligned}\exp(\lambda x) &= 1 + \lambda x + \frac{\lambda^2 x^2}{2} + \frac{\lambda^3 x^3 \exp(\lambda x)}{6}, \\ &\leq 1 + \lambda x + \frac{\lambda^2 x^2}{2} + \frac{\lambda^3 x^3 \exp(4\lambda)}{6}.\end{aligned}$$

Finally this leads to the following inequality which holds almost surely,

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right) | \boldsymbol{W}^{1:T}\right] \leq 1 &+ \lambda \mathbb{E}_{Q,R}\left[\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top) | \boldsymbol{W}^{1:T}\right] \\ &+ \frac{\lambda^2}{2}\mathbb{E}_{Q,R}\left[\left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt})^\top\right)^2 | \boldsymbol{W}^{1:T}\right] \\ &+ \frac{\lambda^3 \exp(4\lambda)}{6}\mathbb{E}_{Q,R}\left[\left(\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top)\right)^3 | \boldsymbol{W}^{1:T}\right].\end{aligned}$$

The conditional symmetry of $Q_{jlt}^\top - R_{jlt}^\top$ around zero ensures that its conditional odd moments are almost surely null and we get,

$$\mathbb{E}_Q\left[\exp\left(\lambda \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right) | \boldsymbol{W}^{1:T}\right] \leq 1 + \frac{\lambda^2}{2}\mathbb{V}_{Q,R}\left[\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top) | \boldsymbol{W}^{1:T}\right].$$

By independence and identical conditional distributions of $Q_{jlt}^\top$ and $R_{jlt}^\top$ we have

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right) | \boldsymbol{W}^{1:T}\right] &\leq 1 + \lambda^2 \mathbb{V}_Q\left[\left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top | \boldsymbol{W}^{1:T}\right], \\ &\leq 1 + \lambda^2.\end{aligned}$$

This ensures the following equalities which hold almost surely for all $\lambda \in \mathbb{R}$ and for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left[\sum_l \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top \geq t | \boldsymbol{W}^{1:T}\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_l \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top(i)\right) \geq \exp(\lambda t) | \boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t)\mathbb{E}\left[\exp\left(\lambda \sum_l \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right) | \boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t) \prod_l \mathbb{E}\left[\exp\left(\lambda \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right) | \boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t) \prod_l \exp\left(\lambda^2\right), \\ &\leq \exp(-\lambda t) \exp\left(N\lambda^2\right).\end{aligned}$$

Choosing $\lambda := \frac{t}{2N}$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to, for all $\epsilon > 0$, for all $j \in [n]$ and for all $t \in [T]$

$$\mathbb{P}\left[\left|\sum_l \left[\epsilon_{im}^+\right]^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right| \leq t\right] \geq 1 - 2\exp\left(-\frac{t^2}{4N}\right).$$

Finally, for all $j \in [n]$, for all $t \in [T]$ and for all $\epsilon > 0$, it comes

$$\mathbb{P}\left[\left|\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{N\epsilon^2}{4}\right).$$

We finally get for all $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\right| < \frac{2\epsilon}{\sqrt{N}}. \tag{5.13}$$

Combining (5.12) and (5.13) leads to, for all $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\right| < \frac{2\min(\sqrt{2}h_{\min}, 1)\epsilon}{\sqrt{N}}.$$

Equivalently, this says that for all $\epsilon > 0$,

$$\mathbb{P}\left[\left|\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{N\max(h_{\min}/2, 1)\epsilon^2}{4}\right).$$

This proves that the variables $(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)$ are SubGaussian. We recall that the SubGaussian norm of a SubGaussian random variable X is defined as

$$\|X\|_{\psi_2} := \inf_{s>0}\left\{\mathbb{E}\left[\frac{X^2}{s^2}\right] \leq 2\right\}.$$

Hence for all $(j,t) \in [n] \times [T]$, the SubGaussian norm of $\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t$ satisfies $\| \left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\|_{\psi_2} \leq 12 \cdot \sqrt{\frac{e}{N\max(h_{\min}/2, 1)\log(2)}}$. Indeed, Proposition 2.5.2 in [130] proves that for a random variable X satisfying, for all $s > 0$, $\mathbb{P}[|X| > s] \leq 2\exp\left(\frac{-s^2}{K_1^2}\right)$ where $K_1 > 0$ is a constant then $\mathbb{E}\left[\frac{X^2}{\left(6K_1\sqrt{e/\log(2)}\right)^2}\right] \leq 2$. This proves the stated result for $K_1^2 = \frac{4}{N\max(h_{\min}/2, 1)}$. In addition we immediately get that $(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2$ is SubExponential, see Lemma 2.7.6 in [130]. We recall that the SubExponential norm of a SubExponential random variable X is defined as

$$\|X\|_{\psi_1} := \inf_{s>0}\left\{\mathbb{E}\left[\frac{|X|}{s}\right] \leq 2\right\}.$$

This Lemma also ensures that its Subexponential norm satisfies $\|(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2\|_{\psi_1} = \|(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)\|_{\psi_2}^2$. Moreover, recalling that a norm is a convex function and using Jensen's inequality provides that its SubExponential norm also satisfies the centering property $\|(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2 - \mathbb{E}\left[(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2\right]\|_{\psi_1} \leq 2\| \left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t\|_{\psi_2}^2 \leq \frac{288 \cdot e}{N\max(h_{\min}/2, 1)\log(2)} := \gamma$. Using Bernstein's inequality for SubExponential random variables, Lemma 1.1.10, conditionally on $\boldsymbol{W}^{1:T}$ leads to, for all $s > 0$ and for an absolute constant $c > 0$,

$$\mathbb{P}\left[\left|\sum_{j=1}^{n}\sum_{t=1}^{T}\left[(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2 - \mathbb{E}\left[(\left[\epsilon_{im}^{+}\right]^{\top} \cdot H^{-1/2} \cdot Z_j^t)^2\right]\right]\right| > nTs|\boldsymbol{W}^{1:T}\right]$$

being bounded from above by $2\exp\left(-cnT\min\left(\frac{s^2}{\gamma^2};\frac{s}{\gamma}\right)\right)$, where $\gamma := \frac{288\cdot e}{N\max(h_{\min}/2,1)\log(2)}$. Then, let us fix $s>0$ and define $\epsilon = \frac{\sqrt{cnT}s}{\gamma}$. This provides that for all $\epsilon>0$,

$$\mathbb{P}\left[\left|\sum_{j=1}^{n}\sum_{t=1}^{T}\left[(\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t)^2 - \mathbb{E}\left[(\left[\epsilon_{im}^{+}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t)^2\right]\right]\right| > \frac{\gamma\epsilon\sqrt{nT}}{\sqrt{c}}|\boldsymbol{W}^{1:T}\right]$$

is bounded from above by $2\exp\left(-\min\left(\epsilon^2;\sqrt{cnT}\epsilon\right)\right)$. Taking on both sides the expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to, for all $\epsilon>0$,

$$|A| < \frac{288\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon\sqrt{nT}}{N\max(h_{\min}/2,1)},$$

with probability at least $1-2\exp\left(-\min\left(\epsilon^2;\sqrt{cnT}\epsilon\right)\right)$.

We now derive an upper bound for the second term $B := \sum_{j=1}^{n}\sum_{t=1}^{T}\mathbb{E}\left[\left(\left[\epsilon_{im}^{-}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t\right)^2\right] - \sum_{j=1}^{n}\sum_{t=1}^{T}\left(\left[\epsilon_{im}^{-}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t\right)^2$. The exact same proof hold as we can again express, for all $(j,l,t)\in[p]\times[N]\times[T]$, the random variables $\left[\epsilon_{im}^{-}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t$ as follows,

$$\left[\epsilon_{im}^{-}\right]^{\top}\cdot H^{-1/2}\cdot Z_j^t = \frac{1}{N}\sum_{l=1}^{N}\left[\epsilon_{im}^{-}\right]^{\top}\cdot H^{-1/2}\cdot\left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)\quad a.s..$$

It follows similarly that, for all $\epsilon>0$, with probability at least $1-2\exp\left(-\min\left(\epsilon^2;\sqrt{cnT}\epsilon\right)\right)$ we have

$$|B| < \frac{288\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon\sqrt{nT}}{N\max(h_{\min}/2,1)}.$$

Finally, for all $\epsilon>0$, for all $(i,m)\in[p]^2$, with probability at least $1-4\exp\left(-\min\left(\epsilon^2;\sqrt{cnT}\epsilon\right)\right)$, we have

$$\left|\frac{[\boldsymbol{Z}^{1:T}]_{i.}^{\top}[\boldsymbol{Z}^{1:T}]_{m.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^{\top}[\boldsymbol{Z}^{1:T}]_{m.}\right]}{\sqrt{h_i\cdot h_m}}\right|$$

bounded from above by

$$\frac{576\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon\sqrt{nT}}{N\max(h_{\min}/2,1)}.$$

We conclude by controlling the probability that this event holds simultaneously for all $(i,m)\in[p]^2$. ■

5.5.4 Proof of Proposition 5.2.7

Proof of Proposition 5.2.7. We start by recalling that, for all $i\in[p]$,

$$[\boldsymbol{M}_*]_{jj} := \frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[A^*]_{ik}\boldsymbol{W}_j^t(k) = \sum_{k=1}^{K}[A^*]_{ik}\left(\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k)\right).$$

Hence under Assumption 7 we have, for all $i \in [p]$, almost surely

$$[\boldsymbol{M}_*]_{jj} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i.$$

By considering $H := \mathsf{diag}(h_1, \dots, h_p)$ follows

$$\|\boldsymbol{M}_*^{-1/2} H^{1/2}\| \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2}.$$

This implies the following results,

$$\begin{aligned}
&\|\boldsymbol{M}_*^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) \boldsymbol{M}_*^{-1/2}\|, \\
=&\|\boldsymbol{M}_*^{-1/2} H^{1/2}\| \|H^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) H^{-1/2}\| \|H^{1/2} \boldsymbol{M}_*^{-1/2}\|, \\
\leq& \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \|H^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) H^{-1/2}\|.
\end{aligned}$$

We denote, for all $\epsilon > 0$, $\mathcal{N}_\epsilon$ an ϵ-net of the Euclidean Sphere of $\mathbb{R}^p$ and $\mathcal{N}(S_{p-1}, \epsilon)$ the smallest possible cardinality of an ϵ-net of S_{p-1}, called the covering number. For any $\epsilon > 0$, Corollary 4.2.13 in [130] ensures that

$$\mathcal{N}(S_{p-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon} \right)^p.$$

In addition Lemma 4.4.1 together with Exercise 4.4.3 in [130] ensure that for any $\epsilon \in (0, 1/2)$ and any symmetric matrix $A \in \mathbb{R}^{p \times p}$,

$$\|A\| \leq \frac{1}{1 - 2\epsilon} \sup_{x \in \mathcal{N}_\epsilon} \left| x^\top A x \right|.$$

We fix $\epsilon = 1/4$ and consider an ϵ-net $\mathcal{N}_\epsilon$ which satisfies $\mathcal{N}(S_{p-1}, \epsilon) \leq 9^p$. The following inequality then holds,

$$\begin{aligned}
&\|H^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) H^{-1/2}\| \\
\leq& 2 \sup_{x \in \mathcal{N}_\epsilon} \left| x^\top H^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) H^{-1/2} x \right|.
\end{aligned}$$

In addition, for any $x \in S_{p-1}$,

$$\begin{aligned}
&x^\top H^{-1/2} \left(\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T} (\boldsymbol{Z}^{1:T})^\top \right] \right) H^{-1/2} x \\
=& \sum_{j=1}^{n} \sum_{t=1}^{T} \left\{ \left(x^\top H^{-1/2} Z_j^t \right)^2 - \mathbb{E}\left[\left(x^\top H^{-1/2} Z_j^t \right)^2 \right] \right\}.
\end{aligned}$$

We fix $x \in S_{p-1}$ and we define here again the random variables (Q_{jl}^t) for $(j, l, t) \in [p] \times [N] \times [T]$ as in (5.7). It allows to express, for all $(j, l, t) \in [p] \times [N] \times [T]$, the random variables $x^\top \cdot H^{-1/2} \cdot Z_j^t$ as follows,

$$x^\top \cdot H^{-1/2} \cdot Z_j^t = \frac{1}{N} \sum_{l=1}^{N} x^\top \cdot H^{-1/2} \cdot \left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t | \boldsymbol{W}^{1:T} \right] \right) \quad a.s..$$

We want to derive an upper bound for the quantity $C := \sum_{j=1}^{n}\sum_{t=1}^{T}\left(x^\top \cdot H^{-1/2}\cdot Z_j^t\right)^2 - \sum_{j=1}^{n}\sum_{t=1}^{T}\mathbb{E}\left[\left(x^\top \cdot H^{-1/2}\cdot Z_j^t\right)^2\right]$. We recall that conditionally on $\boldsymbol{W}^{1:T}$,

$$x^\top \cdot H^{-1/2}\cdot \mathsf{diag}(A\boldsymbol{W}_j^t)\cdot H^{-1/2}\cdot x \leq \|x\|_2^2 \leq 1 \quad a.s.,$$

because we have almost surely, $A\boldsymbol{W}_j^t(i) \leq h_i$ and $A\boldsymbol{W}_j^t(m) \leq h_m$. Hence, given the definition of $\left(Q_{jl}^t\right)_{j,l,t}$ the following equalities hold almost surely for $(j,l,t)\in[p]\times[N]\times[T]$,

$$\begin{aligned}\mathbb{V}\left[x^\top \cdot H^{-1/2}\cdot\left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)|\boldsymbol{W}^{1:T}\right] &\leq 1,\\ \left|x^\top \cdot H^{-1/2}\cdot\left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)\right| &\leq \frac{2}{\sqrt{h_{\min}}}.\end{aligned}$$

Hence applying Hoeffding's inequality, for bounded random variables, Lemma 1.1.8, conditionally on $\boldsymbol{W}^{1:T}$ to $\frac{1}{N}\sum_{l=1}^{N} x^\top \cdot H^{-1/2}\cdot\left(Q_{jl}^t - \mathbb{E}\left[Q_{jl}^t|\boldsymbol{W}^{1:T}\right]\right)$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ gives, for all $(j,t)\in[n]\times[T]$, for all $s>0$,

$$\mathbb{P}\left[\left|x^\top \cdot H^{-1/2}\cdot Z_j^t\right| > s|\boldsymbol{W}^{1:T}\right] \leq 2\exp\left(-\frac{Nh_{\min}s^2}{8}\right). \tag{5.14}$$

On the other hand we adapt the proof of Hoeffding's lemma as follows to control the moment generating function of $x^\top \cdot H^{-1/2}\cdot(Q_{jl}^t(i) - \mathbb{E}[Q_{jl}^t|\boldsymbol{W}^{1:T}])$. It will allow to control the deviation of $x^\top \cdot H^{-1/2}\cdot Z_j^t$ with the conditional variance of Q_{jl}^t. We first consider, for all $(j,t,l,i)\in[n]\times[T]\times[N]\times[p]$ an identical and independent copy of Q_{jl}^t, conditionally on $\boldsymbol{W}^{1:T}$, that we name R_{jl}^t. We then consider their conditionally centered version, namely $Q_{jlt}^\top = Q_{jl}^t - \mathbb{E}[Q_{jl}^t|\boldsymbol{W}^{1:T}]$ and $R_{jlt}^\top = R_{jl}^t - \mathbb{E}[R_{jl}^t|\boldsymbol{W}^{1:T}]$. We first notice that the following equality holds for all $\lambda\in\mathbb{R}$, $\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2}\cdot Q_{jlt}^\top\right)|\boldsymbol{W}^{1:T}\right]$ is equal to $\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - \mathbb{E}_R\left[R_{jlt}^\top\right]\right)\right)|\boldsymbol{W}^{1:T}\right]$ where $\mathbb{E}_Q$ (respectively $\mathbb{E}_R$) denotes the conditional expectation taken *w.r.t* the distribution of $Q_{jlt}^\top$ (respectively $R_{jlt}^\top$). Then applying conditional Jensen's inequality provides, for all $\lambda\in\mathbb{R}$,

$$\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2}\cdot Q_{jlt}^\top\right)|\boldsymbol{W}^{1:T}\right] \leq \mathbb{E}_{Q,R}\left[\exp\left(\lambda x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - R_{jlt}^\top\right)\right)|\boldsymbol{W}^{1:T}\right].$$

We notice that the random variables $x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - R_{jlt}^\top\right)$ are symmetric and centered conditionally on $\boldsymbol{W}^{1:T}$. Indeed the random variables $x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - R_{jlt}^\top\right)$ and $x^\top \cdot H^{-1/2}\cdot\left(R_{jlt}^\top - Q_{jlt}^\top\right)$ share the same distribution conditionally on $\boldsymbol{W}^{1:T}$. This proves that for all $k\in\mathbb{N}$, if k is odd we get $\mathbb{E}_{Q,R}\left[\left(x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - R_{jlt}^\top\right)\right)^k|\boldsymbol{W}^{1:T}\right] = 0$ almost surely. We also note that conditionally on $\boldsymbol{W}^{1:T}$, the variables $x^\top \cdot H^{-1/2}\cdot\left(Q_{jlt}^\top - R_{jlt}^\top\right)$ are bounded almost surely in $[-4/\sqrt{h_{\min}}, 4/\sqrt{h_{\min}}]$. Taylor's theorem ensures that for all $\lambda\in\mathbb{R}$, for all $y\in[-4/\sqrt{h_{\min}}, 4/\sqrt{h_{\min}}]$, there exists $\gamma\in[\min(0,y);\max(0,y)]$ such that

$$\exp(\lambda x) = 1 + \lambda y + \frac{\lambda^2 y^2}{2} + \frac{\lambda^3 y^3\exp(\lambda\gamma)}{6}$$

If y is positive, then y^3 is positive and $\gamma \leq y$. We get $y^3 \exp(\lambda\gamma) \leq y^3 \exp(\lambda y)$. If y is negative, then y^3 is negative and $\gamma \geq y$. We get $y^3 \exp(\lambda\gamma) \leq y^3 \exp(\lambda y)$. Finally this leads to

$$\begin{aligned}\exp(\lambda y) &= 1 + \lambda x + \frac{\lambda^2 y^2}{2} + \frac{\lambda^3 y^3 \exp(\lambda y)}{6}, \\ &\leq 1 + \lambda y + \frac{\lambda^2 y^2}{2} + \frac{\lambda^3 y^3 \exp(4\lambda)}{6}.\end{aligned}$$

Finally this leads to the following inequality which holds almost surely,

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right)|\boldsymbol{W}^{1:T}\right] \leq 1 &+ \lambda \mathbb{E}_{Q,R}\left[x^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top)|\boldsymbol{W}^{1:T}\right] \\ &+ \frac{\lambda^2}{2}\mathbb{E}_{Q,R}\left[\left(x^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt})^\top\right)^2|\boldsymbol{W}^{1:T}\right] \\ &+ \frac{\lambda^3 \exp(4\lambda)}{6}\mathbb{E}_{Q,R}\left[\left(x^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top)\right)^3|\boldsymbol{W}^{1:T}\right].\end{aligned}$$

The conditional symmetry of $Q_{jlt}^\top - R_{jlt}^\top$ around zero ensures that its conditional odd moments are almost surely null and we get,

$$\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right)|\boldsymbol{W}^{1:T}\right] \leq 1 + \frac{\lambda^2}{2}\mathbb{V}_{Q,R}\left[x^\top \cdot H^{-1/2} \cdot (Q_{jlt}^\top - R_{jlt}^\top)|\boldsymbol{W}^{1:T}\right].$$

By independence and identical conditional distributions of $Q_{jlt}^\top$ and $R_{jlt}^\top$ we have

$$\begin{aligned}\mathbb{E}_Q\left[\exp\left(\lambda x^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top\right)|\boldsymbol{W}^{1:T}\right] &\leq 1 + \lambda^2 \mathbb{V}_Q\left[x^\top \cdot H^{-1/2} \cdot Q_{jlt}^\top|\boldsymbol{W}^{1:T}\right], \\ &\leq 1 + \lambda^2.\end{aligned}$$

This ensures the following equalities which hold almost surely for all $\lambda \in \mathbb{R}$ and for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left[\sum_l x^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top \geq t|\boldsymbol{W}^{1:T}\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_l x^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top(i)\right) \geq \exp(\lambda t)|\boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t)\mathbb{E}\left[\exp\left(\lambda \sum_l x^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right)|\boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t)\prod_l \mathbb{E}\left[\exp\left(\lambda x^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right)|\boldsymbol{W}^{1:T}\right], \\ &\leq \exp(-\lambda t)\prod_l \exp\left(\lambda^2\right), \\ &\leq \exp(-\lambda t)\exp\left(N\lambda^2\right).\end{aligned}$$

Choosing $\lambda := \frac{t}{2N}$ and taking the conditional expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to, for all $\epsilon > 0$, for all $j \in [n]$ and for all $t \in [T]$

$$\mathbb{P}\left[\left|\sum_l x^\top \cdot H^{-1/2} \cdot Q_{jtl}^\top\right| \leq t\right] \geq 1 - 2\exp\left(-\frac{t^2}{4N}\right).$$

Finally, for all $j \in [n]$, for all $t \in [T]$ and for all $\epsilon > 0$, it comes

$$\mathbb{P}\left[\left|x^\top \cdot H^{-1/2} \cdot Z_j^t\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{N\epsilon^2}{4}\right).$$

We finally get for all $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|x^\top \cdot H^{-1/2} \cdot Z_j^t\right| < \frac{2\epsilon}{\sqrt{N}}. \tag{5.15}$$

Combining (5.14) and (5.15) leads to, for all $\epsilon > 0$, with probability at least $1 - 2\exp(-\epsilon^2)$,

$$\left|x^\top \cdot H^{-1/2} \cdot Z_j^t\right| < \frac{2\min(\sqrt{2}h_{\min}, 1)\epsilon}{\sqrt{N}}.$$

Equivalently, this says that for all $\epsilon > 0$,

$$\mathbb{P}\left[\left|x^\top \cdot H^{-1/2} \cdot Z_j^t\right| < \epsilon\right] \geq 1 - 2\exp\left(-\frac{N\max(h_{\min}/2, 1)\epsilon^2}{4}\right).$$

This proves that the variables $(x^\top \cdot H^{-1/2} \cdot Z_j^t)$ are SubGaussian. We recall that the SubGaussian norm of a SubGaussian random variable X is defined as

$$\|X\|_{\psi_2} := \inf_{s>0}\left\{\mathbb{E}\left[\frac{X^2}{s^2}\right] \leq 2\right\}.$$

Hence for all $(j,t) \in [n] \times [T]$, the SubGaussian norm of $x^\top \cdot H^{-1/2} \cdot Z_j^t$ satisfies $\|x^\top \cdot H^{-1/2} \cdot Z_j^t\|_{\psi_2} \leq 12 \cdot \sqrt{\frac{e}{N\max(h_{\min}/2, 1)\log(2)}}$. Indeed, Proposition 2.5.2 in [130] proves that for a random variable X satisfying, for all $s > 0$, $\mathbb{P}[|X| > s] \leq 2\exp\left(\frac{-s^2}{K_1^2}\right)$ where $K_1 > 0$ is a constant then $\mathbb{E}\left[\frac{X^2}{\left(6K_1\sqrt{e/\log(2)}\right)^2}\right] \leq 2$. This proves the stated result for $K_1^2 = \frac{4}{N\max(h_{\min}/2, 1)}$. In addition we immediately get that $(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2$ is SubExponential, see Lemma 2.7.6 in [130]. This Lemma also ensures that its Subexponential norm satisfies $\|(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2\|_{\psi_1} = \|(x^\top \cdot H^{-1/2} \cdot Z_j^t)\|_{\psi_2}^2$. Moreover, recalling that a norm is a convex function and using Jensen's inequality provides that its SubExponential norm also satisfies the centering property $\|(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2 - \mathbb{E}\left[(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2\right]\|_{\psi_1} \leq 2\|x^\top \cdot H^{-1/2} \cdot Z_j^t\|_{\psi_2}^2 \leq \frac{288 \cdot e}{N\max(h_{\min}/2, 1)\log(2)} := \gamma$. Using Bernstein's inequality for SubExponential random variables, Lemma 1.1.10, conditionally on $\boldsymbol{W}^{1:T}$ leads to, for all $s > 0$ and for an absolute constant $c > 0$,

$$\mathbb{P}\left[\left|\sum_{j=1}^n \sum_{t=1}^T \left[(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2 - \mathbb{E}\left[(x^\top \cdot H^{-1/2} \cdot Z_j^t)^2\right]\right]\right| > nTs | \boldsymbol{W}^{1:T}\right] \leq 2\exp\left(-cnT\min\left(\frac{s^2}{\gamma^2}; \frac{s}{\gamma}\right)\right),$$

where $\gamma := \frac{288 \cdot e}{N \max(h_{\min}/2, 1) \log(2)}$. Considering $\epsilon > 0$, choosing $s = \frac{\gamma \epsilon}{\sqrt{cnT}}$ and taking on both sides the expectation *w.r.t* $\boldsymbol{W}^{1:T}$ leads to

$$|C| < \frac{288 \cdot e}{\log(2)} \cdot \frac{\sqrt{nT}\epsilon}{N\sqrt{c}\max(h_{\min}/2, 1)},$$

with probability at least $1 - 2\exp\left(-\min\left(\epsilon^2, \sqrt{cnT}\epsilon\right)\right)$. Using a union bound over the ϵ-net we get that with probability at least $1 - 2\exp\left(p\log(9) - \min\left(\epsilon^2, \sqrt{cnT}\epsilon\right)\right)$ we have

$$\begin{aligned}&\frac{1}{2}\|H^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)H^{-1/2}\| \\ <&\frac{288 \cdot e}{\log(2)} \cdot \frac{\sqrt{nT}\epsilon}{N\sqrt{c}\max(h_{\min}/2, 1)}.\end{aligned}$$

■

5.5.5 Proof of Proposition 5.2.8

Proof of Proposition 5.2.8. For all $i \in [p]$, we have

$$[\boldsymbol{M}_*]_{ii} = \frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[A^*]_{ik}\boldsymbol{W}_j^t(k) = \sum_{k=1}^{K}[A^*]_{ik}\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k).$$

Moreover Proposition 4.3.4 ensures that almost surely,

$$\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\boldsymbol{W}_j^t(k) \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}).$$

The stated results follows. ■

5.5.6 Proof of Proposition 5.2.9

Proof of Proposition 5.2.9. We start by recalling that the matrices $\boldsymbol{M}_*$ and H are diagonal and thus commute. Then we re-write the matrix $\boldsymbol{G}_*$ as follows, $\boldsymbol{G}_* = \left(1 - \frac{1}{N}\right) H^{1/2}\boldsymbol{M}_*^{-1/2}H^{-1/2}A^*\boldsymbol{W}^{1:T}\left[A^*\boldsymbol{W}^{1:T}\right]^\top H^{-1/}$

Hence, for any matrix M, denoting $\lambda_{\min}(M)$ its smallest non-zero eigenvalue, we have

$$\lambda_{\min}(\boldsymbol{G}_*) \geq \left(1 - \frac{1}{N}\right)\lambda_{\min}\left(H^{1/2}\boldsymbol{M}_*^{-1/2}\right)\lambda_{\min}\left(H^{-1/2}A^*\boldsymbol{W}^{1:T}\left[A^*\boldsymbol{W}^{1:T}\right]^\top H^{-1/2}\right)\lambda_{\min}\left(\boldsymbol{M}_*^{-1/2}H^{1/2}\right).$$

However $\boldsymbol{M}_*$ and H being diagonal, we obtain almost surely

$$\lambda_{\min}(\boldsymbol{G}_*) \geq \left(1 - \frac{1}{N}\right)\lambda_{\min}\left(\boldsymbol{M}_*^{-1}H\right)\lambda_{\min}\left(H^{-1/2}A^*\boldsymbol{W}^{1:T}\left[A^*\boldsymbol{W}^{1:T}\right]^\top H^{-1/2}\right).$$

Moreover Proposition 5.2.8 ensures that $\lambda_{\min}\left(\boldsymbol{M}_*^{-1}H\right) \geq 1$ almost surely. Hence we get almost surely that

$$\lambda_{\min}(\boldsymbol{G}_*) \geq \left(1-\frac{1}{N}\right)\lambda_{\min}\left(\boldsymbol{\Psi}\right),$$

for $\boldsymbol{\Psi} = H^{-1/2}A^*\boldsymbol{W}^{1:T}\left[A^*\boldsymbol{W}^{1:T}\right]^\top H^{-1/2}$. In addition, Theorem 4.3.3 proves that $\boldsymbol{W}^{1:T}[\boldsymbol{W}^{1:T}]^\top \in \mathbb{R}^{K\times K}$ is symmetric positive definite with high probability. In addition $H^{-1/2}A^* \in \mathbb{R}^{p\times K}$ satisfies $\operatorname{rank}(H^{-1/2}A^*) = K$. Using lemma 5.6.3 and assumption 6 we get the following inequalities holding true almost surely,

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{G}_*) &\geq \left(1-\frac{1}{N}\right)\lambda_{\min}\left([A^*]^\top H^{-1}A^*\right)\lambda_{\min}\left(\boldsymbol{W}^{1:T}[\boldsymbol{W}^{1:T}]^\top\right),\\
&\geq \left(1-\frac{1}{N}\right)nT\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}),\\
&\geq \left(1-\frac{1}{N}\right)nT\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 \geq \left(1-\frac{1}{N}\right)nTc_2^2.
\end{aligned}$$

To bound from above almost surely the largest singular value of $\boldsymbol{G}_*$ we recall that for any matrices U and V we have $\lambda_1(UV) = \lambda_1(VU)$. In addition the spectral norm is sub-multiplicative and by defining $\boldsymbol{\Omega} := H^{1/2}\boldsymbol{M}_*^{-1/2}$ we get almost surely

$$\lambda_1\left(\boldsymbol{G}_*\right) \leq \left(1-\frac{1}{N}\right)\lambda_1(\boldsymbol{\Psi})\lambda_1(\boldsymbol{\Omega}^\top\boldsymbol{\Omega}).$$

However the following equalities hold true almost surely : $\lambda_1(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top) = \lambda_1(\boldsymbol{\Omega}^\top\boldsymbol{\Omega}) = \lambda_1(\boldsymbol{M}_*^{-1}H) \leq \frac{1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} \leq \frac{1}{c_2}$.

Similarly there is almost surely $\lambda_1(\boldsymbol{\Psi}) \leq \lambda_1\left([A^*]^\top H^{-1}A^*\right)\lambda_1\left(\boldsymbol{W}^{1:T}[\boldsymbol{W}^{1:T}]^\top\right)$. Finally this leads, under assumption 7 to, almost surely,

$$\lambda_1\left(\boldsymbol{G}_*\right) \leq \left(1-\frac{1}{N}\right)\frac{nT}{c_2}\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}).$$

We conclude using Proposition 4.3.1.

To prove the last inequality, we start by noting that $\boldsymbol{G}_*$ and $\left(1-\frac{1}{N}\right)nT\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top\boldsymbol{M}_*^{-1}A^*\right)$ share almost surely the same eigenvalues. Thus

$$\lambda_1(\boldsymbol{G}_*) - \lambda_2(\boldsymbol{G}_*) = \left(1-\frac{1}{N}\right)nT\left[\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top\boldsymbol{M}_*^{-1}A^*\right)) - \lambda_2\left(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top\boldsymbol{M}_*^{-1}A^*\right)\right)\right]$$

Assumption 7 ensures that

$$|\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top\boldsymbol{M}_*^{-1}A^*\right)) - \lambda_2(\Sigma_{\boldsymbol{W}}^{1:T}\left([A^*]^\top\boldsymbol{M}_*^{-1}A^*\right))| \geq c_3.$$

Finally this leads to the following inequality holding true almost surely, for all $k \geq 2$,

$$\left(1-\frac{1}{N}\right)nTc_3 + \lambda_2(\boldsymbol{G}_*) \leq \lambda_1(\boldsymbol{G}_*).$$

■

5.5.7 Proof of Proposition 5.2.10

Proof of Proposition 5.2.10. We start by proving that there exists a non singular matrix $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ such that almost surely there are

$$\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1} = [A^*]^\top \boldsymbol{M}_*^{-1} A^*,$$
$$U = \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{B}.$$

We recall that the matrix $\boldsymbol{\Pi}_* := \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}$ is almost surely of rank K and its SVD is defined as $\boldsymbol{\Pi}_* = U\Sigma V$. Hence by definition $U^\top U = VV^\top = I_K$ and Σ is diagonal and invertible. Hence the following equalities hold almost surely,

$$\begin{aligned} U &= (U\Sigma V)V^\top\Sigma^{-1}, \\ &= \boldsymbol{\Pi}_* V^\top \Sigma^{-1}, \\ &= \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T} V^\top \Sigma^{-1}. \end{aligned}$$

Defining $\boldsymbol{B} := \boldsymbol{W}^{1:T} V^\top \Sigma^{-1}$ proves that $U = \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{B}$ almost surely. In addition $U^\top \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{B} = U^\top U = I_K$ almost surely and thus $\boldsymbol{B}$ is uniquely defined and almost surely non-singular. Finally $U^\top U = \boldsymbol{B}^\top [A^*]^\top \boldsymbol{M}_*^{-1} A^* \boldsymbol{B} = I_K$ almost surely. Hence $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{B}\boldsymbol{B}^\top [A^*]^\top \boldsymbol{M}_*^{-1} A^* \boldsymbol{B}\boldsymbol{B}^\top$ almost surely. This proves the stated results and thus for each $i \in [p]$ we have almost surely

$$U_{i.} = [\boldsymbol{M}_*^{-1/2}]_{ii} \boldsymbol{B} A^*_{i.}$$

Proposition 5.2.8 ensures that almost surely we have

$$\begin{aligned} \|U_{i.}\|_2 \leq \frac{\|\boldsymbol{B}\|_{op}\|A^*_{i.}\|_2}{\sqrt{[\boldsymbol{M}_*]_{ii}}} &\leq \frac{\lambda_{min}^{-1/2}\left([A^*]^\top \boldsymbol{M}_*^{-1} A^*\right)\|A^*_{i.}\|_1}{\sqrt{[\boldsymbol{M}_*]_{ii}}} \\ &\leq \frac{\lambda_{min}^{-1/2}\left([A^*]^\top H^{-1} A^*\right) h_i}{\sqrt{[\boldsymbol{M}_*]_{ii}}} \\ &\leq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2} h_i}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i}} \leq \frac{\sqrt{h_i}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} \end{aligned}$$

Finally we get

$$\|U_{i.}\|_2 \leq \sqrt{K}\|U_{i.}\|_1 \leq \frac{\sqrt{K h_i}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}.$$

■

5.5.8 Proof of Theorem 5.2.11

Theorem 5.5.1 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6 and 7. Then for all $i \in [p]$ and for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2K\exp(-\epsilon_2^2) -$*

$2pK\exp(-\epsilon_3^2)-4p\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$, *the quantity* $h_i^{-1/2}\left\|e_i^\top\left(\hat{G}-\boldsymbol{G}_*\right)\right\|_2$ *is bounded from above by*

$$2\epsilon_1\sqrt{\frac{nTp}{N}}\frac{\sqrt{p}}{Nc_1K}\cdot\left(c_2-2\epsilon_1\sqrt{\frac{1}{h_{\min}NnT}}\right)^{-1}+2\sqrt{\frac{nTp}{N}}\frac{\epsilon_3+\epsilon_2\sqrt{K/c_1}}{c_2\sqrt{1-\frac{2\epsilon_1}{c_2}\sqrt{\frac{1}{NnTh_{\min}}}}}\left(1+\frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)$$
$$+2\epsilon_4\sqrt{\frac{nTp}{N}}\cdot\frac{288\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\sqrt{p}}{\sqrt{Nc_1K}c_2}\left(1+\frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)^2+2\epsilon_1K\sqrt{\frac{nTp}{N}}\left(c_2-\frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-2}$$
$$+2\epsilon_1\sqrt{\frac{nTp}{N}}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\left(c_2-\frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.$$

where c *is an absolute constant appearing in Lemma 1.1.10 and for all* $i\in[p]$,

$$\boldsymbol{\Delta}_i:=c_2h_i-2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}},\quad \xi_i:=\left(\boldsymbol{\Delta}_i^{-3/2}\sqrt{h_i\min(2,h_i)}\right).$$

Remark 5.5.1 *Theorem 5.5.1 improves the result presented in Lemma F.4 in [84]. Specifically, by setting*

$$\epsilon_1^2=\log(p)+5\log(nT),\quad \epsilon_2^2=\log(K)+5\log(nT),\quad \epsilon_3^2=\log(pK)+5\log(nT),$$
$$\epsilon_4^2=\log(p)+5\log(nT),$$

it establishes that with probability at least $1-10(nT)^{-5}$ *if* $c\geq\dfrac{\log(p)+5\log(nT)}{nT}$ *and with probability at least* $1-2\exp\left(-\sqrt{cnT\left(\log(p)+5\log(nT)\right)}\right)-6(nT)^{-5}$ *if* $c\leq\dfrac{\log(p)+5\log(nT)}{nT}$ *we have, for all* $i\in[p]$,

$h_i^{-1/2} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_* \right) \right\|_2$ *bounded from above by*

$$
\begin{aligned}
& 2\sqrt{\frac{nTp(\log(p)+5\log(nT))}{N}} \frac{\sqrt{p}}{Nc_1K} \cdot \left(c_2 - 2\sqrt{\frac{p(\log(p)+5\log(nT))}{c_1KNnT}} \right)^{-1} \\
& +2\sqrt{\frac{nTp}{N}} \frac{\sqrt{\log(pK)+5\log(nT)} + \sqrt{K(\log(K)+5\log(nT))/c_1}}{c_2\sqrt{1-\frac{2\epsilon_1}{c_2}\sqrt{\frac{1}{NnTh_{\min}}}}} \left(1 + \frac{\epsilon_1 \max\limits_{i\in[p]} \xi_i}{\sqrt{NnT}} \right) \\
& +2\sqrt{\frac{nTp(\log(p)+5\log(nT))}{N}} \cdot \frac{288 \cdot e}{\log(2)\sqrt{c}} \cdot \frac{\sqrt{p}}{\sqrt{Nc_1K}c_2} \left(1 + \frac{\epsilon_1 \max\limits_{i\in[p]} \xi_i}{\sqrt{NnT}} \right)^2 \\
& +2K\sqrt{\frac{nTp(\log(p)+5\log(nT))}{N}} \left(c_2 - 2\sqrt{\frac{p(\log(p)+5\log(nT))}{c_1KNnT}} \right)^{-2} \\
& + 2\sqrt{\frac{nTp(\log(p)+5\log(nT))}{N}} \frac{\sqrt{K}}{\sqrt{c_1c_2}} \left(c_2 - 2\sqrt{\frac{p(\log(p)+5\log(nT))}{c_1KNnT}} \right)^{-3/2}.
\end{aligned}
$$

Notably, unlike Lemma F.4 in [84], Theorem 5.5.1 does not require any assumption on either the number nT of documents or the size of the number of words per documents N compared to the vocabulary size p. Moreover, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided. Focusing on the asymptotic behaviour of this upper bound we have, when nT goes to infinity and assuming K, p, N remain fixed, for all $i \in [p]$, with probability at least $1 - o_{nT\to\infty}\left((nT)^{-3}\right)$, $h_i^{-1/2} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_* \right) \right\|_2$ bounded from above by

$$
\left(C_1 \frac{\sqrt{p}}{N} + C_2 + C_3\sqrt{\frac{p}{N}} + C_4 \right) \sqrt{\frac{nTp\log(nT)}{N}},
$$

where

$$
\begin{aligned}
& C_1 = \frac{10}{c_1Kc_2}, \quad C_2 = 10(1+\sqrt{K/c_1}), \quad C_3 = \frac{2880 \cdot e}{\log(2)c_2\sqrt{c_1Kc}}, \\
& C_4 = \frac{10K}{c_2^2} + \frac{10\sqrt{K}}{c_2^2\sqrt{c_1}}.
\end{aligned}
$$

It is finally noteworthy to state that with probability at least $1 - o_{nT\to\infty}\left((nT)^{-3}\right)$, for all $i \in [p]$, the following inequality is asymptotically holding true,

$$
\frac{\| e_i^\top \left(\hat{G} - \boldsymbol{G}_* \right) \|_2}{\sqrt{h_i}} \leq C(1 + N^{-1/2}p^{1/2} + N^{-1}p^{1/2})\sqrt{\frac{nTp\log(nT)}{N}}
$$

where $C = 2\max(C_1, C_2, C_3, C_4)$. This improves the result presented under an asymptotic framework in Lemma F.4 in [84].

Proof of Theorem 5.5.1. We start by considering the matrix $\boldsymbol{Z}^{1:T} := \boldsymbol{Y}^{1:T} - A^*\boldsymbol{W}^{1:T}$ and we recall that $\boldsymbol{\Pi}^{1:T} := A^*\boldsymbol{W}^{1:T}$ and that $\hat{G} - \boldsymbol{G}_* = \hat{M}^{-1/2}\boldsymbol{Y}^{1:T}\left(\boldsymbol{Y}^{1:T}\right)^\top \hat{M}^{-1/2} - \frac{nT}{N} I_p - \left(1 - \frac{1}{N}\right) \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}\left(\boldsymbol{\Pi}^{1:T}\right)^\top \boldsymbol{M}_*^{-1/2}$. Hence according to model (4.1) under the constraints defined in (4.2) and (4.3), the conditional distribution of $\boldsymbol{Z}^{1:T}$ given $\boldsymbol{W}^{1:T}$ allows to derive its conditional covariance matrix as follows. First notice that for each $j \in [n]$ and for all $t \in [T]$, $Z_j^t \in \mathbb{R}^p$ and $\mathbb{V}\left[Z_j^t | \boldsymbol{W}_j^t\right] = \frac{1}{N}\left[\mathsf{diag}(A^*\boldsymbol{W}_j^t) - (A^*\boldsymbol{W}_j^t)(A^*\boldsymbol{W}_j^t)^\top\right]$. Hence we derive the following equalities,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top | \boldsymbol{W}^{1:T}\right] &= \frac{1}{N}\sum_{j=1}^{n}\sum_{t=1}^{T}\left[\mathsf{diag}(A^*\boldsymbol{W}_j^t) - (A^*\boldsymbol{W}_j^t)(A^*\boldsymbol{W}_j^t)^\top\right], \\ &= \frac{1}{N}\left[\boldsymbol{M}_* - (\boldsymbol{\Pi}^{1:T})(\boldsymbol{\Pi}^{1:T})^\top\right].\end{aligned}$$

We then rewrite the matrix $\hat{G} - \boldsymbol{G}_*$ as a sum of quantities which we can control. We define

$$\begin{aligned}R_1 &:= \frac{nT}{N}\hat{M}^{-1/2}\left(\boldsymbol{M}_* - \hat{M}\right)\hat{M}^{-1/2}, \\ R_2 &:= \hat{M}^{-1/2}\left(\boldsymbol{\Pi}^{1:T}(\boldsymbol{Z}^{1:T})^\top + \boldsymbol{Z}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\right)\hat{M}^{-1/2}, \\ R_3 &:= \hat{M}^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\hat{M}^{-1/2}, \\ R_4 &:= \left(1 - \frac{1}{N}\right)\left(\hat{M}^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\boldsymbol{M}_*^{-1/2}\right).\end{aligned}$$

We recall that the matrices $\hat{M}$ and $\boldsymbol{M}_*$ are diagonal. Hence one can verify that these quantities can be expanded as follows,

$$\begin{aligned}R_1 &= \frac{nT}{N}\hat{M}^{-1}\boldsymbol{M}_* - \frac{nT}{N}I_p, \\ R_2 &= \hat{M}^{-1/2}\left(\boldsymbol{\Pi}^{1:T}(\boldsymbol{Y}^{1:T})^\top - \boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top + \boldsymbol{Y}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top - \boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\right)\hat{M}^{-1/2}, \\ R_3 &= \hat{M}^{-1/2}\left(\boldsymbol{Y}^{1:T}(\boldsymbol{Y}^{1:T})^\top - \boldsymbol{\Pi}^{1:T}(\boldsymbol{Y}^{1:T})^\top - \boldsymbol{Y}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top + \boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\right)\hat{M}^{-1/2} \\ &\quad - \frac{nT}{N}\hat{M}^{-1}\boldsymbol{M}_* + \frac{1}{N}\hat{M}^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top M^{-1/2}, \\ R_4 &= \left(1 - \frac{1}{N}\right)\left(\hat{M}^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\boldsymbol{M}_*^{-1/2}\right).\end{aligned}$$

Hence it is easy to verify that $\hat{G} - \boldsymbol{G}_* = \sum_{s=1}^{4} R_s$. We consider $(e_1, \ldots, e_p)$ the canonical basis of $\mathbb{R}^p$. This gives that for all $i \in [p]$,

$$\|e_i^\top(\hat{G} - \boldsymbol{G}_*)\|_2 \leq \sum_{s=1}^{4}\|e_i^\top R_s\|_2.$$

We now aim to bound each $\|e_i^\top R_s\|_2$ with high probability. We start with R_1. For all $i \in [p]$ we have

$$\|e_i^\top R_1\|_2 = [R_1]_{ii} = \frac{nT}{N}[\hat{M}^{-1}]_{ii}\left([\boldsymbol{M}_*]_{ii} - [\hat{M}]_{ii}\right).$$

Proposition 5.2.1 ensures that for all $i \in [p]$, for all $\epsilon_1 > 0$ with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$, we have

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| < 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}.$$

Moreover Proposition 5.2.8 gives that almost surely for all $i \in [p]$,

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

Hence we obtain that with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$ we have, for all $i \in [p]$,

$$\begin{aligned}
[\hat{M}]_{ii} &> [\boldsymbol{M}_*]_{ii} - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}, \\
&> \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}.
\end{aligned}$$

This leads to, for all $i \in [p]$, with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$,

$$\begin{aligned}
\|e_i^\top R_1\|_2 &\leq \frac{2nT\epsilon_1}{N} \frac{\sqrt{\min(2, h_i)}}{\sqrt{NnT}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}\right)^{-1}, \\
&\leq \frac{2\epsilon_1 \sqrt{nT \min(2, h_i)}}{N^{3/2}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon \sqrt{\frac{\min(2, h_i)}{NnT}}\right)^{-1}, \\
&\leq \frac{2\epsilon_1 \sqrt{nTh_i}}{N^{3/2}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon \sqrt{\frac{h_i}{NnT}}\right)^{-1}, \\
&\leq \frac{2\epsilon_1 \sqrt{nTh_i}}{N^{3/2} h_i} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon \sqrt{\frac{1}{NnTh_i}}\right)^{-1}, \\
&\leq \frac{2\epsilon_1 \sqrt{nT}}{N\sqrt{Nh_i}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon \sqrt{\frac{1}{h_i NnT}}\right)^{-1}.
\end{aligned}$$

We now consider R_2 and we note that $\boldsymbol{\Pi}^{1:T} := A^* \boldsymbol{W}^{1:T} = \sum_{k=1}^{K} [A^*]_{.k} \left([\boldsymbol{W}^{1:T}]_{k.}\right)^\top$. Hence we get

$$R_2 = \sum_{k=1}^{K} \left(\hat{M}^{-1/2} [A^*]_{.k} \left(\hat{M}^{-1/2} \boldsymbol{Z}^{1:T} [\boldsymbol{W}^{1:T}]_{k.} \right)^\top + \hat{M}^{-1/2} \boldsymbol{Z}^{1:T} [\boldsymbol{W}^{1:T}]_{k.} \left(\hat{M}^{-1/2} [A^*]_{.k} \right)^\top \right).$$

From this result we derive that for all $i \in [p]$, we have

$$\begin{aligned}
\|e_i^\top R_2\|_2 \leq &\sum_{k=1}^{K} [A^*]_{ik} [\hat{M}^{-1/2}]_{ii} \|\hat{M}^{-1/2} \boldsymbol{Z}^{1:T} [\boldsymbol{W}^{1:T}]_{k.}\|_2 \\
&+ \sum_{k=1}^{K} |[\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.}| [\hat{M}^{-1/2}]_{ii} \|\hat{M}^{-1/2} [A^*]_{.k}\|_2
\end{aligned}$$

First Proposition 5.2.3 ensures that for all $i \in [p]$ and for all $k \in [K]$, for all $\epsilon_2 > 0$ with probability at least $1 - 2\exp(-\epsilon_2^2)$ we have

$$\left| [\boldsymbol{Z}^{1:T}]_{i.}^\top [\boldsymbol{W}^{1:T}]_{k.} \right| < 2\epsilon_2 \sqrt{\frac{h_i nT}{N}}.$$

Moreover Corollary 5.2.5 ensures that for all for all $\epsilon_3 > 0$ and for all $k \in [K]$, with probability at least $1 - 2p\exp(-\epsilon_3^2)$, we have

$$\left\| \boldsymbol{M}_*^{-1/2} \boldsymbol{Z}^{1:T} [\boldsymbol{W}^{1:T}]_{k.} \right\|_2 \leq 2\epsilon_3 \sqrt{\frac{pnT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}.$$

In addition $\sum_{k=1}^K [A^*]_{ik} = h_i$ and we recall that for all $i \in [p]$, and for all $\epsilon_1 > 0$ with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$[\hat{M}^{-1/2}]_{ii} < \left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-1/2}.$$

We also recall that for all $i \in [p]$, Proposition 5.2.8 ensures that

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

Furthermore, the function $x \mapsto x^{-1/2}$ is convex and Lemma 5.6.5 ensures that for all $(x, y) \in \mathbb{R}^2$ such that $x > y$, we have

$$(x - y)^{-1/2} \leq x^{-1/2} + \frac{1}{2}(x - y)^{-3/2} y.$$

In addition we recall the previously proved result, that for all $i \in [p]$ and for all $\epsilon_1 > 0$ with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$[\hat{M}]_{ii} \geq \left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right). \tag{5.16}$$

This provides especially with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$[\hat{M}^{-1/2}]_{ii} < [\boldsymbol{M}_*^{-1/2}]_{ii} + \epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-3/2}.$$

Hence we get that for all $i \in [p]$, with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$\begin{aligned}
[\hat{M}^{-1/2}]_{ii} [\boldsymbol{M}_*^{1/2}]_{ii} &\leq 1 + [\boldsymbol{M}_*^{1/2}]_{ii} \epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-3/2}, \\
&\leq 1 + \sqrt{h_i} \epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-3/2}
\end{aligned}$$

Moreover $\|\hat{M}^{-1/2} \boldsymbol{M}_*^{1/2}\|_{op} = \max_{i \in [p]} \left([\hat{M}^{-1/2}]_{ii} [\boldsymbol{M}_*^{1/2}]_{ii} \right)$ which leads to, with probability at least $1 - 2p\exp(-\epsilon_1^2)$,

$$\|\hat{M}^{-1/2} \boldsymbol{M}_*^{1/2}\|_{op} \leq 1 + \max_{i \in [p]} \left(\sqrt{h_i} \epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-3/2} \right).$$

In addition we have, for all $k \in [K]$,

$$\begin{aligned}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2^2 &= \sum_{i=1}^{p}[\boldsymbol{M}_*^{-1/2}]_{ii}^2[A^*]_{ik}^2, \\ &\leq \sum_{i=1}^{p}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}h_i^{-1}[A^*]_{ik}^2.\end{aligned}$$

Hence we deduce from the definition of the quantities h_i and by recalling that $A^* \in \mathbb{R}_+^{p\times K}$ has columns summing to one the following inequality,

$$\begin{aligned}\sum_{k=1}^{K}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2^2 &\leq \sum_{k=1}^{K}\sum_{i=1}^{p}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}h_i^{-1}[A^*]_{ik}^2, \\ &\leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\sum_{k=1}^{K}\sum_{i=1}^{p}[A^*]_{ik} \leq \frac{K}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}.\end{aligned}$$

Then Cauchy-Schwarz inequality ensures that

$$\left(\sum_{k=1}^{K}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2\right)^2 \leq K\sum_{k=1}^{K}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2^2 \leq \frac{K^2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}.$$

For notation simplicity, we denote, for all $i \in [p]$,

$$\begin{aligned}\boldsymbol{\Delta}_i &:= \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}, \\ \xi_i &:= \left(\boldsymbol{\Delta}_i^{-3/2}\sqrt{h_i\min(2,h_i)}\right).\end{aligned}$$

We have especially established with (5.16) that for all $i \in [p]$, $[\hat{M}^{-1/2}]_{ii} \leq \boldsymbol{\Delta}_i^{-1/2}$ with probability at least $1-2\exp(-\epsilon_1^2)$. The previously proved results provide, for all $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)$,

$$\begin{aligned}\|e_i^\top R_2\|_2 \leq h_i\boldsymbol{\Delta}_i^{-1/2}\cdot\left(1+\max_{i\in[p]}\left(\sqrt{h_i}\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\boldsymbol{\Delta}_i^{-3/2}\right)\right)\cdot 2\epsilon_3\sqrt{\frac{pnT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}} \\ + 2\epsilon_2\sqrt{\frac{h_inT}{N}}\boldsymbol{\Delta}_i^{-1/2}\cdot\left(1+\max_{i\in[p]}\left(\sqrt{h_i}\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\boldsymbol{\Delta}_i^{-3/2}\right)\right)\cdot\frac{K}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}.\end{aligned}$$

Hence, for all $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)$,

$$\begin{aligned}\|e_i^\top R_2\|_2 &\leq 2\boldsymbol{\Delta}_i^{-1/2}\sqrt{\frac{nT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\left(1+\frac{\epsilon_1\max_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)\left[h_i\epsilon_3\sqrt{p}+K\epsilon_2\sqrt{h_i}\right], \\ &\leq 2\left(1-\frac{2\epsilon_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i}\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-1/2}\sqrt{\frac{nT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2}}\left(1+\frac{\epsilon_1\max_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)\left[\epsilon_3\sqrt{h_ip}+K\epsilon_2\right].\end{aligned}$$

We now consider $R_3 := \hat{M}^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\hat{M}^{-1/2}$. Hence the following results come, holding for all $i \in [p]$,

$$\begin{aligned}
\|e_i^\top R_3\|_2^2 &= \sum_{s=1}^{p}[R_3]_{is}^2, \\
&= \sum_{s=1}^{p}\frac{\left([\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.}\right]\right)^2}{[\hat{M}]_{ii}[\hat{M}]_{ss}}, \\
&= \sum_{s=1}^{p}\frac{\left([\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.}\right]\right)^2[\boldsymbol{M}_*]_{ii}[\boldsymbol{M}_*]_{ss}}{[\hat{M}]_{ii}[\hat{M}]_{ss}[\boldsymbol{M}_*]_{ii}[\boldsymbol{M}_*]_{ss}}, \\
&= \sum_{s=1}^{p}\left([\hat{M}^{-1/2}]_{ii}[\boldsymbol{M}_*^{1/2}]_{ii}\right)^2\left([\hat{M}^{-1/2}]_{ss}[\boldsymbol{M}_*^{1/2}]_{ss}\right)^2\frac{\left([\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.}\right]\right)^2}{[\boldsymbol{M}_*]_{ii}[\boldsymbol{M}_*]_{ss}}, \\
&\leq \|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\|_{op}^4\sum_{s=1}^{p}\frac{\left([\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.}\right]\right)^2}{[\boldsymbol{M}_*]_{ii}[\boldsymbol{M}_*]_{ss}}.
\end{aligned}$$

Moreover, Proposition 5.2.6 ensures that for all $(i,s) \in [p]^2$, for all $\epsilon_4 > 0$, with probability at least $1 - 4\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$ and for an absolute constant $c > 0$, we have

$$\left|[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.} - \mathbb{E}\left[[\boldsymbol{Z}^{1:T}]_{i.}^\top[\boldsymbol{Z}^{1:T}]_{s.}\right]\right| < \frac{576\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon_4\sqrt{h_i\cdot h_s}\sqrt{nT}}{N\max(h_{\min}/2,1)}.$$

Furthermore, we proved that for all $\epsilon_1 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2)$,

$$\|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\|_{op} \leq \left(1 + \frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right).$$

In addition, Proposition 5.2.8 ensures that for all $i \in [p]$,

$$[\boldsymbol{M}_*]_{ii} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i.$$

Hence, for all $i \in [p]$, for all $\epsilon_1, \epsilon_4 > 0$, with probability at least $1 - 4p\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right) - 2p\exp(-\epsilon_1^2)$ and for an absolute constant $c > 0$ we have

$$\|e_i^\top R_3\|_2^2 \leq \left(1 + \frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)^4 p\cdot\left(\frac{576\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon_4\sqrt{nT}}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\max(h_{\min}/2,1)}\right)^2.$$

Thus, for all $i \in [p]$, for all $\epsilon_1, \epsilon_4 > 0$, with probability at least $1 - 4p\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right) - 2p\exp(-\epsilon_1^2)$ and for an absolute constant $c > 0$ we have

$$\|e_i^\top R_3\|_2 \leq \left(1 + \frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)^2\cdot\frac{576\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon_4\sqrt{nTp}}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\max(h_{\min}/2,1)}.$$

We now consider $R_4 := \left(1-\frac{1}{N}\right)\left(\hat{M}^{-1/2}\mathbf{\Pi}^{1:T}(\mathbf{\Pi}^{1:T})^\top\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}(\mathbf{\Pi}^{1:T})^\top\boldsymbol{M}_*^{-1/2}\right)$. Moreover, $\mathbf{\Pi}^{1:T} := A^*\boldsymbol{W}^{1:T} = \sum_{k=1}^{K}[A^*]_{.k}\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top$. Hence we can re-write R_4 as follows,

$$
\begin{aligned}
R_4 &= \left(1-\frac{1}{N}\right)\hat{M}^{-1/2}\sum_{k=1}^{K}[A^*]_{.k}\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top(\sum_{l=1}^{K}[A^*]_{.l}\left[\boldsymbol{W}^{1:T}\right]_{l.}^\top)^\top\hat{M}^{-1/2} \\
&- \left(1-\frac{1}{N}\right)\boldsymbol{M}_*^{-1/2}\sum_{k=1}^{K}[A^*]_{.k}\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top(\sum_{l=1}^{K}[A^*]_{.l}\left[\boldsymbol{W}^{1:T}\right]_{l.}^\top)^\top\boldsymbol{M}_*^{-1/2}, \\
&= \left(1-\frac{1}{N}\right)\hat{M}^{-1/2}\sum_{k=1}^{K}\sum_{l=1}^{K}[A^*]_{.k}\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}[A^*]_{.l}^\top\hat{M}^{-1/2} \\
&- \left(1-\frac{1}{N}\right)\boldsymbol{M}_*^{-1/2}\sum_{k=1}^{K}\sum_{l=1}^{K}[A^*]_{.k}\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}[A^*]_{.l}^\top\boldsymbol{M}_*^{-1/2}, \\
&= \left(1-\frac{1}{N}\right)\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\hat{M}^{-1/2}[A^*]_{.k}[A^*]_{.l}^\top\hat{M}^{-1/2} \\
&- \left(1-\frac{1}{N}\right)\sum_{l=1}^{K}\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\boldsymbol{M}_*^{-1/2}[A^*]_{.k}[A^*]_{.l}^\top\boldsymbol{M}_*^{-1/2}, \\
&= \left(1-\frac{1}{N}\right)\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\hat{M}^{-1/2}[A^*]_{.k}[A^*]_{.l}^\top\left(\hat{M}^{-1/2}-\boldsymbol{M}_*^{-1/2}\right) \\
&+ \left(1-\frac{1}{N}\right)\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\left(\hat{M}^{-1/2}-\boldsymbol{M}_*^{-1/2}\right)[A^*]_{.k}[A^*]_{.l}^\top\boldsymbol{M}_*^{-1/2}.
\end{aligned}
$$

This leads to, for all $i \in [p]$,

$$
\begin{aligned}
\|e_i^\top R_4\|_2 \leq nT\left[\hat{M}^{-1/2}\right]_{ii}\sum_{k=1}^{K}[A^*]_{ik}\sum_{l=1}^{K}\left\|\left(\hat{M}^{-1/2}-\boldsymbol{M}_*^{-1/2}\right)[A^*]_{.l}\right\|_2 \\
+ nT\left(\left[\hat{M}^{-1/2}\right]_{ii}-\left[\boldsymbol{M}_*^{-1/2}\right]_{ii}\right)\sum_{k=1}^{K}\sum_{l=1}^{K}[A^*]_{ik}\left\|\boldsymbol{M}_*^{-1/2}[A^*]_{.l}\right\|_2.
\end{aligned}
$$

First we already proved that for all $i \in [p]$,

$$
[\boldsymbol{M}_*]_{ii} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i.
$$

Proposition 5.2.1 ensures that for all $i \in [p]$, for all $\epsilon_1 > 0$, with probability at least $1-2\exp(-\epsilon_1^2)$,

$$
\left|[\hat{M}]_{ii}-[\boldsymbol{M}_*]_{ii}\right| < 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}.
$$

This proves that for all $i \in [p]$, for all $\epsilon_1 > 0$, with probability at least $1-2\exp(-\epsilon_1^2)$,

$$
\left[\hat{M}\right]_{ii} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}.
$$

Using the mean value theorem we get that for all $i \in [p]$, for all $\epsilon_1 > 0$, with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$\begin{aligned}\left|[\hat{M}^{-1/2}]_{ii} - [\boldsymbol{M}_*^{-1/2}]_{ii}\right| &\leq 2\epsilon_1 \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-3/2}\sqrt{\frac{\min(2,h_i)}{NnT}},\\ &\leq \frac{2\epsilon_1}{h_i\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{\min(2/h_i,1)}{NnTh_i}}\right)^{-3/2},\\ &\leq \frac{2\epsilon_1}{h_i\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_i}}\right)^{-3/2}.\end{aligned}$$

Hence, reminding that by definition, for all $i \in [p]$ we have $h_i = \sum_{k=1}^{K}[A^*]_{ik}$ leads to, for all $l \in [K]$, for all $\epsilon_1 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2)$,

$$\begin{aligned}\left\|\left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\right)[A^*]_{.l}\right\|_2 &\leq \left[\sum_{i=1}^{p}\left(\frac{2\epsilon_1}{h_i\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_i}}\right)^{-3/2}[A^*]_{il}\right)^2\right]^{1/2},\\ &\leq \frac{2\epsilon_1}{\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}\left[\sum_{i=1}^{p}\left(h_i^{-1}[A^*]_{il}\right)^2\right]^{1/2},\\ &\leq \frac{2\epsilon_1\sqrt{p}}{\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.\end{aligned}$$

Moreover we also proved the following inequality

$$\sum_{l=1}^{K}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.l}\|_2 \leq \frac{K}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}.$$

These inequalities lead to, for all $i \in [p]$, for all $\epsilon > 0$, with probability at least $1 - 2pK\exp\left(-\epsilon_1^2\right)$,

$$\begin{aligned}\|e_i^\top R_4\|_2 \leq{}& \frac{2\epsilon_1 nTh_iK\sqrt{p}}{\sqrt{NnT}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{1/2}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}\\ &+ \frac{2\epsilon_1 nTh_iK}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\frac{1}{h_i\sqrt{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_i}}\right)^{-3/2}.\end{aligned}$$

Finally, for all $i \in [p]$, for all $\epsilon_1 > 0$, with probability at least $1 - 2pK\exp\left(-\epsilon_1^2\right)$,

$$\begin{aligned}\|e_i^\top R_4\|_2 \leq{}& 2\epsilon_1 K\sqrt{\frac{h_inTp}{N}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-2}\\ &+ \frac{2\epsilon_1 K\sqrt{nT}}{\sqrt{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.\end{aligned}$$

We combine all the previously obtained results and get that, for all $i \in [p]$ and for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2K\exp(-\epsilon_2^2) - 2pK\exp(-\epsilon_3^2) - 4p\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$,

$$
\begin{aligned}
\|e_i^\top(\hat{G} - \boldsymbol{G}_*)\|_2 \leq{}& \frac{2\epsilon_1\sqrt{nT}}{N\sqrt{Nh_i}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{h_i NnT}}\right)^{-1} \\
&+ 2\left(1 - \frac{2\epsilon_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i}\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-1/2} \sqrt{\frac{nT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2}} \left(1 + \frac{\epsilon_1 \max_{i\in[p]} \xi_i}{\sqrt{NnT}}\right) \left[\epsilon_3\sqrt{h_i p} + K\epsilon_2\right] \\
&+ \left(1 + \frac{\epsilon_1 \max_{i\in[p]} \xi_i}{\sqrt{NnT}}\right)^2 \cdot \frac{576 \cdot e}{\log(2)\sqrt{c}} \cdot \frac{\epsilon_4\sqrt{nTp}}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\max(h_{\min}/2, 1)} \\
&+ 2\epsilon_1 K\sqrt{\frac{h_i nTp}{N}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-2} \\
&+ \frac{2\epsilon_1 K\sqrt{nT}}{\sqrt{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.
\end{aligned}
$$

We then divide by $\sqrt{h_i}$ and remind that for all $i \in [p]$, $h_i \geq h_{\min}$, to get under the same conditions,

$$
\begin{aligned}
h_i^{-1/2}\|e_i^\top(\hat{G} - \boldsymbol{G}_*)\|_2 \leq{}& \frac{2\epsilon_1\sqrt{nT}}{Nh_{\min}\sqrt{N}} \cdot \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{h_{\min} NnT}}\right)^{-1} \\
&+ 2\frac{\epsilon_3\sqrt{p} + Kh_{\min}^{-1/2}\epsilon_2}{\sqrt{1 - \frac{2\epsilon_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{1}{NnTh_{\min}}}}} \sqrt{\frac{nT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2}} \left(1 + \frac{\epsilon_1 \max_{i\in[p]} \xi_i}{\sqrt{NnT}}\right) \\
&+ \left(1 + \frac{\epsilon_1 \max_{i\in[p]} \xi_i}{\sqrt{NnT}}\right)^2 \cdot \frac{576 \cdot e}{\log(2)\sqrt{c}} \cdot \frac{\epsilon_4 h_{\min}^{-1/2}\sqrt{nTp}}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\max(h_{\min}/2, 1)} \\
&+ 2\epsilon_1 K\sqrt{\frac{nTp}{N}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-2} \\
&+ \frac{2\epsilon_1 K\sqrt{nT}}{\sqrt{Nh_{\min}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.
\end{aligned}
$$

Assumption 6, stating that $h_{\min} \geq \frac{c_1 K}{p}$, ensures the following inequality, holding true under the same

conditions as previously stated,

$$
\begin{aligned}
h_i^{-1/2}\|e_i^\top(\hat{G}-\boldsymbol{G}_*)\|_2 \leq{}& \frac{2\epsilon_1 p\sqrt{nT}}{Nc_1K\sqrt{N}}\cdot\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})-2\epsilon_1\sqrt{\frac{1}{h_{\min}NnT}}\right)^{-1}\\
&+2\frac{\epsilon_3\sqrt{p}+\epsilon_2\sqrt{Kp/c_1}}{\sqrt{1-\frac{2\epsilon_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{1}{NnTh_{\min}}}}}\sqrt{\frac{nT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2}}\left(1+\frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)\\
&+\left(1+\frac{\epsilon_1\max\limits_{i\in[p]}\xi_i}{\sqrt{NnT}}\right)^2\cdot\frac{576\cdot e}{\log(2)\sqrt{c}}\cdot\frac{\epsilon_4 p\sqrt{nT}}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\sqrt{c_1K}}\\
&+2\epsilon_1K\sqrt{\frac{nTp}{N}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})-\frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-2}\\
&+\frac{2\epsilon_1\sqrt{KnTp}}{\sqrt{Nc_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})-\frac{2\epsilon_1}{\sqrt{NnTh_{\min}}}\right)^{-3/2}.
\end{aligned}
$$

Finally we have, for all $s\in[p]$, $h_s\leq K$ and

$$
\begin{aligned}
\xi_s &:= \frac{\sqrt{h_s\min(2,h_s)}}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_s-2\epsilon\sqrt{\frac{\min(2,h_s)}{NnT}}\right)^{3/2}}\\
&\leq \frac{\sqrt{2K}}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})-2\epsilon\sqrt{\frac{p}{NnTc_1K}}\right)^{3/2}}.
\end{aligned}
$$

Moreover c is an absolute constant appearing in Lemma 1.1.10. ■

Proof of Theorem 5.2.11. Use the result stated in Theorem 5.5.1 and notice that $NnT\geq\epsilon_1^2\max\left(\frac{16}{c_2^2h_{\min}};\ \frac{32}{c_2^2};\ \frac{9K^2}{c_2^3h_{\min}^3}\right)$ ensures $NnT\geq\epsilon_1^2\max_i\xi_i^2$. ■

5.5.9 Proof of Theorem 5.2.12

Theorem 5.5.2 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6 and 7. Then for all $i\in[p]$ and for all $\epsilon_1,\epsilon_3,\epsilon_4>0$, with probability at least* $1-2p\exp\left(-\epsilon_1^2\right)-2pK\exp\left(-\epsilon_3^2\right)-$

$2 \cdot 9^p \exp\left(-\min\left(\epsilon_4^2, \sqrt{cnT}\epsilon_4\right)\right)$, *the quantity* $\left\|\left(\hat{G} - \boldsymbol{G}_*\right)\right\|_{op}$ *is bounded from above by*

$$\frac{2\epsilon_1\sqrt{nTp}}{Nc_2\sqrt{Nc_1K}} \cdot \left(1 - \frac{2\epsilon_1\sqrt{p}}{c_2\sqrt{NnTc_1K}}\right)^{-1} + \frac{4\epsilon_3 K\sqrt{nTp}}{c_2\sqrt{N}} \left(1 + 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\left(c_2 - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{-2}\right)$$
$$+\frac{2\epsilon_4\sqrt{nT}}{N} \cdot \frac{288 \cdot e}{c_2\log(2)\sqrt{c}} \cdot \left(1 + 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\left(c_2 - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{-2}\right)$$
$$+\frac{4\epsilon_1\sqrt{nTp}K^2}{c_2^2\sqrt{N}} \left(1 + \epsilon_1 \frac{\sqrt{K}}{\sqrt{2NnT}\left(c_2 - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{3/2}}\right) \left(1 - \frac{2\epsilon_1\sqrt{p}}{c_2\sqrt{NnTc_1K}}\right)^{-3/2}.$$

where c *is an absolute constant appearing in Lemma 1.1.10.*

Remark 5.5.2 *Theorem 5.5.2 improves the result presented in Lemma F.5 in [84]. Specifically, by setting*

$$\epsilon_1^2 = \log(p) + 5\log(nT), \quad \epsilon_3^2 = \log(pK) + 5\log(nT), \quad \epsilon_4^2 = p\log(9) + 5\log(nT),$$

it establishes that with probability at least $1 - 6(nT)^{-5}$ *if* $c \geq \dfrac{p\log(9) + 5\log(nT)}{nT}$ *and with probability at least* $1 - 2\exp\left(-\sqrt{cnT\left(p\log(9) + 5\log(nT)\right)}\right) - 4(nT)^{-5}$ *if* $c \leq \dfrac{p\log(9) + 5\log(nT)}{nT}$ *we have* $\left\|\hat{G} - \boldsymbol{G}_*\right\|_{op}$ *bounded from above by*

$$\frac{2\sqrt{nTp\left(\log(p) + 5\log(nT)\right)}}{Nc_2\sqrt{Nc_1K}} \cdot \left(1 - \frac{2\sqrt{p\left(\log(p) + 5\log(nT)\right)}}{c_2\sqrt{NnTc_1K}}\right)^{-1}$$
$$+\frac{4K\sqrt{nTp\left(\log(pK) + 5\log(nT)\right)}}{c_2\sqrt{N}} \cdot \left(1 + 2\sqrt{\frac{p\left(\log(p) + 5\log(nT)\right)}{NnTc_1K}}\left(c_2 - 2\sqrt{\frac{p\left(\log(p) + 5\log(nT)\right)}{NnTc_1K}}\right)^{-2}\right)$$
$$+\frac{2\sqrt{nT\left(p\log(9) + 5\log(nT)\right)}}{N}\frac{288e}{c_2\log(2)\sqrt{c}}\left(1 + 2\sqrt{\frac{p\left(\log(p) + 5\log(nT)\right)}{NnTc_1K}}\left(c_2 - 2\sqrt{\frac{p\left(\log(p) + 5\log(nT)\right)}{NnTc_1K}}\right)^{-2}\right)$$
$$+\frac{4\sqrt{nTp\left(\log(p) + 5\log(nT)\right)}K^2}{c_2^2\sqrt{N}}\left(1 + \frac{\sqrt{K\left(\log(p) + 5\log(nT)\right)}/\sqrt{2NnT}}{\left(c_2 - 2\sqrt{\frac{p\left(\log(p) + 5\log(nT)\right)}{NnTc_1K}}\right)^{3/2}}\right)$$
$$\cdot\left(1 - \frac{2\sqrt{p\left(\log(p) + 5\log(nT)\right)}}{c_2\sqrt{NnTc_1K}}\right)^{-3/2}.$$

Notably, unlike Lemma F.5 in [84], Theorem 5.5.2 does not require any assumption on either the number nT *of documents or the size of the number of words per documents* N *compared to the vocabulary size* p*. Moreover, the probability of the stated event is controlled non-asymptotically, and the constants are*

explicitly provided. Focusing on the asymptotic behaviour of this upper bound we have, when nT goes to infinity and assuming K, p, N remain fixed, with probability at least $1 - o_{nT\to\infty}\left((nT)^{-3}\right)$, $\left\|\hat{G} - \boldsymbol{G}_*\right\|_{op}$ bounded from above by

$$C_1\sqrt{\frac{nTp\log(nT)}{N}}\frac{1}{N} + C_2\sqrt{\frac{nTp\log(nT)}{N}} + C_3\sqrt{\frac{nT\log(nT)}{N}}\sqrt{\frac{1}{N}} + C_4\sqrt{\frac{nTp\log(nT)}{N}},$$

where

$$C_1 = \frac{10}{\sqrt{c_1 K}c_2}, \quad C_2 = \frac{20K}{c_2}, \quad C_3 = \frac{2880\cdot e}{\log(2)c_2\sqrt{c}}, \quad C_4 = \frac{20K^2}{c_2^2}.$$

It is finally noteworthy to state that with probability at least $1 - o_{nT\to\infty}\left((nT)^{-3}\right)$ the following inequality is asymptotically holding true,

$$\|\hat{G} - \boldsymbol{G}_*\|_{op} \leq C(1 + N^{-1/2}p^{-1/2} + N^{-1})\sqrt{\frac{nTp\log(nT)}{N}}$$

where $C = 2\max(C_1, C_2, C_3, C_4)$. This improves the result presented under an asymptotic framework in Lemma F.5 in [84].

Proof of Theorem 5.5.2. We follow the proof structure of Theorem 5.5.1 and will use the same notations. Especially, for all $i \in [p]$,

$$\begin{aligned}\boldsymbol{\Delta}_i &:= \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon\sqrt{\frac{\min(2,h_i)}{NnT}},\\ \xi_i &:= \left(\boldsymbol{\Delta}_i^{-3/2}\sqrt{h_i\min(2,h_i)}\right).\end{aligned}$$

We remind that $\hat{G} - \boldsymbol{G}_* = \sum_{s=1}^{4} R_s$ which leads to

$$\|(\hat{G} - \boldsymbol{G}_*)\|_{op} \leq \sum_{s=1}^{4}\|R_s\|_{op}.$$

We now aim to bound each $\|R_s\|_{op}$ with high probability. We start with R_1 which is diagonal. Hence we get, for all $\epsilon_1 > 0$, with probability at least $1 - 2p\exp\left(-\epsilon_1^2\right)$,

$$\|R_1\|_{op} = \max_{i\in[p]}\|e_i^\top R_1\|_2 \leq \frac{2\epsilon_1\sqrt{nT}}{N\sqrt{Nh_{\min}}}\cdot\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{h_{\min}NnT}}\right)^{-1}.$$

We now consider R_2 and notice that

$$\begin{aligned}\|R_2\|_{op} &\leq 2\sum_{k=1}^{K}\left\|\hat{M}^{-1/2}[A^*]_{.k}\right\|_2\left\|\hat{M}^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2,\\ &\leq 2\|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\|_{op}^2\sum_{k=1}^{K}\left\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\right\|_2\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2.\end{aligned}$$

Hence, using the previously derived upper bound on $\left\|\boldsymbol{M}_*^{-1/2}\boldsymbol{Z}^{1:T}[\boldsymbol{W}^{1:T}]_{k.}\right\|_2$, for all $k \in [K]$, we get that for all $\epsilon_3 > 0$, with probability at least $1 - 2p\exp(-\epsilon_3^2)$,

$$\begin{aligned}\|R_2\|_{op} &\leq 4\epsilon_3 \max_{i\in[p]}\left([\hat{M}^{-1}]_{ii}[\boldsymbol{M}_*]_{ii}\right)\sqrt{\frac{pnT}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\sum_{k=1}^{K}\left\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\right\|_2,\\ &\leq 4\epsilon_3 \max_{i\in[p]}\left([\hat{M}^{-1}]_{ii}[\boldsymbol{M}_*]_{ii}\right)\sqrt{\frac{nTp}{N}}\frac{K}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})},\end{aligned}$$

where the second inequality is due to the following result,

$$\sum_{k=1}^{K}\|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2 \leq \frac{K}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}.$$

Reminding that the function $x \mapsto x^{-1}$ is convex, Lemma 5.6.5 guarantees that for all $(x, y) \in \mathbb{R}^2$,

$$(x-y)^{-1} \leq x^{-1} + (x-y)^{-2}y.$$

In addition we recall that for all $i \in [p]$, for all $\epsilon_1 > 0$with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$[\hat{M}]_{ii} \geq \left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right).$$

This provides especially with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$[\hat{M}^{-1}]_{ii} < [\boldsymbol{M}_*^{-1}]_{ii} + 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-2}.$$

Hence we get that for all $i \in [p]$, with probability at least $1 - 2\exp(-\epsilon_1^2)$,

$$\begin{aligned}[\hat{M}^{-1}]_{ii}[\boldsymbol{M}_*]_{ii} &\leq 1 + [\boldsymbol{M}_*]_{ii}2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\left([\boldsymbol{M}_*]_{ii} - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-2},\\ &\leq 1 + 2h_i\epsilon_1\sqrt{\frac{h_i}{NnT}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i - 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^{-2},\\ &\leq 1 + 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\right)^{-2}.\end{aligned}$$

Finally using an union bound to control the max and the inequalities $h_i \geq h_{\min} \geq \frac{c_1 K}{p}$ holding for all $i \in [p]$, leads to, for all $\epsilon_1, \epsilon_3 > 0$ and with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2pK\exp(-\epsilon_3^2)$,

$$\begin{aligned}\|R_2\|_{op} &\leq \frac{4\epsilon_3 K}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{nTp}{N}}\left(1 + 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\right)^{-2}\right),\\ &\leq \frac{4\epsilon_3 K\sqrt{nTp}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\sqrt{N}}\left(1 + 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{-2}\right).\end{aligned}$$

We now consider R_3. Following the definition of R_3, we have

$$\begin{aligned}
\|R_3\|_{op} &\leq \left\|\hat{M}^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\hat{M}^{-1/2}\right\|_{op}, \\
&\leq \left\|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\right\|_{op}\left\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\right\|_{op}\left\|\boldsymbol{M}_*^{1/2}\hat{M}^{-1/2}\right\|_{op}, \\
&\leq \left\|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\right\|_{op}^2\left\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\right\|_{op}, \\
&\leq \left\|\hat{M}^{-1}\boldsymbol{M}_*\right\|_{op}\left\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\right\|_{op}.
\end{aligned}$$

Hence with probability at least $1 - 2p\exp(-\epsilon_1^2)$, we have $\|R_3\|_{op}$ bounded from above by

$$\left(1 + 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\right)^{-2}\right)\left\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\right\|_{op}.$$

Moreover, Proposition 5.2.7 ensures that for all $\epsilon_4 > 0$, with probability at least $1-2\exp\left(p\log(9) - \min\left(\epsilon_4^2, \sqrt{cnT}\epsilon_4\right)\right)$, where $c > 0$ is an absolute constant, we have

$$\begin{aligned}
&\|\boldsymbol{M}_*^{-1/2}\left(\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top - \mathbb{E}\left[\boldsymbol{Z}^{1:T}(\boldsymbol{Z}^{1:T})^\top\right]\right)\boldsymbol{M}_*^{-1/2}\|_{op} \\
&\leq \frac{576\cdot e}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\log(2)}\cdot\frac{\sqrt{nT}\epsilon_4}{N\sqrt{c}\max(h_{\min}/2, 1)}.
\end{aligned}$$

Finally, for all $\epsilon_1, \epsilon_4 > 0$, with probability at least $1 - 2\exp\left(p\log(9) - \min\left(\epsilon_4^2, \sqrt{cnT}\epsilon_4\right)\right) - 2p\exp\left(-\epsilon_1^2\right)$, we have

$$\|R_3\|_{op} \leq \frac{576\cdot e}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\log(2)\sqrt{c}}\cdot\frac{\sqrt{nT}\epsilon_4}{N\max(h_{\min}/2, 1)}\left(1 + 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{1}{NnTh_i}}\right)^{-2}\right).$$

We now consider $R_4 := \left(1 - \frac{1}{N}\right)\left(\hat{M}^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\boldsymbol{\Pi}^{1:T}(\boldsymbol{\Pi}^{1:T})^\top\boldsymbol{M}_*^{-1/2}\right)$. As detailed in the proof of Theorem 5.5.1, R_4 can be written as follows,

$$\begin{aligned}
R_4 = &\left(1 - \frac{1}{N}\right)\sum_{k=1}^K\sum_{l=1}^K\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\hat{M}^{-1/2}\left[A^*\right]_{.k}\left[A^*\right]_{.l}^\top\left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\right) \\
&+ \left(1 - \frac{1}{N}\right)\sum_{k=1}^K\sum_{l=1}^K\left(\left[\boldsymbol{W}^{1:T}\right]_{k.}^\top\left[\boldsymbol{W}^{1:T}\right]_{l.}\right)\left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\right)\left[A^*\right]_{.k}\left[A^*\right]_{.l}^\top\boldsymbol{M}_*^{-1/2}.
\end{aligned}$$

Hence we get that

$$\begin{aligned}
\|R_4\|_{op} \leq &nT\sum_{k=1}^K\sum_{l=1}^K\left\|\hat{M}^{-1/2}M_*^{1/2}\right\|_{op}\left\|M_*^{-1/2}\left[A^*\right]_{.k}\right\|_2\cdot\left\|\left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\right)\left[A^*\right]_{.l}\right\|_2 \\
&+ nT\sum_{k=1}^K\sum_{l=1}^K\left\|\left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2}\right)\left[A^*\right]_{.k}\right\|_2\left\|\boldsymbol{M}_*^{-1/2}\left[A^*\right]_{.l}\right\|_2.
\end{aligned}$$

The proof of Theorem 5.5.1 provides the following results. For all $l \in [K]$, for all $\epsilon > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2)$, we have

$$\left\| \left(\hat{M}^{-1/2} - \boldsymbol{M}_*^{-1/2} \right) [A^*]_{.l} \right\|_2 \leq \frac{2\epsilon_1\sqrt{p}}{\sqrt{NnT}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}} \right)^{-3/2}.$$

In addition, we have

$$\sum_{k=1}^{K} \|\boldsymbol{M}_*^{-1/2}[A^*]_{.k}\|_2 \leq \frac{K}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}.$$

Furthermore, we proved that with probability at least $1 - 2p\exp(-\epsilon_1^2)$,

$$\|\hat{M}^{-1/2}\boldsymbol{M}_*^{1/2}\|_{op} \leq \left(1 + \frac{\epsilon_1 \max\limits_{i\in[p]} \xi_i}{\sqrt{NnT}} \right).$$

Hence we deduce that with probability at least $1 - 2p\exp\left(-\epsilon_1^2\right)$, we have

$$\|R_4\|_{op} \leq \left(2 + \frac{\epsilon_1 \max\limits_{i\in[p]} \xi_i}{\sqrt{NnT}} \right) \frac{nTK^2}{\sqrt{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}} \frac{2\epsilon_1\sqrt{p}}{\sqrt{NnT}} \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - \frac{2\epsilon_1}{\sqrt{NnTh_{\min}}} \right)^{-3/2}.$$

Finally we have, for all $s \in [p]$, $h_s \leq K$ and

$$\begin{aligned} \xi_s &:= \frac{\sqrt{h_s \min(2, h_s)}}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_s - 2\epsilon_1\sqrt{\frac{\min(2,h_s)}{NnT}} \right)^{3/2}} \\ &\leq \frac{\sqrt{2K}}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}} \right)^{3/2}}. \end{aligned}$$

We combine all the previously proved results and get the stated inequality. ■

Proof of Theorem 5.2.12. We use the result stated in Theorem 5.5.2 and the given bound on the sample size NnT to get the result. ■

5.5.10 Proof of Theorem 5.2.13

Theorem 5.5.3 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6 and 7. For all $l \in [5]$, the quantities $C_l(K, p, N, n, T, \boldsymbol{W}, h_{\min}, \epsilon)$, defined here under, converge towards a fixed constant when either N, n or T goes to infinity. Under this setup, there exists a matrix $\Omega = diag(\omega, \Omega_{2:K}) \in \mathbb{R}^{K\times K}$ where $\omega \in \{-1, 1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ is an orthogonal matrix such that for all $i \in [p]$ and for all $\alpha > 0$ satisfying*

$$\alpha \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \frac{\min\left(c_3, \lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\right)}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})} < 1,$$

for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ *satisfying*

$$\max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \leq \alpha \frac{(1-1/N)\sqrt{nT}\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{6\left(\frac{C_1\sqrt{p}}{N} + C_2\sqrt{p} + \frac{C_3}{\sqrt{N}} + C_4\sqrt{p}\right)},$$

with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)-2\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$ *the quantity* $\left\|\Omega[\hat{U}]_{i.} - [U]_{i.}\right\|_2$ *is bounded from above by*

$$\frac{C_{tot}(p,N)}{\alpha}\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\sqrt{\frac{Kh_ip}{nT(N-2)}},$$

where

$$\begin{aligned} C_{tot}(p,N) :=& \frac{20}{c_2^2}\left(\frac{C_1}{c_2N} + \frac{C_2}{c_2} + \frac{C_3}{c_2\sqrt{pN}} + \frac{C_4}{c_2} + C_5 + \frac{C_1\sqrt{p}}{Nc_1K} + \frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}} + C_1^2 + C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\right), \\ =& D_1(K) + \frac{D_2}{\sqrt{pN}} + \frac{D_3 + D_4(K)\sqrt{p}}{N} + D_5(K)\sqrt{\frac{p}{N}}. \end{aligned}$$

with $D_1(K) := \frac{20}{c_2^2}\left(\frac{C_2+C_4}{c_2} + C_5 + C_1^2 + C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\right)$, $D_2 := \frac{C_3}{c_2}$, $D_3 := \frac{C_1}{c_2}$, $D_4(K) := \frac{C_1}{c_1K}$, $D_5(K) := \frac{C_3}{\sqrt{c_1K}}$ *converge towards constants when* NnT *grows. The quantities* $(C_l)_{l\in[5]}$ *are defined as follows,*

$$\begin{aligned} C_1 :=& \left(c_2 - \frac{2\epsilon_1\sqrt{p}}{\sqrt{NnTc_1K}}\right)^{-1}, \\ C_2 :=& \frac{K}{c_2}\left(1 + 2\epsilon\sqrt{\frac{p}{NnTc_1K}}\left(c_2 - 2\epsilon\sqrt{\frac{p}{NnTc_1K}}\right)^{-2}\right), \\ C_3 :=& \frac{288\cdot e}{c_2\log(2)\sqrt{c}}\cdot\left(1 + 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\left(c_2 - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{-2}\right), \\ C_4 :=& \frac{2K^2}{c_2^2}\left(1 + \epsilon_1\frac{\sqrt{K}}{\sqrt{2NnT}\left(c_2 - 2\epsilon_1\sqrt{\frac{p}{NnTc_1K}}\right)^{3/2}}\right)\left(1 - \frac{2\epsilon_1\sqrt{p}}{c_2\sqrt{NnTc_1K}}\right)^{-3/2}, \\ C_5 =& \frac{1+\sqrt{K/c_1}}{c_2\sqrt{1 - \frac{2\epsilon_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{1}{NnTh_{\min}}}}}\left(1 + \frac{\epsilon_1\max_{i\in[p]}\xi_i}{\sqrt{NnT}}\right), \end{aligned}$$

with ξ_i *defined in Theorem 5.5.1.*

Remark 5.5.3 *Theorem 5.5.3 improves the result presented in Theorem 3.1 in [84]. Specifically, by setting*

$$\begin{aligned} &\epsilon_1^2 = \log(p) + 5\log(nT), \quad \epsilon_2^2 = \log(K) + 5\log(nT), \quad \epsilon_3^2 = \log(pK) + 5\log(nT), \\ &\epsilon_4^2 = \log(2p + 9^p) + 5\log(nT), \end{aligned}$$

it establishes that with probability at least $1-8(nT)^{-5}$ *if* $c \geq \dfrac{\log(2p+9^p)+5\log(nT)}{nT}$ *and with probability at least* $1-2\exp\left(-\sqrt{cnT\left(\log(2p+9^p)+5\log(nT)\right)}\right)-6(nT)^{-5}$ *if* $c \leq \dfrac{\log(2p+9^p)+5\log(nT)}{nT}$ *we havewe have for all* $i \in [p]$, $\left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2$ *bounded from above by*

$$\frac{20\sqrt{\log(2p+9^p)+5\log(nT)}}{\alpha(N-1)\lambda_K(\Sigma_A)c_2}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{Nc_2}+\frac{C_2}{c_2}+\frac{C_3}{\sqrt{pN}c_2}+\frac{C_4}{c_2}\right),$$
$$+\frac{20\sqrt{\log(2p+9^p)+5\log(nT)}}{\alpha(N-1)\lambda_K(\Sigma_A)c_2}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K}+C_5+\frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}}+C_1^2+C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\right).$$

Notably, unlike Theorem 3.1 in [84], Theorem 5.5.3 does not require any assumption on either the number nT *of documents or the value of* $\log(nT)$ *compared to* $\min(N,p)$ *or the asymptotic behaviour of* $\dfrac{p\log(nT)}{NnT}$*. Moreover, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided. It is finally noteworthy to state that with probability at least* $1-o_{nT\to\infty}\left((nT)^{-3}\right)$ *the following inequality is asymptotically holding true,*

$$\left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2 \leq C\left(1+\frac{1}{N}+\frac{\sqrt{p}}{N}+\sqrt{\frac{p}{N}}\right)\sqrt{\frac{h_ip\log(nT)}{NnT}}$$

where $C=2\dfrac{20K\sqrt{5}}{\alpha c_1\lambda_K(\Sigma_A)c_2^2}\max(C_1,C_2,C_3,C_4)^{3/2}$*. This improves the result presented under an asymptotic framework in Theorem 3.1 in [84].*

Proof of Theorem 5.5.3. We first recall that for any vector $v \in \mathbb{R}^d$, for all $j \in [d]$, $v(j)$ denotes the j^{th} entry of v. We then define $\hat{U}_{2:K}=\left[\hat{u}_2,\ldots,[\hat{U}]_{.K}\right]$ and $U_{2:K}$ its population counterpart. We recall that $\hat{U}=\left[[\hat{U}]_{.1},\ldots,[\hat{U}]_{.K}\right]$ contains the first K left singular vectors of the noisy quantity $\hat{\Pi}$. Their population counterparts are denoted respectively U and $\mathbf{\Pi}_*$. Then for any matrix $\Omega=(\omega,\Omega_{2:K})$, where $\omega \in \{+1,-1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ is orthogonal, we have

$$\left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2 \leq |\omega[\hat{U}]_{.1}(i)-[U]_{.1}(i)|+\left\|\Omega_{2:K}[\hat{U}_{2:K}]_{i.}-[U_{2:K}]_{i.}\right\|_2.$$

Proposition 5.2.9 proves that

$$\left(1-\frac{1}{N}\right)nT\frac{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} \geq \|\boldsymbol{G}_*\|_{op} \geq \lambda_K(\boldsymbol{G}_*) \geq \left(1-\frac{1}{N}\right)nT\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}).$$

In addition, proposition 5.2.9 also ensures that

$$\|\boldsymbol{G}_*\|_{op} \geq \left(1-\frac{1}{N}\right)nTc_3+\max_{k\geq 2}\lambda_K(\boldsymbol{G}_*).$$

In addition, Theorem 5.5.2 states that for all $\epsilon_1,\epsilon_3,\epsilon_4>0$, with probability at least $1-2p\exp\left(-\epsilon_1^2\right)-2pK\exp\left(-\epsilon_3^2\right)-2\cdot 9^p\exp\left(-\min\left(\epsilon_4^2,\sqrt{cnT}\epsilon_4\right)\right)$,

$$\left\|\left(\hat{G}-\boldsymbol{G}_*\right)\right\|_{op} \leq \frac{2\epsilon_1\sqrt{nTp}}{N\sqrt{N}}\cdot C_1+\frac{2\epsilon_3\sqrt{nTp}}{\sqrt{N}}\cdot C_2+\frac{2\epsilon_4\sqrt{nT}}{N}\cdot C_3+\frac{2\epsilon_1\sqrt{nTp}}{\sqrt{N}}\cdot C_4,$$

where for all $l \in [4]$, the quantities $C_l(K, p, N, n, T, \boldsymbol{W}, h_{\min}, \epsilon)$ converge towards a fixed constant when either N, n or T goes to infinity while the other quantities remain fixed. Let us denote A the quantity $\left[\frac{C_1\sqrt{p}}{N} + C_2\sqrt{p} + \frac{C_3}{\sqrt{N}} + C_4\sqrt{p}\right]$. Consider two groups for the eigenvalues of $\boldsymbol{G}_*$, namely : $\{\lambda_1(\boldsymbol{G}_*)\}$ and $\{\lambda_2(\boldsymbol{G}_*), \dots, \lambda_{\min}(\boldsymbol{G}_*)\}$. The gap between two eigenvalues lying in two different groups is at least $\left(1 - \frac{1}{N}\right) nTc_3$ as proved in proposition 5.2.9. Next consider $\alpha \in \mathbb{R}_+$. Proposition 5.2.9 ensures that if

$$\alpha \leq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})c_3}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})}, \quad \text{then} \quad \lambda_1(\boldsymbol{G}_*) - \lambda_2(\boldsymbol{G}_*) \geq \alpha \|\boldsymbol{G}_*\|_{op}.$$

In addition if

$$\alpha \leq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2\lambda_K(\Sigma_A)}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})} < 1, \quad \text{then} \quad \lambda_{\min}(\boldsymbol{G}_*) \geq \alpha \|\boldsymbol{G}_*\|_{op}.$$

Hence if

$$\alpha \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \frac{\min\left(c_3, \lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\right)}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})} < 1 \quad \text{then} \quad \min\{\lambda_1(\boldsymbol{G}_*) - \lambda_2(\boldsymbol{G}_*), \lambda_{\min}(\boldsymbol{G}_*)\} \geq \alpha \|\boldsymbol{G}_*\|_{op}.$$

Moreover, if

$$\max(\epsilon_1, \epsilon_3, \epsilon_4) \leq \alpha \frac{(1 - 1/N)\sqrt{nT}\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{6A} \quad \text{then} \quad 2\sqrt{\frac{nT}{N}} \max(\epsilon_1, \epsilon_3, \epsilon_4) A \leq \frac{\alpha}{3}(1 - \frac{1}{N}) nT\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}),$$

which leads to, with probability at least $1 - 2p\exp\left(-\epsilon_1^2\right) - 2pK\exp\left(-\epsilon_3^2\right) - 2 \cdot 9^p \exp\left(-\min\left(\epsilon_4^2, \sqrt{cnT}\epsilon_4\right)\right)$,

$$\left\| \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_{op} \leq \frac{\alpha}{3} \|\boldsymbol{G}_*\|_{op}.$$

Finally, conditions required to apply Lemma 5.6.6 are fulfilled. Applying this Lemma with respectively $s = k = 1$ and $s = 2,\ k = K$ gives the existence of $\omega \in \{+1, -1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ orthogonal such that,

$$\left\| \omega[\hat{U}]_{\cdot 1}(i) - [U]_{\cdot 1}(i) \right\| \leq \frac{5}{\alpha \|\boldsymbol{G}_*\|_{op}} \left(\left\| \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_{op} \|[U]_{i\cdot}\|_2 + \sqrt{K} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_2 \right),$$

$$\left\| \Omega_{2:K}[\hat{U}_{2:K}]_{i\cdot} - [U_{2:K}]_{i\cdot} \right\|_2 \leq \frac{5}{\alpha \|\boldsymbol{G}_*\|_{op}} \left(\left\| \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_{op} \|[U]_{i\cdot}\|_2 + \sqrt{K} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_2 \right).$$

In addition, Proposition 5.2.10 ensures that under the same conditions,

$$\left\| \omega[\hat{U}]_{\cdot 1}(i) - [U]_{\cdot 1}(i) \right\| \leq \frac{5}{\alpha \|\boldsymbol{G}_*\|_{op}} \left(\left\| \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_{op} \frac{\sqrt{K}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} \sqrt{h_i} + \sqrt{K} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_2 \right),$$

$$\left\| \Omega_{2:K}[\hat{U}_{2:K}]_{i\cdot} - [U_{2:K}]_{i\cdot} \right\|_2 \leq \frac{5}{\alpha \|\boldsymbol{G}_*\|_{op}} \left(\left\| \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_{op} \frac{\sqrt{K}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} \sqrt{h_i} + \sqrt{K} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_2 \right).$$

Theorem 5.5.1 states that for all $i \in [p]$ and for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p\exp(-\epsilon_1^2) - 2K\exp(-\epsilon_2^2) - 2pK\exp(-\epsilon_3^2) - 4p\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$, the quantity $h_i^{-1/2} \left\| e_i^\top \left(\hat{G} - \boldsymbol{G}_*\right) \right\|_2$ is bounded from above by

$$2\sqrt{\frac{nTp}{N}} \left[C_1'\epsilon_1 \frac{\sqrt{p}}{N} + C_2'(\epsilon_2 + \epsilon_3) + C_3'\epsilon_4\sqrt{\frac{p}{N}} + C_4'\epsilon_1 + C_5'\epsilon_1 \right]$$

where for all $l \in [5]$, the quantities $C_l'(K, p, N, n, T, \boldsymbol{W}, h_{\min}, \epsilon)$ converge towards a fixed constant when either N, n or T goes to infinity. Let us denote B the quantity $\left[\frac{C_1'\sqrt{p}}{N} + C_2' + C_3'\sqrt{\frac{p}{N}} + C_4' + C_5'\right]$. Then Theorems 5.5.1 and 5.5.2 provide, under the conditions on $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ and α previously stated, with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)-2\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$,

$$\left\|\omega[\hat{U}]_{.1}(i) - [U]_{.1}(i)\right\| \leq \frac{10\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha\left\|\boldsymbol{G}_*\right\|_{op}}\left(\frac{\sqrt{nTKh_i}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\sqrt{N}}A + \sqrt{\frac{Kh_inTp}{N}}B\right),$$

$$\left\|\Omega_{2:K}[\hat{U}_{2:K}]_{i.} - [U_{2:K}]_{i.}\right\|_2 \leq \frac{10\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha\left\|\boldsymbol{G}_*\right\|_{op}}\left(\frac{\sqrt{nTKh_i}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\sqrt{N}}A + \sqrt{\frac{Kh_inTp}{N}}B\right).$$

This provides

$$\left\|\Omega[\hat{U}]_{i.} - [U]_{i.}\right\|_2 \leq \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha\left\|\boldsymbol{G}_*\right\|_{op}}\sqrt{\frac{Kh_inTp}{N}}\left(\frac{A}{\sqrt{p}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + B\right).$$

Using the lower bound on $\left\|\boldsymbol{G}_*\right\|_{op}$ provided by Proposition 5.2.9 gives

$$\left\|\Omega[\hat{U}]_{i.} - [U]_{i.}\right\|_2 \leq \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{A}{\sqrt{p}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + B\right).$$

The conclusion comes by noticing $\frac{\sqrt{N}}{N-1} \leq \sqrt{\frac{1}{N-2+1/N}}$. ■

Proof of Theorem 5.2.13. Consider the statement of Theorem 5.5.3 and notice that $NnT \geq \epsilon_1^2\max\left(\frac{36K}{c_2^3};\ \frac{64p}{c_2^4c_1K};\ \frac{16}{c_2^2h_{\min}};\ \max_{i\in[p]}\xi_i^2\right)$ ensures :

$$C_1 \leq 2/c_2;\ C_2 \leq 2K/c_2;\ C_3 \leq \frac{576e}{c_2\log(2-\sqrt{c})};\ C_4 \leq \frac{4K^2}{c_2^2};\ C_5 \leq \frac{4}{c_2}\left(1+\sqrt{K/c_1}\right).$$

In addition, $1 - 1/N \leq 0.5$ and we have almost surely

$$\sqrt{K} \geq \lambda_K(\Sigma_A) \geq c_2, \quad \sqrt{K} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq c_2,$$
$$\sqrt{K} \geq \lambda_1(\Sigma_A) \geq K^{-1/2}, \quad \sqrt{K} \geq \lambda_1(\Sigma_{\boldsymbol{W}}^{1:T}) \geq K^{-1/2}.$$

This allows to consider $\alpha = c_2\frac{\min\left(c_3, c_2^2\right)}{K}$ which implies the stated bound on $\max_{i\in[4]}\epsilon_i$ and the value of $C_{tot}(p, N)$. Finally notice that by definition, $\max_{i\in[p]}\xi_i \leq \frac{3K}{c_2^{3/2}h_{\min}^{3/2}}$ if $NnT \geq \epsilon_1^2\frac{32}{c_2^2}$. Thus under this condition, $\max_{i\in[p]}\xi_i^2 \leq \frac{9K^2}{c_2^3h_{\min}^3}$. ■

5.5.11 Proof of Proposition 5.2.14

Proof of Proposition 5.2.14. In the proof outlined in Proposition 5.2.10, it has been established that there exists a non singular matrix $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ such that almost surely the following relationships hold :

$$\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1} = [A^*]^\top\boldsymbol{M}_*^{-1}A^*,$$
$$U = \boldsymbol{M}_*^{-1/2}A^*\boldsymbol{B}.$$

For all $k \in [K]$ and for all $l \in [K-1]$, we introduce the matrix $N \in \mathbb{R}^{K \times (K-1)}$ with elements defined as follows :

$$[N]_{kl} = \frac{[\boldsymbol{B}]_{k(l+1)}}{[\boldsymbol{B}]_{k1}}.$$

This allows us to express the matrix $\boldsymbol{B}$ in terms of N as $\boldsymbol{B} = \text{diag}\left([\boldsymbol{B}]_{.1}\right)[1_K, N]$. By employing these results, we arrive at the following expression for $R \in \mathbb{R}^{p \times (K-1)}$:

$$[1_p, R] = \text{diag}(u_1)^{-1} \boldsymbol{M}_*^{-1/2} A^* \text{diag}\left([\boldsymbol{B}]_{.1}\right)[1_K, N].$$

Furthermore, following Lemma D.2 in [84] we demonstrate that the first column of $\boldsymbol{B}$, denoted $[\boldsymbol{B}]_{.1}$, is an eigenvector of $\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*$. Indeed Let us denote $\sigma_1(\boldsymbol{\Pi}_*), \ldots, \sigma_K(\boldsymbol{\Pi}_*)$ the singular values of $\boldsymbol{\Pi}_*$. By the definition of the singular values and recalling that $\boldsymbol{\Pi}_*$ has nonnegative entries, we have, for all $k \in [K]$,

$$\boldsymbol{\Pi}_* (\boldsymbol{\Pi}_*)^\top u_k = \sigma_k (\boldsymbol{\Pi}_*)^2 u_k.$$

Combining that $\boldsymbol{\Pi}_* := \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T}$ and $U = \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{B}$ leads to

$$\left(\boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T} \left(\boldsymbol{W}^{1:T}\right)^\top (A^*)^\top \boldsymbol{M}_*^{-1/2}\right) \boldsymbol{M}_*^{-1/2} A^* [\boldsymbol{B}]_{.k} = \sigma_k (\boldsymbol{\Pi}_*)^2 \boldsymbol{M}_*^{-1/2} A^* [\boldsymbol{B}]_{.k}.$$

Left multiplying both sides by $\left((A^*)^\top \boldsymbol{M}_*^{-1} A^*\right)^{-1} \left((A^*)^\top \boldsymbol{M}_*^{-1/2}\right)$ ensures that

$$\boldsymbol{W}^{1:T} \left(\boldsymbol{W}^{1:T}\right)^\top (A^*)^\top \boldsymbol{M}_*^{-1} A^* [\boldsymbol{B}]_{.k} = \sigma_k (\boldsymbol{\Pi}_*)^2 [\boldsymbol{B}]_{.k}.$$

Recall that $\Sigma_{\boldsymbol{W}}^{1:T} := (nT)^{-1} \boldsymbol{W}^{1:T} \left(\boldsymbol{W}^{1:T}\right)^\top$ finally gives, for all $k \in [K]$,

$$\Sigma_{\boldsymbol{W}}^{1:T} (A^*)^\top \boldsymbol{M}_*^{-1} A^* [\boldsymbol{B}]_{.k} = (nT)^{-1} \sigma_k (\boldsymbol{\Pi}_*)^2 [\boldsymbol{B}]_{.k}.$$

By applying Perron Frobenius theorem and establishing that $\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*$ is a strictly positive matrix we conclude that the entries of $[\boldsymbol{B}]_{.1}$ have the same sign. In addition, we have for all $i \in [p]$, $[U]_{.1}(i) = \left[\boldsymbol{M}_*^{-1/2}\right]_{ii} [A^*]_{i.} [\boldsymbol{B}]_{.1}$. Given that the entries of A^* are nonnegative by design, and all entries of $[\boldsymbol{B}]_{.1}$ are either all positive or all negative, it follows that the entries of u_1 are either all positive or all negative. Notably U and $\boldsymbol{B}$ are defined up to a sign flip, allowing us to choose U in a manner that u_1 becomes a positive vector. Finally $\boldsymbol{M}_*$ is a diagonal matrix with positive entries, establishing that the rows of R are convex combinations of the rows of N. ■

5.5.12 Proof of Theorem 5.2.15

Theorem 5.5.4 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6 and 7. Consider the matrices R and $\hat{R}$ defined in the Post-SVD Normalization step. Then, for all $i \in [p]$, for all $\alpha > 0$ such that*

$$\alpha \leq c_2 \frac{\min(c_3, \lambda_K(\Sigma_A) c_2)}{\lambda_1(\Sigma_A) \lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})} < 1,$$

for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ such that

$$\max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \leq \alpha \frac{(1 - 1/N)\sqrt{nT} \lambda_K(\Sigma_A) c_2}{6\left(\frac{C_1\sqrt{p}}{N} + C_2\sqrt{p} + \frac{C_3}{\sqrt{N}} + C_4\sqrt{p}\right)},$$

with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)-2\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$, *there exists* $\Omega_{2:K}\in\mathbb{R}^{(K-1)\times(K-1)}$, *an orthogonal matrix, such that*

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2\leq Z\left[\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|\right)^{-1}+Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-Z\right)^{-2}\right]\left(2+\max_{k\in[K]}\|\eta_k\|_2\right),$$

where

$$\left(\min_{k\in[K]}[\boldsymbol{B}]_{k1}\right)^{-1}\leq\frac{p}{c_2^{9/2}c_1K},\quad\max_{k\in[K]}\|\eta_k\|_2\leq\frac{p}{c_2^5c_1K},$$
$$Z:=\frac{C_{tot}(p,N)}{\alpha}\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\sqrt{\frac{Kh_ip}{nT(N-2)}},$$

with $C_{tot}(p,N)$ *defined in Theorem 5.5.3.*

Remark 5.5.4 *Theorem 5.5.4 improves the result presented in Theorem 3.2 in [84]. Specifically, by setting*

$$\epsilon_1^2=\log(p)+5\log(nT),\quad\epsilon_2^2=\log(K)+5\log(nT),\quad\epsilon_3^2=\log(pK)+5\log(nT),$$
$$\epsilon_4^2=\log(2p+9^p)+5\log(nT),$$

it establishes that with probability at least $1-8(nT)^{-5}$ *if* $c\geq\dfrac{\log(2p+9^p)+5\log(nT)}{nT}$ *and with probability at least* $1-2\exp\left(-\sqrt{cnT\left(\log(2p+9^p)+5\log(nT)\right)}\right)-6(nT)^{-5}$ *if* $c\leq\dfrac{\log(2p+9^p)+5\log(nT)}{nT}$ *we have for all* $i\in[p]$, $\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2$ *bounded from above by*

$$Z\left[\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|\right)^{-1}+Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-Z\right)^{-2}\right]\left(2+\max_{k\in[K]}\|\eta_k\|_2\right)$$

where Z *is bounded from above by*

$$\frac{20\sqrt{\log(2p+9^p)+5\log(nT)}}{\alpha(N-1)c_2^2}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{Nc_2}+\frac{C_2}{c_2}+\frac{C_3}{\sqrt{pN}c_2}+\frac{C_4}{c_2}\right),$$
$$+\frac{20\sqrt{\log(2p+9^p)+5\log(nT)}}{\alpha(N-1)c_2^2}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K}+C_5+\frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}}+C_1^2+C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\right).$$

Notably, unlike Theorem 3.2 in [84], Theorem 5.5.4 does not require any assumption on either the number nT *of documents or the value of* $\log(nT)$ *compared to* $\min(N,p)$ *or the asymptotic behaviour of* $\dfrac{p\log(nT)}{NnT}$. *Moreover, the probability of the stated event is controlled non-asymptotically, and the constants are explicitly provided. It is finally noteworthy to state that with probability at least* $1-o_{nT\to\infty}\left((nT)^{-3}\right)$ *the following inequality is asymptotically holding true,*

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2\leq Cp\left(1+\frac{1}{N}+\frac{\sqrt{p}}{N}+\sqrt{\frac{p}{N}}\right)\sqrt{\frac{p\log(nT)}{NnT}}$$

where $C=2\dfrac{20\sqrt{5}}{\alpha c_1^2c_2^{15/2}}\cdot\left(2+\dfrac{p}{c_2^5c_1K}\right)\max(C_1,C_2,C_3,C_4)^{3/2}$.

Proof of Theorem 5.5.4. First, examine the matrix Ω as defined in Theorem 5.5.3. It is worth noting that the normalized eigenvectors are unique only up to a sign. Consequently, u_1 and $[\hat{U}]_{.1}$ can be selected in a manner that sets their first coordinate to be positive, thus fixing $\omega = 1$. Recall that for all $i \in [p]$,

$$\begin{pmatrix} 1 \\ [R]_{i.} \end{pmatrix} = [U]_{.1}(i)^{-1}[U]_{i.} \qquad \text{and} \qquad \begin{pmatrix} 1 \\ \Omega_{2:K}\left[\hat{R}\right]_{i.} \end{pmatrix} = [\hat{U}]_{.1}(i)^{-1}\Omega\left[\hat{U}\right]_{i.}.$$

This provides, for all $i \in [p]$,

$$\begin{aligned} \left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 &= \left\|[U]_{.1}(i)^{-1}[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}\Omega\left[\hat{U}\right]_{i.}\right\|_2, \\ &= \left\|[U]_{.1}(i)^{-1}[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}[U]_{i.} + [\hat{U}]_{.1}(i)^{-1}[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}\Omega\left[\hat{U}\right]_{i.}\right\|_2. \end{aligned}$$

Factoring by $[\hat{U}]_{.1}(i)$ on one side and by $[U]_{i.}$ on the other yields :

$$\begin{aligned} \left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 &= \left\|[U]_{.1}(i)^{-1}[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}\left(\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right)\right\|_2, \\ &= \left\|\left([U]_{.1}(i)^{-1} - [\hat{U}]_{.1}(i)^{-1}\right)[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}\left(\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right)\right\|_2. \end{aligned}$$

Bringing the terms to a common denominator and applying the previously mentioned inequalities results in :

$$\begin{aligned} \left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 &= \left\|\left(\frac{[\hat{U}]_{.1}(i) - [U]_{.1}(i)}{[U]_{.1}(i)[\hat{U}]_{.1}(i)}\right)[U]_{i.} - [\hat{U}]_{.1}(i)^{-1}\left(\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right)\right\|_2, \\ &= \left\|\left(\frac{[\hat{U}]_{.1}(i) - [U]_{.1}(i)}{[\hat{U}]_{.1}(i)}\right)\begin{pmatrix} 1 \\ [R]_{i.} \end{pmatrix} - [\hat{U}]_{.1}(i)^{-1}\left(\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right)\right\|_2, \end{aligned}$$

The triangle inequality ultimately guarantees the following inequality :

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 \leq \left|[\hat{U}]_{.1}(i)^{-1}\right|\left(\left\|\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right\|_2 + \left(1 + \|[R]_{i.}\|_2\right)\left|[\hat{U}]_{.1}(i) - [U]_{.1}(i)\right|\right)$$

Additionally, for all $i \in [p]$, the quantity $\left|[\hat{U}]_{.1}(i) - [U]_{.1}(i)\right|$ is upper-bounded by $\left|\Omega[\hat{U}]_{i.} - [U]_{i.}\right|_2$. This ensures that

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 \leq \left|[\hat{U}]_{.1}(i)^{-1}\right|\left\|\Omega\left[\hat{U}\right]_{i.} - [U]_{i.}\right\|_2\left(2 + \|[R]_{i.}\|_2\right)$$

Moreover, Theorem 5.5.3 ensures that for all $i \in [p]$, for all $\alpha > 0$ such that

$$\alpha \leq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})} < 1,$$

and for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ such that

$$\max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \leq \alpha\frac{(1 - 1/N)\sqrt{nT}\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{6\left(\frac{C_1\sqrt{p}}{N} + C_2\sqrt{p} + \frac{C_3}{\sqrt{N}} + C_4\sqrt{p}\right)},$$

with probability at least $1-2p\exp(-\epsilon_1^2)-2K\exp(-\epsilon_2^2)-2pK\exp(-\epsilon_3^2)-2\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$ the quantity $\left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2$ is bounded from above by

$$\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_3}{\sqrt{pN}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_4}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\right),$$
$$+\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K}+C_5+\frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}}+C_1^2+C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\right).$$

Moreover, the matrix R is constructed in such a way that for all $i\in[p]$, the $[R]_{i.}\in\mathbb{R}^{(K-1)}$ lies in G_η. Thus, for all $i\in[p]$, $\|[R]_{i.}\|_2\leq\max_{k\in[K]}\|\eta_k\|_2$. We proceed to bound $\max_{k\in[K]}\|\eta_k\|_2$ from above using a non-random constant. To this end, we recall that the following statement is established as part of the proof of Proposition 5.2.10 : there exists a non-singular matrix $\boldsymbol{B}\in\mathbb{R}^{K\times K}$ such that almost surely there are

$$\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1}=[A^*]^\top\boldsymbol{M}_*^{-1}A^*,$$
$$U=\boldsymbol{M}_*^{-1/2}A^*\boldsymbol{B}.$$

Then, observe that $\boldsymbol{B}$ is non-singular, which establishes that $[1_k,N]$ is also non-singular. Indeed, $\boldsymbol{B}=\text{diag}\left([\boldsymbol{B}]_{.1}\right)[1_k,N]$, as demonstrated in the proof of Proposition 5.2.14. Furthermore, for all $k\in[K]$,

$$\begin{pmatrix}1\\ \eta_k\end{pmatrix}=[1_k,N]^\top e_k,$$

where e_k is the k^{th} canonical vector of $\mathbb{R}^K$. We define $P=[1_k,N]^\top$ leading to, for all $k\in[K]$,

$$\|\eta_k\|_2\leq\|P\|_{op}.$$

Recalling that $P=\text{diag}\left([\boldsymbol{B}]_{.1}\right)^{-1}\boldsymbol{B}$ leads to $PP^\top=\text{diag}\left([\boldsymbol{B}]_{.1}\right)^{-1}\boldsymbol{B}\boldsymbol{B}^\top\text{diag}\left([\boldsymbol{B}]_{.1}\right)^{-1}$. The submultiplicativity of the operator norm guarantees that

$$\|\eta_k\|_2\leq\left\|PP^\top\right\|_{op}^{1/2}\leq\left\|[\boldsymbol{B}]_{.1}^{-1}\right\|_\infty\left\|\boldsymbol{B}\boldsymbol{B}^\top\right\|_{op}^{1/2},$$

where $[\boldsymbol{B}]_{.1}^{-1}$ denotes the vector whose entries are the inverses of the entries of $[\boldsymbol{B}]_{.1}$. We first control $\left\|\boldsymbol{B}\boldsymbol{B}^\top\right\|_{op}$. Proposition 5.2.8 states that for all $i\in[p]$,

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i\leq[\boldsymbol{M}_*]_{ii}\leq h_i.$$

Hence $(A^*)^\top\left(\boldsymbol{M}_*^{-1}-H^{-1}\right)A^*$ is a positive semi-definite symmetric matrix. It follows that the smallest eigenvalue of $(A^*)^\top\boldsymbol{M}_*^{-1}A^*$ is above the smallest eigenvalue of $(A^*)^\top H^{-1}A^*:=\Sigma_A$. Similarly $(A^*)^\top\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}H^{-1}-\boldsymbol{M}_*^{-1}\right)A^*$ is a positive semi-definite symmetric matrix. It follows that the highest eigenvalue of $(A^*)^\top\boldsymbol{M}_*^{-1}A^*$ is below the highest eigenvalue of $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\Sigma_A$. Under the Assumption 7 the smallest eigenvalue of Σ_A is bounded from below by $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})$. In addition, the columns of A are probability vectors which guarantees that the highest eigenvalue of Σ_A is bounded from above by 1. Finally we have

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\leq\lambda_{\min}\left((A^*)^\top\boldsymbol{M}_*^{-1}A^*\right)\leq\lambda_1\left((A^*)^\top\boldsymbol{M}_*^{-1}A^*\right)\leq\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

Using that $\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1} = [A^*]^\top \boldsymbol{M}_*^{-1} A^*$ finally provides

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \leq \lambda_{\min}\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \leq \lambda_1\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

In the proof of Proposition 5.2.14 it has been demonstrated that the entries of $[\boldsymbol{B}]_{\cdot 1}$ have the same sign, which can be chosen to be positive. Subsequently, $\left\|[\boldsymbol{B}]_{\cdot 1}^{-1}\right\|_\infty$ is bounded from above by the inverse of the minimum value of the first column of $\boldsymbol{B}$, denoted as $\left(\min_{k\in[K]}[\boldsymbol{B}]_{k1}\right)^{-1}$. This yields the inequality :

$$\|\eta_k\|_2 \leq \left\|PP^\top\right\|_{op}^{1/2} \leq \left(\min_{k\in[K]}[\boldsymbol{B}]_{k1}\right)^{-1} \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2}.$$

Finally, as mentionned in the proof of Proposition 5.2.14, $[\boldsymbol{B}]_{\cdot 1}$ is an eigenvector of $\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*$ associated with the eigenvalue $(nT)^{-1}\sigma_1\left(\boldsymbol{\Pi}_*\right)^2$. Therefore, for all $k \in [K]$, we have

$$[\boldsymbol{B}]_{k1} = nT\sigma_1\left(\boldsymbol{\Pi}_*\right)^{-2} \sum_{l=1}^{K} \left[\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{kl} [\boldsymbol{B}]_{l1}\,.$$

However, for all $(k,l) \in [K]^2$, the entry $\left[\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{kl}$ can be expanded as follows :

$$\left[\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{kl} = \sum_{r=1}^{K} \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kr} \left[(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{rl}.$$

Under Assumption 7 the entries of $\Sigma_A := [A^*]^\top H^{-1} A^*$ are bounded from below by $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})$. In addition, the entries of $\boldsymbol{M}_*$ are bounded from below by the ones of $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})H$. Hence the entries of $(A^*)^\top \boldsymbol{M}_*^{-1} A^*$ are bounded from below by $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2$. This guarantees that

$$\left[\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{kl} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 \sum_{r=1}^{K} \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kr}.$$

Moreover, as proven in Proposition 4.3.4, diagonal entries of a positive definite matrix cannot be smaller than the smallest eigenvalue and Assumption 7 states that $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})$, the smallest eigenvalue of $\Sigma_{\boldsymbol{W}}^{1:T}$ is positive. This provides

$$\begin{aligned}
\left[\Sigma_{\boldsymbol{W}}^{1:T}(A^*)^\top \boldsymbol{M}_*^{-1} A^*\right]_{kl} &\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 \left(\sum_{r\neq k} \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kr} + \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk}\right), \\
&\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 \left[\Sigma_{\boldsymbol{W}}^{1:T}\right]_{kk}, \\
&\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^3.
\end{aligned}$$

Finally, this leads to, for all $k \in [K]$,

$$(nT)^{-1}\sigma_1\left(\boldsymbol{\Pi}_*\right)^2 [\boldsymbol{B}]_{k1} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^3 \sum_{l=1}^{K} [\boldsymbol{B}]_{l1}\,.$$

However, entries of $[\boldsymbol{B}]_{.1}$ are positive and thus $\sum_{l=1}^{K} [\boldsymbol{B}]_{l1} = \| [\boldsymbol{B}]_{.1} \|_1$. The $\mathbb{L}_1$-$\mathbb{L}_2$ inequality then ensures

$$(nT)^{-1} \sigma_1 (\boldsymbol{\Pi}_*)^2 [\boldsymbol{B}]_{k1} \geq \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T})^3 \| [\boldsymbol{B}]_{.1} \|_2 .$$

Finally, notice that $\| [\boldsymbol{B}]_{.1} \|_2^2 = \sum_{k=1}^{K} [\boldsymbol{B}]_{k1}^2$. Moreover, for all $(l, m) \in [K]^2$, we have $\left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{lm} = \sum_{k=1}^{K} [\boldsymbol{B}]_{kl} [\boldsymbol{B}]_{km}$. Hence we have

$$\| [\boldsymbol{B}]_{.1} \|_2^2 = \left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{11} .$$

Recalling that diagonal entries of a positive definite matrix cannot be smaller than the smallest eigenvalue and that eigenvalues of a matrix and its transpose are equal ensures that

$$\left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{11} \geq \lambda_{\min} \left(\boldsymbol{B}^\top \boldsymbol{B}\right) = \lambda_{\min} \left(\boldsymbol{B} \boldsymbol{B}^\top\right).$$

With the result previously demonstrated, this provides

$$\| [\boldsymbol{B}]_{.1} \|_2^2 \geq \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T}).$$

Finally, we have for all $k \in [K]$,

$$[\boldsymbol{B}]_{k1} \geq nT \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T})^{7/2} \sigma_1 (\boldsymbol{\Pi}_*)^{-2} ,$$

which leads to

$$\left(\min_{k \in [K]} [\boldsymbol{B}]_{k1} \right)^{-1} \leq (nT)^{-1} \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T})^{-7/2} \sigma_1 (\boldsymbol{\Pi}_*)^2 ,$$

finally providing for all $k \in [K]$,

$$\|\eta_k\|_2 \leq (nT)^{-1} \sigma_1 (\boldsymbol{\Pi}_*)^2 \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T})^{-4}.$$

However, $\boldsymbol{\Pi}_*$ is a random matrix, making $\sigma_1 (\boldsymbol{\Pi}_*)$ itself a random variable. To control this quantity, we invoke the submultiplicativity of the operator norm. This leads to the following sequence of inequalities :

$$\begin{aligned}
\sigma_1 (\boldsymbol{\Pi}_*) &= \sigma_1 \left(\boldsymbol{M}_*^{-1/2} A^* \boldsymbol{W}^{1:T} \right), \\
&\leq \sigma_1 \left(\boldsymbol{M}_*^{-1/2} \right) \sigma_1 \left(A^* \boldsymbol{W}^{1:T} \right), \\
&\leq \lambda_K (\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2} h_{\min}^{-1/2} \sigma_1 \left(A^* \boldsymbol{W}^{1:T} \right), \\
&\leq \sqrt{\frac{p}{\lambda_K (\Sigma_{\boldsymbol{W}}^{1:T}) c_1 K}} \sigma_1 \left(A^* \boldsymbol{W}^{1:T} \right).
\end{aligned}$$

According to Definition 5.6.1, A^* and $\boldsymbol{W}^{1:T}$ are left stochastic matrices. By Lemma 5.6.7 their product $A^* \boldsymbol{W}^{1:T}$ remains left stochastic. Subsequently, Lemma 5.6.8 guarantees that $\left\| A^* \boldsymbol{W}^{1:T} \right\|_1 = 1$ almost surely and the following bounds on the spectrum of $A^* \boldsymbol{W}^{1:T} \in \mathbb{R}^{p \times nT}$ hold true almost surely :

$$\sqrt{\frac{1}{p}} \leq \sigma_1 \left(A^* \boldsymbol{W}^{1:T} \right) \leq \sqrt{nT}.$$

Combining these results yields a non-random upper bound on $\sigma_1(\mathbf{\Pi}_*)$:

$$\sigma_1(\mathbf{\Pi}_*) \leq \sqrt{\frac{nTp}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})c_1K}}.$$

This outcome establishes a non-random upper bound on $\left(\min_{k\in[K]}[\boldsymbol{B}]_{k1}\right)^{-1}$.

Next, the equality $U = \boldsymbol{M}_*^{-1/2}A^*\boldsymbol{B}$ also ensures that for all $i \in [p]$ and for all $k \in [K]$, $u_k(i) = [\boldsymbol{M}_*]_{ii}^{-1/2}[A^*]_{i.}[\boldsymbol{B}]_{.k}$. Hence, the following inequality holds true :

$$|[U]_{.1}(i)| \geq [\boldsymbol{M}_*]_{ii}^{-1/2}\|[A^*]_{i.}\|_1 \min_{k\in[K]}|[\boldsymbol{B}]_{k1}|.$$

Proposition 5.2.8 ensures that for all $i \in [p]$,

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

In addition, for all $i \in [p]$, $\|[A^*]_{i.}\|_1 = h_i$. This leads to, for all $i \in [p]$,

$$|[U]_{.1}(i)| \geq \sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|.$$

Finally, for all $i \in [p]$, we have $[\hat{U}]_{.1}(i) \geq [U]_{.1}(i) - \left|[\hat{U}]_{.1}(i) - [U]_{.1}(i)\right|$. This provides, under the conditions previously stated that

$$\begin{aligned}
[\hat{U}]_{.1}(i) \geq & \sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}| \\
& - \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_3}{\sqrt{pN}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_4}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\right), \\
& - \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K} + C_5 + \frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}} + C_1^2 + C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\right).
\end{aligned}$$

Finally, Lemma 5.6.5 ensures that for all $(x, y) \in \mathbb{R}^2$ such that $x > y$ we have

$$(x-y)^{-1} \leq x^{-1} + y(x-y)^{-2}.$$

Let us define the quantity A as follows,

$$\begin{aligned}
A := & \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_3}{\sqrt{pN}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})} + \frac{C_4}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\right), \\
& + \frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K} + C_5 + \frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}} + C_1^2 + C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\right).
\end{aligned}$$

Hence this provides

$$\left|[\hat{U}]_{.1}(i)\right|^{-1} \leq \left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|\right)^{-1} + A\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}| - A\right)^{-2}.$$

Combining these results leads to

$$\begin{aligned}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 &\leq \left|[\hat{U}]_{.1}(i)^{-1}\right|\left\|\Omega\left[\hat{U}\right]_{i.}-[U]_{i.}\right\|_2\left(2+\max_{k\in[K]}\|\eta_k\|_2\right),\\ &\leq A\left[\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|\right)^{-1}+A\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-A\right)^{-2}\right]\left(2+\max_{k\in[K]}\|\eta_k\|_2\right).\end{aligned}$$

The previously demonstrated inequalities provide the stated result, namely

$$\begin{aligned}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 \leq\ & A\left({h_i}^{-1/2}(nT)^{-1}\sigma_1\left(\boldsymbol{\Pi}_*\right)^2\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-7/2}\right)\left(2+(nT)^{-1}\sigma_1\left(\boldsymbol{\Pi}_*\right)^2\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-4}\right)\\ &+A^2\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-A\right)^{-2}\left(2+(nT)^{-1}\sigma_1\left(\boldsymbol{\Pi}_*\right)^2\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-4}\right).\end{aligned}$$

Using the non-random upper bound on $\sigma_1\left(\boldsymbol{\Pi}_*\right)$ leads to :

$$\begin{aligned}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 \leq\ & A\left({h_i}^{-1/2}\frac{p}{c_1K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-9/2}\right)\left(2+\frac{p}{c_1K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-5}\right)\\ &+A^2\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-A\right)^{-2}\left(2+\frac{p}{c_1K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-5}\right).\end{aligned}$$

■

Proof of Theorem 5.2.15. The same results as the ones stated in Theorem 5.2.13. hold. In addition, $(N-2)nT \geq C_{tot}(p,N)^2\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)^2\frac{p^3}{c_2^9c_1^2K^2}$ ensures

$$\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-Z \geq \sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}/|\,2.$$

Thus the stated result holds by derivation of the bounds stated in Theorem 5.5.4. ■

5.5.13 Proof of Theorem 5.2.16

Theorem 5.5.5 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6 and 7. Let $\hat{A}$ be the estimator of A^* defined in* (5.1)*. Let $\mathcal{D}_K$ be the set of matrices $\Omega = diag(\omega, \Omega_{2:K}) \in \mathbb{R}^{K\times K}$ where $\omega \in \{-1,1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ is an orthogonal matrix. Let us denote*

$$\begin{aligned}\Theta_1 &:= \max_{i\in[p]} h_i^{-1/2}\left|[\hat{M}]_{ii}-[\boldsymbol{M}_*]_{ii}\right|,\\ \Theta_2 &:= \min_{\Psi\in\mathcal{D}_K}\max_{i\in[p]} h_i^{-1/2}\left\|\Psi[\hat{U}]_{i.}-[U]_{i.}\right\|_2.\end{aligned}$$

Then, up to a permutation of columns of $\hat{A}$ we have

$$\max_{i\in[p]}\left(\frac{\left\|\ [\hat{A}]_{i.}-[A^*]_{i.}\right\|_1}{h_i}\right) \leq 2\frac{\kappa}{\left(c_2^{9/2}c_1K-\kappa\right)},$$

where

$$\begin{aligned}\kappa \leq & \frac{K^2}{c_2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\frac{2\max_{i\in[p]}\left\| \Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2}{\left(\frac{c_2}{\sqrt{K}}-KC_{VH}\max_{i\in[p]}\left\| \Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2\right)} \\ & +\frac{K^{7/2}}{c_2^2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\frac{2C_{VH}\max_{i\in[p]}\left\| \Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2}{\left(\frac{c_2}{\sqrt{K}}-KC_{VH}\max_{i\in[p]}\left\| \Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2\right)}\max_{x\in\mathcal{G}_\eta}\|x\|_2 \\ & +K\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\Theta_2 \\ & +\frac{K^{3/2}}{c_2}\Theta_1,\end{aligned}$$

Proof of Theorem 5.5.5. For notation simplicity we omit the permutation $\sigma \in \mathcal{S}_K$ in the defintion of Θ_3. From the definitions of Θ_1, Θ_2 and Θ_3 there exists $\omega \in \{-1,+1\}$ and $\Omega_{2:K} \in \mathbb{R}^{(K-1)\times(K-1)}$ an orthogonal matrix such that for all $(i,k) \in [p]\times[K]$,

$$\begin{aligned}\left|[\hat{M}]_{ii}-[\boldsymbol{M}_*]_{ii}\right| &\leq \Theta_1 h_i, \\ \left\|\Omega[\hat{U}]_{i.}-[U]_{i.}\right\|_2 &\leq \Theta_2\sqrt{h_i}, \\ \Theta_3 &= \left\|\Omega_{2:K}\hat{\eta}_k-\eta_k\right\|_2.\end{aligned}$$

In addition, Perron-Frobenius's theorem guarantees that u_1 does not possess any null entry and the proof of Theorem 5.5.4 contains the following inequality holding true for all $i \in [p]$:

$$|[U]_{.1}(i)| \geq \sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|,$$

where $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ is the non-singular matrix satisfying :

$$\begin{aligned}\left(\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1} &= [A^*]^\top \boldsymbol{M}_*^{-1}A^*, \\ U &= \boldsymbol{M}_*^{-1/2}A^*\boldsymbol{B}.\end{aligned}$$

Moreover $\left|\omega[\hat{U}]_{.1}(i)-[U]_{.1}(i)\right|$ is upper-bounded by $\left|\Omega[\hat{U}]_{i.}-[U]_{i.}\right|_2$ which is itself bounded from above by $\Theta_2\sqrt{h_i}$. Note that fixing $\omega = 1$ ensures that $\sum_{i=1}^{p}\omega[\hat{U}]_{.1}(i) > 0$. In addition, if N, n and/or T is sufficiently large such that $\Theta_2 < \min_{k\in[K]}|[\boldsymbol{B}]_{k1}|$ then for all $i \in [p]$ we have $[\hat{U}]_{.1}(i) > 0$. Next, Theorem 4.3.2 ensures that

$$A^* = \mathcal{N}_{col}\left(\boldsymbol{M}_*^{1/2}\mathsf{diag}(u_1)\mathcal{P}_{round}\left(\hat{\Lambda}\right)\right).$$

Let us recall that $\hat{\Lambda} = [\hat{\lambda}_1,\ldots,\hat{\lambda}_p]^\top \in \mathbb{R}^{p\times K}$ is defined as solving the following linear system for all $i \in [p]$:

$$\begin{pmatrix}1 & \ldots & 1\\ \hat{\eta}_1 & \ldots & \hat{\eta}_K\end{pmatrix}\hat{\lambda}_i = \begin{pmatrix}1\\ \hat{r}_i\end{pmatrix}.$$

This implies

$$\begin{pmatrix} 1 & \dots & 1 \\ \Omega_{2:K}\hat{\eta}_1 & \dots & \Omega_{2:K}\hat{\eta}_K \end{pmatrix} \hat{\lambda}_i = \begin{pmatrix} 1 \\ \Omega_{2:K}\hat{r}_i \end{pmatrix}.$$

For any matrix M we denote $M^\dagger$ its Moore-Penrose inverse. Then consider $\hat{T}_K := \begin{pmatrix} 1 & \dots & 1 \\ \Omega_{2:K}\hat{\eta}_1 & \dots & \Omega_{2:K}\hat{\eta}_K \end{pmatrix} \in \mathbb{R}^{K\times K}$. If $\operatorname{rank}\left(\hat{T}_K\right) = K-1$, then theere is one vector $\hat{\eta}_l$ which is a linear combination of the vectors $(\hat{\eta}_k)_{k\neq l}$. Let e_K be a vector completing $(\hat{\eta}_k)_{k\neq l}$ in a basis of $\mathbb{R}^K$. For any $\epsilon > 0$, we define $\hat{\eta}_l^\epsilon$ as $\hat{\eta}_l + \epsilon e_K$ and for $k \neq l$, $\hat{\eta}_k^\epsilon = \hat{\eta}_k$. Hence, $(\hat{\eta}_k^\epsilon)_k$ is a basis of $\mathbb{R}^K$ and Assumption 8 remains true. Indeed for all $k \in [K]\backslash\{l\}$

$$\| \hat{\eta}_k^\epsilon - \eta_k\|_2 = \| \hat{\eta}_k - \eta_k\|_2 ,$$

and

$$\| \hat{\eta}_l^\epsilon - \eta_l\|_2 \leq \| \hat{\eta}_l - \eta_l\|_2 + \epsilon \|e_K - \eta_l\|_2 .$$

Then for $\epsilon > 0$ small enough we have

$$\max_{k\in[K]} \| \Omega_{2:K}\hat{\eta}_k^\epsilon - \eta_k\|_2 \leq C_{VH} \max_{i\in[p]} \left\| \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 .$$

The same holds by induction if $\operatorname{rank}\left(\hat{T}_K\right) < K-1$. For the remainder of the proof, $\epsilon > 0$ is chosen so that the previously stated condition is satisfied and $\hat{\eta}$ and $\hat{\eta}^\epsilon$ are used interchangeably. Hence $\hat{T}_K$ can be assumed to be invertible and for all $i \in [p]$, $\hat{\lambda}_i$ is defined as follows :

$$\hat{\lambda}_i = \hat{T}_K^{-1} \begin{pmatrix} 1 \\ \Omega_{2:K}\hat{r}_i \end{pmatrix}.$$

It follows that

$$\hat{\Lambda} = \left[1_p, \Omega_{2:K}\hat{R}\right] \left(\hat{T}_K^\top\right)^{-1},$$

where $\left[1_p, \Omega_{2:K}\hat{R}\right] = \Omega_{2:K}[\mathsf{diag}([\hat{U}]_{.1})]^{-1}\hat{U}$. Similarly, the population counterparts satisfy the following equality holding true for all $i \in [p]$:

$$\lambda_i = T_K^\dagger \begin{pmatrix} 1 \\ r_i \end{pmatrix},$$

where $T_K = \begin{pmatrix} 1 & \dots & 1 \\ \eta_1 & \dots & \eta_K \end{pmatrix} \in \mathbb{R}^{K\times K}$. It follows that

$$\Lambda = [1_p, R] \left(T_K^\top\right)^\dagger \in \mathbb{R}_+^{p\times K},$$

where $[1_p, R] = [\mathsf{diag}(u_1)]^{-1}U$. This implies that

$$[1_p, R] = [\mathsf{diag}(u_1)]^{-1}\boldsymbol{M}_*^{-1/2}A^*\boldsymbol{B}.$$

Let us define $N \in \mathbb{R}^{K\times(K-1)}$ as follows :

$$\forall (k,s) \in [K]\times[K-1],\ [N]_{ks} = \frac{[\boldsymbol{B}]_{k,(s+1)}}{[\boldsymbol{B}]_{k1}}.$$

This ensures that $\boldsymbol{B} = \text{diag}([\boldsymbol{B}]_{.1})[1_K, N]$ and thus we have

$$[1_p, R] = [\text{diag}(u_1)]^{-1}\boldsymbol{M}_*^{-1/2}A^*\text{diag}([\boldsymbol{B}]_{.1})[1_K, N].$$

This equality can be equivalently written as

$$\begin{aligned} 1_p &= [\text{diag}(u_1)]^{-1}\boldsymbol{M}_*^{-1/2}A^*\text{diag}([\boldsymbol{B}]_{.1})1_K, \\ R &= [\text{diag}(u_1)]^{-1}\boldsymbol{M}_*^{-1/2}A^*\text{diag}([\boldsymbol{B}]_{.1})N. \end{aligned}$$

This ensures that the rows of R are convex combinations of the rows of N and thus we have

$$\forall k \in [K],\ \eta_k := [N]_{k.} \in \mathbb{R}^{(K-1)}.$$

This implies that $T_K^\top = [1_K, N] = \text{diag}([\boldsymbol{B}]_{.1})^{-1}\boldsymbol{B}$. Moreover, $\boldsymbol{B}$ is non-singular. In the proof of Proposition 5.2.14 is proven that $[\boldsymbol{B}]_{.1}$ has positive entries and thus $\text{diag}([\boldsymbol{B}]_{.1})[1_K, N]$ is non singular. This proves that T_K is non singular and thus $T_K^\dagger = T_K^{-1}$. Globally, it implies that for all $i \in [p]$,

$$\hat{\lambda}_i - \lambda_i = \hat{T}_K^{-1}\begin{pmatrix} 1 \\ \Omega_{2:K}\hat{r}_i \end{pmatrix} - \hat{T}_K^{-1}\begin{pmatrix} 1 \\ r_i \end{pmatrix} + \hat{T}_K^{-1}\begin{pmatrix} 1 \\ r_i \end{pmatrix} - T_K^{-1}\begin{pmatrix} 1 \\ r_i \end{pmatrix}.$$

From this expansion is deduced that for all $i \in [p]$,

$$\left\| \hat{\lambda}_i - \lambda_i \right\|_2 \leq \left\| \hat{T}_K^{-1} \right\|_{op} \left\| \ \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2 + \left\| \hat{T}_K^{-1} - T_K^{-1} \right\|_{op} \| [R]_{i.} \|_2 .$$

In addition, T_K^{-1} can be expressed as $\text{diag}([\boldsymbol{B}]_{.1})(\boldsymbol{B}^\top)^{-1}$. Recalling that for any matrix $A \in \mathbb{R}^{n\times m}$ we have $\|A\|_{op}^2 = \left\|AA^\top\right\|_{op} = \left\|A^\top A\right\|_{op}$ it comes :

$$\|T_K^{-1}\|_{op}^2 = \left\| \left(T_K^\top T_K\right)^{-1} \right\|_{op} = \left\| T_K^{-1}\left(T_K^\top\right)^{-1} \right\|_{op} = \left\| \text{diag}([\boldsymbol{B}]_{.1})(\boldsymbol{B}^\top)^{-1}\ \boldsymbol{B}^{-1}\text{diag}([\boldsymbol{B}]_{.1}) \right\|_{op}.$$

Hence by definition of $(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}$ we have

$$\|T_K^{-1}\|_{op}^2 = \left\| \text{diag}([\boldsymbol{B}]_{.1})(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}\text{diag}([\boldsymbol{B}]_{.1}) \right\|_{op} = \left\| \ \text{diag}([\boldsymbol{B}]_{.1})[A^*]^\top\boldsymbol{M}_*^{-1}A^*\text{diag}([\boldsymbol{B}]_{.1}) \right\|_{op}.$$

Finally, the operator norm being submultiplicative, the following inequality holds true :

$$\|T_K^{-1}\|_{op}^2 \leq \|\ \text{diag}([\boldsymbol{B}]_{.1})\ \|_{op}^2 \left\| [A^*]^\top\boldsymbol{M}_*^{-1}A^* \right\|_{op}.$$

Proposition 5.2.8 then ensures that

$$\left\| [A^*]^\top\boldsymbol{M}_*^{-1}A^* \right\|_{op} \leq \lambda_{\min}\left(\Sigma_{\boldsymbol{W}}^{1:T}\right)^{-1} \left\| [A^*]^\top H^{-1}A^* \right\|_{op}.$$

By definition of the matrix H presented in Assumption 7, we have

$$\left\| [A^*]^\top H^{-1}A^* \right\|_{op} = \max_{k\in[K]} \sum_{l=1}^{K}\sum_{i=1}^{p} h_i^{-1}[A^*]_{ik}[A]_{il}.$$

In addition, for all $i \in [p]$ and for all $k \in [K]$, we have $h_i^{-1}[A^*]_{ik} = \dfrac{[A^*]_{ik}}{\| [A^*]_{i.} \|_1} \leq 1$. Hence

$$\left\| [A^*]^\top H^{-1} A^* \right\|_{op} \leq \max_{k \in [K]} \sum_{l=1}^{K} \sum_{i=1}^{p} [A^*]_{il} = K.$$

Moreover, the proof of Theorem 5.5.4 contains the following inequality :

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \leq \lambda_{\min}\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \leq \lambda_1\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

Noticing that $\| [\boldsymbol{B}]_{.1} \|_2^2 = \sum_{k=1}^{K} [\boldsymbol{B}]_{k1}^2$ and that, for all $(l,m) \in [K]^2$, we have $\left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{lm} = \sum_{k=1}^{K} [\boldsymbol{B}]_{kl} [\boldsymbol{B}]_{km}$ leads to :

$$\| [\boldsymbol{B}]_{.1} \|_2^2 = \left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{11}.$$

Recalling that diagonal entries of a positive definite matrix cannot be above the biggest eigenvalue and that eigenvalues of a matrix and its transpose are equal ensures that

$$\left[\boldsymbol{B}^\top \boldsymbol{B}\right]_{11} \leq \lambda_1\left(\boldsymbol{B}^\top \boldsymbol{B}\right) = \lambda_1\left(\boldsymbol{B}\boldsymbol{B}^\top\right).$$

Then it comes :

$$\| [\boldsymbol{B}]_{.1} \|_2^2 \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

This ensures

$$\| \operatorname{diag}([\boldsymbol{B}]_{.1}) \|_{op}^2 = \max_{k \in [K]} \left| [\boldsymbol{B}]_{k1}^2 \right| \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

Finally

$$\left\| T_K^{-1} \right\|_{op}^2 \leq K \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-2}.$$

Lemma 5.6.8 ensures that for any matrix $M \in \mathbb{R}^{n \times m}$,

$$\frac{1}{\sqrt{n}} \|M\|_1 \leq \|M\|_{op} \leq \sqrt{m} \|M\|_1 \quad \text{and} \quad \|M\|_1 = \max_{j \in [m]} \sum_{i=1}^{n} \left| [M]_{ij} \right|.$$

Hence

$$\left\| \hat{T}_K - T_K \right\|_{op} \leq \sqrt{K} \left\| \hat{T}_K - T_K \right\|_1 = \sqrt{K} \max_{l \in [K]} \sum_{k=1}^{K} \left| \left[\hat{T}_K - T_K \right]_{kl} \right|.$$

Moreover, for all $l \in [K]$, we have $\sum_{k=1}^{K} \left| \left[\hat{T}_K - T_K \right]_{kl} \right| \leq \| \Omega_{2:K} \hat{\eta}_l - \eta_l \|_1$. Recalling that for any $x \in \mathbb{R}^d$ we have $\|x\|_1 \leq \sqrt{d} \|x\|_2$ which leads

$$\left\| \hat{T}_K - T_K \right\|_{op} \leq \sqrt{K}\sqrt{K-1} \max_{l \in [K]} \| \Omega_{2:K} \hat{\eta}_l - \eta_l \|_2 .$$

Assumption 8 then ensures that

$$\left\| \hat{T}_K - T_K \right\|_{op} \leq K C_{VH} \max_{i \in [p]} \left\| \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2 .$$

Next, notice that

$$\left\|\hat{T}_K^{-1} - T_K^{-1}\right\|_{op} = \left\|\hat{T}_K^{-1}\left(T_K - \hat{T}_K\right) T_K^{-1}\right\|_{op}.$$

The operator norm being submultiplicative, we get :

$$\left\|\hat{T}_K^{-1} - T_K^{-1}\right\|_{op} \leq \left\|\hat{T}_K^{-1}\right\|_{op}\left\|T_K - \hat{T}_K\right\|_{op}\left\|T_K^{-1}\right\|_{op}.$$

Hence we have

$$\left\|\hat{T}_K^{-1} - T_K^{-1}\right\|_{op} \leq K^{3/2}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\max_{l\in[K]}\|\,\Omega_{2:K}\hat{\eta}_l - \eta_l\|_2\left\|\hat{T}_K^{-1}\right\|_{op}.$$

The last step is to control $\left\|\hat{T}_K^{-1}\right\|_{op}$. For any matrix $M \in \mathbb{R}^{K\times K}$, for all $k \in [K]$, $\sigma_k(M)$ define the k^{th} largest singular value of M and $\lambda_k(MM^\top)$ define the k^{th} largest eigenvalue of $MM^\top$ which is symmetric. First note that $\left\|\hat{T}_K^{-1}\right\|_{op} = \sigma_{\min}\left(\hat{T}_K\right)^{-1}$. Weyl's inequality, see Lemma 1.1.13, ensures that

$$\sigma_{\min}\left(\hat{T}_K\right) \geq \sigma_{\min}\left(T_K\right) - \sigma_{\max}\left(\hat{T}_K - T_K\right).$$

Moreover, $T_K = \boldsymbol{B}^\top \text{diag}\left([\boldsymbol{B}]_{.1}\right)^{-1}$ and Lemma 5.6.4 ensures that

$$\sigma_{\min}\left(T_K\right) \geq \min_{k\in[K]}\left(([\boldsymbol{B}]_{k1})^{-1}\right)\sigma_{\min}(\boldsymbol{B}).$$

By definition of singular values and using a previously stated result, for all $k \in [K]$,

$$\sigma_k\left(\boldsymbol{B}\right)^2 = \lambda_k\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}).$$

In addition, entries of $[\boldsymbol{B}]_{.1}$ being positive, we have

$$\min_{k\in[K]}\left(([\boldsymbol{B}]_{k1})^{-1}\right) = \|[\boldsymbol{B}]_{.1}\|_\infty^{-1}.$$

Moreover there are

$$\|[\boldsymbol{B}]_{.1}\|_\infty \leq \|[\boldsymbol{B}]_{.1}\|_1 \leq \sqrt{K}\,\|[\boldsymbol{B}]_{.1}\|_2,$$

and

$$\|\,[\boldsymbol{B}]_{.1}\|_2^2 \leq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

Hence

$$\|[\boldsymbol{B}]_{.1}\|_\infty \leq \sqrt{K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2}.$$

Thus,

$$\|[\boldsymbol{B}]_{.1}\|_\infty^{-1} \geq \sqrt{\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{K}}.$$

Finally,

$$\sigma_{\min}\left(T_K\right) \geq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}}.$$

Hence we derive :

$$\sigma_{\min}\left(\hat{T}_K\right) \geq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - K \max_{l\in[K]} \| \Omega_{2:K}\hat{\eta}_l - \eta_l\|_2 \,.$$

Assumption 8 then ensures that $\max_{l\in[K]} \| \Omega_{2:K}\hat{\eta}_l - \eta_l\|_2 \leq C_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2$. Hence

$$\sigma_{\min}\left(\hat{T}_K\right) \geq \frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2 .$$

Hence we have

$$\left\|\hat{T}_K^{-1}\right\|_{op} \leq \left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)^{-1}.$$

Finally,

$$\left\|\hat{T}_K^{-1} - T_K^{-1}\right\|_{op} \leq \frac{K^{3/2}C_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)}.$$

From these results is deduced that for all $i \in [p]$,

$$\begin{aligned}\left\| \hat{\lambda}_i - \lambda_i\right\|_2 \leq & \frac{\left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2}{\left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)} \\ & + \frac{K^{3/2}C_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)} \| [R]_{i.} \|_2 \,.\end{aligned}$$

Moreover, (5.1) ensures that $\hat{A} = \mathcal{N}_{col}\left(\hat{M}^{1/2}\mathsf{diag}([\hat{U}]_{.1})\mathcal{P}_{round}\left(\hat{\Lambda}\right)\right)$. Let us denote $\tilde{\Lambda} = \mathcal{P}_{round}\left(\hat{\Lambda}\right) \in \mathbb{R}_+^{p\times K}$. Hence for all $i \in [p]$, $\tilde{\lambda}_i = \frac{\left(\hat{\lambda}_i\right)_+}{\left\|\left(\hat{\lambda}_i\right)_+\right\|_1}$ where for any vector $x \in \mathbb{R}^d$, for all $s \in [d]$, $(x)_+(s) = \max(x(s), 0)$. Thus the following inequalities hold true for all $i \in [p]$ and are deduced using the triangle inequality and the definition of $\tilde{\lambda}_i$:

$$\begin{aligned}\left\| \tilde{\lambda}_i - \lambda_i\right\|_1 &\leq \left\| \tilde{\lambda}_i - \left(\hat{\lambda}_i\right)_+\right\|_1 + \left\| \left(\hat{\lambda}_i\right)_+ - \lambda_i\right\|_1 , \\ &\leq \left\| \tilde{\lambda}_i - \tilde{\lambda}_i \left\|\left(\hat{\lambda}_i\right)_+\right\|_1\right\|_1 + \left\| \left(\hat{\lambda}_i\right)_+ - \lambda_i\right\|_1 , \\ &\leq \left|1 - \left\|\left(\hat{\lambda}_i\right)_+\right\|_1\right| + \left\| \left(\hat{\lambda}_i\right)_+ - \lambda_i\right\|_1 .\end{aligned}$$

By definition $\|\lambda_i\|_1 = \left\|\hat{\lambda}_i\right\|_1 = 1$. This implies that $\left|1 - \left\|\left(\hat{\lambda}_i\right)_+\right\|_1\right| = \left|\|\lambda_i\|_1 - \left\|\left(\hat{\lambda}_i\right)_+\right\|_1\right|$. Using the reverse triangle inequality leads to

$$\left\| \tilde{\lambda}_i - \lambda_i \right\|_1 \leq 2 \left\| \left(\hat{\lambda}_i\right)_+ - \lambda_i \right\|_1 .$$

Moreover, the entries of λ_i are non negative which gives

$$\left\| \tilde{\lambda}_i - \lambda_i \right\|_1 \leq 2 \left\| \hat{\lambda}_i - \lambda_i \right\|_1 .$$

Finally, using the $\mathbb{L}_1$-$\mathbb{L}_2$ inequality provides

$$\left\| \tilde{\lambda}_i - \lambda_i \right\|_1 \leq 2\sqrt{K} \left\| \hat{\lambda}_i - \lambda_i \right\|_2 .$$

Hence we have

$$\begin{aligned}\left\| \tilde{\lambda}_i - \lambda_i \right\|_2 \leq & \frac{2\sqrt{K} \left\| \ \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2}{\left(\frac{\lambda_K(\Sigma^{1:T}_{\boldsymbol{W}})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \ \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2 \right)} \\ & + \frac{2K^2 C_{VH} \max_{i\in[p]} \left\| \ \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2}{\lambda_K(\Sigma^{1:T}_{\boldsymbol{W}}) \left(\frac{\lambda_K(\Sigma^{1:T}_{\boldsymbol{W}})}{\sqrt{K}} - KC_{VH} \max_{i\in[p]} \left\| \ \Omega_{2:K} \left[\hat{R}\right]_{i.} - [R]_{i.} \right\|_2 \right)} \| [R]_{i.} \|_2 .\end{aligned}$$

The last step is to control $\| [R]_{i.} \|_2$. Proposition 5.2.14 ensures that $[R]_{i.}$ is in the convex hull of $\eta_1, \ldots, \eta_K$ and thus $\| [R]_{i.} \|_2 \leq \max_{x\in\mathcal{G}_\eta} \|x\|_2$. Theorem 5.5.4 then concludes to control $\left\| \tilde{\lambda}_i - \lambda_i \right\|_2$.

Next, consider the step of estimating $\boldsymbol{\Delta}^* := A^* \mathsf{diag}([\boldsymbol{B}]_{.1}) = \boldsymbol{M}_*^{1/2} \mathsf{diag}([U]_{.1})\Lambda \in \mathbb{R}^{p\times K}$ with

$$\tilde{A} := \hat{M}^{1/2} \mathsf{diag}([\hat{U}]_{.1})\tilde{\Lambda} \in \mathbb{R}^{p\times K}.$$

For all $i \in [p]$, using the triangle inequality leads to the following inequality :

$$\begin{aligned}\left\| [\tilde{A}]_{i.} - [\boldsymbol{\Delta}^*]_{i.} \right\|_1 &= \left\| \ [\hat{M}]_{ii}^{1/2} [\hat{U}]_{.1}(i)\tilde{\lambda}_i - \ [\boldsymbol{M}_*]_{ii}^{1/2} [U]_{.1}(i)\lambda_i \right\|_1 , \\ &= \left\| \ [\hat{M}]_{ii}^{1/2} [\hat{U}]_{.1}(i) \left(\tilde{\lambda}_i - \ \lambda_i\right) + \left([\hat{M}]_{ii}^{1/2} [\hat{U}]_{.1}(i) - [\boldsymbol{M}_*]_{ii}^{1/2} [U]_{.1}(i) \right) \lambda_i \right\|_1 , \\ &\leq \left| [\hat{M}]_{ii}^{1/2} [\hat{U}]_{.1}(i) \right| \left\| \ \tilde{\lambda}_i - \ \lambda_i \right\|_1 + \left\| [\hat{M}]_{ii}^{1/2} \left([\hat{U}]_{.1}(i) - [U]_{.1}(i) \right) \lambda_i + [U]_{.1}(i) \left([\hat{M}]_{ii}^{1/2} - [\boldsymbol{M}_*]_{ii}^{1/2} \right) \lambda_i \right\|_1 , \\ &\leq \left| [\hat{M}]_{ii}^{1/2} [\hat{U}]_{.1}(i) \right| \left\| \ \tilde{\lambda}_i - \ \lambda_i \right\|_1 + \left| [\hat{M}]_{ii}^{1/2} \right| \left| [\hat{U}]_{.1}(i) - [U]_{.1}(i) \right| \|\lambda_i\|_1 \\ &\quad + |[U]_{.1}(i)| \left| [\hat{M}]_{ii}^{1/2} - [\boldsymbol{M}_*]_{ii}^{1/2} \right| \|\lambda_i\|_1 .\end{aligned}$$

First notice that

$$\left| [\hat{U}]_{.1}(i) - [U]_{.1}(i) \right| \leq \left\| \tilde{\Omega}[\hat{U}]_{i.} - [U]_{i.} \right\|_2 \leq h_i^{1/2} \Theta_2.$$

Moreover, the equality $U = \boldsymbol{M}_*^{-1/2} A^* \boldsymbol{B}$ also ensures that for all $i \in [p]$ and for all $k \in [K]$, $u_k(i) = [\boldsymbol{M}_*]_{ii}^{-1/2} [A^*]_{i.} [\boldsymbol{B}]_{.k}$. Hence, the following inequality holds true :

$$|[U]_{.1}(i)| \leq [\boldsymbol{M}_*]_{ii}^{-1/2} \|[A^*]_{i.}\|_1 \|[\boldsymbol{B}]_{.1}\|_\infty .$$

Let us remind that for all $i \in [p]$, $h_i = \|[A^*]_{i.}\|_1$, $\|[\boldsymbol{B}]_{.1}\|_\infty$ is bounded from above by $\sqrt{K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1/2}$ and Proposition 5.2.8 ensures that for all $i \in [p]$,

$$\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

We deduce that for all $i \in [p]$,

$$|[U]_{.1}(i)| \leq \sqrt{h_i K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}.$$

Moreover, this leads to the following inequality holding true for all $i \in [p]$,

$$\left|[\hat{U}]_{.1}(i)\right| \leq \left(\sqrt{K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} + \Theta_2\right)\sqrt{h_i}.$$

In addition , for all $i \in [p]$,

$$\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right| \leq h_i^{1/2}\Theta_1.$$

Proposition 5.2.8 then ensures that for all $i \in [p]$,

$$[\hat{M}]_{ii} \leq h_i + h_i^{1/2}\Theta_1.$$

As a result,

$$\begin{aligned}\left\| [\tilde{A}]_{i.} - [\boldsymbol{\Delta}^*]_{i.}\right\|_1 \leq \sqrt{h_i + h_i^{1/2}\Theta_1}\sqrt{h_i K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \left\| \tilde{\lambda}_i - \lambda_i\right\|_1 \\ + \sqrt{h_i + h_i^{1/2}\Theta_1} h_i^{1/2}\Theta_2 \|\lambda_i\|_1 \\ + \sqrt{h_i K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} h_i^{1/2}\Theta_1 \|\lambda_i\|_1\end{aligned}$$

Assumption 6 combined with the definition of λ_i ensuring that for all $i \in [p]$, $\|\lambda_i\|_1 = 1$ provides

$$\begin{aligned}\left\| [\tilde{A}]_{i.} - [\boldsymbol{\Delta}^*]_{i.}\right\|_1 \leq h_i\sqrt{1 + \Theta_1\sqrt{\frac{p}{c_1 K}}}\sqrt{K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \left\| \tilde{\lambda}_i - \lambda_i\right\|_1 \\ + h_i\sqrt{1 + \Theta_1\sqrt{\frac{p}{c_1 K}}}\Theta_2 \\ + h_i\sqrt{K}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\Theta_1.\end{aligned}$$

Finally, consider the step of estimating $A^* \in \mathbb{R}^{p\times K}$, which is equal to $\mathcal{N}_{col}\left(\boldsymbol{M}_*^{1/2}\mathsf{diag}(u_1)\mathcal{P}_{round}(\Lambda)\right)$ according to Theorem 4.3.2, with

$$\hat{A} := \mathcal{N}_{col}\left(\hat{M}^{1/2}\mathsf{diag}([\hat{U}]_{.1})\mathcal{P}_{round}\left(\hat{\Lambda}\right)\right) = \mathcal{N}_{col}\left(\tilde{A}\right).$$

Notice that $\hat{A}$ is the matrix obtained from renormalizing each column of $\tilde{A}$ and by definition $A^* = \mathbf{\Delta}^* \text{diag}([\boldsymbol{B}]_{.1})^{-1}$. It follows that for all $k \in [K]$, for all $i \in [p]$,

$$\begin{aligned}\left[\hat{A}\right]_{ik} &= \left[\tilde{A}\right]_{ik} \left\| \left[\tilde{A}\right]_{.k} \right\|_1^{-1}, \\ [A^*]_{ik} &= [\mathbf{\Delta}^*]_{ik} [\boldsymbol{B}]_{k1}^{-1}.\end{aligned}$$

Hence, for all $k \in [K]$, for all $i \in [p]$,

$$\left| \left[\hat{A}\right]_{ik} - [A^*]_{ik} \right| \leq \left\| \left[\tilde{A}\right]_{.k} \right\|_1^{-1} \left| \left[\tilde{A}\right]_{ik} - [\mathbf{\Delta}^*]_{ik} \right| + \frac{\left| \left\| \left[\tilde{A}\right]_{.k} \right\|_1 - [\boldsymbol{B}]_{k1} \right|}{\left\| \left[\tilde{A}\right]_{.k} \right\|_1} |[A^*]_{ik}| .$$

Moreover, for all $k \in [K]$, $\| [A^*]_{.k} \|_1 = 1$ which ensures that $\| [\mathbf{\Delta}^*]_{.k} \|_1 = [\boldsymbol{B}]_{k1}$. Then the following inequalities hold true for all $k \in [K]$,

$$\begin{aligned}\left| \left\| \left[\tilde{A}\right]_{.k} \right\|_1 - [\boldsymbol{B}]_{k1} \right| &\leq \left| \left\| \left[\tilde{A}\right]_{.k} \right\|_1 - \| [\mathbf{\Delta}^*]_{.k} \|_1 \right|, \\ &\leq \left\| \left[\tilde{A}\right]_{.k} - [\mathbf{\Delta}^*]_{.k} \right\|_1, \\ &\leq \sum_{i=1}^{p} \left| \left[\tilde{A}\right]_{ik} - [\mathbf{\Delta}^*]_{ik} \right|, \\ &\leq \sum_{i=1}^{p} \left\| \left[\tilde{A}\right]_{i.} - [\mathbf{\Delta}^*]_{i.} \right\|_1 .\end{aligned}$$

Then, applying the previously proved inequality on $\left\| \left[\tilde{A}\right]_{i.} - [\mathbf{\Delta}^*]_{i.} \right\|_1$ holding true for all $i \in [p]$ and using that $\sum_{i=1}^{K} h_i = K$ leads to, for all $k \in [K]$,

$$\begin{aligned}\left| \left\| \left[\tilde{A}\right]_{.k} \right\|_1 - [\boldsymbol{B}]_{k1} \right| \leq & K^{3/2} \sqrt{1 + \Theta_1 \sqrt{\frac{p}{c_1 K}}} \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \max_{i \in [p]} \left\| \tilde{\lambda}_i - \lambda_i \right\|_1 + K \sqrt{1 + \Theta_1 \sqrt{\frac{p}{c_1 K}}} \Theta_2 \\ & + K^{3/2} \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \Theta_1.\end{aligned}$$

It can then be deduced that for all $k \in [K]$,

$$\left\| \left[\tilde{A}\right]_{.k} \right\|_1 \geq [\boldsymbol{B}]_{k1} - \left| \left\| \left[\tilde{A}\right]_{.k} \right\|_1 - [\boldsymbol{B}]_{k1} \right| .$$

Moreover, in the proof of Theorem 5.5.4 is proven the following inequalities holding true almost surely and for all $k \in [K]$:

$$\begin{aligned}[\boldsymbol{B}]_{k1} &\geq nT \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{7/2} \sigma_1 (\mathbf{\Pi}_*)^{-2}, \\ \sigma_1 (\mathbf{\Pi}_*) &\leq \sqrt{\frac{nTp}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) c_1 K}}.\end{aligned}$$

These ensure that for all $k \in [K]$,

$$[\boldsymbol{B}]_{k1} \geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2} c_1 K.$$

Hence

$$\begin{aligned}\left\| \left[\tilde{A}\right]_{.k}\right\|_1 \geq & \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - K^{3/2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\max_{i\in[p]}\left\|\tilde{\lambda}_i - \ \lambda_i\right\|_1 \\ & - K\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\Theta_2 - K^{3/2}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\Theta_1.\end{aligned}$$

Let us define

$$\kappa := K^{3/2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\max_{i\in[p]}\left\|\tilde{\lambda}_i - \ \lambda_i\right\|_1 + K\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\Theta_2 + K^{3/2}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\Theta_1.$$

These results lead to, for all $i \in [p]$,

$$\begin{aligned}\left\|\left[\hat{A}\right]_{i.} - [A^*]_{i.}\right\|_1 \leq & \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}\left\|\left[\tilde{A}\right]_{i.} - [\boldsymbol{\Delta}^*]_{i.}\right\|_1 \\ & + \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}\|[A^*]_{i.}\|_1\max_{k\in[K]}\left|\left\|\left[\tilde{A}\right]_{.k}\right\|_1 - [\boldsymbol{B}]_{k1}\right|.\end{aligned}$$

Using that for all $i \in [p]$, $\|[A^*]_{i.}\|_1 = h_i$ brings

$$\begin{aligned}\left\|\left[\hat{A}\right]_{i.} - [A^*]_{i.}\right\|_1 \leq & \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}\left\|\left[\tilde{A}\right]_{i.} - [\boldsymbol{\Delta}^*]_{i.}\right\|_1 \\ & + \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}h_i\max_{k\in[K]}\left|\left\|\left[\tilde{A}\right]_{.k}\right\|_1 - [\boldsymbol{B}]_{k1}\right|.\end{aligned}$$

Finally, noticing that for all $i \in [p]$, $\left\|\left[\tilde{A}\right]_{i.} - [\boldsymbol{\Delta}^*]_{i.}\right\|_1 \leq \kappa h_i K^{-1}$ ensure that for all $i \in [p]$,

$$\begin{aligned}\left\|\left[\hat{A}\right]_{i.} - [A^*]_{i.}\right\|_1 \leq & \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}h_i\kappa K^{-1} \\ & + \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}h_i\kappa, \\ \leq & 2\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)^{-1}h_i\kappa.\end{aligned}$$

Using the upper bound derived on $\left\|\ \tilde{\lambda}_i - \lambda_i\right\|_2$ for all $i \in [p]$ leads to

$$\begin{aligned}\kappa \leq & \frac{K^2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\frac{2\max_{i\in[p]}\left\|\ \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2}{\left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH}\max_{i\in[p]}\left\|\ \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)} \\ & + \frac{K^{7/2}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\frac{2C_{VH}\max_{i\in[p]}\left\|\ \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2}{\left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH}\max_{i\in[p]}\left\|\ \Omega_{2:K}\left[\hat{R}\right]_{i.} - [R]_{i.}\right\|_2\right)}\max_{x\in\mathcal{G}_\eta}\|x\|_2 \\ & + K\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\Theta_2 \\ & + \frac{K^{3/2}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\Theta_1.\end{aligned}$$

The conclusion follows. ■

Proof of Theorem 5.2.16. Proposition 5.2.1 demonstrates that with probability $1-2p\exp(-\epsilon_1^2)$ we have

$$\Theta_1 \leq \frac{2\epsilon_1}{\sqrt{NnT\max(h_i/2,1)}} \leq \frac{2\epsilon_1}{\sqrt{NnT}}.$$

If $NnT \geq \frac{4\epsilon_1^2 p}{c_1 K}$ then $\Theta_1\sqrt{\frac{p}{c_1 K}} \leq 1$. Theorem 5.2.13 ensures that under the stated conditions on $\max(\epsilon_i)$ and N, n and T we have with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2p^2K\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$:

$$\Theta_2 \leq C_{tot}(p,N)\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\sqrt{\frac{p}{nT(N-2)}}.$$

Theorem 5.2.15 guarantees that under the stated conditions on $\max(\epsilon_i)$ and N, n and T we have with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2p^2K\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$:

$$\max_{i\in[p]}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 \leq \left(2\frac{C_{tot}(p,N)\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{c_1c_2^{9/2}K}\frac{p^{3/2}}{\sqrt{nT(N-2)}}\right)\left(2+\frac{p}{c_2^5c_1K}\right).$$

Moreover, $(N-2)nT \geq C_{tot}(p,N)^2(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)^2p^3\frac{4K^2C_{VH}^2}{c_2^2}\left(2+\frac{p}{c_2^5c_1K}\right)^2$ ensures

$$KC_{VH}\max_{i\in[p]}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 \leq \frac{c_2}{2\sqrt{K}}.$$

Then the quantity κ introduced in Proposition 5.5.5 is bounded from above as follows :

$$\kappa \leq 2KC_{tot}(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{\frac{p}{nT(N-2)}}\left[1+\left(2+\frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}}+\frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right]+\frac{4K^{3/2}\epsilon_1}{c_2\sqrt{NnT}}.$$

In addition, if

$$\sqrt{NnT} \geq \frac{4}{c_2^{9/2}c_1}\left[C_{tot}(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{p}\left[1+\left(2+\frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}}+\frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right]+\frac{2\sqrt{K}\epsilon_1}{c_2}\right],$$

we have $c_2^{9/2}c_1K-\kappa \geq c_2^{9/2}c_1K/2$. This concludes. ■

5.5.14 Proof of Theorem 5.2.17

Theorem 5.5.6 *Consider the Dynamic Topic Model, see definition 5.1.1 and assumptions 6, 7 and 8. Let $\hat{A}$ be the estimator of A^* defined in* (5.1)*. Then for all $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4>0$ satisfying the conditions of Theorem 5.5.3, with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$, we have*

$$\sum_{i=1}^{p}\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1 \leq 2K\frac{\kappa}{\left(c_2^{9/2}c_1K-\kappa\right)},$$

with κ is defined in Proposition 5.5.5. In addition, Θ_1 is bounded from above by $\frac{2\epsilon_1\sqrt{2p}}{\sqrt{NnTc_1K}}$, Θ_2 is bounded from above by Z, defined in Theorem 5.5.4, and $\max_{i\in[p]}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2$ is bounded from above by

$$Z\left[\frac{p^{3/2}}{c_2^{9/2}c_1^{3/2}K^{3/2}}+Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-Z\right)^{-2}\right]\left(2+\frac{p}{c_2^5c_1K}\right).$$

Proof of Theorem 5.5.6. Firstly, notice that

$$\sum_{i=1}^{p}\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1=\left(\sum_{i=1}^{p}h_i\right)\max_{i\in[p]}\left(\frac{\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1}{h_i}\right).$$

Using the equality $\sum_{i=1}^{p}h_i=\sum_{i=1}^{p}\sum_{k=1}^{K}[A^*]_{ik}=K$ leads to

$$\sum_{i=1}^{p}\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1\leq K\max_{i\in[p]}\left(\frac{\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1}{h_i}\right).$$

Then, Corollary 5.2.2 ensures that for all $\epsilon_1>0$, with probability at least $1-2p\exp\left(-\epsilon_1^2\right)$, we have

$$\max_{i\in[p]}h_i^{-1/2}\left|[\hat{M}]_{ii}-[\boldsymbol{M}_*]_{ii}\right|<\frac{2\sqrt{2}\epsilon_1}{\sqrt{NnTh_i}}.$$

Using Assumption 6 leads to, for all $\epsilon_1>0$, with probability at least $1-2p\exp\left(-\epsilon_1^2\right)$,

$$\Theta_1:=\max_{i\in[p]}h_i^{-1}\left|[\hat{M}]_{ii}-[\boldsymbol{M}_*]_{ii}\right|<\frac{2\epsilon_1\sqrt{2p}}{\sqrt{NnTc_1K}}.$$

Let us consider

$$\alpha:=\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\frac{\min\left(c_3,\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\right)}{\lambda_1(\Sigma_A)\lambda_1(\Sigma_{\boldsymbol{W}}^{1:T})}<1.$$

Then consider $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4>0$ satisfying

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\leq\alpha\frac{(1-1/N)\sqrt{nT}\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{6\left(\frac{C_1\sqrt{p}}{N}+C_2\sqrt{p}+\frac{C_3}{\sqrt{N}}+C_4\sqrt{p}\right)}.$$

Theorem 5.5.3 ensures that with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$,

$$\Theta_2:=\min_{\Psi\in\mathcal{D}_K}\max_{i\in[p]}h_i^{-1/2}\left\|\Psi[\hat{U}]_{i.}-[U]_{i.}\right\|_2$$

is bounded from above by

$$\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_3}{\sqrt{pN}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_4}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\right),$$
$$+\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K}+C_5+\frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}}+C_1^2+C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\right).$$

We denote $\Omega = \text{diag}\,(w, \Omega_{2:K})$ the matrix which attains the minimum.

Theorem 5.5.4 and Assumption 6 finally provide, with the same lower bound on the probability, the following upper bound on $\max_{i\in[p]}\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2$:

$$\left\|\Omega_{2:K}\left[\hat{R}\right]_{i.}-[R]_{i.}\right\|_2 \leq Z\left[\frac{p^{3/2}}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1^{3/2}K^{3/2}}+Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}|-Z\right)^{-2}\right]\left(2+\frac{p}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^5c_1K}\right),$$

where

$$\left(\min_{k\in[K]}[\boldsymbol{B}]_{k1}\right)^{-1}\leq\frac{p}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K}.$$

and

$$Z:=\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1}{N\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_2}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_3}{\sqrt{pN}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}+\frac{C_4}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\right),$$
$$+\frac{20\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)}{\alpha(N-1)\lambda_K(\Sigma_A)\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}\sqrt{\frac{NKh_ip}{nT}}\left(\frac{C_1\sqrt{p}}{Nc_1K}+C_5+\frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}}+C_1^2+C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}}\right).$$

■

Proof of Theorem 5.2.17.

As detailed in the proof of Theorem 5.5.6,

$$\sum_{i=1}^{p}\left\|[\hat{A}]_{i.}-[A^*]_{i.}\right\|_1\leq K\max_{i\in[p]}\left(\frac{\left\|\ [\hat{A}]_{i.}-[A^*]_{i.}\right\|_1}{h_i}\right).$$

We then use Theorem 5.2.16 to conclude. ■

5.5.15 Proof of Theorem 5.3.1

Theorem 5.5.7 (Estimation of the realizations W_j^t) *For every $t\in[T]$ and for every $j\in[n]$, for every $(\epsilon_i)_{i\in[5]}\in(\mathbb{R}_+^*)^5$ satisfying the conditions of Theorem 5.5.3, with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)$, we have $\left\|\hat{W}_j^t-\boldsymbol{W}_j^t\right\|_1$ bounded from above by :*

$$\frac{2K^{3/2}}{\left(c_2-\left\|\hat{\Phi}-\boldsymbol{\Phi}^*\right\|_{op}\right)}\left[\max(\epsilon_s)_{s\in[4]}\frac{2\sqrt{K}\gamma_3\kappa}{\left(c_2^{9/2}c_1K-\gamma_3\kappa\right)}\sqrt{\frac{p}{nT(N-2)}}\max_{i\in[p]}\left(\frac{(h_i+\xi_i)}{(c_2h_i-\xi_i)^2}h_i^{3/2}\right)+\epsilon_5\frac{\sqrt{2}}{c_2\sqrt{N}}+\frac{\left\|\hat{\Phi}-\boldsymbol{\Phi}^*\right\|_{op}}{c_2^2}\right],$$

with

$$\Theta_1 := \max_{i\in[p]} h_i^{-1}\left|[\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii}\right|, \quad \Theta_2 := \min_{\Psi\in\mathcal{D}_K}\max_{i\in[p]} h_i^{-1/2}\left\|\Psi[\hat{U}]_{i\cdot} - [U]_{i\cdot}\right\|_2,$$

$$\gamma_0 := \left(\frac{K}{c_2} + \frac{C_{VH}K^{5/2}}{c_2^2}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right), \quad \gamma_1 := c_2^{9/2}c_1^{3/2}K^{3/2}, \quad \gamma_2 := c_2^5 c_1 K, \quad \gamma_3 := \sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}},$$

$$Z := \frac{C_{tot}(p,N)}{\alpha}\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\sqrt{\frac{Kh_ip}{nT(N-2)}}, \quad \xi_i := 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}},$$

$$\Delta_1 := \max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\frac{2K\gamma_3\kappa}{\left(c_2^{9/2}c_1K-\kappa\right)}\sqrt{\frac{Kp}{nT(N-2)}}, \quad \Delta_2 := 2\epsilon_1 c_2^{-1}\sqrt{\frac{p}{c_1KNnT}},$$

$$\begin{aligned} C_{tot}(p,N) := {} & \frac{20}{\lambda_K(\Sigma_A)c_2}\left(\frac{C_1}{c_2N} + \frac{C_2}{c_2} + \frac{C_3}{c_2\sqrt{pN}} + \frac{C_4}{c_2} + C_5\right) \\ & + \frac{20}{\lambda_K(\Sigma_A)c_2}\left(\frac{C_1\sqrt{p}}{Nc_1K} + \frac{C_3}{\sqrt{c_1K}}\sqrt{\frac{p}{N}} + C_1^2 + C_1^{3/2}\frac{\sqrt{K}}{\sqrt{c_1c_2}}\right), \end{aligned}$$

$$\kappa := K^{3/2}\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\max_{i\in[p]}\left\|\tilde{\lambda}_i - \ \lambda_i\right\|_1 + K\sqrt{1+\Theta_1\sqrt{\frac{p}{c_1K}}}\Theta_2 + K^{3/2}\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\Theta_1,$$

$$\left\|\hat{\Phi} - \boldsymbol{\Phi}^*\right\|_{op} \leq \left[(1+\Delta_1)\sqrt{K}+1\right]\frac{\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)2K^{7/2}\gamma_3\kappa\sqrt{p}}{c_2(1-\Delta_2)\left(c_2^{9/2}c_1K-\kappa\right)\sqrt{nT(N-2)}} + \frac{2K^{3/2}\epsilon_1}{c_2^2(1-\Delta_2)\sqrt{NnT}}.$$

where $\alpha, C_1, C_2, C_3, C_4, C_5$ are defined in Theorem 5.5.3 and Θ_1, Θ_2 and κ are defined in Proposition 5.5.5.

Proof of Theorem 5.5.7. By definition, for all $j\in[n]$ and for all $t\in[T]$, we have

$$\tilde{W}_j^t = \left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left(\hat{A}^\top\hat{M}^{-1}\boldsymbol{Y}_j^t\right) \quad \text{and} \quad \boldsymbol{W}_j^t = \left((A^*)^\top\boldsymbol{M}_*^{-1}A^*\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right).$$

Using the triangle inequality leads to

$$\begin{aligned} \left\|\tilde{W}_j^t - \boldsymbol{W}_j^t\right\|_1 \leq {} & \left\|\left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left(\hat{A}^\top\hat{M}^{-1}\boldsymbol{Y}_j^t\right) - \left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right)\right\|_1 \\ & + \left\|\left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right) - \left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 \\ & + \left\|\left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right) - \left((A^*)^\top\boldsymbol{M}_*^{-1}A^*\right)^{-1}\cdot\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1. \end{aligned}$$

Let us recall that for any matrix $M\in\mathbb{R}^{q\times r}$, the maximum absolute column sum of M is defined as the matrix norm of M induced by the vector $\mathbb{L}_1$ norm : $\|M\|_1 := \sup_{x\neq 0}\frac{\|Mx\|_1}{\|x\|_1} = \max_{j\in[r]}\sum_{i=1}^{q}\left|[M]_{ij}\right|$. The

matrix $\mathbb{L}_1$-norm being an operator norm we derive the following inequality :

$$\begin{aligned}\left\|\tilde{W}_j^t - \boldsymbol{W}_j^t\right\|_1 \leq & \left\|\left(\hat{A}^\top \hat{M}^{-1}\hat{A}\right)^{-1}\right\|_1 \left\|\left(\hat{A}^\top \hat{M}^{-1}\boldsymbol{Y}_j^t\right) - \left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right)\right\|_1 \\ & + \left\|\left(\hat{A}^\top \hat{M}^{-1}\hat{A}\right)^{-1}\right\|_1 \left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right) - \left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 \\ & + \left\|\left(\hat{A}^\top \hat{M}^{-1}\hat{A}\right)^{-1} - \left((A^*)^\top \boldsymbol{M}_*^{-1}A^*\right)^{-1}\right\|_1 \left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 .\end{aligned}$$

Let us recall that we denote $\boldsymbol{\Phi}^* := (A^*)^\top \boldsymbol{M}_*^{-1}A^*$ and $\hat{\Phi} := \hat{A}^\top \hat{M}^{-1}\hat{A}$. We start by bounding from above $\left\|(\hat{\Phi})^{-1} - (\boldsymbol{\Phi}^*)^{-1}\right\|_1 \left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1$. The focus is firstly set on bounding from above $\left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1$. Proposition 5.2.8 ensures that for all $i \in [p]$,

$$c_2 h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i.$$

It follows that $(A^*)^\top \left(\boldsymbol{M}_*^{-1} - H^{-1}\right) A^*$ and $(A^*)^\top \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}H^{-1} - \boldsymbol{M}_*^{-1}\right) A^*$ are two positive semi-definite matrices almost surely. In addition $(A^*)^\top \boldsymbol{M}_*^{-1}A^*$, $(A^*)^\top H^{-1}A^*$ and $(A^*)^\top \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}H^{-1}A^*$ are symmetric with real entries and are thus diagonalizable. We deduce the following bounds on the spectrum of $\boldsymbol{\Phi}^*$ holding almost surely,

$$\begin{aligned}\lambda_K\left(\boldsymbol{\Phi}^*\right) &\geq \lambda_K\left((A^*)^\top H^{-1}A^*\right) = \lambda_K(\Sigma_A), \\ \lambda_1\left(\boldsymbol{\Phi}^*\right) &\leq \lambda_1\left((A^*)^\top \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}H^{-1}A^*\right) = \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1}\lambda_1(\Sigma_A).\end{aligned}$$

Lemma 5.6.8 ensures that $\lambda_1(\Sigma_A) \leq \sqrt{K}\left\|\Sigma_A\right\|_1$. Moreover,
$\left\|\Sigma_A\right\|_1 = \max_{k\in[K]} \sum_{i=1}^{p} |[\Sigma_A]_{ik}| = \max_{k\in[K]} \sum_{i=1}^{p}\sum_{l=1}^{K} [A^*]_{il}\left[H^{-1}\right]_{ii}[A^*]_{ik}$. However, for all $i \in [p]$, $\sum_{l=1}^{K}[A^*]_{il} = h_i = [H]_{ii}$. Hence $\left\|\Sigma_A\right\|_1 = \max_{k\in[K]} \sum_{i=1}^{p}[A^*]_{ik} = \max_{k\in[K]}\left\|[A^*]_{.k}\right\|_1 = 1$ and then $\lambda_1(\Sigma_A) \leq \sqrt{K}$. In addition, Assumption 7 ensures that $\lambda_K(\Sigma_A) \geq c_2$. It can then be deduced the following inequalities

$$\lambda_K\left(\boldsymbol{\Phi}^*\right) \geq c_2 \quad \text{and} \quad \lambda_1\left(\boldsymbol{\Phi}^*\right) \leq \sqrt{K}c_2^{-1}.$$

Then, from the definition of $\boldsymbol{\Pi}^{1:T}$ we deduce

$$\left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 = \left\|\left(\boldsymbol{\Phi}^*\boldsymbol{W}_j^t\right)\right\|_1 .$$

By definition of the matrix norm induced by the vector $\mathbb{L}_1$ norm :

$$\left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 \leq \left\|\boldsymbol{\Phi}^*\right\|_1 \left\|\boldsymbol{W}_j^t\right\|_1 .$$

Lemma 5.6.8 ensures that $\left\|\boldsymbol{\Phi}^*\right\|_1 \leq \sqrt{K}\lambda_1\left(\boldsymbol{\Phi}^*\right)$. For all $j \in [n]$ and for all $t \in [T]$, $\boldsymbol{W}_j^t$ is almost surely in the simplex S_{K-1}. Hence, the following inequality holds almost surely

$$\left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1 \leq Kc_2^{-1}.$$

Next, the objective is to bound from above $\left\| \left(\hat{\Phi}\right)^{-1} - (\mathbf{\Phi}^*)^{-1} \right\|_1$. First, let us expand the quantity $\left\| \hat{\Phi} - \mathbf{\Phi}^* \right\|_1$ as follows

$$\begin{aligned}
\left\| \hat{\Phi} - \mathbf{\Phi}^* \right\|_1 &= \max_{1\leq k\leq K} \left\{ \sum_{l=1}^{K} \left| \sum_{i=1}^{p} \frac{[\hat{A}]_{ik}[\hat{A}]_{il}}{[\hat{M}]_{ii}} - \frac{[A^*]_{ik}[A^*]_{il}}{[\boldsymbol{M}_*]_{ii}} \right| \right\}, \\
&\leq \max_{1\leq k\leq K} \left\{ \sum_{l=1}^{K} \sum_{i=1}^{p} \frac{[\hat{A}]_{ik} \left| [\hat{A}]_{il} - [A^*]_{il} \right|}{[\hat{M}]_{ii}} \right\} + \max_{1\leq k\leq K} \left\{ \sum_{l=1}^{K} \sum_{i=1}^{p} \frac{[A^*]_{il} \left| [\hat{A}]_{ik} - [A^*]_{ik} \right|}{[\hat{M}]_{ii}} \right\} \\
&\quad + \max_{1\leq k\leq K} \left\{ \sum_{l=1}^{K} \sum_{i=1}^{p} \frac{[A^*]_{ik}[A^*]_{il} \left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right|}{[\hat{M}]_{ii}[\boldsymbol{M}_*]_{ii}} \right\}.
\end{aligned}$$

Then, consider $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ satisfying the conditions of Theorem 5.5.3. Proposition 5.2.1 and Proposition 5.5.5 provide that, with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$, we have for all $i\in[p]$,

$$\begin{aligned}
&\left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \leq h_i \max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \frac{2K\gamma_3\kappa}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \gamma_3\kappa\right)} \sqrt{\frac{Kh_ip}{nT(N-2)}}, \\
&\left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right| < 2\epsilon_1 \sqrt{\frac{\min(2,h_i)}{NnT}},
\end{aligned}$$

where

$$\begin{aligned}
\kappa \leq & \frac{2\gamma_0 C_{tot}(p,N) \left[\frac{p^{3/2}}{\gamma_1} + Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}| - Z\right)^{-2} \right] \left(2 + \frac{p}{\gamma_2}\right)}{\alpha \left(\frac{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})}{\sqrt{K}} - KC_{VH}Z \left[\frac{p^{3/2}}{\gamma_1} + Z\left(\sqrt{h_i}\min_{k\in[K]}|[\boldsymbol{B}]_{k1}| - Z\right)^{-2} \right] \left(2 + \frac{p}{\gamma_2}\right) \right)} \\
& + \frac{C_{tot}(p,N)}{\alpha} + \frac{2\epsilon_1\sqrt{2}}{\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)\gamma_3\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})\sqrt{h_ic_1K}}.
\end{aligned}$$

Applying the reverse triangle inequality also leads, with the same upper bound on the probability, to, for all $i\in[p]$,

$$\left\| [\hat{A}]_{i.} \right\|_1 \leq \|[A^*]_{i.}\|_1 + h_i \max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \frac{2K\gamma_3\kappa}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)} \sqrt{\frac{Kh_ip}{nT(N-2)}}.$$

Hence

$$\left\| [\hat{A}]_{i.} \right\|_1 \leq h_i \left(1 + \max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \frac{2K\gamma_3\kappa}{\left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{9/2}c_1K - \kappa\right)} \sqrt{\frac{Kh_ip}{nT(N-2)}} \right).$$

Similarly, with the same upper bound on the probability, we have for all $i\in[p]$,

$$[\hat{M}]_{ii} \geq [\boldsymbol{M}_*]_{ii} - 2\epsilon_1\sqrt{\frac{h_i}{NnT}}.$$

Proposition 5.2.8 and Assumption 6 then ensure that with the same upper bound on the probability, for all $i \in [p]$,

$$\begin{aligned}
[\hat{M}]_{ii} &\geq \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) h_i - 2\epsilon_1 \sqrt{\frac{h_i}{NnT}}, \\
&\geq h_i \left(\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) - 2\epsilon_1 \sqrt{\frac{1}{h_i NnT}} \right), \\
&\geq h_i \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \left(1 - 2\epsilon_1 \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^{-1} \sqrt{\frac{p}{c_1 KNnT}} \right).
\end{aligned}$$

From these results, the following inequalities are holding true with the same upper bound on the probability :

$$\begin{aligned}
\left\| \hat{\Phi} - \boldsymbol{\Phi}^* \right\|_1 \leq & \sum_{i=1}^{p} \frac{\left\| [\hat{A}]_{i.} \right\|_1 \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{[\hat{M}]_{ii}} + \sum_{i=1}^{p} \frac{\| [A^*]_{i.} \|_1 \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{[\hat{M}]_{ii}} \\
& + \sum_{i=1}^{p} \frac{\| [A^*]_{i.} \|_1^2 \left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right|}{[\hat{M}]_{ii} [\boldsymbol{M}_*]_{ii}}, \\
\leq & \sum_{i=1}^{p} \frac{h_i^{3/2}(1+\Delta_1) \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{h_i \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})(1-\Delta_2)} + \sum_{i=1}^{p} \frac{h_i \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{h_i \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})(1-\Delta_2)} \\
& + \sum_{i=1}^{p} \frac{h_i^2 \left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right|}{h_i^2 \lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 (1-\Delta_2)}, \\
\leq & \left[(1+\Delta_1) \max_{s \in [p]} \sqrt{h_s} + 1 \right] \sum_{i=1}^{p} \frac{\left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})(1-\Delta_2)} + \sum_{i=1}^{p} \frac{\left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right|}{\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T})^2 (1-\Delta_2)}.
\end{aligned}$$

Using the bounds on $\left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1$ and $\left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right|$, holding true for all $i \in [p]$, leads to, with probability at least $1 - 2p^2 \exp(-\epsilon_1^2) - 2pK \exp(-\epsilon_2^2) - 2Kp^2 \exp(-\epsilon_3^2) - 2p \cdot (2p + 9^p) \exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4 \right) \right)$:

$$\begin{aligned}
\left\| \hat{\Phi} - \boldsymbol{\Phi}^* \right\|_1 \leq & \left[(1+\Delta_1) \max_{s \in [p]} \sqrt{h_s} + 1 \right] \sum_{i=1}^{p} \frac{h_i^{3/2} \max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) 2K\gamma_3 \kappa \sqrt{Kp}}{c_2(1-\Delta_2) \left(c_2^{9/2} c_1 K - \kappa \right) \sqrt{nT(N-2)}} \\
& + \sum_{i=1}^{p} \frac{2h_i \epsilon_1}{c_2^2 (1-\Delta_2) \sqrt{NnT}}, \\
\leq & \left[(1+\Delta_1) \max_{s \in [p]} \sqrt{h_s} + 1 \right] \frac{\max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) 2K\gamma_3 \kappa \sqrt{Kp}}{c_2(1-\Delta_2) \left(c_2^{9/2} c_1 K - \kappa \right) \sqrt{nT(N-2)}} \sum_{i=1}^{p} h_i^{3/2} \\
& + \frac{2\epsilon_1}{c_2^2 (1-\Delta_2) \sqrt{NnT}} \sum_{i=1}^{p} h_i.
\end{aligned}$$

Then, using that $\sum_{i=1}^{p} h_i = K$ leads to $\max_{s \in [p]} h_s \leq K$. Hence we deduce $\max_{s \in [p]} \sqrt{h_s} \leq \sqrt{K}$. Then, we deduce

that $\sum_{i=1}^{p} h_i^{3/2} \leq \left(\max_{s\in[p]} \sqrt{h_s}\right) \sum_{i=1}^{p} h_i \leq K^{3/2}$. It follows that with probability at least $1 - 2p^2\exp(-\epsilon_1^2) - 2pK\exp(-\epsilon_2^2) - 2Kp^2\exp(-\epsilon_3^2) - 2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right)$:

$$\left\|\hat{\Phi} - \mathbf{\Phi}^*\right\|_1 \leq \left[(1+\Delta_1)\sqrt{K}+1\right] \frac{\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) 2K^3\gamma_3\kappa\sqrt{p}}{c_2(1-\Delta_2)\left(c_2^{9/2}c_1K - \kappa\right)\sqrt{nT(N-2)}} + \frac{2K\epsilon_1}{c_2^2(1-\Delta_2)\sqrt{NnT}}.$$

Next, we notice that Lemma 5.6.8 ensures :

$$\begin{aligned}
\left\| (\hat{\Phi})^{-1} - (\mathbf{\Phi}^*)^{-1}\right\|_1 \leq& \sqrt{K}\left\| (\hat{\Phi})^{-1} - (\mathbf{\Phi}^*)^{-1}\right\|_{op}, \\
\leq& \sqrt{K}\left\| (\hat{\Phi})^{-1}\left(\hat{\Phi} - \mathbf{\Phi}^*\right)(\mathbf{\Phi}^*)^{-1}\right\|_{op}, \\
\leq& \sqrt{K}\left\| (\hat{\Phi})^{-1}\right\|_{op}\left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\left\|(\mathbf{\Phi}^*)^{-1}\right\|_{op}.
\end{aligned}$$

Next, we notice that Weyl's inequality, Lemma 1.1.13, ensures that

$$\lambda_K(\mathbf{\Phi}^*) - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op} \leq \lambda_K(\hat{\Phi}) \leq \lambda_K(\mathbf{\Phi}^*) + \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}.$$

Hence we deduce that for N, n or T sufficiently large, $\lambda_K(\mathbf{\Phi}^*) - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op} \geq 0$ and then :

$$\left\| (\hat{\Phi})^{-1}\right\|_{op} = \lambda_K(\hat{\Phi})^{-1} \leq \left(\lambda_K(\mathbf{\Phi}^*) - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\right)^{-1}.$$

Finally this provides :

$$\begin{aligned}
\left\| (\hat{\Phi})^{-1} - (\mathbf{\Phi}^*)^{-1}\right\|_1 &\leq \sqrt{K}\left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\left(\lambda_K(\mathbf{\Phi}^*) - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\right)^{-1}\left\|(\mathbf{\Phi}^*)^{-1}\right\|_{op}, \\
\left\| (\hat{\Phi})^{-1} - (\mathbf{\Phi}^*)^{-1}\right\|_1 &\leq \sqrt{K}\left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\left(\left\|(\mathbf{\Phi}^*)^{-1}\right\|_{op}^{-1} - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\right)^{-1}\left\|(\mathbf{\Phi}^*)^{-1}\right\|_{op}, \\
\left\| (\hat{\Phi})^{-1} - (\mathbf{\Phi}^*)^{-1}\right\|_1 &\leq \sqrt{K}\left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\left(c_2 - \left\|\left(\hat{\Phi} - \mathbf{\Phi}^*\right)\right\|_{op}\right)^{-1} c_2^{-1}.
\end{aligned}$$

The second step consists of bounding $\left\|\left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\right\|_1 \left\|\left((A^*)^\top \boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right) - \left((A^*)^\top \boldsymbol{M}_*^{-1}\mathbf{\Pi}_j^t\right)\right\|_1$. The DTM model, see Definition 5.1.1 ensures that for every $k \in [K]$, the variables $\left(\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t - \mathbf{\Pi}_j^t\right)\right)_{j,t}$ are real-valued and independent conditionally on $\boldsymbol{W}^{1:T}$. From the definition of the multinomial distribution, they can be expressed, conditionally on $\boldsymbol{W}^{1:T}$, for all $(k,t,j) \in [K]\times[T]\times[n]$, as,

$$\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t - \mathbf{\Pi}_j^t\right) = \frac{1}{N}\sum_{l=1}^{N}\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(Q_{jl}^t - \mathbb{E}[Q_{jl}^t]\right), \tag{5.17}$$

where for all $l \in [N]$ and for all $(t,j) \in [T]\times[n]$, $Q_{jl}^t|\boldsymbol{W}_j^t \sim$ Multinomial$_p\left(1, \boldsymbol{\Pi}_j^t\right)$ and $\mathbb{P}_{(Q_{j1}^1,\dots,Q_{jN}^1,Q_{j1}^2\dots,Q_{jN}^T)|(\boldsymbol{W}_j^1,\dots,\boldsymbol{W}_j^T)} = \bigotimes_{t=1}^{T}\bigotimes_{l=1}^{N}\mathbb{P}_{Q_{jl}^t|\boldsymbol{W}_j^t}$. Then the following equalities hold for all $(k,t,j,l) \in [K]\times[T]\times[n]\times[N]$,

$$\mathbb{E}\left[\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(Q_{jl}^t - \mathbb{E}[Q_{jl}^t]\right)|\boldsymbol{W}^{1:T}\right] = 0 \quad a.s.,$$
$$\mathbb{P}\left[\left|\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(Q_{jl}^t - \mathbb{E}[Q_{jl}^t]\right)\right| > \left\|\boldsymbol{M}_*^{-1}[A^*]_{.k}\right\|_\infty|\boldsymbol{W}^{1:T}\right] = 0 \quad a.s.$$

Then notice that for every $k \in [K]$, $\left\|\boldsymbol{M}_*^{-1}[A^*]_{.k}\right\|_\infty = \max_{i\in[p]}[(M^*)^{-1}]_{ii}[A^*]_{ik}$. Thus Proposition 5.2.8 and the definition of h_i in Assumption 6 ensure that for every $k \in [K]$:

$$\left\|\boldsymbol{M}_*^{-1}[A^*]_{.k}\right\|_\infty \le c_2^{-1}.$$

Hence applying Hoeffding's inequality, Lemma 1.1.8, for every $k \in [K]$ conditionally on $\boldsymbol{W}^{1:T}$, to $\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t(i) - \boldsymbol{\Pi}_j^t(i)\right)$ gives, for all $\epsilon > 0$,

$$\mathbb{P}\left[\left|\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t - \boldsymbol{\Pi}_j^t\right)\right| > \epsilon|\boldsymbol{W}^{1:T}\right] \le 2\exp\left(-\frac{Nc_2^2\epsilon^2}{2}\right) \quad a.s.,$$
$$\mathbb{P}\left[\left|\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t - \boldsymbol{\Pi}_j^t\right)\right| > \epsilon\right] \le 2\mathbb{E}_{\boldsymbol{W}}\left[\exp\left(-\frac{Nc_2^2\epsilon^2}{2}\right)\right].$$

We conclude that for all $k \in [K]$, for all $\epsilon > 0$, with probability $1 - 2\exp(-\epsilon^2)$,

$$\left(\boldsymbol{M}_*^{-1}[A^*]_{.k}\right)^\top\left(\boldsymbol{Y}_j^t - \boldsymbol{\Pi}_j^t\right) \le \sqrt{2}N^{-1/2}c_2^{-1}\epsilon.$$

Using a union bound provides, for all $\epsilon > 0$, with probability $1 - 2K\exp(-\epsilon^2)$,

$$\left\|(A^*)^\top\boldsymbol{M}_*^{-1}\left(\boldsymbol{Y}_j^t - \boldsymbol{\Pi}_j^t\right)\right\|_1 \le \sqrt{2}KN^{-1/2}c_2^{-1}\epsilon.$$

Then, note that Lemma 5.6.8 ensures :

$$\left\|\left(\hat{A}^\top\hat{M}^{-1}\hat{A}\right)^{-1}\right\|_1 = \left\|\hat{\Phi}^{-1}\right\|_1 \le \sqrt{K}\left\|\hat{\Phi}^{-1}\right\|_{op}.$$

Thus we deduce

$$\left\|\left(\hat{\Phi}\right)^{-1}\right\|_1 \le \sqrt{K}\left(\lambda_K(\boldsymbol{\Phi}^*) - \left\|\left(\hat{\Phi} - \boldsymbol{\Phi}^*\right)\right\|_{op}\right)^{-1}.$$

This finally allows to bound $\left\|\left(\hat{\Phi}\right)^{-1}\right\|_1\left\|\left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right) - \left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{\Pi}_j^t\right)\right\|_1$.

The third and final step consists of bounding $\left\|\left(\hat{\Phi}\right)^{-1}\right\|_1\left\|\left(\hat{A}^\top\hat{M}^{-1}\boldsymbol{Y}_j^t\right) - \left((A^*)^\top\boldsymbol{M}_*^{-1}\boldsymbol{Y}_j^t\right)\right\|_1$. The quantity $\left\|\left(\hat{\Phi}\right)^{-1}\right\|_1$ is already as detailed here above. Thus it remains to control $\left\|\left(\hat{A}^\top\hat{M}^{-1} - (A^*)^\top\boldsymbol{M}_*^{-1}\right)\boldsymbol{Y}_j^t\right\|_1$.

We get, for all $t \in [T]$ and $j \in [n]$, the following results by computation :

$$\begin{aligned}\left\| \left(\hat{A}^\top \hat{M}^{-1} - (A^*)^\top \boldsymbol{M}_*^{-1} \right) \boldsymbol{Y}_j^t \right\|_1 &= \sum_{k=1}^K \left| \sum_{i=1}^p \left[\frac{[\hat{A}]_{ik}}{[\hat{M}]_{ii}} - \frac{[A^*]_{ik}}{[\boldsymbol{M}_*]_{ii}} \right] \boldsymbol{Y}_j^t(i) \right|, \\ &\leq \left[\sum_{i=1}^p \boldsymbol{Y}_j^t(i) \right] \max_{i \in [p]} \left(\sum_{k=1}^K \left| \frac{[\hat{A}]_{ik}}{[\hat{M}]_{ii}} - \frac{[A^*]_{ik}}{[\boldsymbol{M}_*]_{ii}} \right| \right), \\ &\leq \max_{i \in [p]} \left(\sum_{k=1}^K \left| \frac{[\hat{A}]_{ik}}{[\hat{M}]_{ii}} - \frac{[A^*]_{ik}}{[\boldsymbol{M}_*]_{ii}} \right| \right),\end{aligned}$$

where we use that the columns of $\boldsymbol{Y}^{1:T}$ are $\mathbb{L}_1$ normalized by definition, providing $\sum_{i=1}^p \boldsymbol{Y}_j^t(i) = 1$. In addition, Proposition 5.2.1 ensures that for all $i \in [p]$, for all $\epsilon_1 > 0$ with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$, we have

$$\left| [\hat{M}]_{ii} - [\boldsymbol{M}_*]_{ii} \right| < 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}.$$

Moreover Proposition 5.2.8 gives that almost surely for all $i \in [p]$,

$$c_2 h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \leq c_2 h_i \leq [\boldsymbol{M}_*]_{ii} \leq h_i \leq h_i + 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}.$$

Hence we obtain that with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$ we have, for all $i \in [p]$,

$$c_2 h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} < [\hat{M}]_{ii} < h_i + 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}}.$$

Thus with probability at least $1 - 2\exp\left(-\epsilon_1^2\right)$:

$$\begin{aligned}\left\| \left(\hat{A}^\top \hat{M}^{-1} - (A^*)^\top \boldsymbol{M}_*^{-1} \right) \boldsymbol{Y}_j^t \right\|_1 &\leq \max_{i \in [p]} \left(\sum_{k=1}^K \left| \frac{[\boldsymbol{M}_*]_{ii} [\hat{A}]_{ik} - [\hat{M}]_{ii} [A^*]_{ik}}{[\boldsymbol{M}_*]_{ii} [\hat{M}]_{ii}} \right| \right), \\ &\leq \max_{i \in [p]} \left(\left(c_2 h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-2} \left(h_i + 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right) \sum_{k=1}^K \left| [\hat{A}]_{ik} - [A^*]_{ik} \right| \right), \\ &\leq \max_{i \in [p]} \left(\left(c_2 h_i - 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right)^{-2} \left(h_i + 2\epsilon_1 \sqrt{\frac{\min(2, h_i)}{NnT}} \right) \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \right).\end{aligned}$$

Finally, we recall that for all $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$, with probability at least $1 - 2p^2 \exp(-\epsilon_1^2) - 2pK \exp(-\epsilon_2^2) - 2Kp^2 \exp(-\epsilon_3^2) - 2p \cdot (2p + 9^p) \exp\left(- \min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4 \right) \right)$, we have for all $i \in [p]$,

$$\left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \leq h_i \max(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \frac{2K\gamma_3\kappa}{\left(c_2^{9/2} c_1 K - \kappa \right)} \sqrt{\frac{Kh_i p}{nT(N-2)}}.$$

Hence, with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)$, we have $\left\|\left(\hat{A}^\top\hat{M}^{-1}-(A^*)^\top\boldsymbol{M}_*^{-1}\right)\boldsymbol{Y}_j^t\right\|_1$ bounded from above by

$$\max(\epsilon_s)_{s\in[4]}\frac{2K^{3/2}\gamma_3\kappa}{\left(c_2^{9/2}c_1K-\kappa\right)}\sqrt{\frac{p}{nT(N-2)}}\max_{i\in[p]}\left(\frac{\left(h_i+2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)}{\left(c_2h_i-2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}\right)^2}h_i^{3/2}\right).$$

Finally, for all $j\in[n]$ and $t\in[T]$, for every $(\epsilon_i)_{i\in[5]}\in(\mathbb{R}_+^*)^5$, with probability at least $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)$, we have $\left\|\tilde{W}_j^t-W_j^t\right\|_1$ bounded from above by :

$$\frac{K^{3/2}}{\left(\lambda_K(\boldsymbol{\Phi}^*)-\left\|\left(\hat{\Phi}-\boldsymbol{\Phi}^*\right)\right\|_{op}\right)}\left[\max(\epsilon_s)_{s\in[4]}\frac{2\sqrt{K}\gamma_3\kappa}{\left(c_2^{9/2}c_1K-\kappa\right)}\sqrt{\frac{p}{nT(N-2)}}\max_{i\in[p]}\left(\frac{(h_i+\xi_i)}{(c_2h_i-\xi_i)^2}h_i^{3/2}\right)+\epsilon_5\frac{\sqrt{2}}{\sqrt{N}c_2}\right]$$
$$+\frac{K^{3/2}\left\|\left(\hat{\Phi}-\boldsymbol{\Phi}^*\right)\right\|_{op}}{c_2^2\left(c_2-\left\|\left(\hat{\Phi}-\boldsymbol{\Phi}^*\right)\right\|_{op}\right)},$$

where $\xi_i:=2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}}$, $\lambda_K(\boldsymbol{\Phi}^*)\geq c_2$ and

$$\left\|\hat{\Phi}-\boldsymbol{\Phi}^*\right\|_{op}\leq\left[(1+\Delta_1)\sqrt{K}+1\right]\frac{\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4)2K^{7/2}\gamma_3\kappa\sqrt{p}}{c_2(1-\Delta_2)\left(c_2^{9/2}c_1K-\kappa\right)\sqrt{nT(N-2)}}+\frac{2K^{3/2}\epsilon_1}{c_2^2(1-\Delta_2)\sqrt{NnT}}.$$

The final part of the proof is to bound $\left\|\hat{W}_j^t-W_j^t\right\|_1$. By definition, $\hat{W}_j^t$ is defined by setting negative entries of $\tilde{W}_j^t$ to zero and normalizing it to have a unit $\mathbb{L}_1$ norm. We start by defining $\check{W}_j^t$ the vector obtained by setting the negative entries of $\tilde{W}_j^t$ to zero. Then $\hat{W}_j^t=\frac{\check{W}_j^t}{\|\check{W}_j^t\|_1}$. Then using the triangle inequality and the definition of $\hat{W}_j^t$ we deduce the following :

$$\begin{aligned}\left\|W_j^t-\hat{W}_j^t\right\|_1&\leq\left\|W_j^t-\check{W}_j^t\right\|_1+\left\|\check{W}_j^t-\hat{W}_j^t\right\|_1,\\&\leq\left\|W_j^t-\check{W}_j^t\right\|_1+\left\|\check{W}_j^t\right\|_1\left|1-\frac{1}{\|\check{W}_j^t\|_1}\right|_1.\end{aligned}$$

We recall that $\|W_j^t\|_1=1$ and we use the reverse triangle inequality to get :

$$\begin{aligned}\left\|W_j^t-\hat{W}_j^t\right\|_1&\leq\left\|W_j^t-\check{W}_j^t\right\|_1+\left|\|\check{W}_j^t\|_1-1\right|_1,\\&\leq\left\|W_j^t-\check{W}_j^t\right\|_1+\left|\|\check{W}_j^t\|_1-\|W_j^t\|_1\right|_1,\\&\leq2\left\|W_j^t-\check{W}_j^t\right\|_1.\end{aligned}$$

Finally, notice that both W_j^t and $\check{W}_j^t$ have non negative entries. Thus we have :

$$\| \check{W}_j^t - W_j^t \|_1 \leq \left\| \tilde{W}_j^t - W_j^t \right\|_1 .$$

This finally leads to

$$\left\| \hat{W}_j^t - W_j^t \right\|_1 \leq 2 \left\| \tilde{W}_j^t - W_j^t \right\|_1 .$$

■

Proof of Theorem 5.3.1. As detailed in the proof of Theorem 5.2.16, under the stated conditions on NnT, with probability at least
$1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)$, we have $\gamma_3 \leq 2$ and

$$\kappa \leq 2KC_{tot}(p,N)\max(\epsilon_i)_{i\in[4]}\sqrt{\frac{p}{nT(N-2)}}\left[1+\left(2+\frac{p}{c_2^5c_1K}\right)\left(\frac{8pK^{1/2}}{c_1c_2^{13/2}}+\frac{8pK^2}{c_1c_2^{15/2}}\max_{x\in\mathcal{G}_\eta}\|x\|_2\right)\right]+\frac{4K^{3/2}\epsilon_1}{c_2\sqrt{NnT}}$$

Under the stated conditions on NnT we also have $c_2^{9/2}c_1K-\kappa \geq c_2^{9/2}c_1K/2$, as detailed in the proof of Theorem 5.2.16. Moreover, $NnT \geq \frac{64p}{c_2^4c_1K} \geq \frac{16p}{c_1Kc_2^2}$ ensures $(1-\Delta_2)^{-1} \leq 2$. Next, consider the constants $C_A(p,N)$ and C_B defined in Theorem 5.2.16 and notice that the previously stated bound ensures $\frac{\kappa}{c_2^{9/2}c_1} \leq \frac{K\max(\epsilon_i)_{i\in[4]}}{\sqrt{nT(N-2)}}\left[C_A(p,N)\sqrt{p}+C_B\right]$.
Thus if $nT(N-2) \geq 8K^{3/2}\max(\epsilon_i^2)_{i\in[4]}\left[C_A(p,N)\sqrt{p}+C_B\right]\sqrt{p}$ we have $\Delta_1 \leq 1$. This leads to, under these conditions,

$$\left\|\hat{\Phi}-\mathbf{\Phi}^*\right\|_{op} \leq 16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s)_{s\in[4]}K^{5/2}}{c_2}\frac{\kappa}{c_2^{9/2}c_1}\sqrt{\frac{p}{nT(N-2)}}+\frac{4K^{3/2}\epsilon_1}{c_2^2\sqrt{NnT}}.$$

Thus we get

$$\left\|\hat{\Phi}-\mathbf{\Phi}^*\right\|_{op} \leq 16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}}{c_2nT(N-2)}\left[C_A(p,N)\sqrt{p}+C_B\right]+\frac{4K^{3/2}\epsilon_1}{c_2^2\sqrt{NnT}}.$$

We look for a condition on NnT ensuring $\left\|\hat{\Phi}-\mathbf{\Phi}^*\right\|_{op} \leq c_2/2$. Let us denote $X := \sqrt{nT(N-2)}$. Then $\left\|\hat{\Phi}-\mathbf{\Phi}^*\right\|_{op} \leq c_2/2$ is ensured if

$$16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}}{c_2X^2}\left[C_A(p,N)\sqrt{p}+C_B\right]+\frac{4K^{3/2}\epsilon_1}{c_2^2X} \leq c_2/2.$$

Thus we get a second degree polynomial inequality. The condition is then ensured if

$$\sqrt{nT(N-2)} \geq \frac{4K^{3/2}\epsilon_1}{c_2^2}+\frac{\sqrt{16K^3\epsilon_1^2/c_2^4+32\left[2\sqrt{K}+1\right]\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\left[C_A(p,N)\sqrt{p}+C_B\right]\sqrt{p}}}{c_2}.$$

Under this condition we get with the stated probability that $\left\|\hat{W}_j^t - \boldsymbol{W}_j^t\right\|_1$ is bounded from above by :

$$\frac{4K^{3/2}}{c_2}\left[\max(\epsilon_s)_{s\in[4]}\frac{4\sqrt{K}\gamma_3\kappa}{c_2^{9/2}c_1K}\sqrt{\frac{p}{nT(N-2)}}\max_{i\in[p]}\left(\frac{(h_i+\xi_i)}{(c_2h_i-\xi_i)^2}h_i^{3/2}\right)+\epsilon_5\frac{\sqrt{2}}{\sqrt{N}c_2}\right]$$
$$+\frac{4K^{3/2}}{c_2^3}\left[16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}}{c_2nT(N-2)}\left[C_A(p,N)\sqrt{p}+C_B\right]+\frac{4K^{3/2}\epsilon_1}{c_2^2\sqrt{NnT}}\right].$$

Next notice that for all $i\in[p]$, $\xi_i := 2\epsilon_1\sqrt{\frac{\min(2,h_i)}{NnT}} \leq 2\epsilon_1\sqrt{\frac{2}{NnT}}$. Hence $NnT \geq \frac{32\epsilon_1^2}{c_2^2h_{\min}^2}$ ensures for all $i\in[p]$ $\xi_i \leq \frac{c_2h_{\min}}{2} \leq \frac{c_2h_i}{2}$. Thus under this condition and using that for all $i\in[p]$, $h_i \leq K$ leads to

$$\max_{i\in[p]}\left(\frac{(h_i+\xi_i)}{(c_2h_i-\xi_i)^2}h_i^{3/2}\right) \leq K^{3/2}\frac{2+c_2}{c_2}.$$

Finally, under the stated conditions,

$$\begin{aligned}
\left\|\hat{W}_j^t - \boldsymbol{W}_j^t\right\|_1 \leq& \frac{4K^{3/2}}{c_2}\left[\max(\epsilon_s)_{s\in[4]}\frac{4\sqrt{K}\gamma_3\kappa}{c_2^{9/2}c_1K}\sqrt{\frac{p}{nT(N-2)}}K^{3/2}\frac{2+c_2}{c_2}+\epsilon_5\frac{\sqrt{2}}{\sqrt{N}c_2}\right]\\
&+\frac{4K^{3/2}}{c_2^3}\left[16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}}{c_2nT(N-2)}\left[C_A(p,N)\sqrt{p}+C_B\right]+\frac{4K^{3/2}\epsilon_1}{c_2^2\sqrt{NnT}}\right],\\
\leq& \frac{4K^{3/2}}{c_2}\left[\max(\epsilon_s^2)_{s\in[4]}\frac{8K^2\sqrt{p}}{nT(N-2)}\left[C_A(p,N)\sqrt{p}+C_B\right]\frac{2+c_2}{c_2}+\epsilon_5\frac{\sqrt{2}}{\sqrt{N}c_2}\right]\\
&+\frac{4K^{3/2}}{c_2^3}\left[16\left[2\sqrt{K}+1\right]\frac{\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}}{c_2nT(N-2)}\left[C_A(p,N)\sqrt{p}+C_B\right]+\frac{4K^{3/2}\epsilon_1}{c_2^2\sqrt{NnT}}\right],\\
\leq& \epsilon_5\frac{4\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}}+\epsilon_1\frac{16K^3}{c_2^5\sqrt{NnT}}\\
&+\frac{32\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right].
\end{aligned}$$

■

5.5.16 Proof of Theorem 5.4.2

Proof of Theorem 5.4.2. On the one hand, we need to bound from above the quantity $\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\langle \hat{W}_j^{t+1} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - \overline{\hat{W}}\right\rangle$ First, we notice that for all $j\in[n]$, for all $t\in[T-1]$, we have :

$$\left\langle \hat{W}_j^{t+1} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - \overline{\hat{W}}\right\rangle = \left\langle \hat{W}_j^{t+1} - W_j^{t+1} + W_j^{t+1} - \overline{W^{+1}} + \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - W_j^t + W_j^t - \overline{W} + \overline{W} - \overline{\hat{W}}\right\rangle.$$

Using the bilinearity of the scalar product, we get that $\left\langle \hat{W}_j^{t+1} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - \overline{\hat{W}} \right\rangle$ is equal to

$$\begin{aligned}
&\langle \hat{W}_j^{t+1} - W_j^{t+1};\ \hat{W}_j^t - W_j^t \rangle + \langle W_j^{t+1} - \overline{W^{+1}};\ \hat{W}_j^t - W_j^t \rangle + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - W_j^t \rangle \\
&+ \langle \hat{W}_j^{t+1} - W_j^{t+1};\ W_j^t - \overline{W} \rangle + \left| \left\langle W_j^{t+1} - \overline{W^{+1}};\ W_j^t - \overline{W} \right\rangle \right| + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ W_j^t - \overline{W} \rangle \\
&+ \langle \hat{W}_j^{t+1} - W_j^{t+1};\ \overline{W} - \overline{\hat{W}} \rangle + \langle W_j^{t+1} - \overline{W^{+1}};\ \overline{W} - \overline{\hat{W}} \rangle + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ \overline{W} - \overline{\hat{W}} \rangle.
\end{aligned}$$

In addition, the proof of Theorem 4.4.2 contains the following equality :

$$\left\langle W_j^{t+1} - \overline{W}^{+1};\ W_j^t - \overline{W} \right\rangle = (1 - c^*) \left\| W_j^t - \overline{W} \right\|_2^2 + c^* \left\langle \Delta_j^t - \overline{\Delta};\ W_j^t - \overline{W} \right\rangle.$$

Moreover we notice that for all $j \in [n]$, for all $t \in [T-1]$,

$$\begin{aligned}
\left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 = &\left\| \hat{W}_j^t - W_j^t \right\|_2^2 + \left\| W_j^t - \overline{W} \right\|_2^2 + \left\| \overline{W} - \overline{\hat{W}} \right\|_2^2 \\
&+ 2 \left\langle \hat{W}_j^t - W_j^t; W_j^t - \overline{W} \right\rangle + 2 \left\langle \hat{W}_j^t - W_j^t; \overline{W} - \overline{\hat{W}} \right\rangle + 2 \left\langle W_j^t - \overline{W}; \overline{W} - \overline{\hat{W}} \right\rangle.
\end{aligned}$$

Hence $\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 \left[\widehat{(1-c)} - (1 - c^*) \right]$ is equal to

$$\begin{aligned}
&\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left[\langle \hat{W}_j^{t+1} - W_j^{t+1};\ \hat{W}_j^t - W_j^t \rangle + \langle W_j^{t+1} - \overline{W^{+1}};\ \hat{W}_j^t - W_j^t \rangle + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ \hat{W}_j^t - W_j^t \rangle \right] \\
&+ \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left[\langle \hat{W}_j^{t+1} - W_j^{t+1};\ W_j^t - \overline{W} \rangle + c^* \left\langle \boldsymbol{\Delta}_j^t - \overline{\boldsymbol{\Delta}};\ W_j^t - \overline{W} \right\rangle + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ W_j^t - \overline{W} \rangle \right] \\
&+ \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left[\langle \hat{W}_j^{t+1} - W_j^{t+1};\ \overline{W} - \overline{\hat{W}} \rangle + \langle W_j^{t+1} - \overline{W^{+1}};\ \overline{W} - \overline{\hat{W}} \rangle + \langle \overline{W^{+1}} - \overline{\hat{W}^{+1}};\ \overline{W} - \overline{\hat{W}} \rangle \right] \\
&- \frac{1 - c^*}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left[\left\| \hat{W}_j^t - W_j^t \right\|_2^2 + \left\| \overline{W} - \overline{\hat{W}} \right\|_2^2 + 2 \left\langle \hat{W}_j^t - W_j^t; W_j^t - \overline{W} \right\rangle + 2 \left\langle \hat{W}_j^t - W_j^t; \overline{W} - \overline{\hat{W}} \right\rangle \right] \\
&- \frac{1 - c^*}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} 2 \left[\left\langle W_j^t - \overline{W}; \overline{W} - \overline{\hat{W}} \right\rangle \right].
\end{aligned}$$

Cauchy-Schwarz inequality ensures that $\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 \left| \widehat{(1-c)} - (1 - c^*) \right|$ is boun-

ded from above by

$$\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left[\left\|\hat{W}_j^{t+1}-W_j^{t+1}\right\|_2\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2+\left\|W_j^{t+1}-\overline{W^{+1}}\right\|_2\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2+\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2\right]$$
$$+\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left[\left\|\hat{W}_j^{t+1}-W_j^{t+1}\right\|_2\left\|W_j^{t}-\overline{W}\right\|_2+c^*\left|\langle\boldsymbol{\Delta}_j^t-\overline{\boldsymbol{\Delta}};\ W_j^t-\overline{W}\rangle\right|+\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\left\|W_j^{t}-\overline{W}\right\|_2\right]$$
$$+\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left[\left\|\hat{W}_j^{t+1}-W_j^{t+1}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2+\left\|W_j^{t+1}-\overline{W^{+1}}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2+\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2\right]$$
$$+\frac{2(1-c^*)}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left[\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2\left\|W_j^{t}-\overline{W}\right\|_2+\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2+\left\|W_j^{t}-\overline{W}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2\right].$$

Theorem 5.3.1 combined with the triangle inequality and the $\mathbb{L}_1$-$\mathbb{L}_2$ inequality provide upper bounds, converging towards zero, on the quantities $\left\|\hat{W}_j^{t+1}-W_j^{t+1}\right\|_2$, $\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2$, $\left\|\overline{W}-\overline{\hat{W}}\right\|_2$ and $\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2$ with high probability. In addition, notice that the quantities $\left\|W_j^{t}-\overline{W}\right\|_2$ and $\left\|W_j^{t+1}-\overline{W^{+1}}\right\|_2$ are bounded from above by one almost surely. Let us consider $(\epsilon_i)_{i\in[6]}\in(\mathbb{R}_+^*)^6$ satisfying the conditions of Theorem 5.3.1 and assume the conditions on the sample size satisfied. Then using a union bound we get that

$$\frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left[\left\|\hat{W}_j^{t+1}-W_j^{t+1}\right\|_2\left(\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2+\left\|W_j^{t}-\overline{W}\right\|_2+\left\|\overline{W}-\overline{\hat{W}}\right\|_2\right)\right]$$
$$+\frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left[\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\left\|W_j^{t}-\overline{W}\right\|_2\right]$$
$$+\frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left[\left\|\overline{W}-\overline{\hat{W}}\right\|_2\left(\left\|W_j^{t+1}-\overline{W^{+1}}\right\|_2+\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\right)\right]$$
$$+\frac{1}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left[\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2\left(\left\|W_j^{t+1}-\overline{W^{+1}}\right\|_2+\left\|\overline{W^{+1}}-\overline{\hat{W}^{+1}}\right\|_2\right)\right]$$
$$+\frac{2(1-c^*)}{n(T-1)}\sum_{j=1}^{n}\sum_{t=1}^{T-1}\left[\left\|\hat{W}_j^{t}-W_j^{t}\right\|_2\left(\left\|W_j^{t}-\overline{W}\right\|_2+\left\|\overline{W}-\overline{\hat{W}}\right\|_2\right)+\left\|W_j^{t}-\overline{W}\right\|_2\left\|\overline{W}-\overline{\hat{W}}\right\|_2\right].$$

is bounded from above by

$$2\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+\frac{\nu_2\epsilon_5}{\sqrt{N}}+\frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)$$
$$\cdot\left((3-c^*)\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+(3-c^*)\frac{\nu_2\epsilon_5}{\sqrt{N}}+(3-c^*)\frac{\nu_3\epsilon_1}{\sqrt{NnT}}+2(2-c^*)\right)$$

with probability larger than $1-2n(T-1)p^2\exp(-\epsilon_1^2)-2n(T-1)pK\exp(-\epsilon_2^2)-2n(T-1)Kp^2\exp(-\epsilon_3^2)-2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2n(T-1)K\exp(-\epsilon_5^2)-2n(T-1)\exp\left(-\epsilon_6^2/4\right)$. In

addition, $c^* \left| \left\langle \boldsymbol{\Delta}_j^t - \overline{\boldsymbol{\Delta}};\ W_j^t - \overline{W} \right\rangle \right|$ is controlled in the proof of Theorem 4.4.2. Using (4.21) and (4.23) we get that for n and T satisfying (4.30), for all $0 < \epsilon_7 < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:

$$\frac{1}{n(T-1)} \left| \sum_{j=1}^{n} \sum_{t=1}^{T-1} \langle \boldsymbol{\Delta}_j^t - \overline{\boldsymbol{\Delta}};\ W_j^t - \overline{W} \rangle \right| \leq \left[\frac{(\epsilon_7+1)^2}{n(T-1)} \left(1 + \frac{1}{\underline{c}\sqrt{T-1}} \right) + \frac{11\epsilon_7}{\sqrt{n(T-1)}} \right],$$

with probability larger than $1 - 7\exp(-\epsilon_7^2/4)$. Finally, under the stated conditions, the quantity $\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 \left| \widehat{(1-c)} - (1-c^*) \right|$ is bounded from above by

$$\begin{aligned}
&2\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}} \right) \\
&\cdot \left((3-c^*)\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + (3-c^*)\frac{\nu_2\epsilon_5}{\sqrt{N}} + (3-c^*)\frac{\nu_3\epsilon_1}{\sqrt{NnT}} + 2(2-c^*) \right) \\
&+c^* \left[\frac{(\epsilon_7+1)^2}{n(T-1)} \left(1 + \frac{1}{\underline{c}\sqrt{T-1}} \right) + \frac{11\epsilon_7}{\sqrt{n(T-1)}} \right],
\end{aligned}$$

with probability larger than $1 - 2n(T-1)p^2\exp(-\epsilon_1^2) - 2n(T-1)pK\exp(-\epsilon_2^2) - 2n(T-1)Kp^2\exp(-\epsilon_3^2) - 2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right) - 2n(T-1)K\exp(-\epsilon_5^2) - 2n(T-1)\exp\left(-\epsilon_6^2/4\right) - 7\exp(-\epsilon_7^2/4)$.

On the other hand, we need to bound from below the quantity $\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2$. We recall that for all $j \in [n]$, for all $t \in [T-1]$:

$$\left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 \geq \left\| W_j^t - \overline{W} \right\|_2^2 - 2\left\| \hat{W}_j^t - W_j^t \right\|_2 \left(\left\| W_j^t - \overline{W} \right\|_2 + \left\| \overline{W} - \overline{\hat{W}} \right\|_2 \right) - 2\left\| W_j^t - \overline{W} \right\|_2 \left\| \overline{W} - \overline{\hat{W}} \right\|_2.$$

Using Theorem 5.3.1 and the $(\epsilon_i)_{i\in[7]}$ previously introduced we get that, under the stated conditions,

$$\begin{aligned}
\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| \hat{W}_j^t - \overline{\hat{W}} \right\|_2^2 \geq & \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| W_j^t - \overline{W} \right\|_2^2 \\
& - 2\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}} \right) \\
& \cdot \left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}} + 2 \right),
\end{aligned}$$

with probability larger than $1 - 2n(T-1)p^2\exp(-\epsilon_1^2) - 2n(T-1)pK\exp(-\epsilon_2^2) - 2n(T-1)Kp^2\exp(-\epsilon_3^2) - 2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right) - 2n(T-1)K\exp(-\epsilon_5^2) - 2n(T-1)\exp\left(-\epsilon_6^2/4\right)$. Moreover, the proof of Theorem 4.4.2 contains the following inequality holding true for all $0 < \epsilon_7 < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$:

$$\frac{1}{n(T-1)} \sum_{j=1}^{n} \sum_{t=1}^{T-1} \left\| W_j^t - \overline{W} \right\|_2^2 \geq m\frac{\underline{c}}{2-\underline{c}} - \frac{\left(1+\sqrt{2}\right)\epsilon_7 + 1}{\underline{c}\sqrt{n}(T-1)} - \frac{\left(21+4\sqrt{2}\right)\epsilon_7 + 1}{\sqrt{n(T-1)}} \geq \frac{\underline{c}m}{4},$$

for n and T large enough, see the proof of Theorem 4.4.2, with probability larger than $1 - 6\exp(-\epsilon_7^2/4)$. Thus we get that for n and T large enough, with probability larger than $1-2n(T-1)p^2\exp(-\epsilon_1^2)-2n(T-1)pK\exp(-\epsilon_2^2) - 2n(T-1)Kp^2\exp(-\epsilon_3^2) - 2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right) - 2n(T-1)K\exp(-\epsilon_5^2) - 2n(T-1)\exp\left(-\epsilon_6^2/4\right) - 6\exp(-\epsilon_7^2/4)$:

$$\begin{aligned}\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|\hat{W}_j^t - \overline{\hat{W}}\right\|_2^2 \geq & \frac{\underline{c}m}{4} \\ & - 2\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right) \\ & \cdot\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}} + 2\right).\end{aligned}$$

We then combine both results and assume that N, n and T are large enough to ensure

$$2\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}} + 2\right) \leq \frac{\underline{c}m}{8}.$$

Finally, we also assume that

$$(3-c)\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)^2 \leq 2(2-c)\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right).$$

It is then sufficient to state that

$$\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2\epsilon_5}{\sqrt{N}} + \frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right) \leq 2.$$

This concludes. ■

5.5.17 Proof of Theorem 5.4.3

Proof of Theorem 5.4.3. Following the proof of Theorem 4.4.3, we get :

$$\begin{aligned}|\hat{\alpha} - \alpha^*| \leq & \left|\frac{\hat{c}}{2-\hat{c}}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}} - \frac{c^*}{2-c^*}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}}\right| \\ & + \left|\frac{c^*}{2-c^*}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}} - \frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathcal{V}}\right| \\ & + \left|\frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathcal{V}} - \frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))}\right|.\end{aligned}$$

Then we bound from above the three following quantities :

$$Q_1 := \left|\frac{\hat{c}}{2-\hat{c}} - \frac{c^*}{2-c^*}\right|, \quad Q_2 := \left|\|\hat{\theta}\|_2^2 - \|\tilde{\theta}^*\|_2^2\right| \text{ and } Q_3 := \left|\mathcal{V} - \mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))\right|.$$

We first bound from above Q_1 :

$$\left|\frac{\hat{c}}{2-\hat{c}}-\frac{c^*}{2-c^*}\right| = \left|\frac{1}{2-\hat{c}}\left(\hat{c}-c^*\right)+c^*\left(\frac{1}{2-\hat{c}}-\frac{1}{2-c^*}\right)\right| \leq (1+c^*)\cdot|\hat{c}-c^*|.$$

Using Theorem 5.4.2, we have, for every $(\epsilon_i)_{i\in[7]} \in (\mathbb{R}_+^*)^7$ satisfying $\max(\epsilon_6,\epsilon_7) < \sqrt{nm\frac{\underline{c}}{2-\underline{c}}}/2$ and

$$\max(\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4) \leq \sqrt{nT}\frac{c_2^3\min\left(c_3,c_2^2\right)}{12\sqrt{K}\left(\frac{2\sqrt{p}}{N}+2K\sqrt{p}+\frac{576e}{\log(2)\sqrt{Nc}}+\frac{4K^2}{c_2}\sqrt{p}\right)}:$$

$$\begin{aligned}\left|\frac{\hat{c}}{2-\hat{c}}-\frac{c^*}{2-c^*}\right| \leq & \frac{64(1-c^*)^2}{\underline{c}m}\left(\frac{\nu_1(p,N)\max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)}+\frac{\nu_2\epsilon_5}{\sqrt{N}}+\frac{\nu_3\epsilon_1}{\sqrt{NnT}}\right)\\ & +\frac{8c^*(1-c^*)}{\underline{c}m}\left[\frac{(\epsilon_7+1)^2}{n(T-1)}\left(1+\frac{1}{\underline{c}\sqrt{T-1}}\right)+\frac{11\epsilon_7}{\sqrt{n(T-1)}}\right],\end{aligned}$$

with probability larger than $1-2n(T-1)p^2\exp(-\epsilon_1^2)-2n(T-1)pK\exp(-\epsilon_2^2)-2n(T-1)Kp^2\exp(-\epsilon_3^2)-2n(T-1)p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2n(T-1)K\exp(-\epsilon_5^2)-2n(T-1)\exp\left(-\epsilon_6^2/4\right)-13\exp(-\epsilon_7^2/4)$. We next bound from above Q_2 :

$$\left|\|\hat{\theta}\|_2^2-\|\tilde{\theta}^*\|_2^2\right| = \left|\langle\hat{\theta}-\tilde{\theta}^*;\hat{\theta}+\tilde{\theta}^*\rangle\right| \leq \left\|\hat{\theta}+\tilde{\theta}^*\right\|_2\cdot\left\|\hat{\theta}-\tilde{\theta}^*\right\|_2 \leq 2\left\|\hat{\theta}-\tilde{\theta}^*\right\|_2.$$

Using Theorem 5.4.1 we get :

$$\begin{aligned}\left|\|\hat{\theta}\|_2^2-\|\tilde{\theta}^*\|_2^2\right| \leq & \frac{2(\epsilon_6+1)}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right)+\epsilon_5\frac{8\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}}+\epsilon_1\frac{32K^3}{c_2^5\sqrt{NnT}}\\ & +\frac{64\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right],\end{aligned}$$

with probability larger than
$1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)-2\exp\left(-\epsilon_6^2/4\right)$.

Recalling $\overline{\hat{W}} := \hat{\theta}$, we then bound from above Q_3

$$\begin{aligned}|\mathcal{V}-\mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))| &= \frac{1}{n(T-1)}\left|\sum_{jt}\left\|\hat{W}_j^t-\overline{\hat{W}}\right\|_2^2-\sum_{jt}\left\|\boldsymbol{W}_j^t-\tilde{\theta}^*\right\|_2^2\right|,\\ &= \frac{1}{n(T-1)}\left|2\sum_{jt}\left\langle\hat{W}_j^t;\tilde{\theta}^*-\overline{\hat{W}}\right\rangle+n(T-1)\left(\|\overline{\boldsymbol{W}}\|_2^2-\left\|\tilde{\theta}^*\right\|_2^2\right)\right|,\\ &\leq 2\left\|\overline{\hat{W}}\right\|_2\cdot\left\|\tilde{\theta}^*-\overline{\hat{W}}\right\|_2+\left|\|\overline{\boldsymbol{W}}\|_2^2-\left\|\tilde{\theta}^*\right\|_2^2\right|,\\ &\leq 2\|\overline{\hat{W}}\|_2\cdot\left\|\tilde{\theta}^*-\overline{\hat{W}}\right\|_2+\|\overline{\boldsymbol{W}}-\tilde{\theta}^*\|_2\cdot\left(\|\overline{\boldsymbol{W}}\|_2+\|\tilde{\theta}^*\|_2\right),\\ &\leq 2\left\|\tilde{\theta}^*-\overline{\hat{W}}\right\|_2+2\|\overline{\boldsymbol{W}}-\tilde{\theta}^*\|_2.\end{aligned}$$

Notice that the quantity $\|\overline{\boldsymbol{W}} - \tilde{\theta}^*\|_2$ is bounded by Theorem 4.4.1 and the quantity $\left\|\tilde{\theta}^* - \overline{\hat{W}}\right\|_2$ is bounded by Theorem 5.4.1. This leads to

$$\begin{aligned}\left|\mathcal{V} - \mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))\right| \leq & \frac{4(\epsilon_6+1)}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right) + \epsilon_5\frac{8\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}} + \epsilon_1\frac{32K^3}{c_2^5\sqrt{NnT}} \\ & + \frac{64\max(\epsilon_s^2)_{s\in[4]}K^{7/2}\sqrt{p}\left[C_A(p,N)\sqrt{p}+C_B\right]}{c_2^2 nT(N-2)}\left[(2+c_2)+\frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2}\right],\end{aligned}$$

with probability larger than $1-2p^2\exp(-\epsilon_1^2)-2pK\exp(-\epsilon_2^2)-2Kp^2\exp(-\epsilon_3^2)-2p\cdot(2p+9^p)\exp\left(-\min\left(\epsilon_4^2;\sqrt{cnT}\epsilon_4\right)\right)-2K\exp(-\epsilon_5^2)-2\exp\left(-\epsilon_6^2/4\right)$. Next, we remind that (4.29) ensures the following :

$$\begin{aligned}\mathcal{V} &\geq m\frac{\underline{c}}{2-\underline{c}} - \frac{\epsilon_7+1}{\sqrt{n(T-1)}}\left(\frac{1}{\underline{c}\sqrt{T-1}}+1\right) - \frac{\sqrt{2}\epsilon_7}{\underline{c}\sqrt{n}(T-1)} - \frac{\left(20+4\sqrt{2}\right)\epsilon_7}{\sqrt{n(T-1)}}, \\ &\geq m\frac{\underline{c}}{2-\underline{c}} - \frac{\left(1+\sqrt{2}\right)\epsilon_7+1}{\underline{c}\sqrt{n}(T-1)} - \frac{\left(21+4\sqrt{2}\right)\epsilon_7+1}{\sqrt{n(T-1)}} \geq \frac{\underline{c}m}{4},\end{aligned}$$

for n and T large enough, with probability larger than $1-6\exp(-\epsilon_7^2/4)$. Large enough means we need

$$\frac{2\left(1+\sqrt{2}\right)\epsilon_7+2}{m\underline{c}^2\sqrt{T-1}}(2-\underline{c}) + \frac{2(21+4\sqrt{2})\epsilon_7+2}{m\underline{c}}(2-\underline{c}) \leq \sqrt{n(T-1)}.$$

Finally, it is possible to bound from above the distance between $\hat{\alpha}$ and α^* from above :

$$
\begin{aligned}
|\hat{\alpha} - \alpha^*| =& \left| \frac{\hat{c}}{2-\hat{c}} \cdot \frac{1 - \left\|\hat{\theta}\right\|_2^2}{\mathcal{V}} - \frac{c^*}{2-c^*} \cdot \frac{1 - \left\|\tilde{\theta}^*\right\|_2^2}{\mathsf{Tr}(\mathbb{V}(\boldsymbol{W}_j^t))} \right|, \\
\leq& Q_1 \cdot \frac{1 - \left\|\hat{\theta}\right\|_2^2}{\mathcal{V}} + \frac{c^* Q_2}{\mathcal{V}(2-c^*)} + \frac{1+\alpha^*}{\mathcal{V}} Q_3, \\
\leq& Q_1 \cdot \frac{1 - \left\|\hat{\theta}\right\|_2^2}{\mathcal{V}} + \frac{c^* Q_2}{\mathcal{V}(2-c^*)} + \frac{1+\alpha^*}{\mathcal{V}} Q_3, \\
\leq& \frac{256(1-c^*)^2}{\underline{c}^2 m^2} \left(\frac{\nu_1(p,N) \max(\epsilon_s^2)_{s\in[4]}}{nT(N-2)} + \frac{\nu_2 \epsilon_5}{\underline{c} m \sqrt{N}} + \frac{\nu_3 \epsilon_1}{\underline{c} m \sqrt{NnT}} \right) \\
&+ \frac{32 c^*(1-c^*)}{\underline{c}^2 m^2} \left[\frac{(\epsilon_7+1)^2}{n(T-1)} \left(1 + \frac{1}{\underline{c}\sqrt{T-1}} \right) + \frac{11\epsilon_7}{\sqrt{n(T-1)}} \right] \\
&+ \frac{4c^*}{\underline{c} m (2-c^*)} \left[\frac{2(\epsilon_6+1)}{\sqrt{n(T-1)}} \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1 \right) + \epsilon_5 \frac{8\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}} + \epsilon_1 \frac{32K^3}{c_2^5\sqrt{NnT}} \right] \\
&+ \frac{256 c^* \max(\epsilon_s^2)_{s\in[4]} K^{7/2} \sqrt{p} \left[C_A(p,N)\sqrt{p} + C_B \right]}{c_2^2 \underline{c} m (2-c^*) nT(N-2)} \left[(2+c_2) + \frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2} \right] \\
&+ \frac{1+\alpha^*}{\underline{c} m} \frac{16(\epsilon_6+1)}{\sqrt{n(T-1)}} \left(\frac{1}{\underline{c}\sqrt{T-1}} + 1 \right) + \epsilon_5 \frac{32\sqrt{2}K^{3/2}}{c_2^2\sqrt{N}} + \epsilon_1 \frac{128K^3}{c_2^5\sqrt{NnT}} \\
&+ \frac{1+\alpha^*}{\underline{c} m} \frac{256 \max(\epsilon_s^2)_{s\in[4]} K^{7/2} \sqrt{p} \left[C_A(p,N)\sqrt{p} + C_B \right]}{c_2^2 nT(N-2)} \left[(2+c_2) + \frac{2K^{3/2}(2\sqrt{K}+1)}{c_2^2} \right],
\end{aligned}
$$

with probability larger than $1 - 2n(T-1)p^2 \exp(-\epsilon_1^2) - 2n(T-1)pK\exp(-\epsilon_2^2) - 2n(T-1)Kp^2\exp(-\epsilon_3^2) - 2n(T-1)p \cdot (2p+9^p)\exp\left(-\min\left(\epsilon_4^2; \sqrt{cnT}\epsilon_4\right)\right) - 2n(T-1)K\exp(-\epsilon_5^2) - 2n(T-1)\exp\left(-\epsilon_6^2/4\right) - 19\exp(-\epsilon_7^2/4)$. ■

5.6 Auxiliary results

Lemma 5.6.1 (Vector Bernstein Inequality) *Let $X_1, \ldots, X_n$ be independent vector-valued centered random variables with common dimension K. Let $N := \left\| \sum_{i=1}^n X_i \right\|_2$ and $V := \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^2 \right]$. Then, for any $0 < \epsilon < V / \max_{i\in[n]}(\|X_i\|_2)$:*

$$\mathbb{P}\left[N \geq \epsilon + \sqrt{V} \right] \leq \exp\left(-\frac{\epsilon^2}{4V} \right).$$

This implies for any $0 < \epsilon < \sqrt{V} / \max_{i\in[n]}(\|X_i\|_2)$:

$$\mathbb{P}\left[N \geq (\epsilon+1)\sqrt{V} \right] \leq \exp\left(-\frac{\epsilon^2}{4} \right).$$

Proof. The proof of this theorem is given in Lemma 12 in [69]. ■

Lemma 5.6.2 (Smallest eigenvalue of the sum of Hermitian matrices) *Consider A and B two full rank Hermitian matrices in $\mathbb{R}^{K\times K}$. Then*

$$\lambda_K(A+B) \geq \lambda_K(A) + \lambda_K(B)$$

Proof. Matrices A and B are Hermitian and the spectral theorem ensures that they are both diagonalizable and that their eigenvalues are real valued. We denote $(\lambda_1(A),\dots,\lambda_K(A))$ and $(\lambda_1(B),\dots,\lambda_K(B))$ the eigenvalues of A and of B, respectively.

We have, for all $x \in \mathbb{R}^K$,

$$\langle (A+B)x, x\rangle = \langle Ax, x\rangle + \langle Bx, x\rangle \geq (\lambda_K(A) + \lambda_K(B)) \cdot \|x\|_2^2.$$

Finally, we note that $A+B$ is also a Hermitian matrix and if we replace x by the eigenvector $x_{\min}(A+B)$ of $A+B$ associated to the smallest eigenvalue $\lambda_K(A+B)$, we get that $\langle (A+B)x_{\min}(A+B), x_{\min}(A+B)\rangle = \lambda_K(A+B)$. This finishes the proof. ■ For a given diagonalizable matrix M such that $\mathrm{rank}(M) = r$, the smallest non-zero eigenvalue of M, $\lambda_r(M)$, is defined by

$$\lambda_r(M) = \min_{x\notin Ker(M)} \frac{\langle x, Mx\rangle}{\|x\|^2}.$$

Lemma 5.6.3 *Consider $p > K$ two integers, a matrix $A \in \mathbb{R}^{p\times K}$ such that $\mathrm{rank}(A) = \kappa \leq K$ and a symmetric positive definite matrix $B \in \mathbb{R}^{K\times K}$. Then we have*

$$\lambda_\kappa(ABA^\top) \geq \lambda_K(B)\lambda_\kappa(A^\top A).$$

Proof. First, we have $Im(A^\top) \subseteq \mathbb{R}^K$. Hence, B being symmetric, B is diagonalizable by the spectral theorem, and B being positive definite, B has positive eigenvalues. Thus, by the variational characterisation of the eigenvalues, we have :

$$\lambda_K(B) = \min_{x\in\mathbb{R}^K} \frac{\langle x, Bx\rangle}{\|x\|^2} \leq \min_{z\in Im(A^\top)} \frac{\langle z, Bz\rangle}{\|z\|^2} = \min_{y\in\mathbb{R}^p} \frac{\langle A^\top y, BA^\top y\rangle}{\|A^\top y\|^2}.$$

The matrix $ABA^\top$ is also symmetric and the spectral theorem ensures that it is diagonalizable. However, $ABA^\top \in \mathbb{R}^{p\times p}$ and $\mathrm{rank}(ABA^\top) = \kappa \leq K$. Hence the variational characterisation of the eigenvalues ensures that :

$$\lambda_\kappa(ABA^\top) = \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(ABA^\top)}} \frac{\langle x, ABA^\top x\rangle}{\|x\|^2}.$$

Moreover, $Ker(A^\top) \subset Ker(ABA^\top)$. Hence

$$\lambda_\kappa(ABA^\top) \geq \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \frac{\langle x, ABA^\top x\rangle}{\|x\|^2}.$$

Then we can bound from below the smallest eigenvalue of $ABA^\top$ as follows,

$$\begin{aligned}
\lambda_\kappa(ABA^\top) &\geq \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \frac{\langle A^\top x, BA^\top x\rangle}{\|x\|^2} \\
&\geq \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \frac{\langle A^\top x, BA^\top x\rangle}{\|A^\top x\|^2} \cdot \frac{\langle A^\top x, A^\top x\rangle}{\|x\|^2} \\
&\geq \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \left(\frac{\langle A^\top x, BA^\top x\rangle}{\|A^\top x\|^2} \cdot \frac{\langle x, AA^\top x\rangle}{\|x\|^2}\right) \\
&\geq \min_{z\in Im(A^\top)} \left(\frac{\langle z, Bz\rangle}{\|z\|^2}\right) \cdot \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \left(\frac{\langle x, AA^\top x\rangle}{\|x\|^2}\right) > 0.
\end{aligned}$$

To conclude, notice that $Ker(A^\top) = Ker(AA^\top)$. Indeed, $Ker(A^\top) \subset Ker(AA^\top)$. Moreover, for all $v \in Ker(AA^\top)$ we have :

$$v^\top AA^\top v = \left\|A^\top v\right\|_2^2 = 0.$$

Hence $Ker(AA^\top) \subset Ker(A^\top)$. We deduce from this equality :

$$\min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(A^\top)}} \left(\frac{\langle x, AA^\top x\rangle}{\|x\|^2}\right) = \min_{\substack{x\in\mathbb{R}^p;\\ x\notin Ker(AA^\top)}} \left(\frac{\langle x, AA^\top x\rangle}{\|x\|^2}\right) = \lambda_\kappa(AA^\top).$$

Finally, $AA^\top$ and $A^\top A$ have the same non-zero eigenvalues and thus $\lambda_\kappa(AA^\top) = \lambda_\kappa(A^\top A)$.

■

Lemma 5.6.4 (Smallest singular value of a product) *Suppose $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times p}$. We denote $\sigma_{\min}(A)$ and $\sigma_{\min}(B)$ the smallest singular value of A and B, respectively. Then,*

$$\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B).$$

Proof. By definition of the smallest eigenvalue, see Theorem C.3 in [64],

$$\sigma_{\min}(AB) = \min_{x\in\mathbb{R}^p\setminus\{0\}} \frac{\|ABx\|_2}{\|x\|_2}.$$

Then the following inequalities are easily deduced

$$\begin{aligned}
\sigma_{\min}(AB) &= \min_{x\in\mathbb{R}^p\setminus\{0\}} \frac{\|ABx\|_2}{\|Bx\|_2}\frac{\|Bx\|_2}{\|x\|_2}, \\
&\geq \min_{y\in\mathbb{R}^m\setminus\{0\}} \frac{\|Ay\|_2}{\|y\|_2} \min_{x\in\mathbb{R}^p\setminus\{0\}} \frac{\|Bx\|_2}{\|x\|_2}, \\
&\geq \sigma_{\min}(A)\sigma_{\min}(B).
\end{aligned}$$

■

Lemma 5.6.5 (First order characterization of convex functions) *Suppose f is a differentiable convex function from an open domain of $\mathbb{R}^n$ to $\mathbb{R}$. Then for all (x, y) in the domain of f we have*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

Proof. The proof can be found in section 3.1.3 of [30] ■

Lemma 5.6.6 (A row-wise perturbation bound for eigenvector) *Let G_0 and G be $p \times p$ symmetric matrices with $\operatorname{rank}(G_0) = K$. Write $Y = G - G_0 = [y_1, y_2, \ldots, y_p]$. For $1 \leq k \leq K$, let δ_k^0 and δ_k be the respective k-th largest eigenvalue of G_0 and G, and let u_k^0 and u_k be the respective k-th eigenvector of G_0 and G. Fix $1 \leq s \leq k \leq K$. For some $c \in (0, 1)$, suppose (by default, if $s = 1, \delta_{s-1}^0 - \delta_s^0 = \infty$)*

$$\min\left\{\delta_{s-1}^0 - \delta_s^0, \delta_k^0 - \delta_{k+1}^0, \min_{1\leq\ell\leq K}\left|\delta_\ell^0\right|\right\} \geq c\left\|G_0\right\|, \quad \|Y\| \leq (c/3)\left\|G_0\right\|$$

Write $U_0 = \left[u_s^0, u_{s+1}^0, \ldots, u_k^0\right], U = [u_s, u_{s+1}, \ldots, u_k]$ and $U_0^* = \left[u_1^0, u_2^0, \ldots, u_K^0\right]$. There exists an orthogonal matrix O such that

$$\left\|e_i'(UO - U_0)\right\| \leq \frac{5}{c\left\|G_0\right\|}\left(\|Y\|\left\|e_i' U_0^*\right\| + \sqrt{K}\left\|y_i\right\|\right), \quad \textit{for all } 1 \leq i \leq p.$$

Proof. See Lemma F.1 in [84]. ■

Definition 5.6.1 (Left stochastic matrices) *A real-valued matrix M of size $n \times m$ is said to be left stochastic if all its columns consist of non-negative entries and form probability vectors. Namely, for all $j \in [m]$,*

$$\forall i \in [n],\ [M]_{ij} \geq 0, \quad \textit{and } \sum_{i=1}^{n} [M]_{ij} = 1.$$

Lemma 5.6.7 (Stability of the set of left stochastic matrices) *Let $M_1 \in \mathbb{R}^{n\times m}$ and $M_2 \in \mathbb{R}^{m\times p}$ be two left stochastic matrices. Then $M_1 M_2$ is a left stochastic matrix.*

Proof. Consider $M_1 M_2 \in \mathbb{R}^{n\times p}$. Then for all $(i, j) \in [n] \times [p]$ we have

$$[M_1 M_2]_{ij} = \sum_{k=1}^{m} [M_1]_{ik} [M_2]_{kj}.$$

Hence for all $(i, j) \in [n] \times [p]$, $[M_1 M_2]_{ij} \geq 0$. Additionally, since M_1 and M_2 are left stochastic matrices, for all $j \in [p]$,

$$\sum_{i=1}^{n} [M_1 M_2]_{ij} = \sum_{i=1}^{n}\sum_{k=1}^{m} [M_1]_{ik} [M_2]_{kj} = \sum_{k=1}^{m} [M_2]_{kj} = 1.$$

■

Lemma 5.6.8 (Spectral norm inequalities) *Let $M \in \mathbb{R}^{n\times m}$. Then, its 1-norm and ∞-norm satisfy :*

$$\|M\|_1 = \max_{j\in[m]} \sum_{i=1}^{n} \left|[M]_{ij}\right| \quad \textit{and} \quad \|M\|_\infty = \max_{i\in[n]} \sum_{j=1}^{m} \left|[M]_{ij}\right|.$$

Finally, the following inequalities are verified :

$$\sqrt{\frac{1}{m}}\|M\|_\infty \leq \sigma_1(M) \leq \sqrt{n}\|M\|_\infty \quad \textit{and} \quad \sqrt{\frac{1}{n}}\|M\|_1 \leq \sigma_1(M) \leq \sqrt{m}\|M\|_1.$$

Proof. See section 2.3.2 in [68] ■

Lemma 5.6.9 (Perron-Frobenius theorem) *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix such that for all $(i,j) \in [n]^2$, $[M]_{ij} > 0$. Then the largest eigenvalue of M, $\lambda_1(M)$ is positive and is non degenerate, meaning it is a simple root of the characteristic polynomial or equivalently that its associated eigenspace is one dimensional. In addition there exists a corresponding eigenvector with positive entries. Moreover, other eigenvalues satisfy :*

$$\forall k \in [n]\backslash\{1\}, \quad \lambda_1(M) > |\lambda_k(M)|.$$

Proof. See [106]. ■

Chapitre 6

Introduction en français

La motivation principale de ce manuscrit est d'approfondir notre compréhension des phénomènes comportant une composante temporelle. La plupart des algorithmes d'apprentissage automatique et des modèles statistiques en grande dimension sont largement étudiés sous des hypothèses d'indépendance des observations. En effet, il existe moins d'outils, et ceux-ci sont techniquement plus exigeants, pour la concentration des mesures dans ce contexte. Cela rend le contrôle non asymptotique des déviations plus difficile dans ce cadre où la dépendance entre les observations est considérée. Très souvent, une évolution temporelle est évidente dans le modèle sous-jacent, mais n'est pas toujours prise en compte dans les méthodes proposées et dans les résultats d'inférence.

Cette thèse explore divers problèmes d'inférence non paramétrique et d'inférence en grande dimension. En particulier, nous étudions des tests d'hypothèses sur des matrices de covariance et l'estimation de leur support, la prédiction bilatérale en régression matricielle et l'estimation de topiques-modèles dynamiques combinant la factorisation matricielle et un processus autorégressif. Bien qu'ils partagent une motivation commune, les chapitres présentés dans cette thèse peuvent être lus et compris séparément car ils se concentrent sur des problèmes spécifiques et indépendants.

Évaluer la qualité des algorithmes de prévision est crucial dans diverses applications allant des phénomènes naturels comme les modèles météorologiques et les événements sismiques aux variables économiques telles que la prédiction du prix des actions ou de la demande future en énergie. Un indicateur clé de la performance des algorithmes de prédiction est la qualité des résidus, représentant la différence entre les valeurs observées et celles prédites. Ainsi, plus les résidus se rapprochent d'une distribution de bruit blanc, plus le modèle est performant. Dans le chapitre 2, nous étudions les problèmes de test et d'estimation de support d'une matrice de covariance en grande dimension issue d'une série temporelle stationnaire. Plus précisément, nous considérons $X_1, \dots, X_n$ des vecteurs gaussiens indépendants de dimension p avec une matrice de covariance Σ. Lorsque les vecteurs X_i proviennent d'un processus stationnaire, la matrice de covariance Σ a une structure de Toeplitz, c'est-à-dire que ses éléments diagonaux sont tous constants. Comme mentionné dans [46], les séries temporelles stationnaires sont utilisées comme approximations des séries temporelles géométriquement ergodiques. Ce contexte est motivé par l'observation suivante : étant donné une série temporelle de longueur T avec $T \gg p$, il est possible de considérer des vecteurs de longueur p suffisamment éloignés pour supposer qu'ils sont des vecteurs indépendants de dimension p. Le but est alors de tester si la distribution est proche d'un bruit blanc. Pour ce faire, nous testons si la matrice de covariance Σ est la matrice identité I_p ou s'il existe un nombre s d'éléments de covariance qui sont significativement positifs ou significativement différents de zéro. Nous fournissons des procédures de test avec des bornes supérieures

non asymptotiques sur les risques de test maximaux pour des structures de covariance modérément parcimonieuses et grandement parcimonieuses. Si le test est rejeté, il est intéressant de retrouver les entrées non nulles dans Σ, indiquant où l'information peut être perdue dans le processus de modélisation. Nous définissons ensuite une procédure de sélection de ce support et fournissons une borne supérieure non asymptotique sur son risque.

Ensuite, nous introduisons un nouveau modèle de régression matricielle où les corrélations dans la matrice de sortie sont expliquées par deux paramètres matriciels qui multiplient la matrice de prédiction respectivement par la gauche et la droite. Nous supposons que la matrice de bruit a des entrées σ^2-sous gaussiennes indépendantes. Ce modèle général de régression matricielle est largement non identifiable sans hypothèse supplémentaire forte. Ainsi seuls des résultats de prédiction sont fournis. Les prédicteurs sont d'abord définis comme solutions d'un problème de minimisation du risque de prédiction pour la norme de Frobenius au carré sous une contrainte de rang maximal fixe. En utilisant la décomposition en valeurs singulières (SVD) de la matrices cible et de la matrice de prédiction, nous fournissons des solutions à ce problème d'optimisation ainsi qu'une borne supérieure non asymptotique sur le risque de prédiction. Nous montrons que cette borne supérieure peut être décomposée en une somme d'un terme de biais et d'un terme stochastique. Nous dérivons ensuite une procédure de sélection de modèle pour estimer le vrai rang commun des matrices de paramètres, d'abord sous l'hypothèse que le paramètre de bruit σ est disponible. Nous examinons la performance non asymptotique de cette procédure et nous adaptons le problème de minimisation initial en fixant la contrainte de rang maximal à ce rang estimé. Cela conduit à de nouveaux prédicteurs adaptatifs au rang. Nous fournissons à nouveau une borne supérieure non asymptotique sur le risque de prédiction adaptatif au rang dans ce cadre de sélection de modèle. Ensuite, nous adaptons la procédure pour la rendre adaptative au rang et indépendante du paramètre de bruit σ. Nous fournissons à nouveau une borne supérieure non asymptotique sur son risque de prédiction. Enfin, nous reconsidérons le problème de minimisation initial en étudiant la relaxation convexe de la pénalisation par le rang. Nous fournissons des solutions explicites à ce problème et à nouveau une borne non asymptotique sur le risque de prédiction. Des résultats numériques sont fournis pour illustrer les résultats théoriques.

Enfin, nous considérons les topiques-modèles. Nous supposons que nous recueillons un lot de documents et avons accès aux fréquences de chaque mot du vocabulaire pour chaque document. Les colonnes de cette matrice de fréquence mot-document Y sont modélisées comme des réalisations de distributions multinomiales centrées sur des vecteurs de probabilité mot-document. Dans des exemples réels, peu de sujets différents sont abordés dans les corpus de documents. Cela suggère que la matrice de probabilité mot-document Π présente une structure de faible rang. L'objectif est de factoriser cette matrice de probabilité mot-document Π par le produit d'une matrice de probabilité mot-sujet A et d'une matrice de probabilité sujet-document W, c'est-à-dire $\Pi = AW$. Dans ce contexte, ces trois matrices Π, A et W sont toutes stochastiques à gauche, c'est-à-dire que leurs entrées sont positives et que leurs colonnes somment à un. Sous des hypothèses précises, que nous supposons, l'identifiabilité de A et W peut être établie. Nous rappelons également l'algorithme de [84] qui permet de retrouver les termes de cette factorisation. Dans cette thèse, nous supposons une temporalité dans la collecte de documents et modélisons l'évolution dans le temps de la matrice de probabilité sujet-document W par un processus autorégressif stationnaire. Ainsi la matrice W devient dans ce contexte une matrice aléatoire dépendant du temps $\boldsymbol{W}_t$. Plus précisément, à chaque étape temporelle t, la distribution des sujets donnés un document est une combinaison linéaire de la distribution précédente et d'un bruit suivant une distribution de Dirichlet, qui dirige l'évolution temporelle des topiques. Nous supposons en particulier que les paramètres de bruit sont inconnus, c'est-à-dire le paramètre de la distribution de Dirichlet. Une attention particulière est accordée à garantir que ce modèle autorégressif conserve la propriété

que les colonnes de la matrice de probabilité sujet-document somment à un. Nous étudions d'abord un cas oracle où la matrice de probabilité mot-document $(\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_T)$ est disponible. Nous fournissons d'abord des bornes non asymptotiques sur le spectre de la matrice de covariance empirique de $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_T)$. Nous adaptons ensuite l'algorithme de [84] pour récupérer la matrice de probabilité mot-sujet A. Cela permet de récupérer $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_T)$ par projection. Nous proposons ensuite des estimateurs des paramètres autorégressifs conduisant l'évolution de $\boldsymbol{W}_t$. Nous fournissons des bornes supérieures non asymptotiques sur les risques d'estimation. Ensuite, nous adaptons cette procédure au cas réel où seule la matrice complète de fréquence mot-document $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ est disponible. Dans la procédure d'estimation de A, nous donnons des bornes supérieures plus explicites que [84] jusqu'aux facteurs log. Nous fournissons également la dépendance sur toutes les dimensions des matrices apparaissant. Enfin, nous montrons que le bruit dû à la distribution multinomiale des décomptes de mots et le bruit Dirichlet de la distribution stationnaire des sujets donnés les documents publiés dans le temps s'ajoutent dans les vitesses d'estimation finales des paramètres autorégressifs. En particulier, lorsque le nombre de mots par document augmente, c'est-à-dire lorsque le bruit multinomial diminue, nous retrouvons les vitesses oracles.

Historiquement, l'analyse des séries temporelles est généralement effectuée dans un cadre asymptotique. L'analyse asymptotique des séries temporelles à valeurs réelles et vectorielles est bien comprise depuis la publication de [71], [62], [99] et [31]. Il s'agit toujours d'un domaine de recherche actif tant d'un point de vue théorique, [79, 50, 91, 117, 51, 59] que comme outil pour l'étude des propriétés des algorithmes, [142].

Récemment, l'étude des séries temporelles à valeurs matricielles et plus globalement des séries temporelles à valeurs tensorielles a émergé. Les études sont encore principalement menées dans un cadre asymptotique, [47, 49, 44, 96]. Cependant, l'analyse non asymptotique des séries temporelles gagne en importance, [16, 15, 58, 135]. Cette thèse s'inscrit dans cette dynamique de recherche et tous les problèmes étudiés sont conduits dans un cadre non asymptotique. En abordant ces défis et en explorant des méthodologies innovantes dans chaque chapitre, cette thèse contribue à l'avancement de la théorie statistique dans l'analyse des données à valeurs vectorielles et matricielles dans des contextes de grande dimension.

La première partie de l'introduction sert de présentation exhaustive des outils techniques nécessaires à la compréhension des principaux chapitres de cette thèse. Ensuite, dans la deuxième partie, nous présentons les configurations et les détails des résultats.

Problèmes étudiés et contributions

Cette section est consacrée à la présentation des problèmes statistiques étudiés dans les principaux chapitres de la thèse. Nous détaillons d'abord le problème des tests d'hypothèse, qui est essentiel à la compréhension du chapitre 2. Nous explorons ensuite le problème de la régression et en particulier la régression linéaire multivariée pour laquelle le chapitre 3 fournit une extension. Nous présentons ensuite le problème du modèle thématique, pour lequel une extension dynamique est étudiée dans les chapitres 4 et 5.

Test d'hypothèse : décider où se trouve une matrice de covariance

Dans tous les domaines, de l'expérimentation scientifique à la vie quotidienne, nous sommes amenés à prendre des décisions sur des activités risquées à partir de résultats d'expériences ou d'observations de phénomènes dans un contexte incertain. Le problème de décision consiste à trancher, sur la base d'observations, entre une hypothèse dite nulle, notée H_0, et une autre hypothèse dite alternative, notée H_1. Un test d'hypothèse est donc une procédure de décision permettant de déterminer si l'hypothèse nulle peut être rejetée en faveur de l'hypothèse alternative compte tenu des données observées. Nous supposons que les observations sont des réalisations des variables aléatoires $(X_1, \ldots, X_n)$ prenant des valeurs dans $(E, \mathcal{E})$.

Definition 6.0.1 (Procédure de test) *Un test Δ_n est une fonction mesurable des observations prenant ses valeurs dans $\{0, 1\}$:*

$$\Delta_n : E^n \to \{0, 1\}.$$

Δ_n sépare alors l'ensemble des résultats possibles d'un événement aléatoire en deux ensembles contigus, H_0 est rejeté chaque fois que $\Delta_n = 1$ et n'est pas rejeté chaque fois que $\Delta_n = 0$.

Nous considérons dans le chapitre 2 l'observation de n vecteurs aléatoires *i.i. d* vecteurs aléatoires $(X_1, \ldots, X_n)$ définis sur $\mathbb{R}^p$ avec une matrice de covariance commune $\Sigma \in \mathcal{S}_p^{++}$, où $\mathcal{S}_p^{++}$ représente l'ensemble des matrices symétriques définies positives de taille $p \times p$. Le problème de test considéré est le suivant

$$H_0 : \Sigma = \{I_p\}, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}_p,$$

où $\mathcal{F}_p \subset \mathcal{S}_p^{++}$ est un ensemble de matrices de Toeplitz éparses. Nous considérons deux hypothèses alternatives différentes : soit il existe un nombre s d'éléments de covariance qui sont significativement positifs (l'alternative unilatérale $\mathcal{F}_p = \mathcal{F}_+(s, S, \sigma)$) ou significativement différents de zéro *i.e.* (l'alternative bilatérale $\mathcal{F}_p = \mathcal{F}_+(s, S, \sigma)$). (l'alternative bilatérale $\mathcal{F}_p = \mathcal{F}(s, S, \sigma)$). Les classes d'hypothèse d'alternative sont présentées dans la Définition 2.2.1.

Dans un problème de décision, deux types d'erreur sont possibles. Une erreur de type I se produit lorsque nous décidons que H_1 est vrai, *i.e.* observant $\Delta_n = 1$, alors que H_0 est en fait vrai. Une erreur de type II se produit lorsque nous ne parvenons pas à rejeter H_0, *i.e.* observant $\Delta_n = 0$, alors que H_1 est vrai. Les conséquences de ces deux erreurs peuvent être plus ou moins importantes. Chaque décision a donc une probabilité d'être juste et une probabilité d'être fausse. La probabilité d'erreur de type I, c'est-à-dire la pire "chance" de rejeter à tort l'hypothèse nulle, est notée α et est appelée niveau de signification du test. La probabilité d'erreur de type II, c'est-à-dire la pire "chance" de ne pas rejeter l'hypothèse nulle, est notée $1 - \beta$. Ainsi, β est la probabilité de rejeter correctement l'hypothèse nulle et est appelée la puissance du test.

Definition 6.0.2 (Erreurs de type I et de type II) *Considérons la procédure de test Δ_n pour le problème de test $H_0 : \Sigma = I_p$, vs. $H_1 : \Sigma \in \mathcal{F}_p$. La probabilité d'erreur de type I de Δ_n est alors définie comme suit :*

$$\alpha := \mathbb{P}_{I_p}(\Delta_n = 1).$$

De même, la probabilité d'erreur de type II de Δ_n est définie comme suit :

$$1 - \beta := \sup_{\Sigma \in \mathcal{F}_p} \mathbb{P}_{\Sigma}(\Delta_n = 0).$$

Pour définir une procédure de test, l'idéal serait évidemment de trouver celle qui minimise les deux risques d'erreur en même temps. Malheureusement, on peut montrer qu'ils varient dans des directions opposées, *i.e.* toute procédure qui diminue α augmentera généralement $1-\beta$ et vice versa. Il existe donc essentiellement deux façons de définir une procédure de test optimale. La première est la procédure de test optimale de Neyman-Pearson. Dans ce cadre, nous considérons que l'une des deux erreurs est plus importante que l'autre et nous essayons d'éviter cette erreur. En général, nous choisissons H_0 et H_1 de manière à ce que l'erreur que nous essayons d'éviter soit l'erreur de type I. Remarquez que le test idéal ne rejetterait alors presque jamais à tort H_0. Cependant, dans les cas habituels, le seul test ayant $\alpha = 0$ est le test trivial $\Delta_n = 0$. Nous devons donc laisser l'autre erreur se produire. Par exemple, dans le cas d'un procès, nous faisons généralement tout notre possible pour éviter de condamner un innocent, même si cela implique de prendre le risque d'acquitter un coupable. Mathématiquement, on fixe une valeur pour le niveau $\alpha \in [0, 1]$. Plus les conséquences de l'erreur de type I sont graves, plus α sera petit. Toutefois, pour le même problème de décision, il peut exister plusieurs tests dont la probabilité d'erreur de type I est inférieure à α. Dans ce cas, le meilleur de ces tests est celui qui minimise la probabilité de l'erreur de type II, *i.e.* celui qui maximise la puissance β parmi les tests dont le niveau est au plus α.

Definition 6.0.3 (Procédure de test optimale de Neyman-Pearson) *Désignons par Δ^α l'ensemble de toutes les procédures de test dont le niveau est au plus égal à α. Le test optimal de Neyman-Pearson, noté Δ_{NP}, est alors un test de niveau α qui résout la question suivante :*

$$\forall \Sigma \in \mathcal{F}_p, \quad \mathbb{P}_\Sigma[\Delta_{NP} = 0] = \inf_{\Delta \in \Delta^\alpha} \mathbb{P}_\Sigma[\Delta = 0].$$

S'il existe, Δ_{NP} est appelé test uniformément le plus puissant.

Comme le problème que Δ_{NP} doit résoudre n'a pas toujours de solution, la notion d'optimalité définie par la procédure de test optimal de Neyman-Pearson n'est pas universelle. Il est donc nécessaire d'adopter une approche plus générale pour trouver une procédure de test optimale. Comme décrit précédemment, il n'est pas possible de trouver un test qui minimise le niveau α et maximise la puissance β car α et $1 - \beta$ évoluent dans des directions opposées. Cependant, il est possible de minimiser la somme des probabilités d'erreur de type I et de type II. Un rôle égal est donc accordé à H_0 et H_1. Ce critère est décrit comme l'approche minimax.

Definition 6.0.4 (Risque de test maximal) *Considérons une procédure de test Δ et définissons $R(\Delta)$ son risque de test maximal :*

$$R(\Delta, \mathcal{F}_p) := \mathbb{P}_{I_p}(\Delta = 1) + \sup_{\Sigma \in \mathcal{F}_p} \mathbb{P}_\Sigma(\Delta = 0).$$

On dit alors qu'un test est optimal minimax s'il minimise le risque de test maximal parmi toutes les procédures de test. Son risque de test maximal est alors appelé risque de test minimax.

Definition 6.0.5 (Risque de test minimax) *Le risque de test minimax est défini comme suit*

$$R^*(\mathcal{F}_p) := \inf_\Delta R(\Delta, \mathcal{F}_p).$$

Si elle existe, la procédure de test qui permet d'obtenir le risque de test minimax, noté Δ_*, est appelée test minimax.

Un autre point important à mentionner est que la classe d'hypothèse nulle est un singleton, à savoir la matrice d'identité. L'objectif de la procédure est donc de déterminer s'il est possible ou non de rejeter avec une forte probabilité l'hypothèse selon laquelle Σ est la matrice identité. En outre, nous avons choisi comme classes d'hypothèses alternatives un sous-ensemble de matrices de Toeplitz peu denses, $\mathcal{F}_p = \mathcal{F}_+(s, S, \sigma)$ ou $\mathcal{F}_p = \mathcal{F}(s, S, \sigma)$. Essentiellement, on peut se demander pourquoi un tel problème de test ne prend pas la forme plus générale suivante :

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{S}_p^{++} \backslash \{I_p\}.$$

Dans ce scénario, on remarque que pour tout choix standard de distance sur $\mathcal{S}_p^{++}$, *i.e.* la distance dérivée de la norme de Frobenius, dénotée par $\| \ \|_F$, on a

$$\inf_{\Sigma \in \mathcal{S}_p^{++} \backslash \{I_p\}} \|I_p - \Sigma\|_F = 0.$$

Il n'est donc pas possible de séparer l'hypothèse nulle de l'hypothèse alternative. Le risque de test minimax est donc égal à un et le test de supposition aléatoire devient optimal. Par conséquent, dans ce problème de test d'adéquation, il est obligatoire que la classe d'hypothèses alternatives soit bien séparée du singleton d'hypothèses nulles. Ainsi, pour un $\epsilon > 0$ fixé, nous devons définir $\mathcal{F}_p^{(\epsilon)}$ de telle sorte que

$$\inf_{\Sigma \in \mathcal{F}_p^{(\epsilon)}} \| I_p - \Sigma \|_F \geq \epsilon.$$

La définition de nos classes alternatives montre que $\mathcal{F}_+(s, S, \sigma)$ et $\mathcal{F}(s, S, \sigma)$ sont bien séparés du singleton $\{I_p\}$. Enfin, le choix optimal du rayon de séparation ϵ est discuté dans la littérature et peut être défini comme le rayon de séparation minimax. Cela dépasse le cadre de cette thèse. Cependant, les lecteurs intéressés peuvent consulter [95] et [82] pour plus de détails sur les procédures de test minimax.

Chapter 2 : Test de la matrice de covariance et récupération du support . Nous considérons $(X_i)_{i=1,\dots,n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$ où Σ a une structure de Toeplitz. Nous notons ensuite $\sigma_{|i-j|}$ la covariance $\mathrm{Cov}(X^i, X^j)$ pour $i, j \in \{1, \dots, p\}$. Tout d'abord, nous testons si la matrice de covariance Σ est la matrice identité I_p par rapport à l'alternative unilatérale $\mathcal{F}_+(s, S, \sigma)$ ou l'alternative bilatérale $\mathcal{F}(s, S, \sigma)$, voir Definition 2.2.1. D'un point de vue asymptotique, s peut tendre vers l'infini lorsque p tend vers l'infini, ce qui autorise un modèle non paramétrique, c'est-à-dire que le nombre de paramètres peut augmenter. De tels modèles n'ont été considérés que dans l'estimation non paramétrique de la densité spectrale de séries temporelles stationnaires, voir [89]. Nous définissons tout d'abord φ_A la fonctionnelle linéaire de la matrice de covariance Σ associée à la matrice A appartenant à $\mathcal{S}_p$ comme $\varphi_A(\Sigma) := \mathrm{Tr}(A\Sigma)$. La matrice de covariance de l'échantillon est notée Σ_n. Ainsi, l'élément de covariance σ_j, $j \geq 1$, peut être écrit comme suit

$$\sigma_j = \mathbb{E}[X^T A_j X] = \mathrm{Tr}(A_j \Sigma) = \varphi_{A_j}(\Sigma), \quad \text{avec } [A_j]_{k\ell} = \frac{1}{2(p-j)} \mathbb{1}(|k - \ell| = j);$$

où A_j est une matrice qui a 0 pour élément sauf sur les $j$eme diagonales supérieure et inférieure. De même, l'estimateur empirique de σ_j peut être défini comme $\varphi_{A_j}(\Sigma_n)$.

Dans le cas modérément clairsemé, la somme de toutes les valeurs de S permettra d'effectuer le test, alors que dans le cas très clairsemé, une recherche sur des sous-ensembles de taille s sera nécessaire. C'est ce qu'on appelle une procédure de balayage, qui est très rapide pour les vecteurs. Il convient de noter que, si la densité s est inconnue, une deuxième recherche sur différentes valeurs

possibles de s produira une procédure agrégée, exempte de s. Dans le cas modérément clairsemé où l'hypothèse alternative est $\mathcal{F}_+(s,S,\sigma)$, nous considérons pour un certain seuil $t_{n,p}^{MS+}$ la statistique de test Δ_n^{MS+} définie dans (2.5). Lorsque l'hypothèse alternative est $\mathcal{F}(s,S,\sigma)$, nous considérons pour un certain seuil $t_{n,p}^{MS}$ la statistique de test Δ_n^{MS+} définie dans (2.6). Les bornes supérieures de leurs risques de test maximaux sont dérivées respectivement dans le Théorème 2.3.1 et le Théorème 2.3.1. Dans le cas très peu dense, lorsque l'hypothèse alternative est $\mathcal{F}_+(s,S,\sigma)$, nous considérons pour un certain seuil $t_{n,p}^{MS+}$ la statistique de test Δ_n^{HS+} définie dans (2.7). Lorsque l'hypothèse alternative est $\mathcal{F}(s,S,\sigma)$, nous considérons pour un certain seuil $t_{n,p}^{MS}$ la statistique de test Δ_n^{HS+} définie dans (2.8). Les tests Δ_n^{HS+} et Δ_n^{HS} essaient successivement tous les ensembles possibles $\mathcal{C}$ de s diagonales parmi les premières S. Si l'un de ces tests décide de rejeter H_0, alors Δ_n^{HS+} rejette également H_0. Les bornes supérieures de leurs risques maximaux de test sont dérivées respectivement dans le Théorème 2.3.1 et le Théorème 2.3.1.

Pour limiter par le haut les risques de test maximaux des procédures mentionnées, nous donnons une nouvelle variante de l'inégalité de concentration pour les formes quadratiques des grands vecteurs gaussiens et ces limites sont spécifiées pour les matrices de covariance qui sont Toeplitz avec peu de diagonales non nulles dans le Théorème 2.3.1. Ces bornes sont spécifiées pour les matrices de covariance qui sont des Toeplitz avec quelques diagonales non nulles dans le Corollaire 2.2.4.

Théorème 2.3.1 en français La variable aléatoire $\varphi_A(\Sigma_n - \Sigma)$ est centrée et sous-exponentielle avec des paramètres
$\left(\nu^2 = \frac{2||A\Sigma||_F^2}{n(1-K)}, b = \frac{2||A\Sigma||_\infty}{nK}\right)$, pour un K arbitraire dans $]0,1[$. Par conséquent, pour tout $u > 0$:

$$\mathbb{P}[\varphi_A\left(\Sigma_n - \Sigma\right) \geq \max\left\{\sqrt{u}\frac{||A\Sigma||_F}{\sqrt{n(1-K)}}, u\frac{||A\Sigma||_\infty}{nK}\right\}] \leq \exp\left(-\frac{u}{4}\right).$$

Des inégalités de concentration ont déjà été données pour de telles fonctionnelles. La plus proche de notre cas est l'inégalité de concentration de type chi-carré dans [121] pour les vecteurs gaussiens standardisés et généralisée aux vecteurs sous-gaussiens. Mentionnons également [65] qui a donné une inégalité de Bernstein pour l'élément de covariance empirique d'un processus gaussien centré stationnaire et l'a généralisée aux processus gaussiens localement stationnaires.

Nous proposons également une méthode pour identifier les éléments diagonaux σ_j, $j = 1, ..., S$, avec des entrées non nulles dans σ, en indiquant où l'information peut être perdue dans le processus de modélisation. L'objectif est de sélectionner correctement les coefficients de corrélation non nuls. On peut définir un problème de sélection de retard comme l'estimation de η, un vecteur avec des entrées $\eta_j = \mathbf{1}(|\varphi_{A_j}(\Sigma)| > 0)$. L'objectif est de trouver un sélecteur $\widehat{\eta}$ avec $\widehat{\eta_j} = \mathbb{1}(|\varphi_{A_j}(\Sigma_n)| > \tau_n)$ qui soit cohérent au sens où le risque $R^{LS}(\widehat{\eta}, \mathcal{F}) = \sum_{j=1}^{S} \mathbb{E}_\Sigma[|\widehat{\eta_j} - \eta_j|]$ stays bounded. Nous fournissons dans le Théorème 2.4.1 une valeur explicite de τ_n telle que le risque $R^{LS}(\widehat{\eta}, \mathcal{F})$ reste limité par une quantité décroissante en S.

Regression multivariée

L'analyse de régression est une méthode statistique fondamentale utilisée pour explorer et quantifier la relation entre une ou plusieurs variables indépendantes (les prédicteurs) et une variable dépendante (la cible). L'objectif de l'analyse de régression est de développer un modèle prédictif capable d'estimer la valeur de la cible en fonction des valeurs des variables prédictives. Ce problème est au cœur du chapitre 3.

Nous observons un ensemble de données composé de $T \subset \mathbb{N}^*$ réponses Y_t et T caractéristiques correspondantes X_t. L'objectif est de développer un modèle capable de prédire la réponse Y_{T+1} sur la base d'une nouvelle caractéristique X_{T+1}. Nous écrivons notre modèle comme suit :

$$\forall t \in [T], \quad Y_t = f^*(X_t) + \epsilon_t,$$

où ϵ_t englobe les erreurs de mesure et les facteurs qui font que Y dépend d'autres facteurs que le seul X considéré. La véritable fonction f^* est inconnue, ce qui nous amène à rechercher une fonction f appropriée qui prédit avec précision les valeurs Y aux nouveaux points $X = x$. Une fonction f performante permet d'identifier les composantes de X qui sont significatives pour expliquer Y et celles qui ne le sont pas. Au cours de la collecte des données, il peut arriver que de nombreuses caractéristiques partagent la même valeur, par exemple $X_i = X_j = x$ avec $i \neq j$. Malgré cela, nous pouvons observer $Y_i \neq Y_j$, ce qui indique que ϵ_i et ϵ_j représentent des erreurs irréductibles dans notre modèle. Même avec une fonction optimale f, prédire Y_t en utilisant f à chaque $X_t = x$ peut toujours donner lieu à des erreurs car $f(x)$ ne représente qu'une valeur parmi une distribution de valeurs potentielles de Y_t. Une approche consiste à considérer que la fonction f^* évaluée sur x produit la moyenne des valeurs observées Y_t correspondant à $X_t = x$. Cela conduit à modéliser la fonction de régression f^* comme $f^*(x) = \mathbb{E}[Y|X = x]$. La fonction de régression f^* est le prédicteur optimal de Y en ce qui concerne l'erreur quadratique moyenne :

$$f^* \in \operatorname{argmin}_g \mathbb{E}\left[(Y - g(X))^2 \,|X = x\right].$$

De plus, pour toute estimation $\hat{f}$ de f^*, on a

$$\mathbb{E}[\left(Y - \hat{f}(X)\right)^2 |X = x] = \left(f^*(x) - \hat{f}(x)\right)^2 + \mathbb{V}(\epsilon).$$

Cela montre qu'il existe une erreur irréductible que nous ne pouvons pas réduire, à savoir $\mathbb{V}(\epsilon)$, même si nous connaissons la vraie fonction f^*. Nous sommes particulièrement intéressés par les modèles linéaires, c'est-à-dire lorsque f^* est une fonction linéaire. Nous appelons ce problème le problème de la régression linéaire.

Cible à valeur vectorielle

Dans le cadre de la régression conventionnelle, les variables cibles Y_t sont scalaires. Cependant, dans diverses applications, l'objectif n'est pas de prédire une variable scalaire mais plutôt un vecteur $Y_{T+1} \in \mathbb{R}^m$. Nous considérons toujours que les prédicteurs sont à valeur vectorielle, à savoir pour $t \in [\![1, T]\!]$, $X_t \in \mathbb{R}^p$. Par conséquent, la fonction de régression $f^*(x) = \mathbb{E}[Y|X = x]$ prend des arguments dans $\mathbb{R}^p$ et produit des valeurs dans $\mathbb{R}^m$. Sans hypothèse supplémentaire, f^* peut être estimé indépendamment pour chaque coordonnée, ce qui conduit à des régressions linéaires indépendantes avec des cibles à valeur réelle. En effet, l'hypothèse de linéarité sur f permet de réécrire le modèle comme suit :

$$Y = XB^* + E, \tag{6.1}$$

où $Y \in \mathbb{R}^{T \times m}$ est la matrice cible, $X \in \mathbb{R}^{T \times p}$ est la matrice prédicteur et $B^* \in \mathbb{R}^{p \times m}$ est le paramètre et $E \in \mathbb{R}^{T \times m}$ est la matrice de bruit, généralement supposée avoir *i.i.d.* σ^2-sousGaussiennes. On remarque que pour tout $j \in [\![1, m]\!]$, la j^{ieme} colonne de Y, dénotée $[Y]_{.j}$ ne dépend que de la j^{ieme} colonne $[B^*]_{.j}$ de B^* et pour tout $i \in [\![1, T]\!]$, la i^{ieme} ligne de Y, notée $[Y]_{i.}$, ne dépend que de la

i^{ieme} ligne $[X]_{i.}$ de X. Nous pouvons donc considérer ce problème comme p problèmes de régression linéaire indépendants avec des cibles à valeurs réelles :

$$\forall j \in [\![1, p]\!], \quad [Y]_{.j} = X[B^*]_{.j} + [E]_{.j}.$$

Ce problème est une instance de l'apprentissage multitâche, qui est fortement étudié dans la littérature [107, 101, 5, 119, 55, 9, 143]. En particulier, un estimateur de XB^* peut être dérivé en résolvant p problèmes de moindres carrés ordinaires. Notons $X\hat{B}$ l'estimateur correspondant. Si E a des entrées indépendantes σ^2-sous-Gaussiennes, on déduit de l'analyse standard des MCO, voir [115], l'existence d'une constante positive C telle que :

$$\frac{1}{T}\mathbb{E}\left[\left\|X\hat{B} - XB^*\right\|_F^2\right] \leq C\sigma^2\frac{pm}{T}.$$

Ce résultat prouve que dans un cadre de grande dimension, c'est-à-dire lorsque $T < pm$, l'erreur quadratique moyenne de prédiction de $\hat{B}$ n'est pas nulle. Il est donc naturel de se demander si un autre estimateur de B^* peut être dérivé pour résoudre ce problème. Malheureusement, le corollaire 4.13 de [115] prouve que l'estimateur des moindres carrés atteint la vitesse d'estimation minimax dans le modèle de séquence gaussienne univariée. Cela implique que l'estimateur des moindres carrés est optimal parmi tous les estimateurs sans aucune connaissance préalable sur la structure de B^*. Puisque cette borne est optimale, on pourrait penser qu'il n'y a aucun espoir de résoudre ce problème statistique de grande dimension.

Heureusement, on constate souvent que les données à grande dimension présentent une faible complexité inhérente. Lorsque les structures de basse dimension sont bien définies, l'analyse revient à des statistiques de basse dimension plus conventionnelles. Toutefois, les données à grande dimension posent des problèmes en raison des structures sous-jacentes inconnues à basse dimension. Par conséquent, une tâche fondamentale consiste à identifier ou à approximer ces structures. Dans le cadre de la régression multivariée, il existe souvent des structures partagées entre les coordonnées qui peuvent être exploitées pour améliorer les limites de prédiction. Par exemple, on peut supposer que les colonnes de B^* partagent le même modèle de rareté avec seulement s entrées non nulles. Si chaque tâche est exécutée individuellement, on obtient l'estimateur de groupe-lasso $\hat{B}^{GL}$ étudié dans [98]. Dans ce cadre, il existe une constante positive $C > 0$ telle que l'erreur quadratique moyenne de prédiction de $\hat{B}^{GL}$ devient :

$$\frac{1}{T}\mathbb{E}\left[\left\|X\hat{B}^{GL} - XB^*\right\|_F^2\right] \leq C\sigma^2\frac{sm\log(p)}{T}.$$

Nous rappelons que le facteur logarithmique supplémentaire apparaît en raison du support inconnu des entrées non nulles de B^*. Par conséquent, dans le régime de grande dimension sous cette hypothèse de structure de sparsité, l'erreur quadratique moyenne de prédiction converge vers zéro tant que $T > sm\log(p)$. En outre, nous soulignons que cette hypothèse de structure de rareté imite la structure univariée standard, résolue avec la procédure Lasso et sa variante, voir [124, 22, 114, 33, 19]. Heureusement, des structures plus complexes peuvent être capturées dans le cadre de la régression multivariée. Par exemple, si les colonnes de Y sont corrélées, on peut supposer une structure de faible rang sur B^*. Cela conduit à la régression multivariée de rang faible.

Une solution possible à ce problème est de considérer un estimateur $\hat{B}_\lambda$ de B^* qui peut être défini comme la solution d'une version pénalisée par le rang du problème des moindres carrés ordinaires. Ainsi, pour tout $\lambda > 0$, nous considérons :

$$\hat{B}_\lambda \in \operatorname{argmin}_B \|Y - XB\|_F^2 + \lambda r_B, \tag{6.2}$$

où r_B représente le rang de B. Une première question d'intérêt est la sélection de l'hyperparamètre $\lambda > 0$. Ce problème relève de la catégorie de la sélection de modèles et nous renvoyons le lecteur à [64, 100] pour des introductions complètes. La première étape du calcul de cet estimateur consiste à définir les estimateurs de rangs restreints, c'est-à-dire $\hat{B}^{(k)}$ qui minimise $\|Y - XB\|_F^2$ parmi les matrices B de rang inférieur ou égal à k.

Lemme 6.0.1 (lemme 8.1 dans [64]) *Considérons $P := X\left(X^\top X\right)^+ X^\top$ le projecteur orthogonal sur l'intervalle de X où $\left(X^\top X\right)^+$ désigne le pseudo-inverse de Moore-Penrose de $X^\top X$. Dénote $\sum_{i=1}^{\mathrm{rank}(PY)} \sigma_i u_i v_i^\top$ la SVD de PY. Alors $X\hat{B}^{(k)}$ peut être défini comme $\sum_{i=1}^{k} \sigma_i(PY) u_i v_i^\top$.*

Lorsque le rang de B^* est inconnu, l'estimateur précédent peut être calculé pour toute valeur de $rdans\mathbb{N}^*$, ce qui conduit à $\hat{B}^{(k)}$. La qualité de cet estimateur est donnée dans le lemme suivant.

Lemme 6.0.2 (Limite non asymptotique de l'erreur quadratique de prédiction, Théorème 5 dans [32]) *Il existe une constante positive C telle que pour tout $kdans\mathbb{N}^*$,*

$$\left\| X\hat{B}^{(k)} - XB^* \right\|_F^2 \leq C \left[\sum_{i=r+1}^{\mathrm{rank}(XB^*)} \sigma_i(XB^*)^2 + k\|PE\|_{op}^2 \right].$$

Notons que cette limite, qui présente un compromis biais-variance, tient presque sûrement mais dépend de la plus grande valeur singulière de la projection de la matrice de bruit E sur l'étendue de X. On peut dériver une borne supérieure ne dépendant pas de E en contrôlant le spectre de la matrice aléatoire PE, puis fournir une borne supérieure vraie avec une probabilité élevée. Les bornes ainsi dérivées seront plus ou moins étroites selon les hypothèses que l'on fait sur la distribution de la matrice de bruit E. Le lemme suivant en donne un exemple.

Lemme 6.0.3 (Erreur quadratique moyenne en régression multivariée de rang faible, corollaire 6 dans [32]) *Supposons que la matrice de bruit E ait des entrées gaussiennes centrées indépendantes avec une variance σ^2. Il existe alors une constante positive C telle que pour tout $r \in \mathbb{N}^*$,*

$$\mathbb{E}\left[\left\| X\hat{B}^{(k)} - XB^* \right\|_F^2 \right] \leq C \left[\sum_{i=r+1}^{\mathrm{rank}(XB^*)} \sigma_i(XB^*)^2 + \sigma^2 k(m + r_X) \right],$$

où r_X représente le rang de X.

Le Lemme 6.0.3 montre que l'erreur quadratique moyenne est limitée par une erreur d'approximation et un terme stochastique. L'erreur d'approximation est décroissante en k et disparaît pour $k > \mathrm{rank}(XB^*)$. De plus, l'erreur quadratique moyenne satisfait pour $k > \mathrm{rank}(XB^*)$:

$$\frac{1}{T}\mathbb{E}\left[\left\| X\hat{B} - XB^* \right\|_F^2 \right] \leq C\sigma^2 \frac{k(m + r_X)}{T}.$$

On peut alors remarquer que $\mathrm{rank}(B^*) \geq \mathrm{rank}(XB^*)$ et que dans un cadre à grande dimension avec un rang très faible, $\mathrm{rank}(XB^*)(m + r_X) \ll pm$. Cependant, la valeur de $\mathrm{rank}(XB^*)$ est inconnue et la limite oracle précédemment énoncée ne peut donc pas être atteinte. Une procédure d'adaptation des

données est proposée dans [32] à la fois dans le cas d'un σ^2 connu et d'un σ^2 inconnu, le paramètre du bruit. Les performances obtenues sont similaires à celles obtenues dans le cas de l'oracle.

Par conséquent, si les colonnes de la matrice observée Y sont corrélées et si nous supposons que B^* a une structure de faible rang, un estimateur $\hat{B}_r$ de B^* peut être dérivé avec des garanties non asymptotiques. Cependant, si les lignes de Y sont corrélées, le modèle exposé précédemment ne peut pas le capturer. Cela peut se produire lorsque les prédicteurs et les cibles observés présentent une dépendance sérielle. Ce problème est au cœur du chapitre 3. En conclusion, la généralisation de ces résultats à des tenseurs d'ordre supérieur suscite un intérêt considérable au sein de la communauté des chercheurs. Nous renvoyons le lecteur à [97] et aux références qui y figurent pour une introduction complète.

Chapter 3 : Régression matricielle bilatérale.

Dans ce chapitre, nous étudions un problème de régression multivariée dans lequel les colonnes et les lignes de la quantité cible Y sont supposées être corrélées. Nous observons la matrice cible $Y \in \mathbb{R}^{n\times p}$ et une matrice de prédiction $X \in \mathbb{R}^{m\times q}$ liées par le modèle de régression matricielle bilatérale (2MR). Ce modèle implique deux matrices de paramètres $A^* \in \mathbb{R}^{n\times m}$ et $B^* \in \mathbb{R}^{q\times p}$ et s'exprime comme suit

$$Y = A^* X B^* + E.$$

La matrice de bruit E est supposée avoir des entrées indépendantes centrées $\sigma-$sousGaussiennes. L'objectif est de dériver des prédicteurs $\hat{A}$ et $\hat{B}$ tels que $\hat{A}X\hat{B}$ reste proche du signal A^*XB^*, sous des hypothèses de faible rang sur A^* et B^*.

Bien que ce modèle n'implique pas de dépendance temporelle, les résultats non asymptotiques obtenus ici peuvent améliorer notre compréhension des séries temporelles autorégressives à valeur matricielle : $Y_t = A^* X_t B^* + E_t$ (voir [47]). Le modèle 2MR englobe également des modèles connus tels que la régression matricielle et la factorisation matricielle. Par exemple, si $n = m$ et A^* est la matrice identité, le modèle 2MR se réduit au modèle de régression matricielle unilatérale $Y = XB^* + E$ (voir [108], [32], [104]). De même, si $m = q$ et que la matrice de prédiction X est la matrice identité de rang m inférieur à la fois à n et à p, le modèle 2MR devient un modèle de factorisation du signal $M^* = A^*B^*$ observé avec du bruit.

Une autre représentation du modèle 2MR se présente sous la forme d'un *vector regression model*. En empilant les colonnes des matrices Y, X et E dans vec(Y), vec(X) et vec(E), respectivement, on obtient

$$\text{vec}(Y)^\top = \text{vec}(X)^\top \cdot (A^*)^\top \otimes B^* + \text{vec}(E)^\top,$$

où $\otimes$ représente le produit tensoriel de deux matrices. Selon cette formulation, nous prédisons un vecteur de lignes de taille np en utilisant un vecteur de lignes de taille mq (la matrice de caractéristiques étant de rang 1) par l'intermédiaire d'un paramètre de taille $(mq) \times (np)$. Cette approche est problématique à moins que la structure de A^* et B^* ne soit triviale. Elle ne tient pas compte de la structure matricielle des caractéristiques et des matrices A^* et B^*, ce qui conduit à des résultats sous-optimaux. L'objectif est de construire des prédicteurs explicites $(\hat{A}_r, \hat{B}_r)$ solutions au risque de prédiction de Frobenius au carré sous contrainte de rang maximal, voir (3.3).

Le théorème 3.2.1 fournit, pour un problème équivalent (3.5), des prédicteurs explicites $\hat{A}_{0r}$ et $\hat{B}_{0r}$ avec une borne supérieure non asymptotique sur le risque de prédiction. Nous remarquons en particulier que cette borne peut être décomposée comme la somme d'un terme de biais, qui est la cause du choix du rang r des prédicteurs, potentiellement inférieur au rang des matrices A^* et B^* et d'un

terme stochastique. L'analyse de ce terme stochastique fait principalement appel à la théorie des matrices aléatoires, voir [129]. Ces prédicteurs permettent de dériver $\hat{A}_r$ et $\hat{B}_r$ la solution du problème d'optimisation initial (3.3). Ce résultat est énoncé dans le Corollaire 3.2.2.

Cependant, dans le problème d'optimisation (3.3), la question de la sélection de r se pose. Nous proposons une procédure adaptative au rang pour y répondre. Nous sélectionnons d'abord le rang $\hat{r}$ en résolvant une version pénalisée par le rang du problème de minimisation du carré de Frobenius, (3.8). Nous considérons ensuite les prédicteurs correspondants $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$. Le risque de prédiction de ces prédicteurs est étudié dans le Théorème 3.2.3. La cohérence de la procédure de sélection des rangs (3.8) est également démontrée dans la Proposition 3.2.6. Toutefois, ces deux résultats sont énoncés à la condition que le paramètre sous-gaussien σ^2 des entrées de la matrice de bruit soit connu.

Enfin, nous proposons une procédure adaptative aux rangs guidée par les données, qui permet de sélectionner $\bar{r}$ et de dériver des prédicteurs $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$. Ces prédicteurs présentent des garanties prouvables non asymptotiques sans que la vraie valeur σ soit connue. Pour ce faire, nous modifions le problème de minimisation pénalisé (3.8) en remplaçant le rang r par $r\widehat{\sigma}_r^2$, voir (3.9), où

$$\widehat{\sigma}_r^2 = \frac{1}{np}\|Y - \hat{A}_r X \hat{B}_r\|_F^2.$$

Les performances de cette procédure de prédiction sont détaillées dans le théorème 3.2.7.

Enfin, comme dans le cas de la régression linéaire standard où l'estimateur BIC est remplacé par sa version relâchée convexe, l'estimateur Lasso, nous comparons les performances de prédiction obtenues à l'aide d'une pénalité de rang à celles obtenues à l'aide d'une pénalité de norme nucléaire, qui sert de relâchement convexe de la pénalité de rang. Plus précisément, nous considérons la version pénalisée par la norme nucléaire de la minimisation du risque de prédiction de Frobenius au carré, voir (3.10). Nous fournissons des solutions $\bar{A}$ et $\bar{B}$ à ce problème dans le Théorème 3.3.1 et dérivons une borne supérieure non asymptotique sur le risque de prédiction correspondant $\|A^*XB^* - \bar{A}X\bar{B}\|_F^2$.

Nous concluons en notant que le modèle de régression matricielle bilatérale souffre d'inconvénients liés à l'identifiabilité. En effet, de nombreux couples de matrices (A, B) résolvent l'équation $M = AXB$ pour une matrice M donnée.

Nous ne pouvons espérer identifier les matrices A et B que dans des conditions très restrictives où $X^\top X$ est de plein rang et où la matrice A ou la matrice B est supposée avoir des valeurs singulières connues, *e.g.* comme un projecteur avec des valeurs singulières 1 ou 0. Peu d'autres configurations sont connues pour être identifiables dans la littérature de la factorisation des matrices, *e.g.* la factorisation des matrices non négatives (NMF), voir [54], NMF pour les topiques-modèles [84], [25], [86] ou la factorisation des matrices de covariance [57].

Topiques-Modèles

Cette section est consacrée à la présentation du cadre de modélisation thématique, qui est au cœur des chapitres 4 et 5. Considérons un corpus comprenant n documents textuels écrits dans une langue caractérisée par un dictionnaire de taille p. Pour analyser et exploiter l'information véhiculée dans ces n documents, l'objectif principal est de dériver une représentation vectorielle pour cet ensemble de documents. Cette expression mathématique permettra d'appliquer des outils analytiques afin d'extraire et d'examiner les informations plus efficacement. Compte tenu de la longueur variable des documents, un simple comptage de l'occurrence de chaque mot ne serait pas pertinent. Par conséquent, pour chaque document, l'accent est mis sur la fréquence d'apparition des mots individuels. Chaque document peut

ainsi être représenté comme un point dans le simplexe de $\mathbb{R}^p$. Cela implique que l'ensemble du corpus est représenté comme un ensemble de n points à l'intérieur du simplexe. Il est important de noter que l'ordre des documents n'a pas d'importance dans ce contexte. En outre, nous supposons que ces n points ne sont pas linéairement indépendants mais couvrent un sous-espace de $\mathbb{R}^p$ de dimension $K \ll \min(n, p)$. Interprété comme le nombre de sujets discutés dans le corpus, K joue un rôle crucial dans la capture de la structure sous-jacente. L'objectif principal est de trouver un encastrement de ces n points dans l'espace de dimension inférieure $\mathbb{R}^K$. Par conséquent, il s'agit d'identifier une correspondance entre $\mathbb{R}^p$ et $\mathbb{R}^K$ de telle sorte que les n points initiaux de $\mathbb{R}^p$ puissent être effectivement intégrés dans $\mathbb{R}^K$ par le biais de cette correspondance.

Dans un contexte plus formel, chaque document $jdans[n]$ est modélisé comme une collection de N_j mots tirés d'un dictionnaire de taille p. Chaque document suit une distribution discrète π_j^* sur le simplexe de $\mathbb{R}^p$. Pour chaque document $j \in [n]$, le vecteur de dimension p Y_j des fréquences de mots est observé et supposé suivre une distribution multinomiale centrée sur π_j^* :

$$N_j \boldsymbol{Y}_j \sim \text{Multinomial}_p(N_j, \pi_j^*). \tag{6.3}$$

Cependant, dans les exemples réels, seuls quelques sujets différents sont abordés dans d'énormes corpus de documents. Cela conduit à supposer que la matrice de probabilité mot-document $\Pi^* = (\pi_1^*, \ldots, \pi_n^*) \in \mathbb{R}^{p \times n}$ est de rang $K \ll \min(n, p)$, le nombre de sujets, et peut être factorisée comme suit :

$$\Pi^* = A^* W^*, \tag{6.4}$$

où $A^* \in \mathbb{R}^{p \times K}$ est la matrice de probabilité mot-sujet et $W^* \in \mathbb{R}^{K \times n}$ est la matrice de probabilité sujet-document.

Ce cadre suppose que la probabilité d'occurrence du mot $idans[p]$ dans un document traitant du sujet $kdans[K]$ est indépendante du document lui-même. Plus précisément, le vecteur de probabilité π_j^* du document j, appelé vecteur de probabilité mot-document, est une combinaison convexe de K vecteurs de probabilité mot-sujet avec des poids correspondant à l'attribution de K sujets. D'un point de vue probabiliste, cela peut être exprimé par la formule de la probabilité totale, comme suit :

$$\mathbb{P}(\text{mot } i | \text{document } j) = \sum_{k=1}^{K} \mathbb{P}(\text{mot } i | \text{sujet } k) \mathbb{P}(\text{sujet } k | \text{document } j),$$

L'objectif principal dans le cadre du modèle thématique traditionnel est de récupérer A^* et/ou W^* sur la base des observations $\boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n$ avec ou sans un nombre fixe connu de sujets K. L'estimation des matrices A^* et W^* répond à des objectifs distincts. En effet, l'estimation de la matrice A^* permet de discerner la distribution des mots dans le dictionnaire pour un sujet donné, tandis que l'estimation de la matrice W^* révèle la distribution des sujets pour un document donné.

Il convient de noter qu'en l'absence de bruit, c'est-à-dire lorsque la matrice Π^* est observée, la récupération de A^* et W^* devient un cas de factorisation de matrices non négatives. Le problème de la factorisation de matrices non négatives (NMF) a été largement étudié, les algorithmes attirant l'attention en raison de leur capacité à générer des facteurs avec des contraintes non négatives, ce qui améliore l'interprétabilité. Généralement, la NMF est formulée comme la minimisation d'une fonction de coût régularisée [94, 93, 112], présentant des défis d'optimisation non convexe, en particulier dans les scénarios où de nombreux mots sont absents dans un seul document ($N \ll p$). La principale limitation de la NMF est que la résolution du problème exact de la NMF, c'est-à-dire, en supposant un rang connu K de $\Pi^* \mathbb{R}^{p \times n}$ et en récupérant les matrices $A^* \mathbb{R}^{p \times K}$ et $W^* \mathbb{R}^{K \times n}$ telles que $A^* W^* = \Pi^*$,

sans aucune hypothèse supplémentaire, est NP-hard, voir [127]. Ce résultat implique la nécessité d'hypothèses supplémentaires pour garantir l'existence d'algorithmes rapides capables d'estimer A^* et/ou W^*. De plus, les algorithmes NMF sont confrontés à un problème d'identifiabilité. Il est concevable de trouver différentes matrices non négatives $(A_1^*, W_1^*) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{K\times n}$ et $(A_2^*, W_2^*) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{K\times n}$ tel que $A_1^* W_1^* = A_2^* W_2^*$. Des hypothèses supplémentaires sont nécessaires pour garantir l'unicité de la représentation. La première de ces hypothèses est l'hypothèse de *séparabilité* et a été initialement introduite par [54]. Elle garantit l'unicité de la NMF. Cette hypothèse a ensuite été incorporée dans le cadre du modèle thématique par [8], avec l'interprétation que, pour chaque thème, il existe certains mots qui se produisent exclusivement dans ce thème spécifique. Ces mots sont appelés "mots d'ancrage". L'hypothèse *mot d'ancrage* a ensuite été adoptée dans la plupart des publications sur les modèles thématiques.

Assumption 9 (Anchor word assumption) *Pour chaque sujet $k \in [K]$, il existe au moins un mot j tel que $[A^*]_{jk} > 0$ et $[A^*]_{jl} = 0$ pour $l \in [K]\backslash\{k\}$.*

Le modèle (1.4) suppose que la matrice mot-sujet et la matrice sujet-document sont statiques. En outre, il suppose que les documents sont échangeables au sein de la collection. En effet, le modèle reste le même en cas de permutation des colonnes de la matrice observée $\boldsymbol{Y}$.

Des travaux récents abordent les aspects algorithmiques et donnent des résultats d'inférence sur le problème de l'estimation de la matrice A^* dans un cadre statique sous l'hypothèse *mots d'ancrage*. Par exemple, les auteurs de [84] proposent un estimateur $\hat{A}$ atteignant les vitesses minimax pour A^* dense, *i.e.* non parcimonieuse, avec un K fixe et connu. La procédure de [84] effectue une SVD sur une version normalisée de la matrice Y suivie d'une recherche exhaustive sur un simplexe de dimension p. Pour K inconnu et A^* dense, les auteurs de [24] considèrent $\hat{A}_K$, atteignant les vitesses optimales minimax dans ce cadre. La procédure de [24] commence par la récupération des mots d'ancrage et dérive ensuite un estimateur à partir d'une version normalisée de $YY^\top$. Les auteurs de[25] étudient l'estimation de A^* sous l'hypothèse de parcimonie avec K inconnu, en proposant une procédure d'estimation optimale minimax $\hat{A}_{sparse}$ de A^*. La procédure de [25] se concentre principalement sur l'estimation de la partie de A^* correspondant aux mots non ancrés. Pour s'adapter à la parcimonie de A^*, leur algorithme nécessite également la résolution d'un programme quadratique pour chaque ligne non ancrée. Récemment, plusieurs articles ont également étudié le problème de l'estimation de la matrice W^* statique sous différentes hypothèses. Lorsque A^* est connue et que W^* est supposée parcimonieuse, [23] propose un estimateur du maximum de vraisemblance (MLE) pour W^*. Leur analyse a prouvé que le MLE est à la fois minimax optimal et adaptatif à la parcimonie. Lorsque A^* est inconnue, [23] estime W^* en optimisant la fonction de vraisemblance correspondant à un estimateur plug-in $\hat{A}$ de A^*. Par conséquent, l'erreur d'estimation de W^* dans leur procédure dépend de la qualité de l'estimation de A^* par $\hat{A}$. Lorsque A^* et W^* sont tous deux inconnues et que les colonnes de W^* sont supposées peu nombreuses, K pouvant être grand, [140] propose des procédures computationellement efficaces pour estimer ces deux matrices. En outre, il est possible d'estimer directement W^* en supposant une structure supplémentaire. Ainsi, [86] suppose une autre version de l'hypothèse *mot d'ancrage*, appelée *document d'ancrage*. Cette hypothèse signifie que pour chaque sujet, il existe un document qui ne traite que de ce sujet. Leur procédure, appelée Successive Projection Overlapping Clustering (SPOC), s'inspire de l'algorithme de projection successive (SPA). L'idée est de commencer par la décomposition en valeurs singulières (SVD) de la matrice Y, et de lancer une procédure itérative qui, à chaque étape, choisit la ligne de norme maximale de la matrice composée de vecteurs singuliers. Elle projette ensuite sur le sous-espace linéaire orthogonal à la ligne sélectionnée.

Chapitre 4 : Topiques-modèles dynamique : cas oracle Dans ce chapitre, nous supposons que des lots de n documents sont collectés en T étapes dans le temps. L'objectif est de prendre en compte l'aspect temporel de la collecte de documents et de refléter l'évolution dynamique des thèmes abordés dans les corpus. Nous supposons que la matrice de probabilité sujet-document W^* suit un modèle autorégressif simplex-valué d'ordre un. Par conséquent, la matrice $\boldsymbol{W}^{1:T} := (\boldsymbol{W}^1, \ldots, \boldsymbol{W}^T)$ est maintenant considérée comme aléatoire. Plus précisément, à chaque pas de temps t, la distribution des sujets donnés par un document est une combinaison linéaire de la distribution précédente et d'un bruit distribué par Dirichlet, qui détermine l'évolution temporelle des sujets. Plus précisément, nous considérons que pour tout $t \in [T-1]$:

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t$$

où $c^* \in (0, 1)$, et chaque $\boldsymbol{\Delta}^t$ est une matrice de bruit de taille $K \times n$ telle que les colonnes sont indépendamment et identiquement tirées d'une distribution de Dirichlet $\mathcal{D}(\theta^*)$ ayant pour paramètre $\theta^* \in \mathbb{R}_+^K$. L'objectif de ce chapitre est d'estimer les paramètres de ce modèle autorégressif en supposant que la matrice de probabilité mot-document $\boldsymbol{\Pi}^{1:T} := (\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_T)$ est disponible. Nous appelons ce cadre le cas de l'oracle. Nous commençons par étudier les propriétés spectrales de la matrice de covariance empirique $\Sigma_{\boldsymbol{W}}^{1:T} := \frac{1}{nT} \left(\boldsymbol{W}^{1:T}\right) \left(\boldsymbol{W}^{1:T}\right)^\top$.

En particulier, dans le Théorème 4.3.3, nous fournissons un contrôle sur sa plus petite valeur propre et montrons qu'elle est bornée par des quantités dépendant de c^*, α et $\tilde{\theta}^*$ avec une grande probabilité. Dans la Proposition 4.3.1, nous contrôlons sa plus grande valeur propre en la bornant presque sûrement par des quantités dépendant exclusivement de K. Ces résultats légitiment une hypothèse forte que nous faisons sur le spectre de cette matrice. À la suite du travail effectué dans [84], nous présentons une procédure algorithmique basée sur la SVD qui récupère exactement la matrice de probabilité mot-sujet A^*. La projection de la matrice de probabilité mot-document $\boldsymbol{\Pi}^{1:T}$ sur A^* permet de récupérer exactement la matrice de probabilité sujet-document $\boldsymbol{W}^{1:T}$. Nous estimons ensuite les paramètres $\tilde{\theta}^*$, c^* et α avec les estimateurs définis respectivement dans (4.8), (4.9) et (4.11). Des bornes non asymptotiques sur leurs erreurs d'estimation sont dérivées respectivement dans le Théorème 4.4.1, le Théorème 4.4.2 et le Théorème 4.4.3. En particulier, nous prouvons qu'il existe des constantes absolues C_1, $C_2 > 0$ telles que :

$$\mathbb{P}\left[\max\{\|\hat{\theta} - \tilde{\theta}^*\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\} \leq C_1 \cdot \sqrt{\frac{\log(nT)}{nT}}\right] \geq 1 - \frac{C_2}{nT}.$$

En particulier, la dimension du vecteur θ^*, qui est le nombre K de sujets, n'apparaît pas dans ces bornes grâce aux propriétés du bruit de Dirichlet.

Chapitre 5 : Topique-modèles dynamiques : cas réel Dans ce chapitre, nous considérons le même cadre que dans le chapitre 4 sans que la matrice de probabilité mot-document $\boldsymbol{\Pi}^{1:T}$ ne soit plus disponible. Nous supposons que nous n'avons accès qu'à la matrice de fréquence mot-document $\boldsymbol{Y}^{1:T}$. Ensuite, nous définissons d'abord les versions empiriques des quantités impliquées dans la procédure exposée précédemment, en récupérant A^*. Cette procédure empirique adaptée conduit à un estimateur $\hat{A}$ de A^*. Nous présentons une étude minutieuse de cette procédure d'estimation. Plus précisément, nous donnons des bornes supérieures explicites jusqu'à des facteurs logarithmiques et leur dépendance à l'égard de toutes les dimensions des matrices d'apparition. Nous projetons ensuite la matrice de fréquence mot-document $\boldsymbol{Y}^{1:T}$ sur la matrice mot-sujet estimée $\hat{A}$. Il en résulte une matrice sujet-document estimée $\hat{W}^{1:T}$. Les estimateurs des paramètres autorégressifs, introduits dans le chapitre 4, sont adaptés à ce cadre. Des bornes non asymptotiques sur leur erreur d'estimation sont dérivées respectivement dans Theoreme 5.4.1, Theoreme 5.4.2 et Theoreme 5.4.3. En particulier, nous prouvons

que pour chaque estimateur, il existe des constantes absolues C_1, $C_2 > 0$ et a, $b > 0$ telles que :

$$\mathbb{P}\left[\max\{\|\hat{\theta} - \tilde{\theta}^*\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\} \leq C_1 \cdot K^a p^b \left(\sqrt{\frac{\log(nT)}{nT}} + \sqrt{\frac{\log(nT)}{N}}\right)\right] \geq 1 - \frac{C_2}{nT}.$$

Par conséquent, les taux de convergence obtenus dans le cas réel montrent des contributions additives du bruit Dirichlet, qui détermine la probabilité des sujets pour des documents donnés, et du modèle multinomial des nombres de mots. De plus, pour les documents très longs, c'est-à-dire lorsque $N \gg nT$, les taux de convergence ne sont influencés par le bruit de Dirichlet que par des termes multiplicatifs du nombre de sujets K et de la taille du vocabulaire p.

Bibliographie

[1] Pierre Alquier, Karine Bertin, Paul Doukhan, and Rémy Garnier. High-dimensional var with low-rank transition. *Statistics and Computing*, 30(4) :1139–1153, 2020.

[2] Pierre Alquier and Nicolas Marie. Matrix factorization for multivariate time series analysis. *Electronic Journal of Statistics*, 13(2) :4346 – 4366, 2019.

[3] Pierre Alquier, Nicolas Marie, and Amélie Rosier. Tight risk bound for high dimensional time series completion. *Electronic Journal of Statistics*, 16(1) :3001 – 3035, 2022.

[4] G. Aneiros, R. Cao, R. Fraiman, C. Genest, and P. Vieu. Recent advances in functional data analysis and high-dimensional statistics. *J. Multivar. Anal.*, 170 :3–9, 2019.

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73 :243–272, 2008.

[6] Ery Arias-Castro, Sébastien Bubeck, and Gabor Lugosi. Detecting positive correlations in a multivariate sample. *Bernoulli*, 21(1) :209–241, 02 2015.

[7] Ery Arias-Castro, Sébastien Bubeck, and Gábor Lugosi. Detection of correlations. *Ann. Stat.*, 40(1) :412–435, 02 2012.

[8] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models – going beyond svd. *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10, 2012.

[9] David Azriel and Yosef Rinott. Optimal selection of sample-size dependent common subsets of covariates for multi-task regression prediction. *Electronic Journal of Statistics*, 15(2) :4966 – 5013, 2021.

[10] Francis R Bach. Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 9 :1019–1048, 2008.

[11] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1) :135–171, 2003.

[12] Peiliang Bai, Yue Bai, Abolfazl Safikhani, and George Michailidis. Multiple change point detection in structured var models : the vardetect r package. *arXiv preprint arXiv :2105.11007*, 2021.

[13] Yue Bai and Abolfazl Safikhani. A unified framework for change point detection in high-dimensional linear models. *arXiv preprint arXiv :2207.09007*, 2022.

[14] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[15] Sumanta Basu, Xianqi Li, and George Michailidis. Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5) :1207–1222, 2019.

[16] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4) :1535 – 1567, 2015.

[17] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567) :47–55, 2015.

[18] Pierre C. Bellec. Concentration of quadratic forms under a Bernstein moment assumption. *ArXiv e-prints*, 2019.

[19] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets lasso : Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B) :3603–3642, 2018.

[20] Nayel Bettache and Cristina Butucea. Two-sided matrix regression. *arXiv preprint arXiv :2303.04694*, 2023.

[21] Nayel Bettache, Cristina Butucea, and Marianne Sorba. Fast nonasymptotic testing and support recovery for large sparse toeplitz covariance matrices. *Journal of Multivariate Analysis*, 190 :104883, 2022.

[22] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4) :1705 – 1732, 2009.

[23] Xin Bing, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Likelihood estimation of sparse topic distributions in topic models and its applications to wasserstein document distance calculations. *The Annals of Statistics*, 50(6) :3307–3333, 2022.

[24] Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3) :1765 – 1796, 2020.

[25] Xin Bing, Florentina Bunea, and Marten Wegkamp. Optimal estimation of sparse topic models. *The Journal of Machine Learning Research*, 21(1) :7189–7233, 2020.

[26] Xin Bing and Marten H. Wegkamp. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.*, 47(6) :3157–3184, 2019.

[27] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138 :33–73, 2007.

[28] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[29] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[30] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[31] Peter J Brockwell, Richard A Davis, and Stephen E Fienberg. *Time series : theory and methods : theory and methods*. Springer Science & Business Media, 1991.

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

[33] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1(none) :169 – 194, 2007.

[34] Cristina Butucea and Yuri I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B) :2652–2688, 2013.

[35] Cristina Butucea and Rania Zgheib. Sharp minimax tests for large covariance matrices and adaptation. *Electron. J. Statist.*, 10(2) :1927–1972, 2016.

[36] Cristina Butucea and Rania Zgheib. Sharp minimax tests for large toeplitz covariance matrices with repeated observations. *J. Multivar. Anal.*, 146(C) :164–176, 2016.

[37] T Tony Cai, Zhao Ren, and Harrison H Zhou. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2) :101–143, 2013.

[38] T Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices : Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1) :1–59, 2016.

[39] T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.*, 38(4) :2118–2144, 2010.

[40] T. Tony Cai and Harrison H. Zhou. Minimax estimation of large covariance matrices under l1-norm. *Statistica Sinica*, 22(4) :1319–1349, 2012.

[41] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.*, 106(494) :672–684, 2011.

[42] Tony Cai, Weidong Liu, and Yin Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Stat. Assoc.*, 108(501) :265–277, 2013.

[43] Tony Cai and Zongming Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B) :2359–2388, 11 2013.

[44] Jinyuan Chang, Jing He, Lin Yang, and Qiwei Yao. Modelling matrix time series via a tensor CP-decomposition. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 85(1) :127–148, 01 2023.

[45] Elynn Y Chen, Ruey S Tsay, and Rong Chen. Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, 2019.

[46] Minshuo Chen, Lin Yang, Mengdi Wang, and Tuo Zhao. Dimensionality reduction for stationary time series via stochastic nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Adv. Neural Inf. Process Syst.*, volume 31. Curran Associates, Inc., 2018.

[47] Rong Chen, Han Xiao, and Dan Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1) :539–560, 2021.

[48] Rong Chen, Han Xiao, and Dan Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1) :539–560, 2021.

[49] Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537) :94–116, 2022.

[50] Violetta Dalla, Liudas Giraitis, and Peter M Robinson. Asymptotic theory for time series with changing mean and variance. *Journal of econometrics*, 219(2) :281–313, 2020.

[51] Zinsou-Max Debaly and Lionel Truquet. Multivariate time series models for mixed data. *Bernoulli*, 29(1) :669 – 695, 2023.

[52] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *ArXiv*, abs/1907.05545, 2019.

[53] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.*, 32(3) :962–994, 2004.

[54] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts ? *Advances in neural information processing systems*, 16, 2003.

[55] Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5) :2015 – 2039, 2023.

[56] Farida Enikeeva, Olga Klopp, and Mathilde Rousselot. Change point detection in low-rank var processes. *arXiv preprint arXiv :2305.00311*, 2023.

[57] Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6) :3320, 2011.

[58] Qin Fang, Shaojun Guo, and Xinghao Qiao. Finite sample theory for high-dimensional functional/scalar time series with applications. *Electronic Journal of Statistics*, 16(1) :527–591, 2022.

[59] Zhe Fei and Yi Li. Estimation and inference for high dimensional generalized linear models : A splitting and smoothing approach. *Journal of Machine Learning Research*, 22(58) :1–32, 2021.

[60] Thomas J Fisher. On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size. *J. Stat. Plan. Inference*, 142(1) :312–326, 2012.

[61] Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic factor model : one-sided estimation and forecasting. *Journal of the American statistical association*, 100(471) :830–840, 2005.

[62] Wayne A Fuller. *Introduction to statistical time series*. John Wiley & Sons, 2009.

[63] Christophe Giraud. Low rank multivariate regression. *Electron. J. Stat.*, 5 :775–799, 2011.

[64] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[65] Mihai Giurcanu and Vladimir Spokoiny. Confidence estimation of the covariance function of stationary and locally stationary processes. *Stat. Decis.*, 22(4) :283–300, 2004.

[66] A. Goia and P. Vieu. An introduction to recent advances in high/infinite dimensional statistics. *J. Multivar. Anal.*, 146 :1–6, 2016.

[67] Gene H Golub. Some modified matrix eigenvalue problems. *SIAM review*, 15(2) :318–334, 1973.

[68] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

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

[70] Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

[71] James D Hamilton. *Time series analysis*. Princeton university press, 2020.

[72] Yuefeng Han, Rong Chen, Cun-Hui Zhang, and Qiwei Yao. Simultaneous decorrelation of matrix time series. *Journal of the American Statistical Association*, pages 1–13, 2023.

[73] Edward J Hannan and Laimonis Kavalieris. Multivariate linear time series models. *Advances in Applied Probability*, 16(3) :492–561, 1984.

[74] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17(none) :1 – 6, 2012.

[75] Nan-Jung Hsu, Hsin-Cheng Huang, and Ruey S Tsay. Matrix autoregressive spatio-temporal models. *Journal of Computational and Graphical Statistics*, 30(4) :1143–1155, 2021.

[76] Nan-Jung Hsu, Hsin-Cheng Huang, and Ruey S. Tsay. Matrix autoregressive spatio-temporal models. *J. Comput. Graph. Statist.*, 30(4) :1143–1155, 2021.

[77] Nan-Jung Hsu, Hsin-Cheng Huang, Ruey S Tsay, and Tzu-Chieh Kao. Rank-r matrix autoregressive models for modeling spatio-temporal data. *Statistics and Its Interface*, 17(2) :275–290, 2024.

[78] Nan-Jung Hsu, Hung-Lin Hung, and Ya-Mei Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, 52(7) :3645–3657, 2008.

[79] Marie Hušková, Zuzana Prášková, and Josef G Steinebach. Estimating a gradual parameter change in an ar (1)-process. *Metrika*, 85(7) :771–808, 2022.

[80] Yu. I. Ingster. Adaptive detection of a signal of growing dimension I. *Math. Methods Statist.*, 10(4) :395–421, 2001.

[81] Yu. I. Ingster. Adaptive detection of a signal of growing dimension II. *Math. Methods Statist.*, 11(1) :37–68, 2002.

[82] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

[83] S John. Some optimal multivariate tests. *Biometrika*, 58(1) :123–127, 1971.

[84] Zheng Tracy Ke and Minzhe Wang. Using svd for topic modeling. *Journal of the American Statistical Association*, pages 1–16, 2022.

[85] Olga Klopp, Yu Lu, Alexandre B. Tsybakov, and Harrison H. Zhou. Structured matrix estimation and completion. *Bernoulli*, 25(4B) :3883–3911, 2019.

[86] Olga Klopp, Maxim Panov, Suzanne Sigalla, and Alexandre Tsybakov. Assigning topics to documents by successive projections. *arXiv preprint arXiv :2107.03684*, 2021.

[87] Tonu Kollo. *Advanced multivariate statistics with matrices*. Springer, 2005.

[88] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5) :2302–2329, 2011.

[89] Jens-Peter Kreiss, Efstathios Paparoditis, and Dimitris N. Politis. On the range of validity of the autoregressive sieve bootstrap. *Ann. Stat.*, 39(4) :2103–2130, 08 2011.

[90] Chen Kun, Dong Hongbo, and Chan Kung-Sik. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100 :901–920, 2013.

[91] Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series : Inference for the number of factors. *The Annals of Statistics*, 40(2) :694 – 726, 2012.

[92] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces : isoperimetry and processes*. Springer Science & Business Media, 2013.

[93] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

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

[95] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

[96] Zebang Li and Han Xiao. Multi-linear tensor autoregressive models. *arXiv preprint arXiv :2110.00928*, 2021.

[97] Eric F Lock. Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3) :638–647, 2018.

[98] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4) :2164 – 2204, 2011.

[99] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[100] Pascal Massart. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[101] Andreas Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7 :117–139, 2006.

[102] Hisao Nagao. On some test criteria for covariance matrix. *Ann. Stat.*, pages 700–709, 1973.

[103] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Proceedings of the 27 th International Conference on Machine Learning*, 2011.

[104] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4) :538–557, 2012.

[105] William B. Nicholson, Ines Wilms, Jacob Bien, and David S. Matteson. High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166) :1–52, 2020.

[106] F Ninio. A simple proof of the perron-frobenius theorem for positive symmetric matrices. *Journal of Physics A : Mathematical and General*, 9(8) :1281, 1976.

[107] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1) :1 – 47, 2011.

[108] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1) :1–47, 2011.

[109] Daniel Peña and Víctor J Yohai. A review of outlier detection and robust estimation methods for high dimensional time series data. *Econometrics and Statistics*, 2023.

[110] Liuhua Peng, Song Xi Chen, and Wen Zhou. More powerful tests for sparse high-dimensional covariances matrices. *J. Multivar. Anal.*, 149 :124–143, 2016.

[111] Alfio Quarteroni. The role of statistics in the era of big data : A computational scientist' perspective. *Stat. Probab. Lett.*, 136 :63–67, 2018. The role of Statistics in the era of big data.

[112] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[113] Gregory C Reinsel. *Elements of multivariate time series analysis*. Springer Science & Business Media, 2003.

[114] Philippe Rigollet and Alexandre Tsybakov. Exponential Screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2) :731 – 771, 2011.

[115] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

[116] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2) :887–930, 2011.

[117] Alexis Rosuel, Philippe Loubaton, and Pascal Vallet. On the asymptotic distribution of the maximum sample spectral coherence of gaussian time series in the high dimensional regime. *arXiv preprint arXiv :2107.02891*, 2021.

[118] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *ArXiv e-prints*, June 2013.

[119] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[120] Song Song and Peter J. Bickel. Large vector auto regressions. *arXiv : Machine Learning*, 2011.

[121] V. Spokoiny and M. Zhilova. Sharp deviation bounds for quadratic forms. *Math. Methods Stat.*, 22(2) :100–113, Apr 2013.

[122] Muni S Srivastava. Some tests concerning the covariance matrix in high dimensional data. *J. Japan Stat. Soc.*, 35(2) :251–272, 2005.

[123] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3) :505–563, 1996.

[124] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1) :267–288, 1996.

[125] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2) :1–230, 2015.

[126] Ruey S Tsay. *Multivariate time series analysis : with R and financial applications*. John Wiley & Sons, 2013.

[127] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3) :1364–1377, 2010.

[128] Raja P Velu, Gregory C Reinsel, and Dean W Wichern. Reduced rank models for multiple time series. *Biometrika*, 73(1) :105–118, 1986.

[129] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv :1011.3027*, 2010.

[130] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[131] Martin J Wainwright. *High-Dimensional Dtatistics : A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[132] AT Walden and A Serroukh. Wavelet analysis of matrix–valued time–series. *Proceedings of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences*, 458(2017) :157–179, 2002.

[133] Chong Wang, David M. Blei, and David E. Heckerman. Continuous time dynamic topic models. In *Conference on Uncertainty in Artificial Intelligence*, 2008.

[134] Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv :1909.06359*, 2019.

[135] Di Wang and Ruey S. Tsay. Rate-optimal robust estimation of high-dimensional vector autoregressive models. *The Annals of Statistics*, 51(2) :846 – 877, 2023.

[136] Dong Wang, Xialu Liu, and Rong Chen. Factor models for matrix-valued high-dimensional time series. *Journal of econometrics*, 208(1) :231–248, 2019.

[137] Xuerui Wang and Andrew McCallum. Topics over time : a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[138] Xing Wei, Jimeng Sun, and Xuerui Wang. Dynamic mixture models for multiple time-series. In *International Joint Conference on Artificial Intelligence*, 2007.

[139] Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2) :1124 – 1142, 2020.

[140] Ruijia Wu, Linjun Zhang, and T Tony Cai. Sparse topic modeling : Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, 118(543) :1849–1861, 2023.

[141] H Xiao, Y Han, R Chen, and C Liu. Reduced rank autoregressive models for matrix time series. *Journal of Business and Economic Statistics*, 2022.

[142] Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR, 2023.

[143] Yong-Li Xu, Di-Rong Chen, and Han-Xiong Li. Least Square Regularized Regression for Multitask Learning. *Abstract and Applied Analysis*, 2013(SI32) :1 – 7, 2013.

[144] Jianxin Yin and Hongzhe Li. Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107 :119–140, 2012.

[145] Yi Yu. A review on minimax rates in change point detection and localisation. *arXiv preprint arXiv :2011.01857*, 2020.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Séries temporelles matricielles en grande dimension

Mots clés : Séries temporelles ; Factorisation de matrices ; Régression linéaire matricielle ; statistiques en grande dimension ; matrices aléatoires ; topique-modèles.

Résumé : L'objectif de cette thèse est de modéliser des séries temporelles à valeurs matricielles dans un cadre de grande dimension. Pour ce faire, la totalité de l'étude est présentée dans un cadre non asymptotique. Nous fournissons d'abord une procédure de test capable de distinguer dans le cas de vecteurs ayant une loi centrée stationnaire si leur matrice de covariance est égale à l'identité ou si elle possède une structure de Toeplitz sparse. Dans un second temps, nous proposons une extension de la régression linéaire matricielle de faible rang à une régression à deux paramètres matriciels qui créent des corrélations entre les lignes et les colonnes des observations. Enfin nous introduisons et estimons un topiques-modèle dynamique où l'espérance des observations est factorisée en une matrice statique et une matrice qui évolue dans le temps suivant un processus autorégressif d'ordre un à valeurs dans un simplexe.

Title : Matrix-valued Time Series in High Dimension

Keywords : Time series ; matrix factorisation ; matrix linear regression ; high-dimensional statistics ; random matrices ; topic models.

Abstract : The objective of this thesis is to model matrix-valued time series in a high-dimensional framework. To this end, the entire study is presented in a non-asymptotic framework. We first provide a test procedure capable of distinguishing whether the covariance matrix of centered random vectors with centered stationary distribution is equal to the identity or has a sparse Toeplitz structure. Secondly, we propose an extension of low-rank matrix linear regression to a regression model with two matrix-parameters which create correlations between the rows and he columns of the output random matrix. Finally, we introduce and estimate a dynamic topic model where the expected value of the observations is factorizes into a static matrix and a time-dependent matrix following a simplex-valued auto-regressive process of order one.

Institut Polytechnique de Paris
91120 Palaiseau, France