HYPOTHESIS TEST FOR MANIFOLDS AND NETWORKS

A Dissertation
Presented to
The Academic Faculty

By

Rui Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

May 2021

HYPOTHESIS TEST FOR MANIFOLDS AND NETWORKS

Thesis committee:

Dr. Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Andy Sun
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Alexander Shapiro
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Shihao Yang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Feng Qiu
Energy Systems Division
Argonne National Laboratory

Date approved: April 21, 2021

ACKNOWLEDGMENTS

My sincere thanks go to Professor Yao Xie and Professor Alexander Shapiro for being my advisors. They are very supportive and patient to me all the time. I am really grateful that they spend so much time answering my questions, meeting with me, and explaining every detail. They have taught me a lot about mathematics and statistics and inspired me a lot in my research. They give me many interesting projects, and I really enjoy working with them. It's my great pleasure to have their advice for my pursuit of Ph.D. and my career development.

I would also like to thank Prof. Andy Sun, Prof. Shihao Yang, and Dr. Feng Qiu for serving on my thesis committee. Special thanks go to Dr. Feng Qiu, my mentor during my summer intern at Argonne national laboratory. With his guidance and help, I had a very good time in Argonne with many talented researchers.

Many thanks go to the H.Milton Stewart School of Industrial and Systems Engineering (ISyE). Staffs of ISyE are very helpful and informative. They provide us with a very good work environment.

Last but not least, I would like to thank my wife, Xinyuan Zhao, for her unchanging love and understanding. She is very patient and supportive of me. She has given me so much excellent guidance in writing and communication. Also, I would like to thank my parents for their love and support of my pursuit of Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Statistical inference of high-dimensional data is crucial for science and engineering. Such high-dimensional data are often structured. For example, they can be data from a certain manifold or from a large network. Motivated by the problems that arise in recommendation systems, power systems and social media etc., this dissertation aims to provide statistical modeling for such problems and perform statistical inferences. This dissertation focus on two problems. (i) statistical modeling for smooth manifold and inferences for the corresponding characteristic rank; (ii) detection of change-points for sequential data in a network.

In chapter 2, we study the problem of matrix completion. From a geometric perspective, we address the following questions: (i) what is the minimum achievable rank in the minimum rank matrix completion (MRMC) problem? (ii) Under what conditions, there will be a locally unique solution for MRMC problem? We also provide a statistical model for low rank matrix approximation problems. With such a model, we present a statistical test of the rank. With numerical experiments, we verify our theoretical results and show the performance of the proposed test procedure.

In chapter 3, we generalize the results in chapter 2. We develop a general theory for testing the goodness-of-fit of non-linear models. The observation noise is additive Gaussian. Our main result shows that the "residual" of the model fit (by solving a non-linear least-square problem) follows a (possibly non-central) χ^2 distribution. The natural use of our result is to select the order of a model via a sequential test procedure by choosing between two nested models. We demonstrate the applications of this general theory in the settings of real and complex matrix completion from incomplete and noisy observations, signal source identification, and determining the number of hidden nodes in neural networks.

In chapter 4, we develop an online change-point detection procedure for power system's cascading failure using multi-dimensional measurements over the networks. We incorpo-

rate the cascading failure's characteristic into the detection procedure and model multiple changes caused by cascading failures using a diffusion process over networks. The model captures the property that the risk of component failing increases as more components around it fail. Our change-point detection procedure using the generalized likelihood ratio statistics assuming unknown post-change parameters of the measurements and the true failure time (change-points) at each node. We also provide a fast algorithm to perform the change-points detection. Numerical experiments show that our proposed method demonstrates good performance and can scale up to large systems.

In chapter 5, we proposed a change-point detection procedure by scan score statistics in a multivariate Hawkes network. Our scan score statistics are computationally efficient since we don't need to compute the estimates of the post-change parameters, which is of importance for online detection. We present the theoretical results of our proposed procedure, including the analysis of the false alarm rate (FAR) and average run length (ARL) of the procedure under null hypothesis. We use simulation studies to testify our theoretical results and compare our method with an existing change-point detection procedure with generalized likelihood ratio statistics. We also apply our proposed procedure in real-world data such as memetracker and the stock market, which shows promising results in detecting an abrupt change in the network.

CHAPTER 1
INTRODUCTION

Statistical inference of high-dimensional data is crucial for science and engineering. Such high-dimensional data are often structured. For example, they can be data from a certain manifold or from a large network. Motivated by the problems that arise in recommendation systems, power systems and social media etc., this dissertation aims to provide statistical modeling for such problems and perform statistical inferences. This dissertation focus on two topics. (i) statistical modeling for smooth manifold and inferences for the corresponding characteristic rank; (ii) detection of change-points for sequential data in a network.

A typical problem of first topic is matrix completion problem. Matrix completion (e.g.,[1, 2, 3]) is a fundamental problem in signal processing and machine learning, which studies the recovery of a low-rank matrix from an observation of a subset of its entries. It has attracted a lot attention from researchers and practitioners and there are various motivating real-world applications including recommender systems and the Netflix challenge (see a recent overview in [4]). A popular approach for matrix completion is to find a matrix of minimal rank satisfying the observation constraints. Due to the non-convexity of the rank function, popular approaches are convex relaxation (see, e.g., [5]) and nuclear norm minimization. There is a rich literature, both in establishing performance bounds, developing efficient algorithms and providing performance guarantees. In chapter 2, we study the problem of matrix completion. From a geometric perspective, we address the following questions: (i) what is the minimum achievable rank in the minimum rank matrix completion (MRMC) problem? (ii) Under what conditions, there will be a locally unique solution for MRMC problem? We also provide a statistical model for low rank matrix approximation problems. With such a model, we present a statistical test of the rank. With numerical experiments, we verify our theoretical results and show the performance of the proposed

test procedure.

To extend the problem in chapter 2 to a more general setting, we are interested in model selection for non-linear models. Although much has been done for model selection in linear models, it is unclear how to select models given noisy observations in the non-linear setting, especially when there are underlying manifold structures. Such problems arise very often in machine learning and signal processing applications. For instance, how to select the rank of a low-rank matrix, decide the number of hidden nodes in neural networks, and determine the number of signal sources when observing their mixture. In chapter 3, we develop a general theory for testing the goodness-of-fit of non-linear models. The observation noise is additive Gaussian. Our main result shows that the "residual" of the model fit (by solving a non-linear least-square problem) follows a (possibly non-central) χ^2 distribution. The natural use of our result is to select the order of a model via a sequential test procedure by choosing between two nested models. We demonstrate the applications of this general theory in the settings of real and complex matrix completion from incomplete and noisy observations, signal source identification, and determining the number of hidden nodes in neural networks.

As for the second topic, detection of the change-points for the sequential data is also an important problem. A change-point in such data is the critical time point where the distribution of the data changes. The change-points often represent a transition of the state. For example, in a power system, the change-points may represent the power outage [6]. In river systems, they may represent potential water contaminant hazards [7]. In public health, they may represent contagious outbreaks [8]. Therefore, the goal of change-points detection is to raise alarm of the change-points as soon as possible given the control of false alarm rate. In chapter 4, we develop an online change-point detection procedure for power system's cascading failure using multi-dimensional measurements over the networks. We incorporate the cascading failure's characteristic into the detection procedure and model multiple changes caused by cascading failures using a diffusion process over networks. The model

captures the property that the risk of component failing increases as more components around it fail. Our change-point detection procedure using the generalized likelihood ratio statistics assuming unknown post-change parameters of the measurements and the true failure time (change-points) at each node. We also provide a fast algorithm to perform the change-points detection. Numerical experiments show that our proposed method demonstrates good performance and can scale up to large systems.

One type of sequential data is event data. For event data, the time intervals between two observations have different length. Examples of event data can be extreme events in stock markets, the seismic signals, the activities in social media, etc. Therefore, to detect the change-points of high dimensional event data is also an important topic with widely applications. In chapter 5, we proposed a change-point detection procedure by scan score statistics in a multivariate Hawkes network. Our scan score statistics are computationally efficient since we don't need to compute the estimates of the post-change parameters, which is of importance for online detection. We present the theoretical results of our proposed procedure, including the analysis of the false alarm rate (FAR) and average run length (ARL) of the procedure under null hypothesis. We use simulation studies to testify our theoretical results and compare our method with an existing change-point detection procedure with generalized likelihood ratio statistics. We also apply our proposed procedure in real-world data such as memetracker and the stock market, which shows promising results in detecting an abrupt change in the network.

CHAPTER 2

RANK SELECTION IN MATRIX COMPLETION PROBLEM

2.1 Introduction

In this chapter, we consider the solution of the Minimum Rank Matrix Completion (MRMC) formulation, which leads to a non-convex optimization problem. We address the following questions: (i) Given observed entries arranged according to a (deterministic) pattern, by solving the MRMC problem, what is the minimum achievable rank? (ii) Under what conditions, there will be a unique matrix that is a solution to the MRMC problem? We give a sufficient condition (which we call the *well-posedness condition*) for the local uniqueness of MRMC solutions, and illustrate how such condition can be verified. We also show that such well-posedness condition in a sense is generic.

We also propose a sequential statistical testing procedure to determine the 'true' rank from noisy observed entries. Such statistical approach can be useful for many existing low-rank matrix completion algorithms, which require a pre-specification of the matrix rank, such as the alternating minimization approach to solving the non-convex problem by representing the low-rank matrix as a product of two low-rank matrix factors (see, e.g., [9, 4, 10]).

This chapter is organized as follows. In the next section, we . In Section 2.2 we present the considered setting, some basic definitions and the problem set-up, including the MRMC, LRMA, and convex relaxation formulations. Section 2.3 contains the main theoretical results. A statistical test of rank is presented in Section 2.4. In Section 2.5 we present numerical results related to the developed theory. Finally Section 2.6 concludes this chapter. All proofs are in the Appendix.

We use conventional notations. For $a \in \mathbb{R}$ we denote by $\lceil a \rceil$ the least integer that is

greater than or equal to a. By $A \otimes B$ we denote the Kronecker product of matrices (vectors) A and B, and by $\mathrm{vec}(A)$ column vector obtained by stacking columns of matrix A. We use the following matrix identity for matrices A, B, C of appropriate order

$$\mathrm{vec}(ABC) = (C^\top \otimes A)\mathrm{vec}(B). \tag{2.1}$$

By $\mathbb{S}^p$ we denote the linear space of $p \times p$ symmetric matrices and by writing $X \succeq 0$ we mean that matrix $X \in \mathbb{S}^p$ is positive semidefinite. By $\sigma_i(Y)$ we denote the i-th largest singular value of matrix $Y \in \mathbb{R}^{n_1 \times n_2}$. By I_p we denote the identity matrix of dimension p.

2.2 Matrix completion and problem set-up

Consider the problem of recovering an $n_1 \times n_2$ data matrix of low rank when observing a small number m of its entries, which are denoted as M_{ij}, $(i, j) \in \Omega$. We assume that $n_1 \geq 2$ and $n_2 \geq 2$. Here $\Omega \subset \{1, ..., n_1\} \times \{1, ..., n_2\}$ is an index set of cardinality m. The low-rank matrix completion problem, or matrix completion problem, aims to infer the missing entries, based on the available observations M_{ij}, $(i, j) \in \Omega$, by using a matrix whose rank is as small as possible.

Low-rank matrix completion problem is usually studied under a missing-at-random model, under which the necessary and sufficient conditions for perfect recovery of the true matrix are known [11, 12, 13, 14, 15, 16]. Study of deterministic sampling pattern is relatively rare. This includes the finitely rank-r completability problem in [17], which shows the conditions for the deterministic sampling pattern such that there exists at most *finitely many* rank-r matrices that agrees with its observed entries. In this chapter, we study a related but different problem, i.e., when will the matrix have a unique way to be completed, given a fixed sampling pattern. This is a fundamental problem related to *the identifiability* of a low-rank matrix given an observation pattern Ω.

2.2.1 Definitions

Let us introduce some necessary definitions. Denote by M the $n_1 \times n_2$ matrix with the specified entries M_{ij}, $(i,j) \in \Omega$, and all other entries equal zero. Consider $\Omega^c := \{1,...,n_1\} \times \{1,...,n_2\} \setminus \Omega$, the complement of the index set Ω, and define

$$\mathbb{V}_\Omega := \left\{ Y \in \mathbb{R}^{n_1 \times n_2} : Y_{ij} = 0,\ (i,j) \in \Omega^c \right\}.$$

This linear space represents the set of matrices that are filled with zeros at the locations of the unobserved entries. Similarly define

$$\mathbb{V}_{\Omega^c} := \left\{ Y \in \mathbb{R}^{n_1 \times n_2} : Y_{ij} = 0,\ (i,j) \in \Omega \right\}.$$

By P_Ω we denote the projection onto the space $\mathbb{V}_\Omega$, i.e., $[P_\Omega(Y)]_{ij} = Y_{ij}$ for $(i,j) \in \Omega$ and $[P_\Omega(Y)]_{ij} = 0$ for $(i,j) \in \Omega^c$. By this construction, $\{M + X : X \in \mathbb{V}_{\Omega^c}\}$ is the affine space of all matrices that satisfy the observation constraints. Note that $M \in \mathbb{V}_\Omega$ and the dimension of the linear space $\mathbb{V}_\Omega$ is $\dim(\mathbb{V}_\Omega) = m$, while $\dim(\mathbb{V}_{\Omega^c}) = n_1 n_2 - m$.

We say that a property holds for *almost every* (a.e.) M_{ij}, or almost surely, if the set of matrices $Y \in \mathbb{V}_\Omega$ for which this property does not hold has Lebesgue measure zero in the space $\mathbb{V}_\Omega$.

2.2.2 Minimum Rank Matrix Completion (MRMC)

Since the true rank is unknown, a natural approach is to find the minimum rank matrix that is consistent with the observations. This goal can be written as the following optimization problem referred to as the Minimum Rank Matrix Completion (MRMC),

$$\min_{Y \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(Y) \ \text{ subject to } Y_{ij} = M_{ij},\ (i,j) \in \Omega. \tag{2.2}$$

In general, the rank minimization problem is non-convex and NP-hard to solve. How-

ever, this problem is fundamental to various efficient heuristics derived from here. Largely, there are two categories of approximation heuristics: (i) approximate the rank function with some surrogate function such as the nuclear norm function, (ii) or solve a sequence of rank-constrained problems such as the matrix factorization based method, which we will discuss below. Approach (ii) requires to specify the target rank of the recovered matrix beforehand, which we will present a novel statistical test next.

2.2.3 Low Rank Matrix Approximation (LRMA)

Consider the problem

$$\min_{Y \in \mathbb{R}^{n_1 \times n_2},\, X \in \mathbb{V}_{\Omega^c}} F(M + X, Y) \;\; \text{s.t. } \operatorname{rank}(Y) = r, \tag{2.3}$$

where $M \in \mathbb{V}_\Omega$ is the given data matrix, and $F(A, B)$ is a discrepancy between matrices $A, B \in \mathbb{R}^{n_1 \times n_2}$. For example, let $F(A, B) := \|A - B\|_F^2$ with $\|Y\|_F^2 = \operatorname{tr}(Y^\top Y) = \sum_{i,j} Y_{ij}^2$, being the Frobenius norm. Define the set of $n_1 \times n_2$ matrices of rank r

$$\mathcal{M}_r := \left\{ Y \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(Y) = r \right\} \tag{2.4}$$

Then (Equation 2.3) becomes the least squares problem

$$\min_{Y \in \mathcal{M}_r} \sum_{(i,j) \in \Omega} (M_{ij} - Y_{ij})^2 . \tag{2.5}$$

The least squares approach although is natural, is not the only one possible. For example, in the statistical approach to Factor Analysis the discrepancy function is based on the Maximum Likelihood method and is more involved (e.g., [18]).

2.3 Main theoretical results

To gain insights into the identifiability issue of matrix completion, we aim to answer the following two related questions: (i) what is achievable minimum rank (the optimal value of problem Equation 2.2), and (ii) whether the minimum rank matrix, i.e., the optimal solutions to Equation 2.2, is unique given a problem set-up. These result will also help to gain insights in the tradeoff in the theoretical properties of other matrix completion formulations, including LRMA and SDP formulations, compared with the original MRMC formulation.

We show that given $m = |\Omega|$ observations of an $n_1 \times n_2$ matrix: (i) if the minimal rank r^* is less than $\mathfrak{R}(n_1, n_2, m) := (n_1+n_2)/2-[(n_1+n_2)^2/4-m]^{1/2}$, then the corresponding solution is unstable: an arbitrary small perturbation of the observed values can make this rank unattainable; (ii) if $r^* > \mathfrak{R}(n_1, n_2, m)$, then almost surely the solution is not (even locally) unique (cf., [19]). This indicates that except in rare occasions, the MRMC problem cannot have both properties of possessing unique and stable solutions. Consequently, LRMA approaches (also used in [4, 20]) could be a better alternative to the MRMC formulation.

2.3.1 Rank reducibility

We denote by r^* the optimal value of problem (Equation 2.2). That is, r^* is the minimal rank of an $n_1 \times n_2$ matrix with prescribed elements M_{ij}, $(i, j) \in \Omega$. Clearly, r^* depends on the index set Ω and values M_{ij}. A natural question is what values of r^* can be attained. Recall that Equation 2.2 is a non-convex problem and may have multiple solutions.

In a certain *generic sense* it is possible to give a lower bound for the minimal rank r^*. Let us consider intersection of a set of low-rank matrices and the affine space of matrices satisfying the observation constraints. Define the (affine) mapping $\mathcal{A}_M : \mathbb{V}_{\Omega^c} \to \mathbb{R}^{n_1 \times n_2}$ as

$$\mathcal{A}_M(X) := M + X, \quad X \in \mathbb{V}_{\Omega^c}.$$

As it has been pointed out before, the image $\mathcal{A}_M(\mathbb{V}_{\Omega^c}) = M + \mathbb{V}_{\Omega^c}$ of mapping $\mathcal{A}_M$ defines the space of feasible points of the MRMC problem (Equation 2.2). It is well known that $\mathcal{M}_r$ is a smooth, C^∞, manifold with

$$\dim(\mathcal{M}_r) = r(n_1 + n_2 - r). \tag{2.6}$$

It is said that the mapping $\mathcal{A}_M$ intersects $\mathcal{M}_r$ *transverally* if for every $X \in \mathbb{V}_{\Omega^c}$ either $\mathcal{A}_M(X) \notin \mathcal{M}_r$, or $\mathcal{A}_M(X) \in \mathcal{M}_r$ and the following condition holds

$$\mathbb{V}_{\Omega^c} + \mathcal{T}_{\mathcal{M}_r}(Y) = \mathbb{R}^{n_1 \times n_2}, \tag{2.7}$$

where $Y := \mathcal{A}_M(X)$ and $\mathcal{T}_{\mathcal{M}_r}(Y)$ denotes the tangent space to $\mathcal{M}_r$ at $Y \in \mathcal{M}_r$ (we will give explicit formulas for the tangent space $\mathcal{T}_{\mathcal{M}_r}(Y)$ in equations (Equation 2.14) and (Equation 2.15) below.)

By using a classical result of differential geometry, it is possible to show that for *almost every* (a.e.) M_{ij}, $(i, j) \in \Omega$, the mapping $\mathcal{A}_M$ intersects $\mathcal{M}_r$ transverally (this holds for every r) (see [19] for a discussion of this result). Transversality condition (Equation 2.7) means that the linear spaces $\mathbb{V}_{\Omega^c}$ and $\mathcal{T}_{\mathcal{M}_r}(Y)$ together span the whole space $\mathbb{R}^{n_1 \times n_2}$. Of course this cannot happen if the sum of their dimensions is less than the dimension of $\mathbb{R}^{n_1 \times n_2}$. Therefore transversality condition (Equation 2.7) implies the following dimensionality condition

$$\dim(\mathbb{V}_{\Omega^c}) + \dim(\mathcal{T}_{\mathcal{M}_r}(Y)) \geq \dim(\mathbb{R}^{n_1 \times n_2}). \tag{2.8}$$

In turn the above condition (Equation 2.8) can be written as

$$r(n_1 + n_2 - r) \geq m, \tag{2.9}$$

or equivalently $r \geq \mathfrak{R}(n_1, n_2, mm)$, where

$$\mathfrak{R}(n_1, n_2, m) := (n_1 + n_2)/2 - \sqrt{(n_1 + n_2)^2/4 - m}. \tag{2.10}$$

That is, if $r < \mathfrak{R}(n_1, n_2, m)$, then the transversality condition (Equation 2.7) cannot hold and hence for a.e. M_{ij} it follows that $\mathrm{rank}(M + X) \neq r$ for all $X \in \mathbb{V}_{\Omega^c}$.

Now if $\mathcal{A}_M$ intersects $\mathcal{M}_r$ transverally at $\mathcal{A}_M(X) \in \mathcal{M}_r$ (i.e., condition (Equation 2.7) holds), then the intersection $\mathcal{A}_M(\mathbb{V}_{\Omega^c}) \cap \mathcal{M}_r$ forms a smooth manifold near the point $Y := \mathcal{A}_M(X)$. When $r > \mathfrak{R}(n_1, n_2, m)$, *this manifold has dimension greater than zero and hence the corresponding rank r solution is not (locally) unique*. This leads to the following (for a formal discussion of these results we can refer to [19]).

Theorem 2.3.1 (Generic lower bound and non-uniqueness of solutions)**.** *For any index set Ω of cardinality m and almost every M_{ij}, $(i, j) \in \Omega$, the following holds:* (i) *for every feasible point Y of problem* (Equation 2.2) *it follows that*

$$\mathrm{rank}(Y) \geq \mathfrak{R}(n_1, n_2, m), \tag{2.11}$$

(ii) *if $r^* > \mathfrak{R}(n_1, n_2, m)$, then problem* (Equation 2.2) *has multiple (more than one) optimal solutions.*

It follows from part (i) of Theorem 2.3.1 that $r^* \geq \mathfrak{R}(n_1, n_2, m)$ for a.e. M_{ij}. *Generically* (i.e., almost surely) the following lower bound for the minimal rank r^* holds

$$r^* \geq \mathfrak{R}(n_1, n_2, m), \tag{2.12}$$

and (Equation 2.2) may have unique optimal solution only when $r^* = \mathfrak{R}(n_1, n_2, m)$. Of course such equality could happen only if $\mathfrak{R}(n_1, n_2, m)$ is an integer number. As Example 2.3.1 below shows, for any integer $r^* \leq \lceil \sqrt{m} \rceil$ satisfying (Equation 2.12), there exists an index set Ω such that the corresponding MRMC problem attains the minimal rank r^* for

a.e. M_{ij}. In particular this shows that the lower bound (Equation 2.12) is tight. When we have a square matrix $n_1 = n_2 = n$, it follows that

$$\mathfrak{R}(n, n, m) = n - \sqrt{n^2 - m}. \tag{2.13}$$

For $n_1 = n_2 = n$ and small m/n^2 we can approximate

$$\mathfrak{R}(n, n, m) = n\left(1 - \sqrt{1 - m/n^2}\right) \approx m/(2n).$$

For example, for $n_1 = n_2 = 1000$ and $m = 20000$ we have $\mathfrak{R}(n, n, m) = 10.05$, and hence the bound (Equation 2.12) becomes $r^* \geq 11$. The nuclear norm minimization guarantees to recover a solution of rank $r \leq 199$ [21].

Example 2.3.1 (Tightness of the lower bound for r^*)**.** For $r < \min\{n_1, n_2\}$ consider data matrix M of the following form $M = \left(\begin{smallmatrix} M_1 & 0 \\ M_2 & M_3 \end{smallmatrix}\right)$. Here, the three sub-matrices M_1, M_2, M_3, of the respective order $r \times r$, $(n_1 - r) \times r$ and $(n_1 - r) \times (n_2 - r)$, represent the observed entry values. Cardinality m of the corresponding index set Ω is $r(n_1 + n_2 - r)$, i.e., here $r = \mathfrak{R}(n_1, n_2, m)$. Suppose that the $r \times r$ matrix M_1 is nonsingular, i.e., its rows are linearly independent. Then any row of matrix M_2 can be represented as a (unique) linear combination of rows of matrix M_1. It follows that the corresponding MRMC problem has (unique) solution of rank $r^* = r$. In other words, the rank of the completed matrix will be equal to r (the rank of the sub-matrix M_1) and there will be a unique matrix that achieves this rank. Now suppose that some of the entries of the matrices M_2 and M_3 are not observed, and hence cardinality of the respective index set Ω is less than $r(n_1 + n_2 - r)$, and thus $r > \mathfrak{R}(n_1, n_2, m)$. In that case the respective minimal rank still is r, provided matrix M_1 is nonsingular, although the corresponding optimal solutions are not unique. In particular, if $M = \left(\begin{smallmatrix} M_1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, i.e., only the entries of matrix M_1 are observed, then $m = r^2$ and the minimum rank is r.

2.3.2 Uniqueness of solutions of the MRMC problem

Following Theorem Theorem 2.3.1, for a given matrix $M \in \mathbb{V}_\Omega$ and the corresponding minimal rank $r^* \leq \mathfrak{R}(n_1, n_2, m)$, the question is whether the corresponding solution Y^* of rank r^* is unique. Although, the set of such matrices M is "thin" (in the sense that it has Lebesgue measure zero), this question of uniqueness is important, in particular for the statistical inference of rank (discussed in Section Section 2.4). Available results, based on the so-called Restricted Isometry Property (RIP) for low-rank matrix recovery from linear observations and based on the coherence property for low-rank matrix completion, assert that for certain probabilistic (Gaussian) models such uniqueness holds with high probability. However for a given matrix $M \in \mathbb{V}_\Omega$ it could be difficult to verify whether the solution is unique (some sufficient conditions for such uniqueness are given in [17, Theorem 2], we will comment on this below.)

Let us consider the following concept of local uniqueness of solutions.

Definition 2.3.1. *We say that an $n_1 \times n_2$ matrix $\bar{Y}$ is a* locally *unique solution of problem* (Equation 2.2) *if $P_\Omega(\bar{Y}) = M$ and there is a neighborhood $\mathcal{V} \subset \mathbb{R}^{n_1 \times n_2}$ of $\bar{Y}$ such that $\operatorname{rank}(Y) \neq \operatorname{rank}(\bar{Y})$ for any $Y \in \mathcal{V}$, $P_\Omega(Y) = M$ and $Y \neq \bar{Y}$.*

Note that rank is a lower semicontinuous function of matrix, i.e., if $\{Y_k\}$ is a sequence of matrices converging to matrix Y, then $\liminf_{k\to\infty} \operatorname{rank}(Y_k) \geq \operatorname{rank}(Y)$. Therefore local uniqueness of $\bar{Y}$ actually implies existence of the neighborhood $\mathcal{V}$ such that $\operatorname{rank}(Y) > \operatorname{rank}(\bar{Y})$ for all $Y \in \mathcal{V}$, $Y \neq \bar{Y}$, i.e., that at least locally problem (Equation 2.2) does not have optimal solutions different from $\bar{Y}$. The Definition 2.3.1 is closely related to the *finitely rank-r completability* condition introduced in [17], which assumes that the MRMC problem has a finite number of rank r solutions. Of course if problem (Equation 2.2) has a non locally unique solution of rank r, then the finitely rank-r completability condition cannot hold.

We now introduce some constructions associated with the manifold $\mathcal{M}_r$ of matrices of

rank r. There are several equivalent forms how the tangent space to the manifold $\mathcal{M}_r$ at $Y \in \mathcal{M}_r$ can be represented. Let $Y = VW^\top$ for some matrices $V \in \mathbb{R}^{n_1 \times r}$ and $W \in \mathbb{R}^{n_2 \times r}$ of rank r. Then

$$\mathcal{T}_{\mathcal{M}_r}(Y) = \left\{ (dV)W^\top + V(dW)^\top : dV \in \mathbb{R}^{n_1 \times r},\ dW \in \mathbb{R}^{n_2 \times r} \right\}. \tag{2.14}$$

In an equivalent form this tangent space can be written as

$$\mathcal{T}_{\mathcal{M}_r}(Y) = \left\{ H \in \mathbb{R}^{n_1 \times n_2} : FHG = 0 \right\}, \tag{2.15}$$

where F is an $(n_1 - r) \times n_1$ matrix of rank $n_1 - r$ such that $FY = 0$ (referred to as a *left side complement* of Y) and G is an $n_2 \times (n_2 - r)$ matrix of rank $n_2 - r$ such that $YG = 0$ (referred to as a *right side complement* of Y). We also use the linear space of matrices orthogonal (normal) to $\mathcal{M}_r$ at $Y \in \mathcal{M}_r$, denoted by $\mathcal{N}_{\mathcal{M}_r}(Y)$. By (Equation 2.14) it follows that

$$\mathcal{N}_{\mathcal{M}_r}(Y) = \left\{ Z \in \mathbb{R}^{n_1 \times n_2} : Z^\top Y = 0 \text{ and } YZ^\top = 0 \right\}. \tag{2.16}$$

Definition 2.3.2 (Well-posedness condition)**.** *We say that a matrix $\bar{Y} \in \mathcal{M}_r$ is* well-posed*, for problem* (Equation 2.2)*, if $P_\Omega(\bar{Y}) = M$ and the following condition holds*

$$\mathbb{V}_{\Omega^c} \cap \mathcal{T}_{\mathcal{M}_r}(\bar{Y}) = \{0\}. \tag{2.17}$$

Condition (Equation 2.17) (illustrated in Figure 2.1) is a natural condition having a simple geometrical interpretation. Intuitively, it means that the null space of the observation operator does not have any non-trivial matrix that lies in the tangent space of low-rank matrix manifold. Hence, there cannot be any local deviations from the optimal solution that still satisfy the measurement constraints. This motivates us to introduce the well-

posedness condition that guarantees a matrix to be locally unique solution. Note that this is different from the so-called Null Space Property (NSP) (see, e.g.,[22]). Although both the well-posedness and NSP have a similar geometrical flavor, the NSP is aimed at ensuring uniqueness of solution of a convex problem, while the MRMC is essentially a nonconvex construction. For instance, NSP can be used to guarantee uniqueness of solutions of the optimization problem $\min_{x\in\mathbb{R}^n} \|x\|_1$ subject to $Ax = b$, which can be formulated as a linear programming problem.

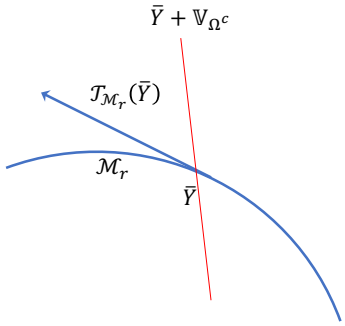

Figure 2.1: Illustration of the well-posedness condition.

Now we can give sufficient conditions for local uniqueness:

Theorem 2.3.2 (Sufficient conditions for local uniqueness)**.** *Matrix $\bar{Y} \in \mathcal{M}_r$ is a locally unique solution of problem* (Equation 2.2) *if $\bar{Y}$ is well-posed for* (Equation 2.2)*.*

2.3.3 Verifiable form of well-posedness condition

Below we present an equivalent form of the well-posedness condition that can be verified algebraically. By Theorem Theorem 2.3.2 we have that if matrix $\bar{Y} \in \mathcal{M}_r$ is well-posed, then $\bar{Y}$ is a locally unique solution of problem (Equation 2.2). Note that condition (Equation 2.17) implies that $\dim(\mathbb{V}_{\Omega^c}) + \dim(\mathcal{T}_{\mathcal{M}_r}(\bar{Y})) \leq n_1 n_2$. That is, condition (Equation 2.17) implies that $r(n_1 + n_2 - r) \leq m$ or equivalently $r \leq \mathfrak{R}(n_1, n_2, m)$. By Theorem Theorem 2.3.1 we have that if $r^* > \mathfrak{R}(n_1, n_2, m)$, then the corresponding optimal solution cannot be locally unique almost surely. Note that since the space $\mathbb{V}_\Omega$ is orthogonal to the space $\mathbb{V}_{\Omega^c}$, by duality arguments condition (Equation 2.17) is equivalent to the

following condition

$$\mathbb{V}_\Omega + \mathcal{N}_{\mathcal{M}_r}(\bar{Y}) = \mathbb{R}^{n_1 \times n_2}. \tag{2.18}$$

By using formula (Equation 2.15) it is also possible to write condition (Equation 2.17) in the following form

$$\{X \in \mathbb{V}_{\Omega^c} : FXG = 0\} = \{0\}, \tag{2.19}$$

where F is a left side complement of $\bar{Y}$ and G is a right side complement of $\bar{Y}$. Recall that $\mathrm{vec}(FXG) = (G^\top \otimes F)\mathrm{vec}(X)$. Column vector of matrix $G^\top \otimes F$ corresponding to component x_{ij} of vector $\mathrm{vec}(X)$, is $g_j^\top \otimes f_i$, where f_i is the i-th column of matrix F and g_j is the j-th row of matrix G. Condition (Equation 2.19) means that the column vectors $g_j^\top \otimes f_i$, $(i,j) \in \Omega^c$, are linearly independent. It could be noted that the left and right side complements are not unique. That is, the left side complement can be changed to QF for an arbitrary $(n_1 - r) \times (n_1 - r)$ nonsingular matrix Q, and similarly the right side complement can be changed to GR for an arbitrary $(n_2 - r) \times (n_2 - r)$ nonsingular matrix R. We have that $(GR)^\top \otimes (QF) = (R^\top \otimes Q)(G^\top \otimes F)$. Therefore the condition for vectors $g_j^\top \otimes f_i$, $(i,j) \in \Omega^c$, to be linearly independent does not depend on a particular choice of the left and right side complements.

We obtain the following verifiable condition for checking the well-posedness of a given solution:

Theorem 2.3.3 (Equivalent condition of well-posedness)**.** *Matrix $\bar{Y} \in \mathcal{M}_r$ satisfies condition* (Equation 2.17) *if and only if for any left side complement F and right side complement G of $\bar{Y}$, the column vectors $g_j^\top \otimes f_i$, $(i,j) \in \Omega^c$, are linearly independent.*

A consequence of Theorem 2.3.3 is that if $\bar{Y} \in \mathcal{M}_r$ is well-posed, then necessarily $(n_1 - r)(n_2 - r) \geq |\Omega^c|$, since vectors $g_j^\top \otimes f_i$ have dimension $(n_1 - r)(n_2 - r)$. Since $|\Omega^c| = n_1 n_2 - m$, this is equivalent to $r(n_1 + n_2 - r) \leq m$. That is, the well-posedness cannot happen if $r > \mathfrak{R}(n_1, n_2, m)$. This of course is not surprising in view of discussion of subsection 2.3.1.

Theorem 2.3.3 also implies the following necessary condition for well-posedness of $\bar{Y} \in \mathcal{M}_r$ in terms of the pattern of the index set Ω, which is related to the completability condition in [17] that each row and each column has at least r observations. If matrix $\bar{Y} \in \mathcal{M}_r$ is well-posed for problem (Equation 2.2), then at each row and each column of $\bar{Y}$ there are at least r elements of the index set Ω. Indeed, suppose that in row $i \in \{1, ..., n_1\}$ there are less than r elements of Ω. This means that the set $\sigma_i := \{j : (i, j) \in \Omega^c\}$ has cardinality greater than $n_2 - r$. Let F be a left side complement of $\bar{Y}$ and G be a right side complement of $\bar{Y}$. Since rows g_j of G are of dimension $1 \times (n_2 - r)$, we have then that vectors g_j, $j \in \sigma_i$, are linearly dependent, i.e., $\sum_{j\in\sigma_i} \lambda_j g_j = 0$ for some λ_j, not all of them zero. Then

$$\textstyle\sum_{j\in\sigma_i} \lambda_j (g_j^\top \otimes f_i) = \big(\sum_{j\in\sigma_i} \lambda_j g_j\big)^\top \otimes f_i = 0. \tag{2.20}$$

This contradicts the condition for vectors $g_j^\top \otimes f_i$, $(i, j) \in \Omega^c$, to be linearly independent. Similar arguments can be applied to the columns of matrix $\bar{Y}$. This necessary condition for well-posedness is not surprising since if there is a row with less than r elements of Ω, then this row in not uniquely defined in the corresponding rank r solution (cf., [17]). However, although necessary, the condition for the index set Ω to have at each row and each column at least r elements is not sufficient to ensure well-posedness as shown by Theorem 2.3.5 below. Note that by definition the matrices F and G are of full rank.

2.3.4 Generic nature of the well-posedness

In a certain sense the well-posedness condition is generic, as we explain below. Denote by $\mathcal{F}_r \subset \mathbb{R}^{n_1\times r}$ and $\mathcal{G}_r \subset \mathbb{R}^{n_2\times r}$ the respective sets of matrices of rank r. Consider the set $\Theta := \mathcal{F}_r \times \mathcal{G}_r \times \mathbb{V}_{\Omega^c}$ viewed as a subset of $\mathbb{R}^{n_1 r+n_2 r+n_1 n_2-m}$, and mapping $\mathfrak{F} : \Theta \to \mathbb{R}^{n_1\times n_2}$ defined as

$$\mathfrak{F}(\theta) := VW^\top + X, \quad \theta = (V, W, X) \in \Theta.$$

Note that the sets $\mathcal{G}_r$ and $\mathcal{F}_r$ are open and connected, and hence the set Θ is open and connected, and the components of mapping $\mathfrak{F}(\cdot)$ are polynomial functions.

Let $\Delta(\theta)$ be the Jacobian of mapping $\mathfrak{F}$. That is, $\Delta(\theta)$ is $(n_1r+n_2r+n_1n_2-m)\times(n_1n_2)$ matrix of partial derivatives of $\mathfrak{F}(\theta)$ taken with respect to a specified order of the components of the corresponding matrices. Let us consider the following concept associated with rank r and index set Ω (cf., [23]).

Definition 2.3.3. *We refer to*

$$\varrho := \max_{\theta\in\Theta} \big\{\mathrm{rank}\big(\Delta(\theta)\big)\big\} \tag{2.21}$$

as the characteristic rank *of mapping $\mathfrak{F}$ and say that $\theta \in \Theta$ is a* regular point *of $\mathfrak{F}$ if $\mathrm{rank}\big(\Delta(\theta)\big) = \varrho$. We say that $(V, W) \in \mathcal{F}_r \times \mathcal{G}_r$ is regular if $\theta = (V, W, X)$ is regular for some $X \in \mathbb{V}_{\Omega^c}$.*

Since $\mathfrak{F}(V, W, \cdot)$ is linear, the Jacobian $\Delta(V, W, X)$ is the same for all $X \in \mathbb{V}_{\Omega^c}$, i.e., $\Delta(V, W, X) = \Delta(V, W, X')$ for any $X, X' \in \mathbb{V}_{\Omega^c}$ and $(V, W) \in \mathcal{F}_r \times \mathcal{G}_r$. Hence if a point $\theta = (V, W, X)$ is regular for some $X \in \mathbb{V}_{\Omega^c}$, then (V, W, X') is regular for any $X' \in \mathbb{V}_{\Omega^c}$. Therefore regularity actually is a property of points $(V, W) \in \mathcal{F}_r \times \mathcal{G}_r$.

Consider $\theta = (V, W, X) \in \Theta$ and $Y = VW^\top$. We have that $\mathrm{rank}\big(\Delta(\theta)\big) = \dim\big(\mathfrak{V}(\theta)\big)$, where $\mathfrak{V}(\theta)$ denotes the image of the differential of $\mathfrak{F}(\theta)$. Since the differential $d\,\mathfrak{F}(\theta) = (dV)W^\top + V(dW)^\top + dX$ and because of (Equation 2.14), the linear space $\mathfrak{V}(\theta)$ is equal to $\mathcal{T}_{\mathcal{M}_r}(Y) + \mathbb{V}_{\Omega^c}$. It follows that $\varrho \le \mathfrak{f}(r, m)$, where

$$\begin{aligned} \mathfrak{f}(r, m) &:= \dim\big(\mathcal{T}_{\mathcal{M}_r}(Y)\big) + \dim\big(\mathbb{V}_{\Omega^c}\big) \\ &= r(n_1 + n_2 - r) + n_1n_2 - m. \end{aligned} \tag{2.22}$$

It also follows that $\mathrm{rank}\big(\Delta(\theta)\big) = \mathfrak{f}(r, m)$ iff condition (Equation 2.17) holds at Y. In other words we have the following result.

Proposition 2.3.1. *Rank of $\Delta(\theta)$ attains the maximal value $\mathfrak{f}(r,m)$ if and only if the corresponding matrix $Y = VW^\top$ is well posed.*

Furthermore we have the following.

Theorem 2.3.4. *The following holds: (i) Almost every point $(V,W) \in \mathcal{F}_r \times \mathcal{G}_r$ is regular. (ii) The set of regular points forms an open subset of $\mathcal{F}_r \times \mathcal{G}_r$. (iii) For any regular point $(V,W) \in \mathcal{F}_r \times \mathcal{G}_r$, the corresponding matrix $Y = VW^\top$ satisfies the well-posedness condition* (Equation 2.17) *if and only if the characteristic rank ϱ is equal to $\mathfrak{f}(r,m)$. (iv) If $\varrho < \mathfrak{f}(r,m)$ and a point $(\bar{V},\bar{W}) \in \mathcal{F}_r \times \mathcal{G}_r$ is regular, then for any $Y \in \mathcal{M}_r$ in a neighborhood of $\bar{Y} = \bar{V}\bar{W}^\top$ there exists $X \in \mathbb{V}_{\Omega^c}$ such that $Y = \bar{Y} + X$.*

The significance of Theorem 2.3.4 is that this shows that for given rank r and index set Ω, either $\varrho = \mathfrak{f}(r,m)$ in which case a.e. $Y \in \mathcal{M}_r$ satisfies the well-posedness condition (Equation 2.17), or $\varrho < \mathfrak{f}(r,m)$ in which case condition (Equation 2.17) does not hold for all $Y \in \mathcal{M}_r$ and generically rank r solutions are not locally unique.

We have that a necessary condition for $\varrho = \mathfrak{f}(r,m)$ is that each row and each column of the considered matrix has at least r observed entries. Another necessary condition is for the index set to be irreducible (see Theorem 2.3.5). Whether these two conditions are sufficient for $\varrho = \mathfrak{f}(r,m)$ to hold remains an open question. Numerical experiments, reported in Section Section 2.5, indicate that in a certain probabilistic sense chances of occurring not well posed solution are negligible when r is slightly less than $\mathfrak{R}(n_1,n_2,m)$.

2.3.5 Global uniqueness of solutions for special cases

In some rather special cases it is possible to give verifiable conditions for global uniqueness of minimum rank solutions. The following conditions are straightforward extensions of well known conditions in Factor Analysis (cf., [24, Theorem 5.1]).

Assumption 2.3.1. *Suppose that:* (i) *for a given index $(k,l) \in \Omega^c$, there exist index sets $\mathcal{I}_1 \subset \{1,...,n_1\} \setminus \{k\}$ and $\mathcal{I}_2 \subset \{1,...,n_2\} \setminus \{l\}$ such that $|\mathcal{I}_1| = |\mathcal{I}_2| = r$, $\mathcal{I}_1 \times \mathcal{I}_2 \subset \Omega$,*

and $\{k\} \times \mathcal{I}_2 \subset \Omega$ *and* $\{l\} \times \mathcal{I}_1 \subset \Omega$, (ii) *the* $r \times r$ *submatrix of* M *corresponding to rows* $i \in \mathcal{I}_1$ *and columns* $j \in \mathcal{I}_2$ *is nonsingular.*

For example, for $r = 1$ part (i) of the above assumption means existence of indexes $k' \neq k$ and $l' \neq l$ such that $(k', l), (k, l'), (k', l') \in \Omega$.

Proposition 2.3.2. *Suppose that Assumption 2.3.1 holds for an index* $(k, l) \in \Omega^c$. *Then the minimum rank* $r^* \geq r$, *and for any matrix* $Y \in \mathcal{M}_r$ *such that* $P_\Omega(Y) = M$ *it follows that* $Y_{kl} = \bar{Y}_{kl}$.

Clearly part (ii) of Assumption 2.3.1 implies that $r^* \geq r$. The other result of the above proposition follows by observing that the $(r+1) \times (r+1)$ submatrix of Y corresponding to rows $\{k\} \cup \mathcal{I}_1$ and columns $\{l\} \cup \mathcal{I}_2$ has rank r and hence zero determinant, and applying Shur complement for the element Y_{kl}. Note that provided the part (i) holds, part (ii) is generic in the sense that it holds for a.e. M_{ij}.

If Assumption 2.3.1 holds for *every* $(k, l) \in \Omega^c$, then the uniqueness of the solution $\bar{Y}$ follows. This is closely related to [17, Theorem 2], but is not the same. It is assumed in [17] that every column of M has $r+1$ observed entries. For example, consider 2×2 matrix with 3 observed entries, $M_{12} = M_{21} = M_{22} = 1$. The only unobserved entry, corresponding to the index $(1, 1)$, satisfies Assumption 2.3.1 and rank one matrix, with all entries equal 1, is the unique solution of the MRMC problem. On the other hand the first column of matrix M has only one observed entry.

Remark 2.3.1. The following example was constructed in Wilson and Worcester [25], of two 6×6 symmetric matrices of rank 3 with the same off-diagonal and different diagonal elements. If we define the index set as $\Omega := \{(i, j) : i \neq j,\ i, j = 1, ..., 6\}$, then this can be viewed as an example of two different locally unique solutions of rank 3. Note that here $m = 30$ and $\mathfrak{R}(6, 6, 30) = 6 - \sqrt{6}$. That is $\mathfrak{R}(6, 6, 30) > 3$ and generically (almost surely) rank cannot be reduced below $r = 4$. We will discuss this example further in Section Section 2.5.

2.3.6 Identifiable Ω

Our results can also be used to determine whether observation patterns Ω is identifiable. First note that uniqueness of the minimum rank solution is invariant with respect to permutations of rows and columns of matrix M. This motivates to introduce the following definition.

Definition 2.3.4. *We say that the index set Ω is* reducible *if by permutations of rows and columns, the set Ω can be represented as the union $\Omega' \cup \Omega''$ of two disjoined sets $\Omega' \subset \{1, ..., k\} \times \{1, ..., l\}$ and $\Omega'' \subset \{k+1, ..., n_1\} \times \{l+1, ..., n_2\}$ for some $1 \leq k < n_1$ and $1 \leq l < n_2$. Otherwise we say that Ω is irreducible.*

Reducibility of the index set Ω means that by permutations of rows and columns, matrix M can be represented in the block diagonal form

$$M = \begin{bmatrix} M' & 0 \\ 0 & M'' \end{bmatrix}, \tag{2.23}$$

where matrices M' and M'' are of order $k \times l$ and $(n_1 - k) \times (n_2 - l)$, respectively, with observed entries M'_{ij}, $(i, j) \in \Omega'$, and M''_{ij}, $(i, j) \in \Omega''$. Some entries of matrices M' and M'' can also be zero if the corresponding entries of matrix M are zeros.

Theorem 2.3.5 (Reducible index set)**.** *If the index set Ω is reducible, then any minimum rank solution $\bar{Y}$ is not locally (and hence globally) unique.*

As it was shown in Theorem Theorem 2.3.2, if $\bar{Y}$ is not locally unique, then it cannot be well-posed. Therefore if the index set Ω is reducible, then any minimum rank solution is not well-posed. Of course even if Ω is reducible, it still can happen that in each row and column there are at least r elements of the index set Ω. That is, the condition of having r elements of the index set Ω in each row and column is not sufficient to ensure the well-posedness property.

Remark 2.3.2. Reducibility/irreducibility of the index set Ω can be verified in the following way. Consider the undirected graph $G = (V, E)$ with the set of vertices $V := \Omega$, and edges between two vertices $(i, j), (i', j') \in \Omega$ if and only if $i = i'$ or $j = j'$. Then Ω is irreducible if and only if G has only one connected component. A connected component of G is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph G. There are algorithms of running time $O(|V| + |E|)$ which can find every vertex that is reachable from a given vertex of G, and hence to determine a connected component of G, e.g., the well known *breadth-first search* algorithm [26, Section 22.2]. Note that the number of vertices in G is $m = |\Omega|$, which could be much smaller than $n_1 n_2$.

2.3.7 Uniqueness of rank one solutions

In this section we discuss uniqueness of rank one solutions of the MRMC problem (Equation 2.2). We show that in case of the minimum rank one, irreducibility of Ω is sufficient for the global uniqueness. We assume that all $M_{ij} \neq 0$, $(i, j) \in \Omega$, and that every row and every column of the matrix M has at least one element M_{ij}. Let $\bar{Y}$ be a solution of rank one of problem (Equation 2.2), i.e., there are nonzero column vectors v and w such that $\bar{Y} = vw^\top$ with $P_\Omega(\bar{Y}) = M$.

Recall that permutations of the components of vector v corresponds to permutations of the rows of the respective rank one matrix, and permutations of the components of vector w corresponds to permutations of the columns of the respective rank one matrix. It was shown in Theorem Theorem 2.3.5 that if the index set Ω is reducible, then solution $\bar{Y}$ cannot be locally unique. In case of rank one solution the converse of that also holds.

Theorem 2.3.6 (Global uniqueness for rank one solution)**.** *Suppose that Ω is irreducible, $M_{ij} \neq 0$ for all $(i, j) \in \Omega$, and every row and every column of the matrix M has at least one element M_{ij}, $(i, j) \in \Omega$. Then any rank one solution is globally unique.*

It could be mentioned that even for $r = 1$ the irreducibility is a weaker condition than

part (i) of Assumption 2.3.1 applied to every $(k, l) \in \Omega^c$. For example, let $n_1 = n_2 = n \geq 3$ and $\Omega = \{(i, j) : i \geq j,\ i, j = 1, ..., n\} \setminus \{(n, 1)\}$. This set Ω irreducible. However for the index $(1, n)$, Assumption 2.3.1(i) does not hold.

2.3.8 LRMA and its properties

We discuss below the LRMA approach (Equation 2.5). Compared with the formulation of exact low rank recovery, the LRMA is more realistic in the presence of noise. By Theorem Theorem 2.3.1 we have that if the minimal rank r^* is less than $\mathfrak{R}(n_1, n_2, m)$, then the corresponding solution is unstable in the sense that an arbitrary small perturbation of the observed values M_{ij} can make this rank unattainable. On the other hand if $r^* > \mathfrak{R}(n_1, n_2, m)$, then almost surely the solution is not (even locally) unique. This indicates that except in rare occasions, problem (Equation 2.2) of exact rank minimization cannot have both properties of possessing unique and stable solutions. Consequently, what makes sense is to try to solve the minimum rank problem approximately.

Proposition 2.3.3 (Necessary condition for LRMA)**.** *The following are necessary conditions for $Y \in \mathcal{M}_r$ to be an optimal solution of problem* (Equation 2.5)

$$(P_\Omega(Y) - M)^\top Y = 0 \ \text{ and } \ Y(P_\Omega(Y) - M)^\top = 0. \tag{2.24}$$

Remark 2.3.3. We can view the least squares problem (Equation 2.5) from the following point of view. Consider function

$$\phi(Y, \Theta) := \tfrac{1}{2}\mathrm{tr}[(P_\Omega(Y) - \Theta)^\top (P_\Omega(Y) - \Theta)], \tag{2.25}$$

with $\Theta \in \mathbb{V}_\Omega$ viewed as a parameter. Define

$$\begin{aligned} f(Y) &:= \tfrac{1}{2} \sum_{(i,j)\in\Omega} (Y_{ij} - M_{ij})^2 \\ &= \tfrac{1}{2}\mathrm{tr}[(P_\Omega(Y) - M)^\top (P_\Omega(Y) - M)], \end{aligned} \tag{2.26}$$

Hence, the problem (Equation 2.5) consists of minimization of $f(Y)$ subject to $Y \in \mathcal{M}_r$. Note that for $\Theta = M$ we have $f(\cdot) = \phi(\cdot, M)$, where $f(\cdot)$ is defined in (Equation 2.26). Let $\bar{Y} \in \mathcal{M}_r$ be such that $\phi(\bar{Y}, \Theta_0) = 0$ for some $\Theta_0 \in \mathbb{V}_\Omega$, i.e., $P_\Omega(\bar{Y}) = \Theta_0$. A sufficient condition for $\bar{Y}$ to be a locally unique solution of problem (Equation 2.2), at $M = \Theta_0$, is

$$\mathrm{tr}\left[P_\Omega(H)^\top P_\Omega(H)\right] > 0, \;\; \forall H \in \mathcal{T}_{\mathcal{M}_r}(\bar{Y}) \setminus \{0\}. \tag{2.27}$$

The above condition means that if $H \in \mathcal{T}_{\mathcal{M}_r}(\bar{Y})$ and $H \neq 0$, then $P_\Omega(H) \neq 0$. In other words this means that the kernel

$$\mathrm{Ker}(P_\Omega) := \{H \in \mathcal{T}_{\mathcal{M}_r}(\bar{Y}) : P_\Omega(H) = 0\}$$

is $\{0\}$. Since $P_\Omega(H) = 0$ for any $H \in \mathbb{V}_{\Omega^c}$, it follows that: *condition* (Equation 2.27) *is equivalent to the sufficient condition* (Equation 2.17) *of Theorem 2.3.2*. That is, condition (Equation 2.27) means that matrix $\bar{Y}$ is well-posed for problem (Equation 2.2).

Assuming that condition (Equation 2.27) (or equivalently condition (Equation 2.17)) holds, by applying the Implicit Function Theorem to the first order optimality conditions of the least squares problem (Equation 2.5) we have the following result.

Proposition 2.3.4. *Let $\bar{Y} \in \mathcal{M}_r$ be such that $P_\Omega(\bar{Y}) = \Theta_0$ for some $\Theta_0 \in \mathbb{V}_\Omega$ and suppose that the well posedness condition* (Equation 2.17) *holds. Then there exist neighborhoods $\mathcal{V}$ and $\mathcal{W}$ of $\bar{Y}$ and Θ_0, respectively, such that for any $M \in \mathcal{W} \cap \mathbb{V}_\Omega$ there exists unique $Y \in \mathcal{V} \cap \mathcal{M}_r$ satisfying the optimality conditions* (Equation 2.24).

The above proposition implies the following. Suppose that we run a numerical pro-

cedure which identifies a matrix $\bar{Y} \in \mathcal{M}_r$ satisfying the (necessary) first order optimality conditions (Equation 2.24). Then if $P_\Omega(\bar{Y})$ is sufficiently close to M (i.e., the fit $\sum_{(i,j)\in\Omega} (Y_{ij} - M_{ij})^2$ is sufficiently small) and condition (Equation 2.17) holds at $\bar{Y}$, then we can say that $f(Y) > f(\bar{Y})$ for all $Y \neq \bar{Y}$ in a neighborhood of $\bar{Y}$. That is, $\bar{Y}$ solves the least squares problem at least locally. Unfortunately it is not clear how to quantify the "sufficiently close" condition, and this does not guarantee global optimality of $\bar{Y}$ unless $\bar{Y}$ is the unique minimum rank solution.

2.4 Statistical test for rank selection

In this section, we propose a statistical test procedure for value of the "true" minimal rank, when the entries of the data matrix M are observed with noise. Such statistical approach can be useful for many existing low-rank matrix completion algorithms, which require a pre-specification of the matrix rank, such as the alternating minimization approach to solving the non-convex problem by representing the low-rank matrix as a product of two low-rank matrix factors (see, e.g., [4]).

Consider this for the LRMA formulation. By the above discussion, it will be natural to take some value of r less than $\mathfrak{R}(n_1, n_2, m)$, since otherwise we will not even have locally unique solution. Can the fit of $Y \in \mathcal{M}_r$ to $X + M$, and hence the choice of r, be tested in some statistical sense?

To proceed we assume the following model with noisy and possibly biased observations of a subset of matrix entries. There is a (population) value Y^* of $n_1 \times n_2$ matrix of rank $r < \mathfrak{R}(n_1, n_2, m)$ and M_{ij} are viewed as observed (estimated) values of Y^*_{ij}, $(i, j) \in \Omega$, based on a sample of size N. The observed values are modeled as

$$M_{ij} = Y^*_{ij} + N^{-1/2}\Delta_{ij} + \varepsilon_{ij},\ (i, j) \in \Omega, \tag{2.28}$$

where $Y^* \in \mathcal{M}_r$ and Δ_{ij} are some (deterministic) numbers. The random errors ε_{ij} are

assumed to be independent of each other and such that $N^{1/2}\varepsilon_{ij}$ converge in distribution to normal with mean zero and variance σ_{ij}^2, $(i,j) \in \Omega$. The additional terms $N^{-1/2}\Delta_{ij}$ in (Equation 2.28) represent a possible deviation of population values from the "true" model and are often referred to as the population drift or a sequence of local alternatives (we can refer to [27] for a historical overview of invention of the local alternatives setting). This is a reasonably realistic model motivated by many real applications.

Definition 2.4.1. *We say that the model is* globally identifiable *(at Y^*) if $\bar{Y} \in \mathbb{R}^{n_1 \times n_2}$ of $\operatorname{rank}(\bar{Y}) \leq r$ and $P_\Omega(\bar{Y}) = P_\Omega(Y^*)$ imply that $\bar{Y} = Y^*$, i.e., Y^* is the unique solution of the respective matrix completion problem. Similarly it is said that the model is* locally identifiable *if this holds for all such $\bar{Y}$ in a neighborhood of Y^*, i.e., Y^* is a locally unique solution.*

Consider the following weighted least squares problem (a generalization of (Equation 2.5)):

$$\min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} w_{ij} \left(M_{ij} - Y_{ij}\right)^2, \tag{2.29}$$

for some weights $w_{ij} > 0$, $(i,j) \in \Omega$. (Of course, if $w_{ij} = 1$, $(i,j) \in \Omega$, then problem (Equation 2.29) coincides with the least squares problem (Equation 2.5).) We have the following standard result about consistency of the least squares estimates.

Proposition 2.4.1. *Suppose that the model is globally identifiable at $Y^* \in \mathcal{M}_r$ and values M_{ij}, $(i,j) \in \Omega$, converge in probability to the respective values Y_{ij}^* as the sample size N tends to infinity. Then an optimal solution $\hat{Y}$ of problem* (Equation 2.29) *converges in probability to Y^* as $N \to \infty$.*

Consider the following weighted least squares test statistic

$$T_N(r) := N \min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} w_{ij} \left(M_{ij} - Y_{ij}\right)^2, \tag{2.30}$$

where $w_{ij} := 1/\hat{\sigma}_{ij}^2$ with $\hat{\sigma}_{ij}^2$ being consistent estimates of σ_{ij}^2 (i.e., $\hat{\sigma}_{ij}^2$ converge in probability to σ_{ij}^2 as $N \to \infty$). Recall that the respective condition of form (Equation 2.17), or equivalently (Equation 2.27), is sufficient for local identifiability of Y^*. The following asymptotic results can be compared with similar results in the analysis of covariance structures (cf., [28]).

Proposition 2.4.2 (Asymptotic properties of test statistic)**.** *Consider the noisy observation model* (Equation 2.28). *Suppose that the model is globally identifiable at $Y^* \in \mathcal{M}_r$ and Y^* is well-posed for problem* (Equation 2.2). *Then as $N \to \infty$, the test statistic $T_N(r)$ converges in distribution to noncentral χ^2 distribution with degrees of freedom $\mathrm{df}_r = m - r(n_1 + n_2 - r)$ and the noncentrality parameter*

$$\delta_r = \min_{H \in \mathcal{T}_{\mathcal{M}_r}(Y^*)} \sum_{(i,j)\in\Omega} \sigma_{ij}^{-2} \left(\Delta_{ij} - H_{ij}\right)^2. \tag{2.31}$$

Note that the optimal (minimal) value of the weighted least squares problem (Equation 2.29) can be approximated by

$$\min_{H \in \mathcal{T}_{\mathcal{M}_r}(Y^*)} \sum_{(i,j)\in\Omega} w_{ij} \left(E_{ij} - H_{ij}\right)^2 + R_N, \tag{2.32}$$

with $E_{ij} := N^{-1/2}\Delta_{ij} + \varepsilon_{ij}$ and the error term $R_N = o\left(\|M - P_\Omega(Y^*)\|^2\right)$ being of stochastic order $R_N = o_p(N^{-1})$. Hence, the noncentrality parameter, given in (Equation 2.31), can be approximated as

$$\delta_r \approx N \min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} w_{ij} \left(Y_{ij}^* + N^{-1/2}\Delta_{ij} - Y_{ij}\right)^2. \tag{2.33}$$

That is, the noncentrality parameter is approximately equal to N times the fit to the "true" model of the alternative population values $Y_{ij}^* + N^{-1/2}\Delta_{ij}$ under small perturbations of order $O(N^{-1/2})$.

Remark 2.4.1. The above asymptotic results are formulated in terms of the "sample size N" suggesting that the observed values are estimated from some data. That is, the given values $\bar{M}_{ij}$, $(i,j) \in \Omega$, are obtained by averaging i.i.d. data points M_{ij}^{ℓ}, $\ell = 1, ..., N$. In that case asymptotic normality of $N^{1/2}\varepsilon_{ij}$ can be justified by application of the Central Limit Theorem, and the corresponding variances σ_{ij}^2 can be estimated from the data in the usual way $\hat{\sigma}_{ij}^2 = (N-1)^{-1}\sum_{\ell=1}^{N}(M_{ij}^{\ell} - \bar{M}_{ij})^2$. *This model allows to formulate mathematically precise convergence results. One can take a more pragmatic point of view that when there is a "small" random noise in the observed values, the respective test statistics properly normalized with respect to magnitude of that noise have approximately a noncentral chi square distribution.*

The asymptotics of the test statistic $T_N(r)$ depends on r and also on the cardinality m of the index set Ω. Suppose now that more observations become available at additional entries of the matrix. That is we are testing now the model with a larger index set Ω', of cardinality m', such that $\Omega \subset \Omega'$. In order to emphasize that the test statistic also depends on the corresponding index set we add the index set in the respective notations. Note that if Y^* is a solution of rank r for both sets Ω and Ω' and the model is globally (locally) identifiable at Y^* for the set Ω, then the model is globally (locally) identifiable at Y^* for the set Ω'. Note also that if the regularity condition (Equation 2.17) holds at Y^* for the smaller model (i.e. for Ω), then it holds at Y^* for the larger model (i.e. for Ω'). The following result can be proved in the same way as Theorem Proposition 2.4.2 (cf., [28]).

Proposition 2.4.3. *Consider index sets $\Omega \subset \Omega'$ of cardinality $m = |\Omega|$ and $m' = |\Omega'|$, and the noisy observation model* (Equation 2.28). *Suppose that the model is globally identifiable at $Y^* \in \mathcal{M}_r$ and condition* (Equation 2.17) *holds at Y^* for the smaller model (and hence for both models). Then the statistic $T_N(r,\Omega') - T_N(r,\Omega)$ converges in distribution to noncentral χ^2 with $\mathrm{df}_{r,\Omega'} - \mathrm{df}_{r,\Omega} = m' - m$ degrees of freedom and the noncentrality parameter $\delta_{r,\Omega'} - \delta_{r,\Omega}$, and $T_N(r,\Omega') - T_N(r,\Omega)$ is asymptotically independent of $T_N(r,\Omega)$.*

For given index set Ω and observed (estimated) values M_{ij}, $(i,j) \in \Omega$, the statistic

$T_N(r)$ can be used for testing the (null) hypothesis that the "true" rank is r. That is the null hypothesis is rejected if $T_N(r)$ is large enough on the scale of the χ^2 distribution with the respective df_r degrees of freedom. It is often observed in practice that such tests reject the null hypothesis even when the fit is reasonable. In that respect the role of values Δ_{ij} in the model is to suggest that the "true" model is true only approximately, and the corresponding noncentrality parameter δ_r gives an indication of the deviation from the exact rank r model. It is a common practice to perform such tests sequentially for increasing values of r, with all deficiencies of such sequential testing.

Such testing procedure assumes that the sample size N is given and the corresponding variances σ_{ij}^2 can be consistently estimated. When the observed values are obtained by averaging N data points, this is available in the straightforward way (see Remark 2.4.1). Otherwise setting $N = 1$ and assuming that all $\sigma_{ij}^2 = \sigma^2$, $(i,j) \in \Omega$, are equal to each other, we need to specify range of σ^2. We will discuss this further in Section Section 2.5.

Remark 2.4.2. It is also possible to give asymptotic distribution of solutions of problem (Equation 2.29). Suppose now that the assumptions of Proposition 2.4.2 hold with all Δ_{ij} in equation (Equation 2.28) being zeros. Let $\hat{Y}_N$ be a solution of problem (Equation 2.29), i.e.,

$$\hat{Y}_N \in \arg\min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} w_{ij} \Big(\underbrace{Y_{ij}^* + \varepsilon_{ij}}_{M_{ij}} - Y_{ij} \Big)^2. \tag{2.34}$$

Consider operator $\mathcal{A} : \mathbb{V}_\Omega \to \mathcal{T}_{\mathcal{M}_r}(Y^*)$ defined as

$$\mathcal{A}(W) := \arg\min_{H \in \mathcal{T}_{\mathcal{M}_r}(Y^*)} \sum_{(i,j)\in\Omega} \sigma_{ij}^{-2} \left(W_{ij} - H_{ij}\right)^2, \tag{2.35}$$

for $W \in \mathbb{V}_\Omega$. Because of the assumption of well posedness (which is equivalent to (Equation 2.27)) the minimizer in (Equation 2.35) is unique and hence $\mathcal{A}(W)$ is well defined. Then

$$\hat{Y}_N = \mathcal{A}(M) + o_p(N^{-1/2}). \tag{2.36}$$

Note that the operator $\mathcal{A}$ is linear.

We have that $Y^* \in \mathcal{T}_{\mathcal{M}_r}(Y^*)$ and hence $\mathcal{A}(P_\Omega(Y^*)) = Y^*$. Thus $\mathcal{A}(M) = Y^* + \mathcal{A}(E)$, where $E \in \mathbb{R}^{n_1 \times n_2}$ is such that $E_{ij} = \varepsilon_{ij}$ for $(i,j) \in \Omega$, and $E_{ij} = 0$ otherwise. Since $N^{1/2}\varepsilon_{ij}$, $(i,j) \in \Omega$, converge in distribution to normal with mean zero and variance σ_{ij}^2 and independent of each over, it follows that $N^{1/2}(\hat{Y}_N - Y^*)$ converges in distribution to the random matrix $\mathcal{A}(Z)$, where $Z \in \mathbb{V}_\Omega$ is a random matrix with entries $Z_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$, $(i,j) \in \Omega$, having normal distribution and independent of each over. Note that since $\mathcal{A}(\cdot)$ is a linear operator, $\mathcal{A}(Z)$ has a multivariate normal distribution with zero means. Since $\mathcal{A}(Z)$ belongs to the linear subspace $\mathcal{T}_{\mathcal{M}_r}(Y^*)$ of $\mathbb{R}^{n_1 \times n_2}$, the multivariate normal distribution of $\mathcal{A}(Z)$ is degenerate.

2.5 Numerical Examples

We present some numerical experiments to illustrate our theory[1]. In this section, without further notification, the nuclear norm minimization is solved by TFOCS [29] in Matlab and LRMA problem is solved by "SoftImpute" [30] (regularization parameter equals to 0) in R.

2.5.1 An example of 6×6 matrix considered in [25]

As pointed in Remark 2.3.1, Wilson and Worcester showed in [25] using analysis that there are two different locally unique solutions of rank $r^* = 3$ for a 6×6 matrix with the index

[1] More discussions can be found in a supplementary material at https://www2.isye.gatech.edu/∼yxie77/Experiment.pdf.

set Ω corresponding to its off-diagonal elements. The matrix M in that example is given by

$$M = \begin{pmatrix} 0 & 0.56 & 0.16 & 0.48 & 0.24 & 0.64 \\ 0.56 & 0 & 0.20 & 0.66 & 0.51 & 0.86 \\ 0.16 & 0.20 & 0 & 0.18 & 0.07 & 0.23 \\ 0.48 & 0.66 & 0.18 & 0 & 0.3 & 0.72 \\ 0.24 & 0.51 & 0.07 & 0.30 & 0 & 0.41 \\ 0.64 & 0.86 & 0.23 & 0.72 & 0.41 & 0 \end{pmatrix}.$$

It can be verified that there are two rank **3** solutions by filling the diagonal entries by $(0.64, 0.85, 0.06, 0.56, 0.50, 0.93)$, and $(0.42, 0.90, 0.06, 0.55, 0.39, 1.00)$, respectively.

This simple test case where we know the ground truth can illustrate the problem. Both the nuclear norm minimization and LRMA fail to recover any of these two local solutions above. The soft-thresholded SVD converges to a completely incorrect solution with off-diagonals far off from those of M, and the nuclear norm minimization produces a rank **4** solution by filling out the diagonal entries by (0.44, 0.76, 0.05, 0.53, 0.19, 0.96). Note that here both optimal solutions satisfy the well-posedness condition, and yet these numerical procedures can not recover either one of them. It is not clear how typical this example, of different locally optimal solutions, is. Recall that generally the nuclear norm minimization problem possesses unique optimal solution. However, it is not clear how well it approximates the "true" minimal rank solution when there is observation noise.

2.5.2 Probability of well-posedness

We show the probability of satisfying the well-posedness condition by generating random cases. For each rank r^*, we generate an $40 \times r^*$ orthonormal matrix V, an $50 \times r^*$ orthonormal matrix W, and an $r^* \times r^*$ diagonal matrix D. Set $Y^* = VDW^\top$. For each instance, we randomly generate the observation pattern Ω such that each entry is observed with probability p. We check the well-posedness condition according to Theorem The-

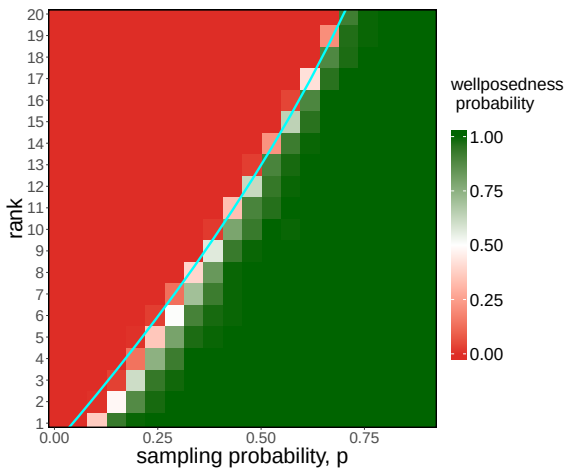

Figure 2.2: Probability that well-posedness is satisfied; random instances for different rank and sampling probability. For each sampling probability and rank, we generate Y^* and Ω. Then, we check the well-posedness condition and compute the probability. Blue curve is the estimated generic bound for the corresponding sampling probability.

orem 2.3.3 and using the verifiable algebraic condition. Then we repeat the above procedure 100 times and compute the percentage of cases that satisfy the well-posedness condition. Figure 2.2 shows the resulted proportion. We also plotted the generic bound $\hat{\mathfrak{R}}(n_1, n_2, p) = (n_1 + n_2)/2 - ((n_1 + n_2)^2/4 - n_1 n_2 p)^{1/2}$. Figure 2.2 shows that the probability that a matrix satisfies the well-posedness condition is not small, when the true rank is less than the generic lower bound. Moreover, the probability converge to 1 quickly, when the rank is 2 or 3 less than the generic bound. This demonstrates that the $\hat{\mathfrak{R}}(n_1, n_2, p)$ is a sharp bound.

2.5.3 Comparison of LRMA and nuclear norm minimization

In this section, we compare the performance of LRMA and matrix completion using standard nuclear norm minimization, when the well-posedness condition is satisfied and when it is violated, respectively. The results show that the well-posedness condition is indeed necessary for good recovery performance. Moreover, our examples show that LRMA performs more stable than nuclear norm minimization in these cases.

We generate Y^*, an $n_1 \times n_2$ matrix of rank r^*, by uniformly generated an $n_1 \times r^*$ matrix

V, an $n_2 \times r^*$ matrix W and an $r^* \times r^*$ diagonal matrix D and setting $Y^* = \tilde{V} D \tilde{W}^\top$, where $\tilde{V}$ and $\tilde{W}$ are orthonormalization of V, W, respectively. We again sample Ω uniformly random with probability p, where $|\Omega| = m$. Observation matrix M is generated by $M_{ij} = Y^*_{ij} + \varepsilon_{ij}, (i,j) \in \Omega$, where $\varepsilon_{ij} \sim N(0, \sigma^2 N^{-1})$. Algorithms stop when either relative change in the Frobenius norm between two successive estimates, $\|Y^{(t+1)} - Y^t\|_F / \|Y^{(t)}\|_F$, is less than some tolerance, denoted as tol or the number of iterations exceeds the maximum it.

Element-wise error for three cases

We first consider three individual instances, when the well-posedness condition is satisfied and violated, respectively:

(1) In Figure 2.3 the well-posedness condition is satisfied. The element-wise reconstruction error for LRMA is much smaller than that of the nuclear norm minimization. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 10$, $m = 1000$, $\sigma = 5$, $N = 50$ and Ω is sampled until the well-posedness condition is satisfied. The parameters are $tol = 10^{-20}$ and $it = 50000$.

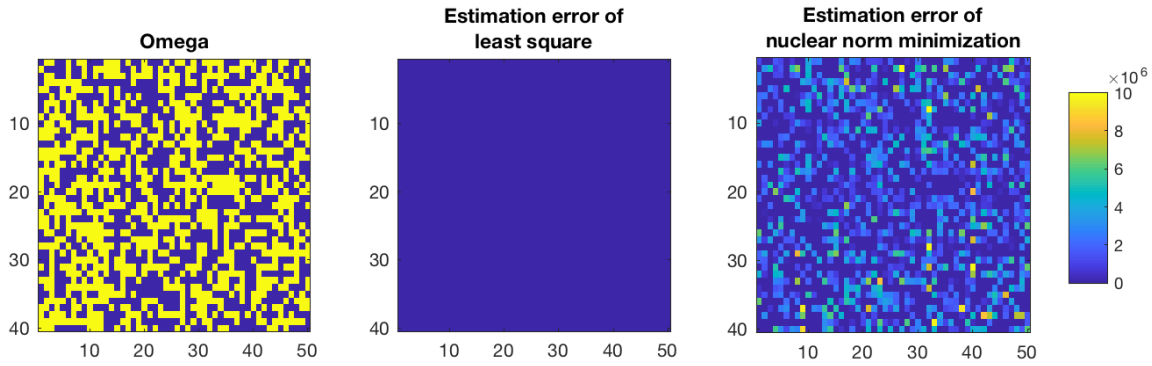

Figure 2.3: When the well-posedness condition is satisfied, the absolute errors at each entries $|Y_{ij} - Y^*_{ij}|$ for the LRMA (middle panel) and the nuclear norm minimization (right panel). The left panel shows the sampling pattern Ω. Here the true matrix $Y^* \in \mathbb{R}^{40 \times 50}$, $\mathrm{rank}(Y^*) = 10$, $|\Omega| = 1000$, $\varepsilon_{ij} \sim N(0, 5^2/50)$ and the observation matrix $M_{ij} = Y^*_{ij} + \varepsilon_{ij}, (i,j) \in \Omega$.

(2) In Figure 2.4, the well-posedness condition is violated. As predicted by our theory, both LRMA and nuclear perform worse, and the errors are especially large at index numbers 3,

6, 30, 46, 50, where the necessary condition for the well-posedness condition is violated. Still, in this situation, the nuclear norm minimization has a larger total recover error than LRMA. In this experiment, $n_1 = 70$, $n_2 = 40$, $r^* = 11$, $m = 1300$, $\sigma = 5$, and $N = 50$. We repeatedly sample Ω until the necessary condition for the well-posedness condition is violated. The parameters $tol = 10^{-16}$ and $it = 50000$.

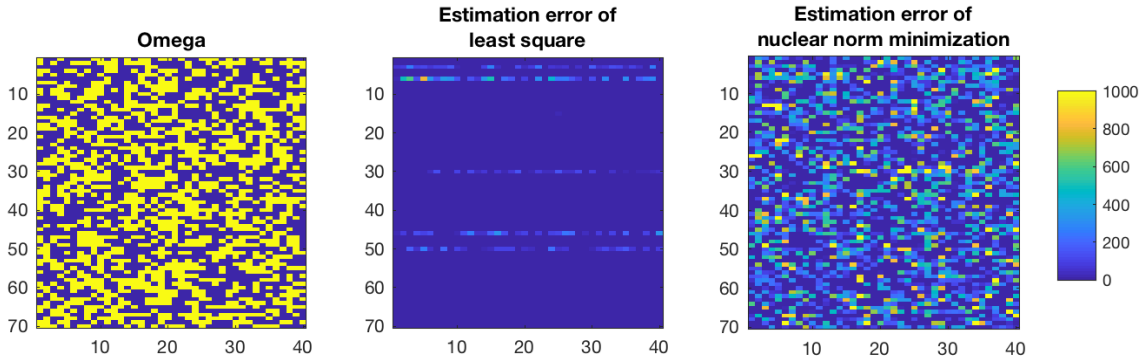

Figure 2.4: When the well-posedness condition is violated, the absolute errors at each entries $|Y_{ij} - Y^*_{ij}|$ for the LRMA (middle panel) and the nuclear norm minimization (right panel). The left panel shows the sampling pattern Ω. Here the true matrix $Y^* \in \mathbb{R}^{70\times 40}$, $\mathrm{rank}(Y^*) = 11$, $|\Omega| = 1300$, $\varepsilon \sim N(0, 5^2/50)$ and the observation matrix $M_{ij} = Y^*_{ij} + \varepsilon_{ij}, (i,j) \in \Omega$. The necessary condition for the well-posedness condition is violated (i.e., the numbers of observations are less than 11) at row with index numbers 3, 6, 30, 46, 50.

(3) In Figure 2.5, Ω is reducible and thus the well-posedness condition is violated. Consistent with our theory, in this situation, both methods fail to recover the true matrix since the necessary condition of local uniqueness is violated. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 10$, $m = 1000$, $\sigma = 5$, $N = 50$ and $\Omega = \{(i,j) \in \{1\cdots 20\} \times \{1\cdots 20\} \cup \{21\cdots 40\} \times \{21\cdots 50\}\}$. The parameters are $tol = 10^{-20}$ and $it = 50000$.

Mean-square-error performance

In this section, we consider the mean-square-error performance, defined by

$$\mathrm{MSE} = \frac{1}{n_1 n_2 K} \sum_{k=1}^{K} \sum_{i,j} (Y^*_{ij,k} - \hat{Y}_{ij,k})^2$$

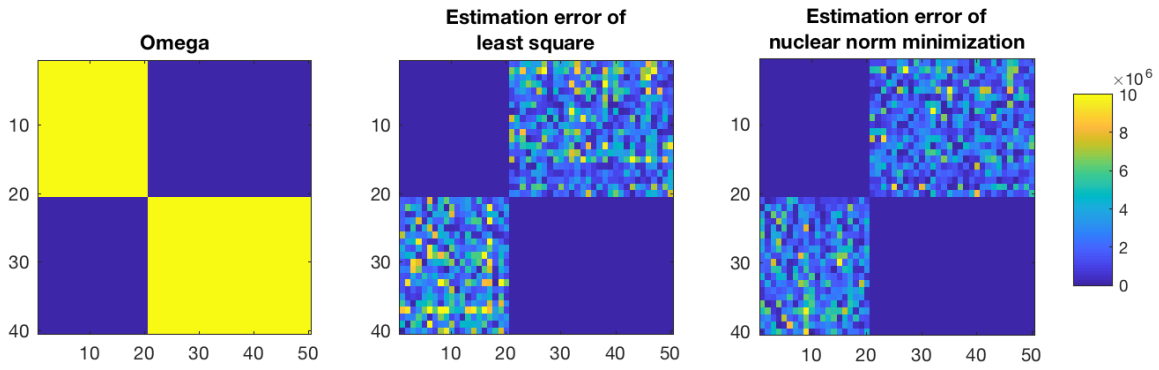

Figure 2.5: When Ω is reducible, the absolute errors at each entries $|Y_{ij} - Y^*_{ij}|$ for the LRMA (middle panel) and the nuclear norm minimization (right panel). The left panel shows the sampling pattern Ω. Here the true matrix $Y^* \in \mathbb{R}^{40\times 50}$, $\mathrm{rank}(Y^*) = 10$, $|\Omega| = 1000$, $\varepsilon_{ij} \sim N(0, \frac{5^2}{50})$ and the observation matrix $M_{ij} = Y^*_{ij} + \varepsilon_{ij}, (i,j) \in \Omega$. Ω is reducible. In this case, only two diagonal block matrices $M_1 \in \mathbb{R}^{20\times 20}$ and $M_2 \in \mathbb{R}^{20\times 30}$ are observed.

where K is the total number of repetitions. Figure 2.6 shows the difference between the *mean square error* of LRMA and the nuclear norm minimization. In this experiment, $n_1 = 40$, $n_2 = 50$, $\sigma = 5$, and we generate 50 random instances to compute the average error. The estimated $\hat{\mathfrak{R}}(n_1, n_2, p)$ is also drawn as the blue curve. Figure 2.6 shows that, indeed, as predicted by our theory, when the true rank is lower than the generic lower bound, the performance of LRMA is much better than that of the nuclear norm minimization.

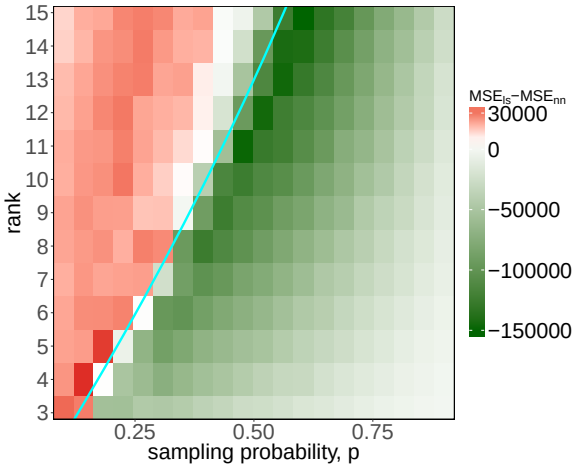

Figure 2.6: Difference between the MSEs of LRMA and the nuclear norm minimization. The blue curve is the generic bound for the corresponding sampling probability.

2.5.4 Testing for true rank

Asymptotic distribution of test statistic

In Section Section 2.4 (see (Equation 2.28)), we show that the asymptotical distribution of the test statistic for the "true" rank is χ^2 distribution, which we will verify numerically here. We generate the true matrix Y^*, an $n_1 \times n_2$ matrix of rank r^*, by uniformly generated an $n_1 \times r^*$ matrix V, an $n_2 \times r^*$ matrix W, and an $r^* \times r^*$ diagonal matrix D and setting $Y^* = \tilde{V} D \tilde{W}^\top$, where $\tilde{V}$ and $\tilde{W}$ are orthonormalization of V, W, respectively. We sample Ω uniformly random, where $|\Omega| = m$. The noisy and repeated observation matrices are generated by $M_{ij}^{(k)} = Y_{ij}^* + \varepsilon_{ij}^{(k)}, (i,j) \in \Omega$, where $\varepsilon_{ij}^{(k)} \sim N(0, \sigma^2 N^{-1})$. When computing the test statistic $T_N^{(k)}(r)$ (Equation 2.30), the least square approximation problem is solved by the soft-thresholded SVD solver. The algorithm stops when either relative change in the Frobenius norm between two successive estimates is less than some tolerance, denoted as tol, or the number of iterations reaches the maximum, denoted as it.

Figure 2.7 shows the Q-Q plot of $\{T_N^{(k)}(r)\}_{k=1}^{200}$ against the corresponding χ^2 distribution. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 11$, $m = 1000$, $\sigma = 5$, $N = 400$ and Ω is sampled until the well-posedness condition is satisfied. The parameters $tol = 10^{-20}$ and $it = 50000$. From the result, we can see $T_N(r)$ follows a central χ^2 distribution with a degree-of-freedom $\mathrm{df}_r = m - r(n_1 + n_2 - r) = 131$, which is consistent with Theorem Proposition 2.4.2.

Figure 2.8 shows the Q-Q plot of $\{T_N^{(k)}(r, \Omega') - T_N^{(k)}(r, \Omega)\}_{k=1}^{200}$ against the corresponding χ^2 distribution. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 11$, $m = 996$, $\sigma = 5$, $N = 50$, $m' = |\Omega'| = 1001$ and Ω is sampled until the well-posedness condition is satisfied. Note that Ω' also satisfied well-posedness condition since $\Omega'^C \subset \Omega^C$. The parameters $tol = 10^{-20}$ and $it = 50000$. From the result, we can see $T_N(r, \Omega') - T_N(r, \Omega)$ follows a central χ^2 distribution with a degree-of-freedom $\mathrm{df}_{r,\Omega'} - \mathrm{df}_{r,\Omega} = m' - m = 5$, which is consistent with Theorem Proposition 2.4.3.

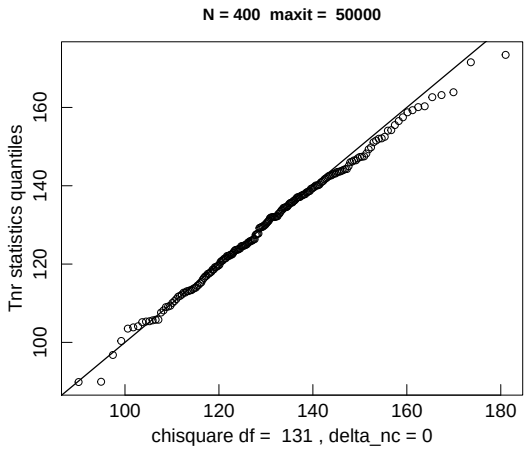

Figure 2.7: Q-Q plot of $T_N(r)$ against quantiles of χ^2 distribution: $Y^* \in \mathbb{R}^{40\times 50}$, $rank(Y^*) = 11$, $|\Omega| = 1000$, the observation matrix M is generated 200 times, $M_{ij}^{(k)} = Y_{ij}^* + \varepsilon_{ij}^{(k)}, (i,j) \in \Omega$, where $\varepsilon_{ij}^{(k)} \sim N(0, 5^2/400)$. For each $M^{(k)}$, $T_N^{(k)}(r)$ is computed as equation Equation 2.30. By Theorem Proposition 2.4.2, $\{T_N^{(k)}(r)\}$ follows central χ^2 distribution with the degree-of-freedom $\mathrm{df}_r = m - r(n_1 + n_2 - r) = 131$.

Table 2.1: p-value for sequential rank test in simulation.

rank	p-value	rank	p-value
1	0.00	7	0.00
2	0.00	8	0.00
3	0.00	**9**	**0.94**
4	0.00	10	0.69
5	0.00	11	0.41
6	0.00	12	0.00

Test for true rank

As discussed in Section 2.4, we can determine the true rank r^* by sequential χ^2 tests. That is, for r ranging from 1 to $\lceil \mathfrak{R}(n_1, n_2, m) \rceil$, we solve the least square approximations and compute $T_N(r)$. According to $T_N(r)$ we can determine which rank can be accepted for a predefined significant level. Table Table 2.1 shows a result of sequential rank test on a simulated data set. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 9$, $m = 1000$, $\sigma = 5$, $N = 100$, and Ω is sampled until well-posedness condition is satisfied. The true rank 9, is the first one accepted for 0.05 significant level.

Figure 2.9 shows the comparison of rank selection between our sequential rank test, the

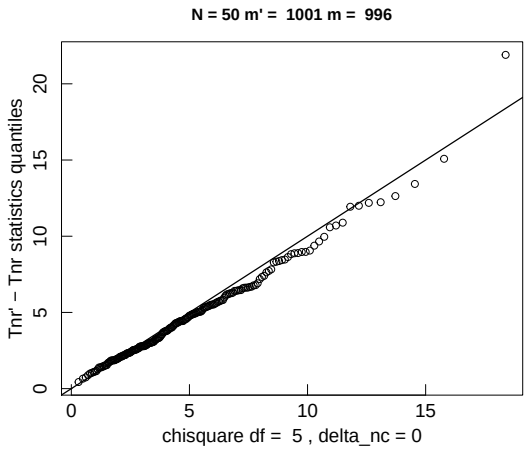

Figure 2.8: Q-Q plot of $T_N(r,\Omega') - T_N(r,\Omega)$ against the quantiles of χ^2 distribution: $Y^* \in \mathbb{R}^{40\times 50}$, rank$(Y^*) = 11$, $|\Omega'| = 1001$, $|\Omega| = 996$, where $\Omega \subset \Omega'$. The observation matrix M' and M are generated 200 times, By Theorem Proposition 2.4.3, $\{T_N^{(k)}(r,\Omega') - T_N^{(k)}(r,\Omega)\}$ follows central χ^2 distribution with the degree-of-freedom $\mathrm{df}_{r,\Omega'} - \mathrm{df}_{r,\Omega} = m' - m = 5$.

nuclear norm minimization and the method suggested in [31] (we refer to it as M^E method in the following). Since the nuclear norm minimization and the M^E method cannot give the exact rank, we choose the rank by thresholding the percentage of the singular value of the recovered matrix in this two methods, i.e. $\hat{r} = \mathrm{argmin}_r \sum_{i=1}^{r} \lambda_{(i)} / \sum_{i=1}^{min(n_1,n_2)} \lambda_{(i)} > b$, where b is a chosen threshold. In this experiment, $n_1 = 100$, $n_2 = 1000$, $\sigma = 5$, $N = 50$ and the sampling probability $p = 0.3$. For each true rank, we generate 100 instances of (Y^*, Ω, M), complete the rank selection with these three methods and compute the median of the error of estimated rank of each method. For the sequential rank test, we choose the first rank accepted with a 0.05 significant level. For the nuclear norm minimization and the M^E method, we choose the threshold that gives us the best results for these two methods. It shows that selection by sequential χ^2 test outperforms the other two methods.

2.6 Conclusion

In this chapter, we have examined the matrix completion from a geometric viewpoint and established a sufficient condition for local uniqueness of solutions. Our characterization assumes deterministic patterns and the results are general. We argue that the exact minimum

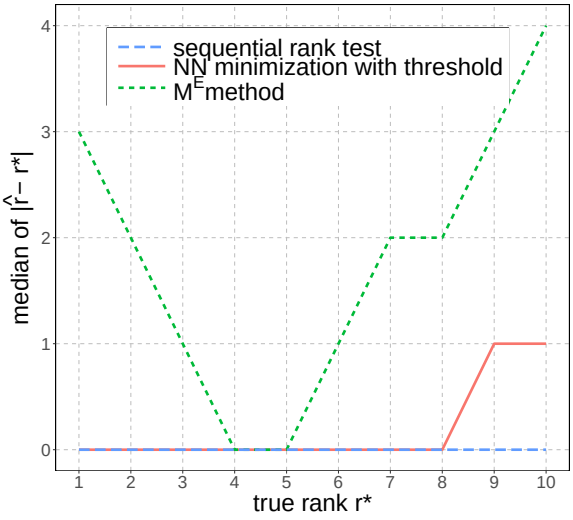

Figure 2.9: Comparison of rank selection between sequential χ^2 test, the nuclear norm minimization and the M^E method, when the sampling probability p=0.3. For each true rank, we compute the median of rank error for 100 experiments. $Y^{*(k)} \in \mathbb{R}^{100\times 1000}$, $M_{ij}^{(k)} = Y_{ij}^{*(k)} + \varepsilon_{ij}^{(k)}, (i,j) \in \Omega$, where $\varepsilon_{ij}^{(k)} \sim N(0, 5^2/50)$. Threshold $b_{nm} = 0.25$, $b_{ME} = 0.13$ for the nuclear norm minimization and the M^E method, respectively.

rank matrix completion (MRMC) leads to either unstable or non-unique solutions and thus the alternative low-rank matrix approximation (LRMA) is a more reasonable approach. We propose a statistical test for rank selection, based on observed entries, which can be useful for practical matrix completion algorithms.

For small values of the "true" rank, when the respective dual of the "true" minimum trace problem has more than one optimal solution, the asymptotic bias of the optimal value of the approximating MT problem is of order $O(N^{-1/2})$ [32]. On the other hand, under the model (Equation 2.28) when the values M_{ij}, $(i,j) \in \Omega$, are computed by averaging N data points having normal distribution (see Remark 2.4.1), the least squares approach corresponds to the Maximum Likelihood method which is an asymptotically efficient estimation procedure. This gives an insight into the relatively poor performance of the nuclear norm approach, as compared with the least squares method, as reported in Section 2.5.

CHAPTER 3

GOODNESS-OF-FIT TEST ON MANIFOLDS

3.1 Introduction

In this chapter, we develop a general theory for testing the goodness-of-fit of non-linear models. In particular, we assume that the observations are noisy samples of a submanifold (defined by a sufficiently smooth non-linear map). The observation noise is additive Gaussian. Our main result shows that the "residual" of the model fit (by solving a non-linear least-square problem) follows a (possibly non-central) χ^2 distribution. The parameters of the χ^2 distribution are related to the model order and dimensions of the problem. A key component of our analysis is the characteristic rank of the Jacobian matrix associated with the non-linear map that defines the submanifold. A natural use of our result is to the select order of a model via a sequential test procedure by choosing between two nested models. We are particularly interested in "nested" models, i.e., one can order the models by their complexity. We demonstrate the applications of this general theory in the settings of real and complex matrix completion from incomplete and noisy observations, signal source identification, and determining the number of hidden nodes in neural networks.

The rest of the chpater is organized as follows. Section 3.2 presents the background knowledge. Section 3.3 contains the main results: the test statistics for model selection on manifolds. Section 3.4 gives several examples to demonstrate the use the general theory in specific settings. Section 3.5 presents numerical experiments. Finally, Section 3.6 concludes the chapter with discussions on future directions.

Our notations are conventional. By $\|x\|_2$ we denote the Euclidean norm of vector $x \in \mathbb{R}^m$. By $\mathrm{lin}(A)$ we denote the linear space generated by columns of the matrix A and by $\mathrm{tr}(A)$ the trace of the square matrix A. For a linear space $\mathcal{L} \subset \mathbb{R}^m$, we denote by

$\mathcal{L}^{\perp} = \{y \in \mathbb{R}^m : y^{\top}x = 0,\ x \in \mathcal{L}\}$ its orthogonal space. All proofs are in the Appendix.

3.2 Background

In this section, we present the general theory, which, in particular, will help to develop subsequent test statistics for determining model orders in Section Section 3.3.

Consider a nonempty set $\Theta \subseteq \mathbb{R}^d$ and a mapping $G : \Theta \to \mathbb{R}^m$. We assume throughout the chapter that the set Θ is *open and connected*. Here, d is the dimension of the parameter space (also referred to as the intrinsic dimension), and m is the dimension of the observation space. Consider a point $\hat{y} \in \mathbb{R}^m$ and the least squares problem:

$$\min_{\theta \in \Theta} \|\hat{y} - G(\theta)\|_2^2. \tag{3.1}$$

Define the image of the mapping G,

$$\mathfrak{M} := \{G(\theta) : \theta \in \Theta\}. \tag{3.2}$$

Then problem (Equation 3.1) can be written as

$$\min_{x \in \mathfrak{M}} \|\hat{y} - x\|_2^2. \tag{3.3}$$

That is, in problem (Equation 3.3), we aim to find a point of the set $\mathfrak{M}$ such that the Euclidean distance is minimized. We deal with situations where the set $\mathfrak{M}$ is a *smooth manifold*; we will discuss this below. By saying that the manifold is *smooth* we mean that it is at least $\mathcal{C}^2$ smooth.

We assume that the map $G(\cdot)$ is at least $\mathcal{C}^2$ smooth, i.e., $G(\cdot) = (g_1(\cdot), \ldots, g_m(\cdot))$ with functions $g_i : \Theta \to \mathbb{R}$, $i = 1, \ldots, m$, being twice continuously differentiable. In some cases we make the stronger assumption that $G(\cdot)$ is *analytic*, i.e., every $g_i(\cdot)$, $i = 1, \ldots, m$, is analytic. Recall that a function is analytic on an open subset of $\mathbb{R}^d$, if it can be expanded

in power series in a neighborhood of every point of this set. For instance, every polynomial function is analytic.

With the mapping $G(\theta)$ is associated the $m \times d$ Jacobian matrix

$$J(\theta) := \partial G(\theta)/\partial\theta, \tag{3.4}$$

whose components are formed by partial derivatives

$$[J(\theta)]_{ij} = \partial g_i(\theta)/\partial\theta_j,\ i = 1, \ldots, m,\ j = 1, \ldots, d.$$

The differential of $G(\cdot)$ at a point $\theta \in \Theta$ is the linear mapping $dG(\theta) : \mathbb{R}^d \to \mathbb{R}^m$ given by $dG(\theta)h = J(\theta)h$.

Remark 3.2.1. It is possible to deal with more general settings where the set Θ is a smooth connected manifold (without boundaries) rather than an open set. In that case, the derivations below can be pushed through by considering the corresponding Jacobian matrices in the local systems of coordinates of Θ.

Definition 3.2.1 (Characteristic rank)**.** *We refer to the maximal rank of the Jacobian matrix,*

$$\mathfrak{r} := \max_{\theta\in\Theta}\{\mathrm{rank}(J(\theta))\}, \tag{3.5}$$

as the characteristic rank *of the mapping* $G(\cdot)$.

The following Proposition 3.2.1 shows that, when $G(\cdot)$ is *analytic*, the characteristic rank in a certain sense is generic. By saying that a property holds for almost every (a.e.) $\theta \in \Theta$, we mean that there is a set $\Upsilon \subset \Theta$ of Lebesgue measure zero such that the property holds for all $\theta \in \Theta \setminus \Upsilon$. Discussions of the following result can be found in [23]; we give its proof in the Appendix.

Proposition 3.2.1. *The following holds:* (i) *The set* $\{\theta \in \Theta : \mathrm{rank}(J(\theta)) = \mathfrak{r}\}$ *is open.* (ii)

If the map $G(\cdot)$ is analytic*, then for a.e. $\theta \in \Theta$ the rank of the Jacobian matrix $J(\theta)$ is equal to the characteristic rank $\mathfrak{r}$.*

If $\mathrm{rank}(J(\theta_0)) = \mathfrak{r}$ for some $\theta_0 \in \Theta$, then there is a neighborhood of θ_0 such that $\mathrm{rank}(J(\theta)) = \mathfrak{r}$ for all θ in that neighborhood. It follows by the Constant Rank Theorem (e.g., [33]) that there is a neighborhood $\mathcal{V}$ of θ_0 such that the set $G(\mathcal{V})$ forms a smooth manifold of dimension $\mathfrak{r}$, in the space $\mathbb{R}^m$, with the tangent space generated by the columns of the Jacobian matrix $J(\theta)$. When the map $G(\cdot)$ is analytic, if we choose a point θ_0 at random, with respect to a continuous distribution on the set Θ, then $\mathrm{rank}(J(\theta_0)) = \mathfrak{r}$ almost surely (with probability one).

Remark 3.2.2. Assuming that the mapping $G(\cdot)$ is $\mathcal{C}^\infty$ smooth, we have by Sard's theorem [34] that the image $\mathfrak{M}$ (of G) has Lebesgure measure zero in the observation space $\mathbb{R}^m$ if and only if $\mathfrak{r} < m$.

Definition 3.2.2 (Regularity [23])**.** *We say that a point $\theta_0 \in \Theta$ is* regular *if rank of the Jacobian matrix $J(\theta_0)$ is equal to the characteristic rank $\mathfrak{r}$ and moreover there exist neighborhoods $\mathcal{V}$ of θ_0 and $\mathcal{W}$ of $G(\theta_0)$ such that $\mathfrak{M} \cap \mathcal{W} = G(\mathcal{V})$.*

The regularity of θ_0 ensures that the local structure of $\mathfrak{M}$ near $x_0 = G(\theta_0)$ is provided by the mapping $G(\cdot)$ defined in a neighborhood of θ_0. Hence, $\mathfrak{M}$ is a smooth manifold of the dimension of the characteristic rank $\mathfrak{r}$, in a neighborhood of x_0. In particular, this implies that if $\theta' \in \Theta$ is such that $G(\theta') = G(\theta_0)$, then there are neighborhoods $\mathcal{V}'$ of θ' and $\mathcal{V}_0$ of θ_0 such that $G(\mathcal{V}') = G(\mathcal{V}_0)$. A result deeper than the one of Proposition 3.2.1(ii) says that when the coordinate mappings $g_i(\cdot)$, $i = 1, \ldots, m$, are analytic (for instance polynomial) and either the set Θ is bounded or $G(\theta) \to \infty$ as $\theta \to \infty$, then a.e. point $\theta_0 \in \Theta$ is regular (e.g., [35, Section 3.4]).

We denote by $\mathcal{T}_{\mathfrak{M}}(x)$ the *tangent space* to $\mathfrak{M}$ at a point $x \in \mathfrak{M}$, provided $\mathfrak{M}$ is a smooth manifold in a neighborhood of x. Let θ_0 be a regular point of $G(\cdot)$ and $x_0 = G(\theta_0)$. Then $\mathcal{T}_{\mathfrak{M}}(x_0) = \mathrm{lin}(J(\theta_0))$ and dimension of $\mathcal{T}_{\mathfrak{M}}(x_0)$ is equal to the rank $\mathfrak{r}$ of $J(\theta_0)$. Also,

$\mathcal{T}_{\mathfrak{M}}(x_0)$ coincides with the image of the differential $dG(\theta_0)$, i.e.,

$$\mathcal{T}_{\mathfrak{M}}(x_0) = \left\{ dG(\theta_0)h : h \in \mathbb{R}^d \right\}. \tag{3.6}$$

3.3 Test statistics on manifold

We view now the mapping $G(\theta)$ as a considered model of the parameter vector $\theta \in \Theta$, and problem (Equation 3.1) as the least squares estimation (LSE) procedure with $\hat{y}$ being a given data point. More specifically, we assume the following model

$$\hat{y} = x_0 + N^{-1/2}\gamma + \varepsilon, \tag{3.7}$$

where $x_0 \in \mathfrak{M}$ is viewed as the population (true) value, vector $\gamma \in \mathbb{R}^m$ is a deterministic bias, and the error vector ε is random. When $\hat{y}$ is estimated from a random sample, the parameter N represents the sample size. In general, N can be viewed as a normalization parameter allowing to formulate rigorous convergence results for N tending to infinity. We assume that the components ε_i, $i = 1, \ldots, m$, of ε are independent of each other and such that $N^{1/2}\varepsilon_i$ converges in distribution, as $N \to \infty$, to normal distribution with mean zero and variance $\sigma^2 > 0$. The term $N^{-1/2}\gamma$ represents systematic deviations form the "true" model and is referred to in statistics literature as the population drift (e.g.,[27]).

We consider the following least squares test statistic to determine the model

$$T_N := N\hat{\sigma}^{-2} \min_{x \in \mathfrak{M}} \|\hat{y} - x\|_2^2, \tag{3.8}$$

where $\hat{\sigma}^2$ is a consistent estimate of σ^2.

3.3.1 Test statistic on manifolds

We now consider the general case defined in (Equation 3.7). We will show that for the problem defined on smooth manifolds, similar results in the form of RSS for linear models

hold.

Remark 3.3.1. For any $\hat{y} \in \mathbb{R}^m$, the generalized least-square problem (Equation 3.3) has an optimal solution which may be not unique. If y_k is a sequence converging to $x_0 \in \mathfrak{M}$ and x_k is an optimal solution of (Equation 3.3), then x_k converges to x_0 (e.g., [36, Theorem 7.23]). Under the model (Equation 3.7) we have that $\hat{y}$ converges to x_0 in probability as $N \to \infty$. It follows that any minimizer $\hat{x}$ in the right hand side of (Equation 3.8) converges in probability to x_0.

Suppose that $\mathfrak{M}$ is a smooth manifold in a neighborhood $\mathcal{W}$ of the point x_0. If $\hat{x} \in \mathcal{W}$ is an optimal solution of the least squares problem (Equation 3.8), then it follows that

$$\hat{y} - \hat{x} \in [\mathcal{T}_{\mathfrak{M}}(\hat{x})]^{\perp}, \tag{3.9}$$

where $\mathcal{T}_{\mathfrak{M}}(\hat{x})$ is the respective tangent space (see (Equation 3.6)). The following result shows that for $\hat{y}$ sufficiently close to x_0, the necessary optimality condition (Equation 3.9) is also sufficient (cf., [32, Proposition III.4]).

Proposition 3.3.1. *Suppose that $\mathfrak{M}$ is a smooth manifold in a neighborhood of $x_0 \in \mathfrak{M}$. Then there exists a neighborhood $\mathcal{W}$ of x_0 such that if $\hat{y} \in \mathcal{W}$ and a point $\hat{x} \in \mathcal{W} \cap \mathfrak{M}$ satisfies condition* (Equation 3.9)*, then $\hat{x}$ is the unique globally optimal solution of the least squares estimation problem* (Equation 3.8)*.*

Since the least-squares problem in (Equation 3.8) is non-convex, standard optimization algorithms are at most guaranteed to converge to a stationary point satisfying first-order optimality conditions of the form (Equation 3.9). The above proposition shows that if the fit is "sufficiently good", then, in fact, the computed stationary point is globally optimal. Of course, this result is of a local nature, and it would be difficult to quantify what fit is good enough. Nevertheless, this tries to explain an empirical observation that for good fits, the problem of *local* optima does not happen too often.

Under the model (Equation 3.7) we have the following asymptotic results, which are counterparts of the properties when $\mathfrak{M}$ is a linear space (cf., [23]).

Theorem 3.3.1 (Asymptotic distribution of test statistic)**.** *Suppose that $\mathfrak{M}$ is a smooth manifold, of dimension $\mathfrak{r}$, in a neighborhood of the point $x_0 \in \mathfrak{M}$. Let P be the orthogonal projection matrix onto the tangent space $\mathcal{T}_{\mathfrak{M}}(x_0)$. Then the following holds as $N \to \infty$:*

(i) *With probability tending to one the least squares problem* (Equation 3.8) *has unique optimal solution $\hat{x}$,*

(ii) *The test statistic T_N in* (Equation 3.8) *converges in distribution to the noncentral χ^2 distribution with $m - \mathfrak{r}$ degrees-of-freedom and the noncentrality parameter $\delta = \sigma^{-2}\|(I_m - P)\gamma\|_2^2$.*

(iii) *The scaled estimator $N^{1/2}(\hat{x} - x_0)$ converges in distribution to a multivariate normal distribution with the mean vector $P\gamma$ and the covariance matrix $\sigma^2 P$.*

(iv) *The scaled error $N^{1/2}e$ converges in distribution to a multivariate normal distribution with the mean vector $(I_m - P)\gamma$ and the covariance matrix $\sigma^2(I_m - P)$, where $e = \hat{y} - \hat{x}$ is a vector of residuals.*

3.3.2 Nested models

Consider now nested models, meaning the setting such that models can be naturally ordered by their complexity. For instance, the linear regression problems, one can sequentially increase or remove the variables being used in the model. Mathematically, this poses a natural order for the parameter space. That is, let $\Theta' \subset \Theta$ be a smooth manifold of dimension d', and let

$$\mathfrak{M}' := \{G(\theta) : \theta \in \Theta'\}.$$

Let $\theta_0 \in \Theta'$ be a regular point of the mapping G. Then $\mathfrak{M}$ is a smooth manifold in a neighborhood of the point $x_0 = G(\theta_0)$. Moreover, $\mathfrak{M}'$ forms a smooth submanifold in a

neighborhood of the point x_0 with the tangent space (compare with (Equation 3.6))

$$\mathcal{T}_{\mathfrak{M}'}(x_0) = \{dG(\theta_0)h : h \in \mathcal{T}_{\Theta'}(\theta_0)\}. \tag{3.10}$$

Note that $\mathcal{T}_{\mathfrak{M}'}(x_0) \subseteq \mathcal{T}_{\mathfrak{M}}(x_0)$ and it could happen that $\mathcal{T}_{\mathfrak{M}'}(x_0) = \mathcal{T}_{\mathfrak{M}}(x_0)$ even when $d' < d$.

Consider now the test statistic

$$T'_N := N\sigma^{-2} \min_{x \in \mathfrak{M}'} \|\hat{y} - x\|_2^2. \tag{3.11}$$

We have the following results (cf., [28]).

Theorem 3.3.2. *Suppose that $\mathfrak{M}$ is a smooth manifold of dimension $\mathfrak{r}$ and $\mathfrak{M}' \subset \mathfrak{M}$ is a smooth manifold of dimension $\mathfrak{r}'$, in a neighborhood of the point $x_0 \in \mathfrak{M}'$. Then the following holds:*

(i) *T'_N converges in distribution to a noncentral χ^2 random variable with $m - \mathfrak{r}'$ degrees-of-freedom and the noncentrality parameter $\delta' = \sigma^{-2}\|(I_m - P')\gamma\|_2^2$, where P' is the orthogonal projection matrix onto the tangent space $\mathcal{T}_{\mathfrak{M}'}(x_0)$.*

(ii) *The difference statistic $T'_N - T_N$ converges in distribution to a noncentral χ^2 random variable with $(m - \mathfrak{r}') - (m - \mathfrak{r}) = \mathfrak{r} - \mathfrak{r}'$ degrees-of-freedom and the noncentrality parameter $\delta' - \delta$.*

(iii) *The statistics $T'_N - T_N$ and T_N are asymptotically independent.*

3.3.3 Decomposable maps

Now we will make additional structural assumptions about the mapping that defines the manifold of our problem. We will make sense of such structural decompositions in specific applications in Section Section 3.4. Consider model defined by the following mapping

$$G(\theta) := \mathcal{G}(\xi) + \mathcal{A}(\zeta), \tag{3.12}$$

where $\Xi \subseteq \mathbb{R}^d$ is a nonempty open connected set, $\mathcal{G} : \Xi \to \mathbb{R}^m$ is a smooth mapping and $\mathcal{A} : \mathbb{R}^k \to \mathbb{R}^m$ is a linear mapping. Note that $G(\cdot)$ inherits smoothness properties of $\mathcal{G}(\cdot)$. In particular, if $\mathcal{G}(\cdot)$ is analytic, then the corresponding mapping $G(\cdot)$ is analytic.

The parameter vector here is $\theta = (\xi, \zeta)$ and the parameter space $\Theta = \Xi \times \mathbb{R}^k$. We assume that $\mathcal{A}(\zeta) = A\zeta$, where A is an $m \times k$ matrix of rank k. Denote by

$$\mathcal{M} := \{\mathcal{G}(\xi) : \xi \in \Xi\} \text{ and } \mathcal{L} := \{\mathcal{A}(\zeta) : \zeta \in \mathbb{R}^k\}$$

the images of the mappings $\mathcal{G}$ and $\mathcal{A}$, respectively. Note that the linear space $\mathcal{L}$ has a dimension k, and $\mathfrak{M} = \mathcal{M} + \mathcal{L}$ is the image of the mapping $G : \Theta \to \mathbb{R}^m$. We denote by $\mathfrak{r}$ the characteristic rank of mapping $G(\cdot)$, and by ρ the *characteristic rank* of $\mathcal{G}(\cdot)$, i.e.,

$$\rho := \max_{\xi \in \Xi} \operatorname{rank}(\mathcal{J}(\xi)), \tag{3.13}$$

where $\mathcal{J}(\xi) = \partial \mathcal{G}(\xi)/\partial \xi$ is the Jacobian of $\mathcal{G}(\cdot)$.

Consider the corresponding least squares problem (Equation 3.3), the model (Equation 3.7) and the least squares test statistic T_N, defined in (Equation 3.8), for the mapping $G(\theta)$ of the form (Equation 3.12).

Remark 3.3.2. Note that the optimal value of least squares problem (Equation 3.3) is not changed if the point $\hat{y}$ is replaced by $\hat{y} + v$ for any $v \in \mathcal{L}$. Therefore the corresponding test statistic T_N can be considered as a function of $\hat{y}' = P_{\mathcal{L}^\perp}\hat{y}$, where $P_{\mathcal{L}^\perp} = I_m - P_{\mathcal{L}}$ is the orthogonal projection onto the linear space orthogonal to $\mathcal{L}$.

Recall that $\mathfrak{M} = \mathcal{M} + \mathcal{L}$. If $\mathfrak{M}$ is a smooth manifold, of dimension $\mathfrak{r}$, in a neighborhood of x_0, then Theorems Theorem 3.3.1 and Theorem 3.3.2 can be applied. In particular, it will follow that the test statistic T_N converges in distribution to a noncentral χ^2 with $m - \mathfrak{r}$ degrees-of-freedom and certain noncentrality parameter.

Note that for $\theta = (\xi, \zeta) \in \Theta$, the differential $dG(\theta) : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^m$ is given by

$$dG(\theta)(h, z) = d\mathcal{G}(\xi)h + Az, \ h \in \mathbb{R}^d, z \in \mathbb{R}^k. \tag{3.14}$$

This implies that the corresponding characteristic rank $\mathfrak{r} \leq \rho + k$.

Definition 3.3.1. *We say that a point $x \in \mathcal{M}$ is* well-posed *if $\mathcal{M}$ is a smooth manifold of dimension ρ in a neighborhood of x and*

$$\mathcal{T}_{\mathcal{M}}(x) \cap \mathcal{L} = \{0\}. \tag{3.15}$$

We say that the model is well-posed *if*

$$\mathfrak{r} = \rho + k. \tag{3.16}$$

For the matrix completion problem the well-posedness condition (at a point) was introduced in [32]. Note that condition (Equation 3.15) means that

$$\dim(\mathcal{T}_{\mathcal{M}}(x) + \mathcal{L}) = \dim(\mathcal{T}_{\mathcal{M}}(x)) + \dim(\mathcal{L}). \tag{3.17}$$

Of course, a necessary condition for (Equation 3.17) to hold is that $\rho + k \leq m$. Note also that assuming the mapping $\mathcal{G}(\cdot)$, and hence the mapping $G(\cdot)$, is analytic we have that the image $\mathfrak{M} = \mathcal{M} + \mathcal{L}$ has Lebesgue measure zero in the observation space $\mathbb{R}^m$ if and only if $\mathfrak{r} < m$ (see Remark 3.2.2).

Proposition 3.3.2. *Suppose that the mapping $\mathcal{G}(\cdot)$ is analytic. Then the following holds. If there exists at least one well-posed point $x \in \mathcal{M}$, then the model is well-posed. Conversely if $\mathcal{M}$ is a smooth manifold of dimension ρ and the model is well-posed, then for a.e. $\xi \in \Xi$, the corresponding point $x = \mathcal{G}(\xi)$ is well-posed.*

Let us make the following observation. By the definition of $\mathfrak{M}$ under the decomposition

(Equation 3.12), we have that the point $x_0 \in \mathfrak{M}$ can be represented as

$$x_0 = x^* + v_0 \text{ for some } x^* \in \mathcal{M}, v_0 \in \mathcal{L}. \tag{3.18}$$

Definition 3.3.2. *We say that the model is* identifiable *at x^* (at x_0) if the representation* (Equation 3.18) *is unique, i.e., if $x_0 = x' + v'$ with $x' \in \mathcal{M}$ and $v' \in \mathcal{L}$, then $x' = x^*$. We say that the model is* locally identifiable *at x^*, if such uniqueness holds locally, i.e., there is a neighborhood $\mathcal{W}$ of x^* such that if $x_0 = x' + v'$ with $x' \in \mathcal{M} \cap \mathcal{W}$ and $v' \in \mathcal{L}$, then $x' = x^*$.*

The following result can be proved in the same way as [32, Theorem III.2].

Proposition 3.3.3. *If a point $x^* \in \mathcal{M}$ is well-posed, then the model is locally identifiable at x^*.*

To verify the (global) identifiability of a nonlinear model is difficult, and often is out of reach. Of course, local identifiability is a necessary condition for global identifiability. When $\mathcal{G}(\cdot)$ is analytic, the well-posedness condition (Equation 3.16) can be verified numerically; it is necessary and sufficient for the local identifiability in the generic sense of Proposition 3.3.2. We argue that the well-posedness condition is a minimal property that should be verified for a considered model.

3.4 Applications of general theory

In this section, we present several examples in signal processing and machine learning to illustrate how to use the general theory, developed in the previous section, to determine the "model order" in the specific setting.

Remark 3.4.1. For some well-structured manifolds, it is possible to give an explicit formula for the characteristic rank. In more complicated settings, we can find the characteristic rank numerically. That is, we compute the Jacobian matrix of the considered mapping at

several randomly generated points of Θ, and subsequently compute its rank. By Proposition 3.2.1, we can expect that this will give us the characteristic rank of the considered mapping. This approach worked quite well in experiments reported in Sections subsection 3.4.2, subsection 3.4.3, and subsection 3.4.6 below.

3.4.1 Noisy matrix completion

We first show that the problem of selecting rank for noisy matrix completion can be addressed using our general theory. Part of the relevant discussion can be found in [32]; here, we generate a conclusion using the framework of our general theory in this chapter.

Consider the noisy matrix completion problem (e.g., [3], [4],[2] and references there in). Suppose we observe a subset of entries of a low-rank matrix with Gaussian noise and aim to recover the matrix. A common approach to solve this problem, is to use a matrix factorization by selecting a rank of the matrix using subjective choice or experiments and cross-validation. However, it is not clear what would be a good statistical procedure to determine the rank of the matrix.

Consider a mapping $G(\theta)$ of the form (Equation 3.12) with the following parameters. Let $\xi = (V, W)$ with $V \in \mathbb{R}^{n_1 \times r}$ and $W \in \mathbb{R}^{n_2 \times r}$, $r \leq \min\{n_1, n_2\}$, and let $\Xi \subset \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r}$ be the set of such ξ with both matrices V and W having full column rank r. Define

$$\mathcal{G}(\xi) := VW^\top \in \mathbb{R}^{n_1 \times n_2},$$

and

$$\mathcal{L} := \{X \in \mathbb{R}^{n_1 \times n_2} : X_{ij} = 0,\ (i,j) \in \Omega\},$$

for an index set $\Omega \subset \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$. Then $\mathcal{M} = \mathcal{M}_r$ forms the set of $n_1 \times n_2$ matrices of rank r. Note that the set Ξ is an open connected subset of $\mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r}$, and

$$\dim(\mathcal{L}) = n_1 n_2 - |\Omega|,$$

where $|\Omega|$ is the cardinality (number of elements) of the index set Ω. The parameter set

$$\Theta = \Xi \times \mathbb{R}^{n_1 n_2 - |\Omega|}.$$

Here the least squares problem of (Equation 3.8), associated with the test statistic T_N, can be written as

$$\min_{X \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} (\hat{Y}_{ij} - X_{ij})^2, \tag{3.19}$$

where $\hat{Y}_{ij}$, $(i,j) \in \Omega$, are observed values of the data matrix. Then the model (Equation 3.7) can be written as

$$\hat{Y}_{ij} = X^*_{ij} + N^{-1/2}\Gamma_{ij} + \varepsilon_{ij},\ (i,j) \in \Omega, \tag{3.20}$$

where $X^* \in \mathcal{M}_r$. Note that here the test statistic T_N is a function of the components $\hat{Y}_{ij}$, $(i,j) \in \Omega$, of the corresponding matrix $\hat{Y}$ (compare with Remark 3.3.2 and (Equation 3.18)).

It is well known that the set $\mathcal{M}_r$, of $n_1 \times n_2$ matrices of rank $r > 0$, is a smooth manifold of dimension $r(n_1+n_2-r)$ in a neighborhood of its every point (excluding origin). Therefore here every $\xi \in \Xi$ is a regular point of the mapping $\mathcal{G}(\cdot)$ with the characteristic rank $\rho = r(n_1 + n_2 - r)$. Thus for the characteristic rank $\mathfrak{r}$ of the corresponding mapping $G(\cdot)$ we have that

$$\mathfrak{r} \leq r(n_1 + n_2 - r) + n_1 n_2 - |\Omega|, \tag{3.21}$$

and that the model is well-posed if and only if the equality holds in (Equation 3.21).

Let us make the following assumption.

(A) The set $\mathfrak{M} = \mathcal{M}_r + \mathcal{L}$ is a smooth manifold, of dimension $\mathfrak{r}$, in a neighborhood of the point X.

Note that if Assumption (A) holds, then $\mathfrak{M}$ is a smooth manifold of dimension $\mathfrak{r}$ in a neighborhood of $X' = X + U$ for any $U \in \mathcal{L}$. Therefore by the discussion of Section Section 3.2, the above assumption (A) holds generically. By Theorem Theorem 3.3.1 we have the following result as N tends to infinity (cf., [32]).

Proposition 3.4.1. *Suppose that Assumption (A) holds. Then the test statistic T_N converges in distribution to a noncentral χ^2 with degrees-of-freedom $n_1 n_2 - \mathfrak{r}$ and the noncentrality parameter*

$$\delta = \sigma^{-2} \min_{H \in \mathcal{T}_{\mathcal{M}_r(X^*)}} \sum_{(i,j)\in\Omega} (\Gamma_{ij} - H_{ij})^2. \tag{3.22}$$

Moreover, applying Proposition 3.3.2, we can conclude the following under the assumption:

(B) The point X^* is well-posed and the model is identifiable at X^*.

Proposition 3.4.2. *Suppose that Assumption (B) holds. Then: (i) the equality holds in* (Equation 3.21)*, (ii) the test statistic T_N converges in distribution to noncentral χ^2 with degrees-of-freedom $|\Omega| - r(n_1 + n_2 - r)$ and the noncentrality parameter δ given in* (Equation 3.22)*, (iii) with probability tending to one, problem* (Equation 3.19) *has a unique optimal solution $\{\hat{X}_{ij}\}_{(i,j)\in\Omega}$.*

The difference test statistic can be applied to the following setting. Consider another index set $\Omega' \subset \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ of cardinality $|\Omega'|$ such that $\Omega \subset \Omega'$ and the corresponding space

$$\mathcal{L}' := \{X \in \mathbb{R}^{n_1 \times n_2} : X_{ij} = 0,\ (i,j) \in \Omega'\}.$$

Clearly, $\mathcal{L}'$ is a subspace of $\mathcal{L}$, and the corresponding set

$$\Theta' = \Xi \times \mathbb{R}^{n_1 n_2 - |\Omega'|},$$

is a linear subspace of the set Θ. By Theorem Theorem 3.3.2 we have the following.

Proposition 3.4.3. *Suppose that Assumption (A) holds and moreover $\mathfrak{M}'$ is a smooth manifold, of dimension $\mathfrak{r}'$, in a neighborhood of $X^* \in \mathfrak{M}'$. Then the difference statistic $T'_N - T_N$ converges in distribution to noncentral χ^2 with degrees-of-freedom $\mathfrak{r} - \mathfrak{r}'$ and the noncen-*

trality parameter $\delta' - \delta$. Moreover, the statistics $T'_N - T_N$ and T_N are asymptotically independent.

The above result can be used to compare the goodness-of-fit of two models.

Remark 3.4.2. An application of Theorem 3.3.2 and Proposition 3.4.3 allows to estimate σ^2 when the variance of the noise is unknown. Specifically, let's assume $N = 1$, $\Gamma_{ij} = 0$ and ε_{ij} follows normal distribution with zero mean and variance σ^2. Denote the set of observation indices as Ω'. By leaving out some observation, we have a new set of observation indices Ω such that $\Omega \subset \Omega'$. Then we can construct the estimate of σ^2 as the following:

$$\begin{aligned}
\tilde{T}'_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j) \in \Omega'} (\hat{Y}_{ij} - X_{ij})^2, \\
\tilde{T}_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j) \in \Omega} (\hat{Y}_{ij} - X_{ij})^2, \\
\hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{|\Omega'| - |\Omega|}.
\end{aligned}$$

By Theorem 3.3.2 and Proposition 3.4.3, we have $\sigma^{-2}(\tilde{T}'_N - \tilde{T}_N)$ follows a χ^2 distribution with degrees-of-freedom $|\Omega'| - |\Omega|$ asymptotically for the true model. Therefore, $\hat{\sigma}^2$ is a consistent estimator of σ^2. This method can be generalized to the other applications in this chapter and more discussion is provided in the Appendix.

3.4.2 Complex noisy matrix completion

In this section, we generalize the results to "complex matrix completion." Here, the observations and underlying low-rank matrices are over the field $\mathbb{C}$ of complex numbers. Consider the matrix completion problem (over complex numbers), where $X \in \mathbb{C}^{n_1 \times n_2}$, $V \in \mathbb{C}^{n_1 \times r}$, $W \in \mathbb{C}^{n_2 \times r}$:

$$\min_{V,W} \|X - VW^\top\|_2^2 \text{ s.t. } X_{ij} = b_{ij}, \ (i,j) \in \Omega. \tag{3.23}$$

This can be formulated in terms of a real numbers problem as follows. Write

$$V = V_1 + \mathfrak{i}V_2,$$

where $\mathfrak{i}^2 = -1$, $V_1 \in \mathbb{R}^{n_1 \times r}$, and $V_2 \in \mathbb{R}^{n_1 \times r}$ are the real and imaginary parts of matrix $V \in \mathbb{C}^{n_1 \times r}$. Similarly, let

$$W = W_1 + \mathfrak{i}W_2, \quad X = X_1 + \mathfrak{i}X_2.$$

Then

$$VW^\top = (V_1W_1^\top - V_2W_2^\top) + \mathfrak{i}(V_1W_2^\top + V_2W_1^\top).$$

Define

$$\mathcal{L}_1 := \{U \in \mathbb{R}^{n_1 \times n_2} : U_{ij} = 0, (i,j) \in \Omega\},$$

and $\mathcal{L} = \mathcal{L}_1 \times \mathcal{L}_1$. Then we can set

$$\theta = (V_1, W_1, V_2, W_2, U_1, U_2),$$

and mapping

$$G(\theta) := (G_1(\theta), G_2(\theta)),$$

where

$$G_1(\theta) = V_1W_1^\top - V_2W_2^\top + U_1,$$

$$G_2(\theta) = V_1W_2^\top + V_2W_1^\top + U_2,$$

and $U_1 \in \mathcal{L}_1$ and $U_2 \in \mathcal{L}_1$. Hence we can write the problem (Equation 3.23) in the following form

$$\begin{aligned} \min_{\theta} \quad & \|X_1 - G_1(\theta)\|_2^2 + \|X_2 - G_2(\theta)\|_2^2 \\ \text{s.t.} \quad & X_{1,ij} = b_{1,ij}, X_{2,ij} = b_{2,ij},\ (i,j) \in \Omega, \\ & X_{1,ij} = X_{2,ij} = 0,\ (i,j) \in \Omega^c. \end{aligned} \tag{3.24}$$

The dimension of the manifold of $n_1 \times n_2$ complex matrices of rank r, in terms of real numbers, is twice the corresponding dimension $r(n_1 + n_2 - r)$ in the real case. That is, the characteristic rank of the respective mapping $\mathcal{G}(\cdot)$ here is

$$\mathfrak{r} = 2r(n_1 + n_2 - r).$$

Note that this differs from the real-value matrix completion case in subsection 3.4.1 by a factor of 2.

3.4.3 Low-rank matrix sensing

Matrix sensing problems [37] is related to matrix completion, where the observations are linear projections of the underlying low-rank matrix. Specifically, denote by $\mathbb{S}^{d\times d}$ the space of $d \times d$ symmetric matrices, and $\langle A, B\rangle := \operatorname{tr}(AB)$ the scalar product of $A, B \in \mathbb{S}^{d\times d}$. Let $X^* \in \mathbb{S}^{d\times d}$ be a positive semidefinite matrix of rank r needed to be recovered. Given measurement matrices $A_i \in \mathbb{S}^{d\times d}$, $i = 1, \ldots, m$, we observe $y \in \mathbb{R}^m$, such that

$$y_i = \langle A_i, X^* \rangle.$$

Then we aim to solve the following least square problem.

$$\min_{U\in\mathbb{R}^{d\times r}} f(U) := \sum_{i=1}^{m} \left(y_i - \langle A_i, UU^\top\rangle\right)^2. \tag{3.25}$$

It is shown in [37] that (Equation 3.25) is the same problem as the problem of fitting one-layer neural networks with quadratic activation in (Equation 3.27), which we discuss next.

3.4.4 One-hidden-layer neural networks

We will show the general theory can be applied to determine the number of hidden nodes. Consider a one-layer neural networks. Let $x_i \in \mathbb{R}^d$ be the inputs and the observation is assume to be generated by:

$$y_i = \mathbf{1}^\top q(U^{*\top} x_i) + \varepsilon_i, \tag{3.26}$$

where $U^* \in \mathbb{R}^{d\times r}$, $\mathbf{1} \in \mathbb{R}^r$ with all entries equal to 1 and ε_i is the Gaussian noise with mean zero and variance σ^2. The activation function can be one of the following,

(i) Quadratic activation:

$$q(z_1, \cdots, z_r) = (z_1^2, z_2^2, \cdots, z_r^2).$$

(ii) Sigmoid activation:

$$q(z_1, \cdots, z_r) = (1/(1+e^{-z_1}), \cdots, 1/(1+e^{-z_r})).$$

A commonly used approach to fit neural networks is to solve the least square problem:

$$\min_{U\in\mathbb{R}^{d\times r}} f(U) := \sum_{i=1}^{m} \big(y_i - \mathbf{1}^\top q(U^\top x_i)\big)^2. \tag{3.27}$$

Define $\Theta = \mathbb{R}^{d\times r}$, for $U \in \Theta$,

$$G(U) = (g_1(U), \ldots, g_m(U)),$$

where $g_i(U) = \mathbf{1}^\top q(U^\top x_i)$. In this setting problem (Equation 3.27) becomes a least

squares problem of the form (Equation 3.1).

It is difficult to evaluate the characteristic rank $\mathfrak{r}$ of the mapping G in a theoretical way. By computing the rank of the corresponding Jacobian matrix (see Remark 3.4.1), we find the following formulas for the characteristic rank fit well in numerical experiments:

$$\mathfrak{r} = dr - r(r-1)/2,$$

for the Quadratic activation function; and $\mathfrak{r} = dr$ for the Sigmoid activation function.

3.4.5 Tensor completion

Next, we consider the problem of determining the rank of a tensor from incomplete and noisy observations to illustrate the role of the general theory.

Consider a tensor $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ of order d over the field of real numbers. It is said that X has rank one if

$$X = a^1 \circ \cdots \circ a^d,$$

where $a^i \in \mathbb{R}^{n_i}$ is $n_i \times 1$ vector, $i = 1, \ldots, d$, and " $\circ$ " denotes the vector outer product. That is, every element of tensor X can be written as the product

$$X_{i_1,\ldots,i_d} = a^1_{i_1} \times \cdots \times a^d_{i_d}.$$

The smallest number r such that tensor X can be represented as a sum $X = \sum_{i=1}^{r} Y_i$ of rank one tensors Y_i is called the rank of X, and the corresponding decomposition is often referred to as the (tensor) rank decomposition, minimal CP decomposition, or Canonical Polyadic Decomposition (CPD).

The *tensor completion problem* can be formulated as the problem of reconstructing tensor of rank r by observing a relatively small number of its entries. The second order tensor (i.e., when $d = 2$) can be viewed as a matrix, and this becomes the matrix completion

problem discussed in Section subsection 3.4.1. Consider now third order tensors $X \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, and denote by $\mathcal{M}_r$ third order tensors of rank r. Without loss of generality, we can assume that $n_1 \geq n_2 \geq n_3$. With tensor $X \in \mathcal{M}_r$ are associated matrices $A \in \mathbb{R}^{n_1 \times r}$, $B \in \mathbb{R}^{n_2 \times r}$, $C \in \mathbb{R}^{n_3 \times r}$ such that

$$X = A \otimes B \otimes C,$$

meaning that

$$X = \sum_{i=1}^{r} a^i \circ b^i \circ c^i,$$

with a^i, b^i, c^i being ith columns of the respective matrices A, B, C.

The above leads to the following parameterization of $\mathcal{M}_r$. For

$$\xi = (A, B, C) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r},$$

consider mapping

$$\mathcal{G}(\xi) := A \otimes B \otimes C.$$

By definition of the tensor rank we have that rank of tensor $X = \mathcal{G}(\xi)$ cannot be larger than r. So we define the parameter set

$$\Xi := \left\{ \xi \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r} : \mathcal{G}(\xi) \in \mathcal{M}_r \right\}. \tag{3.28}$$

We need to verify that the set Ξ is open and connected. Note that it could happen that the complement $(\mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}) \setminus \Xi$ of the set Ξ, has positive (Lebesgue) measure, or even that Ξ has measure zero.

Careful analysis of properties of $\mathcal{M}_r$ is not trivial and is beyond the scope of this chapter. We will make some comments below. Let us consider the following examples. Suppose that $n_3 = 1$. In that case, assuming that the elements of a matrix $C \in \mathbb{R}^{1 \times r}$ are nonzero,

by rescaling columns of the respective matrices A and B, we can assume that all elements of C equal 1. Consequently, essentially, this becomes the matrix completion problem discussed in Section subsection 3.4.1. Thus the characteristic rank of $\mathcal{G}(\xi)$ in that case is $\mathfrak{r} = r(n_1 + n_2 - r)$.

The key question of the tensor rank decomposition is its uniqueness. Clearly the decomposition $X = A \otimes B \otimes C$, of $X \in \mathcal{M}_r$, is invariant with respect to permutations, and rescaling of the columns of matrices A, B, C by factors $\lambda_{1i}, \lambda_{2i}, \lambda_{3i}$, $i = 1, \ldots, r$, such that $\lambda_{1i}\lambda_{2i}\lambda_{3i} = 1$. It is said that the decomposition $X = A \otimes B \otimes C$ is (globally) *identifiable* if it is unique up to the corresponding permutation and rescaling. It is beyond the scope of this chapter to give a careful discussion of the (very nontrivial) problem of tensor rank identifiability. As it was pointed above, for $n_3 = 1$ this becomes the matrix rank problem for which the identifiability never holds for $r > 1$ (e.g., [38, section 3.2]).

Suppose now that $n_3 \geq 2$. In that case the situation is different.

Definition 3.4.1. *It is said that the rank r decomposition is* generically *identifiable if for almost every $(A, B, C) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}$ the corresponding tensor $A \otimes B \otimes C$ has identifiable rank r.*

In particular, the generic identifiability implies that the complement of the parameter set Ξ, defined in (Equation 3.28), has (Lebesgue) measure zero. It is known that for sufficiently small r, the identifiability holds in the generic sense (we refer to [39],[40], and references therein for a discussion of the tensor rank identifiability from a generic point of view).

The identifiability is related to the characteristic rank:

Definition 3.4.2. *We say that $(A, B, C) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}$ is* locally identifiable *if there is a neighborhood $\mathcal{W}$ of (A, B, C) such that $(A', B', C') \in \mathcal{W}$ and $A' \otimes B' \otimes C' = A \otimes B \otimes C$ imply that (A', B', C') can be obtained from (A, B, C) by the corresponding rescaling. We say that model (n_1, n_2, n_3, r) is* generically *locally identifiable if a.e. $(A, B, C) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}$ is locally identifiable.*

Note that local identifiability of $(A, B, C) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}$ is a local property, it could happen that rank of the corresponding tensor $A \otimes B \otimes C$ is less than r. If indeed the rank of tensor $A \otimes B \otimes C$ is r, then its global identifiability implies its local identifiability (note that the permutation invariance does not affect the local identifiability). Note also that the rank of the Jacobian matrix of a mapping $\mathcal{G}(\xi)$ is always less than or equal to $r(n_1 + n_2 + n_3) - 2r$. This follows by counting the number of elements in (A, B, C) and making corrections for the scaling factors. That is, the characteristic rank $\mathfrak{r}$ of map $\mathcal{G}(\cdot)$ cannot be larger than $r(n_1 + n_2 + n_3 - 2)$.

Proposition 3.4.4. *Model (n_1, n_2, n_3, r) is generically locally identifiable if and only if the following formula for the characteristic rank $\mathfrak{r}$ holds,*

$$\mathfrak{r} = r(n_1 + n_2 + n_3 - 2). \tag{3.29}$$

Since the generic (global) identifiability implies generic local identifiability we have the following consequence of the above proposition.

Corollary 3.4.1. *If the rank r decomposition is generically identifiable, then formula* (Equation 3.29) *for the characteristic rank follows.*

3.4.6 Determining number of sources in blind de-mixing problem

De-mixing problem (e.g., [41]) is a fundamental challenge in signal processing, which arises from applications such as ambient noise seismic imaging [42], NMR imaging, etc. In such problems, the goal is to recover the signals by observing their weighted mixture. Blind de-mixing is particularly challenging in which we do not know the waveforms of the signal. Moreover, the number of signals and the magnitudes of the waveforms are also unknown. Such a problem has been addressed using a matrix factorization approach [43]. However, in existing approaches, there is no efficient method to determine the number of signals, which is usually a critical input parameter to algorithms. In this section, we show

how to determine the number of sources in the context of ambient noise imaging using the general theory.

Assume there are N sensors. Define the signal received by the nth sensor as follows:

$$x_n(t) = \sum_{k=1}^{K} s_k(t - \tau_{n,k}), \quad n = 1, \ldots, N. \tag{3.30}$$

Assume the number of signals K and the delays $\tau_{n,k}$ are all unknown. Further assume the signal is a Gaussian function

$$s_k(t) = \rho_k e^{-\alpha_k t^2},$$

where α_k defines the width of the kth source, and ρ_k is the magnitude of the kth source. Here, our goal is to estimate the number of signal sources K from observations of $x_n(t)$ buried in Gaussian noise.

We now derive the observation model. For the ease of presentation, we present the derivation in continuous time (and continuous frequency) domain, and the switch to discrete-time (and discrete frequency) domain later. Let the Fourier transform of the signal to be

$$S_k(f) := \mathcal{F}\{s_k(t)\}(f) = \int_{-\infty}^{\infty} s_k(t) e^{-2\pi \mathrm{i} t f} dt.$$

Recall that the Fourier transform of the delayed signal corresponds to a phase-shift. Hence, for Gaussian signals in (Equation 3.30), it can be shown that

$$\mathcal{F}\{s_k(t - \tau)\}(f) = \rho_k \sqrt{\frac{\pi}{\alpha_k}} e^{-2\pi \mathrm{i} f \tau} e^{-\pi^2 f^2 / \alpha_k}.$$

For continuous function h_1 and h_2, the cross-correlation is defined as:

$$(h_1 \otimes h_2)(s) := \int_{-\infty}^{\infty} h_1(t - s) h_2(t) dt.$$

Here, in this section, $\otimes$ represents the cross-correlation operator. By the duality of convo-

lution in frequency and time, we have

$$\mathcal{F}\{h_1 \otimes h_2\}(f) = \mathcal{F}\{h_1\}^*(f)\mathcal{F}\{h_2\}(f),$$

where $(\cdot)^*$ denotes the conjugate of a complex number.

In ambient noise imaging, the useful "signal" are extracted by performing pairwise cross-correlation between sensors. Define $r_{n,m}(t)$ as the cross-correlation function of the nth and the mth sensors:

$$\begin{aligned} r_{n,m}(t) &= x_n(t) \otimes x_m(t) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K} s_l(t-\tau_{n,l}) \otimes s_k(t-\tau_{m,k}). \end{aligned}$$

Now consider the frequency domain. Denote the Fourier transform operator by $\mathcal{F}$ and frequency by f. Define $R_{n,m}(f)$ as the Fourier transform of $r_{n,m}$ at the frequency f,

$$\begin{aligned} R_{n,m}(f) :=& \mathcal{F}\{r_{n,m}(t)\}(f) \\ =& \sum_{k=1}^{K}\sum_{l=1}^{K} Q_{lk}(f) \cdot e^{2\pi \mathrm{i} f(\tau_{n,l}-\tau_{m,k})}. \end{aligned} \tag{3.31}$$

where

$$Q_{lk}(f) = \mathcal{F}\{s_l(t) \otimes s_k(t)\}(f) = S_l^*(f)S_k(f).$$

The matrix $Q(f)$ depends on unknown signal waveforms $s_k(t)$ as well as the the number of sources K. For Gaussian signals defined in (Equation 3.30), we can write specifically

$$\begin{aligned} R_{n,m}(f) &= \sum_{k=1}^{K}\sum_{l=1}^{K} Q_{lk}(f) \cdot e^{2\pi \mathrm{i} f(\tau_{n,l}-\tau_{m,k})} \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K} \rho_k\rho_l e^{2\pi \mathrm{i} f(\tau_{n,l}-\tau_{m,k})} \pi \sqrt{\frac{1}{\alpha_k\alpha_l}} e^{-\pi^2 f^2(\frac{1}{\alpha_k}+\frac{1}{\alpha_l})}. \end{aligned}$$

Now we can write $R_{n,m}(f)$ in (Equation 3.31) in a compact form and show its low-rank structure. Define a matrix $Q(f) \in \mathbb{C}^{K\times K}$, where the (l,k)th entry of the matrix is $Q_{lk}(f)$. Clearly, $Q(f)$ is a rank-one complex matrix. Define

$$S(f) = [S_1{}^*(f), \ldots, S_K^*(f)]^\top,$$

then

$$Q(f) = S(f)S(f)^H,$$

where $(\cdot)^H$ denote the Hermitian of a complex vector or matrix (i.e., the complex conjugate and transpose). Define

$$\alpha_n = \left[e^{-2\pi \mathrm{i} f\tau_{n,1}}, e^{-2\pi \mathrm{i} f\tau_{n,2}}, \ldots, e^{-2\pi \mathrm{i} f\tau_{n,K}}\right]^\top.$$

We have

$$R_{n,m}(f) = \alpha_n^H Q(f)\alpha_m, \forall f.$$

Define a matrix $A = [\alpha_1, \ldots, \alpha_N] \in \mathbb{C}^{K\times N}$, and a matrix $R(f)$, whose (n,m)th entry is given by $R_{nm}(f)$. We can further write

$$R(f) = A^H Q(f) A, \forall f.$$

Assume our observations are a subset of entries of the tensor R with additive Gaussian noise. The missing data can be due to distance and communication constraints; see [44] for context. Certain pairs of cross-correlations functions are not available. This can happen when sensors far away, and it is impractical for them to communicate information and perform cross-correlation, and only a subset of frequency samples are communicated. This can also happen when the signal-to-noise ratio is too small for a pair of sensors. Denote the indices of the observations as Ω. To recap, our goal is to infer K, from noisy and partial

observations of a complex tensor R, indexed on Ω.

Now we present the form of the non-linear map. Consider discrete-time and frequency samples. Assume the discrete event samples are indexed by $t = 0, \ldots, T-1$. Thus, for discrete Fourier transform, the frequency samples are also indexed by $f = 0, \ldots, T-1$. Define a vector of coefficients in our problem $\xi \in \Xi \subset \mathbb{R}^{2K+NK}$:

$$\xi = (\rho_1, \ldots, \rho_K, \alpha_1, \ldots, \alpha_K, \tau_{1,1}, \tau_{1,2}, \ldots, \tau_{N,K}).$$

Define the set

$$\mathcal{L} = \{M \in \mathbb{R}^{N \times N \times T} : M_{i,j,k} = 0, \forall (i,j,k) \in \Omega\},$$

which can be viewed as the "nullspace" of a given observation index set Ω. Then we set

$$\theta = (\xi, M_1, M_2),$$

where $M_1 \in \mathcal{L}$ and $M_1 \in \mathcal{L}$. Denote the real and imaginary parts of the frequency samples as $\mathcal{R}_{n,m,f} = \mathrm{Re}(R_{n,m}(f))$, and $\mathcal{I}_{n,m,f} = \mathrm{Im}(R_{n,m}(f))$, respectively, and define the corresponding tensors $\mathcal{R}$ and $\mathcal{I}$ (which depend on the parameter vector ξ). The non-linear map (similar to the case the complex matrix completion) is defined by

$$G(\theta) := (\mathcal{R} + M_1, \mathcal{I} + M_2). \tag{3.32}$$

Hence, although the situation is fairly complex here, we can cast it into the format of the general problem and use our result.

Numerical experiments suggest the following formula for the characteristic rank

$$\mathfrak{r} = 2K + NK - 1.$$

This is achieved by evaluating the rank of the Jacobian matrix of the map defined by (Equa-

tion 3.32) (see Remark 3.4.1) and the appendix for the derivation of the Jacobian matrix).

3.5 Numerical Experiments

3.5.1 Complex matrix completion

In this section we consider the complex matrix completion problem (Equation 3.23). To solve the related optimization problem, we use a generalize version the hard thresholding algorithm in [30]. In the experiment, we generate a rank-r complex matrix with size $n_1 \times n_2$, by first generating $V_1, V_2 \in \mathbb{R}^{n_1 \times r}$ and $W_1, W_2 \in \mathbb{R}^{n_2 \times r}$, where each entries are i.i.d $\mathcal{N}(0,1)$, and form $X = (V_1 + \mathrm{i}V_2)(W_1 + \mathrm{i}W_2)^\top$. We numerically verified that the characteristic rank of the manifold $\mathcal{M}_r \subset \mathbb{C}^{n_1 \times n_2}$, of matrices of rank r, is $\rho = 2r(n_1 + n_2 - r)$ for all random instances, which is consistent with the results in Section subsection 3.4.2.

To show the asymptotic distribution of test statistics (Theorem Theorem 3.3.1), we generate a rank-2 true matrix $X^* \in \mathbb{C}^{100 \times 100}$. The observed entries are contaminated with Gaussian noise:

$$Y_{ij} = X^*_{ij} + \varepsilon^{(k)}_{ij} + \eta^{(k)}_{ij},\ (i,j) \in \Omega,$$

where $|\Omega| = 1500$ and the noise $\varepsilon^{(k)}_{ij}, \eta^{(k)}_{ij} \overset{iid}{\sim} \mathcal{N}(0, 5^2)$. The experiments are repeated 400 times, i.e., $k = 1, \ldots, 400$, to demonstrate the empirical distribution of the test statistic. Figure 3.1 shows the QQ-plot of $\{T_N(2)^{(k)}\}_{k=1}^{400}$ against the χ^2 distribution with a degrees-of-freedom equal to 2208. Recall that the characteristic rank of the manifold $\mathcal{M}_r \subset \mathbb{C}^{n_1 \times n_2}$, of matrices of rank r, is $\rho = 2r(n_1 + n_2 - r)$ (see Section subsection 3.4.2). The results in Figure 3.1 show that the χ^2 distribution fits the test statistics reasonably well. Moreover, we show the result of detecting the rank in table Table 3.1, with the same experiment setting. In each experiment, we complete the matrix from rank $r = 1$ to $r = 4$. We choose the smallest r, such that $T_N(r)$ has p-value larger than 0.05. In table Table 3.1, there are the results of 200 experiments for true rank $r^* = 2$ and $r^* = 3$. We can see the power of tests are high when $r < r^*$ since there is no false acceptance and the false rejection rate

is close to the significant level 0.05 when $r = r^*$.

Table 3.1: Result of hypothesis tests for the rank of complex matrix completion: r^* is the true rank. For each r^*, there are 200 experiments. We perform the test from $r = 1$ to $r = 4$ and count the number of determined r with significant level, 0.05; $r = 0$ means tests are rejected for $r = 1, \ldots, 4$.

	$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$	FDR
$r^* = 2$	0	0	**190**	10	0	5%
$r^* = 3$	0	0	0	**193**	7	3.5%

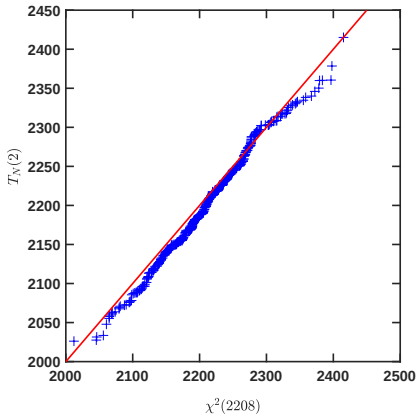

Figure 3.1: QQ-plot of test statistics against χ^2 distribution.

3.5.2 Characteristic rank of third order tensor

To generate third-order tensors of size $n_1 \times n_2 \times n_3$, we form $A \in \mathbb{R}^{n_1 \times r}$, $B \in \mathbb{R}^{n_2 \times r}$, $C \in \mathbb{R}^{n_3 \times r}$, where each entry in A, B, C are *i.i.d.* distributed as standard normal (zero-mean and unit variance). Let $X = A \otimes B \otimes C$ and a^k, b^k, c^k be the kth columns of A, B, C, respectively. To compute the Jacobian matrix, for all $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, $l = 1, \ldots, n_3$ and $k = 1, \ldots, r$, we can show that

$$\frac{\partial X_{ijl}}{\partial a_i^k} = b_j^k c_l^k, \quad \frac{\partial X_{ijl}}{\partial b_j^k} = a_i^k c_l^k, \quad \frac{\partial X_{ijl}}{\partial c_l^k} = a_i^k b_j^k.$$

All the other entries in the Jacobian matrix are zero.

Table Table 3.2 shows the rank (evaluated numerically) of the Jacobian matrices for different (n_1, n_2, n_3, r) values. We note that when r is sufficiently small, the characteristic rank is equal to $r(n_1+n_2+n_3-2)$, as expected. When r is large, the characteristic rank can be less than $r(n_1 + n_2 + n_3 - 2)$. This effect can be explained by Proposition 3.4.4: since in those cases the model is not generically locally identifiable, and hence is not generically identifiable. It is not surprising that when r is large enough (the cases marked with * in the left column), the rank of the Jacobian matrix is equal to $n_1n_2n_3$. The interesting cases are when $r \approx (n_1n_2n_3)/(n_1 + n_2 + n_3 - 2)$. The right column of table Table 3.2 shows some cases in which ranks of the Jacobian matrices are less than $\min\{n_1n_2n_3, r(n_1 + n_2 + n_3 - 2)\}$.

Table 3.2: Rank of the Jacobian matrices for third order tensor. For each combination of (n_1, n_2, n_3, r), the experiments are repeated 100 times and the results are all the same. When r is small, rank$(J) = r(n_1 + n_2 + n_3 - 2)$. When r is large (cases marked with *), rank$(J) < r(n_1 + n_2 + n_3 - 2)$.

n_1	n_2	n_3	r	rank(J)	n_1	n_2	n_3	r	rank(J)
3	4	5	1	10	2	2	4	3	15*
3	4	5	5	50	2	2	5	3	18*
3	4	5	12	60*	2	3	5	4	28*
15	15	15	5	215	3	3	3	4	26*
15	15	15	15	645	3	4	4	5	44
15	15	15	100	3375*	3	5	5	7	74*

3.5.3 Determining the number of signals in blind de-mixing

Consider the ambient noise imaging in a distributed sensor network setting (described in Section subsection 3.4.6), where there are missing values in the observations. Our goal is to determine the number of sources. For this problem, one can show that the characteristic rank is $2K + NK - 1$ for large enough T. Therefore, by identifying the characteristic rank, we can determine the number of sources.

In each experiment, we generate the random instances are follows: $\alpha_k \sim$ Unif$[10, 11]$, $\rho_k \sim$ Unif$[10, 11]$, $\tau_{n,k} \sim$ Unif$[-2.5, 2.5]$, $\forall n = 1, \dots N$ and $k = 1, \dots, K^*$.

First, we want to verify the characteristic rank of the Jacobian matrix predicted using our theory. Let $N = 8, 10, 12$ and $K = 1, \ldots, 5$. For each N and K, we generate parameters and compute the corresponding rank of the Jacobian matrix numerically. In Figure 3.2, each point is the mean of ranks in 100 experiments corresponding to a certain pair of N and K. The lines plotted correspond to $2K + NK - 1$, for $N = 8, 10, 12$. We can see the points are exactly on the lines, which justifies our formulation for the characteristic rank.

Second, we show the result of testing the rank in this problem. The observation noise are normal random variables with zero mean and variance equal to 0.05. Table Table 3.3 is the result of determining source number K^* with α_k, ρ_k and $\tau_{n,k}$ being unknown. We run experiments for $K^* = 1, \ldots, 5$. For each K^*, 100 experiments are run and in each experiment, the test is running from $K = 1$ to $K = 6$ and the significant level is 0.01. In the table, $K = 0$ means all the tests are rejected. We can see our test gives the true number of sources most of the time, except $K^* = 5$. When $K^* = 5$, the algorithm becomes difficult to converge to the optimal solution and therefore leads to a large fitting error.

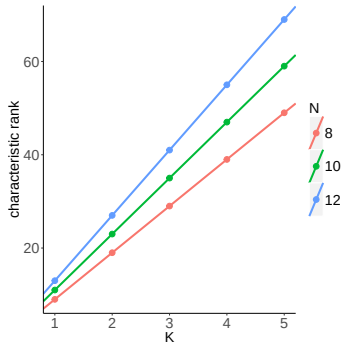

Figure 3.2: The characteristic rank of the problem in Section subsection 3.4.6: K is the number of sources, N is the number of sensors, the points are the rank of the Jacobian matrix of the mapping, and the line is $2K + NK - 1$.

3.5.4 One-hidden-layer neural networks

In this section, we consider the problem of determining the number of hidden units for one-hidden-layer neural networks; the problem described in (Equation 3.27). In the experiment, $x_i \sim \mathcal{N}(0, I_d)$, $U \in \mathbb{R}^{d \times r^*}$, such that $U_{ij} \sim \mathcal{N}(0, 1)$ and $m = 1000$. Consider the activation

Table 3.3: Results of hypothesis tests for the number of sources: K^* is the true number of sources. For each K^*, there are 100 experiments. We perform the test from $K = 1$ to $K = 6$ and count the number of determined K; $K = 0$ means tests are rejected for $K = 1, \ldots, 6$.

	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	FDR
$K^* = 1$	0	**100**	0	0	0	0	0	0
$K^* = 2$	1	0	**98**	0	1	0	0	6%
$K^* = 3$	4	0	0	**94**	2	0	0	3%
$K^* = 4$	9	0	0	0	**91**	0	0	9%
$K^* = 5$	32	0	0	0	0	**67**	1	33%

function to be quadratic activation and sigmoid activation, respectively. Table Table 3.4 and Table Table 3.5 are the ranks of Jacobian matrices for different combinations of (d, r^*). The results justify the formula of characteristic rank of one-hidden-layer neural networks are $dr^* - r^*(r^* - 1)/2$ for quadratic activation and dr^* for sigmoid activation, respectively.

Although we could not provide any theoretical prediction for the characteristic rank when the activation function is a ReLu function, here we provide some numerical examples. We show the performance of our rank test for one-hidden-layer neural networks with a ReLU activation function. In the experiments, $d = 50$, and $\sigma = 0.1$. We perform 100 experiments each from r^*, with the true rank of U being equals to 1 to 6. For each r^*, we perform the test from $r = 1$ to $r = 7$ with significant level 0.05. With this setting, the p-value is computed under the $\chi^2(m - dr)$. The optimization problem involved with fitting the neural networks model is solved using gradient descent (implemented by Pytorch package).

For ReLu activation function, Table Table 3.6 shows the rank determined by our proposed test for each r^*. Here, $r = 0$ means all tests are rejected. Results are similar to what we observed in Table Table 3.3. When the order of the model is small, the test is consistent with the significant level. When the order of the model increase, convergence to the optimal solution becomes more difficult; in this setting, the false discovery rate will increase but is still tolerable. An interesting finding is that our test still gives promising results even

though the ReLU activation is not an analytic function.

Table 3.4: Rank of the Jacobian matrix for one-hidden-layer neural networks with a quadratic activation function. For each combination of (d, r^*), the experiments are repeated 100 times, and the results are all the same. This justifies the formula of the characteristic rank of one-hidden-layer neural networks with quadratic activation is $dr^* - r^*(r^* - 1)/2$.

d	r^*	rank(J)	d	r^*	rank(J)
10	1	10	30	11	275
10	5	40	30	17	374
10	10	55	30	23	473
20	1	20	50	10	455
20	12	174	70	10	655
20	18	207	90	10	855

Table 3.5: Rank of the Jacobian matrix for one-hidden-layer neural networks with sigmoid activation. For each combination of (d, r^*), the experiments are repeated 100 times and the results are all the same. This justifies the formula of the characteristic rank of one-hidden-layer neural networks with sigmoid activation is dr^*.

d	r^*	rank(J)	d	r^*	rank(J)
10	1	10	30	11	330
10	5	50	30	17	510
10	10	100	30	23	690
20	1	20	50	10	500
20	12	240	70	10	700
20	18	360	90	10	900

3.6 Conclusions

We develop a general theory for the goodness-of-fit test to non-linear models, which essentially shows that the parameter-of-interests are related to the characteristic rank of the linear map that defines the manifold structure of our observation. The test statistic has a simple chi-square distribution whose parameters are specified explicitly. Based on this result, it is convenient to implement a test procedure to determine the model order in practice. Our general theory can provide precise answers to several questions, such as determining the rank of (complex) low-rank matrix from noisy and incomplete observations. In some other

Table 3.6: Result of ReLU activation function: r^* is the rank of true U^*. For each r^*, there are 100 experiments. We perform the test from $r = 1$ to $r = 7$ and count the number of determined r. $r = 0$ means tests are rejected for $r = 1, \ldots, 7$.

	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	FDR
$r^*=2$	3	0	**96**	1	0	0	0	0	4%
$r^*=3$	4	0	0	**96**	0	0	0	0	4%
$r^*=4$	4	0	0	0	**94**	1	0	1	6%
$r^*=5$	2	0	0	0	0	**93**	5	0	7%
$r^*=6$	5	0	0	0	0	0	**88**	7	12%

applications, we show that how the general theory can shed light on finding the "model-order-of-interests", such as tensor completion, determining the number of hidden nodes in neural networks, determining the number of sources in blind signal demixing problems, using analysis and simulations.

CHAPTER 4
DETECTION OF CASCADING FAILURES

4.1 Introduction

This chapter develops an online change-point detection procedure for power system's cascading failure using multi-dimensional measurements over the networks. We incorporate the cascading failure's characteristic into the detection procedure and model multiple changes caused by cascading failures using a diffusion process over networks [45]. The model captures the property that the risk of component failing increases as more components around it fail. Our change-point detection procedure using the generalized likelihood ratio statistics assuming unknown post-change parameters of the measurements and the true failure time (change-points) at each node.

The rest of the chapter is organized as follows. In Section 4.2, we present the problem setting, the models of failure propagation and the measurements. Section 4.3 provide our proposed test procedure and the computation of the log likelihood function. In Section 4.4, we introduced our fast algorithm to implement the test procedures and discuss the computation complexity. In Section 4.5, we compare our proposed procedure with other existed methods. Finally, Section 4.6 concludes this chapter.

4.2 Problem Setup

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is formed by a set of nodes $\mathcal{V} = \{1, 2, 3, \ldots, N\}$ and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Here $\mathcal{V}$ corresponds to the set of components in the power network, and $\mathcal{E}$ can be constructed according to the physical network or interaction graph [46, 47]. Assume $\mathcal{G}$ is undirected; $X_{i,t} \in \mathbb{R}$ is the measurement of ith node at time t, $t = 1, \ldots, T$.

We make the following assumptions about the change points, which correspond to the

failure times at each node. Assume the true failure time of the ith node is $\tau_i^* \in \mathbb{R}^+ \cup \{\infty\}$, and $\tau_{(i)}^*$ denotes the corresponding ordered failure time. When $\tau_i^* = \infty$, it means that there is no failure on the ith node. Let $\tau^* = (\tau_1^*, \tau_2^*, \ldots, \tau_N^*)$ denote the vector of all true failure times, which is unknown.

4.2.1 Failure (change-point) propagation model

Consider the following cascading model for the propagation process of the network's failures. We assume that whenever a failure occurs on a node, it increases the neighboring nodes' tendency to fail. Mathematically, we define the influence of node i on node j as $\alpha_{i,j} > 0$. We assume $\alpha_{j,i}$ are known since they can be typically estimated beforehand using historical and simulation data given on the topology of the power grid and power flow. We do not know the distribution for the first failure. After the occurrence of the first failure, the distribution of the subsequent failures is determined by the conditional hazard rate (intensity function) $\lambda_i(t)$:

$$\lambda_i(t) = \begin{cases} \sum_{j:(j,i)\in\mathcal{E}, \tau_j^* < t} \alpha_{j,i}, & \tau_{(1)}^* < t \leq \tau_i^*, \\ 0, & \text{o.w.} \end{cases} \tag{4.1}$$

We assume that the failed nodes can only affect the neighboring nodes in the graph. The influence of the failed nodes is constant over time. Therefore, each node's hazard rate before failure is a piece-wise constant, starting at 0 and jumps when the failure affects its neighboring nodes. Figure 4.1 shows an example of the failure propagation process.

Remark 4.2.1. *Define the history (filtration) up to time t as*

$$\mathcal{H}(t) = \{\tau_i^* \leq t, i = 1, \ldots, N\}.$$

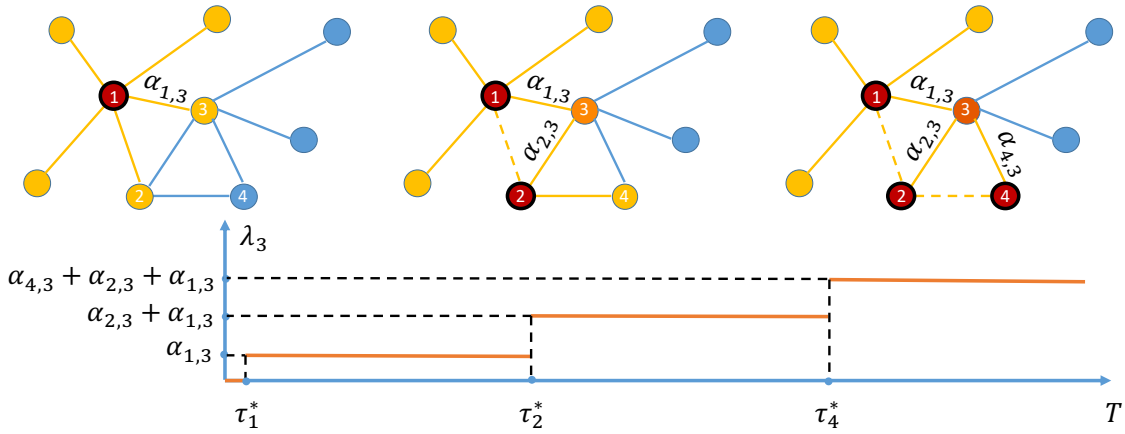

Figure 4.1: Example of how cascading failure propagates over networks. The failure initiate at node one, then all the neighbors of node one are affected, and node two and node four fail eventually. As the failure propagates, node three is surrounded by more and more failed nodes, and its hazard rate continues to increase. Here, red circles correspond to failed nodes, solid yellow lines are possible paths for failures to diffusion, dashed yellow lines correspond to paths with failed nodes at both ends, yellow circles are nodes affected by failed neighbors.

Then, given the $\mathcal{H}(t)$, the conditional intensity function of the ith node is defined as:

$$\lambda_i(t) \triangleq \lim_{\Delta t \to 0} \frac{\mathbb{P}\{\tau_i^* \in [t + \Delta t] | \mathcal{H}(t), \tau_i^* > t\}}{\Delta t}.$$

The distribution of τ^* is uniquely defined by the conditional hazard rate [45].

4.2.2 Measurement model

To simplify the study, we assume that the measurements at each node are independent, conditioned on the failure time. Before a change, they follow an $i.i.d.$ standard normal distribution, and after a change they follow an $i.i.d.$ normal distribution with an unknown mean and variance. That is,

$$X_{i,t} \overset{\text{i.i.d}}{\sim} \begin{cases} \mathcal{N}(0,1), & t < \tau_i^*, \\ \mathcal{N}(\mu_i, \sigma_i^2), & t \geq \tau_i^*. \end{cases} \tag{4.2}$$

Since we can typically use a certain length of data as a warm start to estimate the sample mean and variance of the pre-change distribution, therefore we assume that the pre-change distribution is known and can be standardized.

4.2.3 Likelihood function

According to the above models, in a time window $[0, T]$, given measurements and failure times, Proposition 4.2.1 is the likelihood function.

Proposition 4.2.1. *According to the model defined by Equation (Equation 4.1) and Equation (Equation 4.2), the likelihood function for a given τ^* and $X_{i,t}$ in $[0, T]$ can be expressed as the following:*

$$\begin{aligned} & f(\tau_i^*, X_{i,t}, \forall i = 1, \ldots, N, t = 1, \ldots, T) \\ = & \underbrace{\prod_{i=2}^{N} f(\tau_{(i)}^* | \tau_{(1)}^* \cdots \tau_{(i-1)}^*)}_{(a)} \cdot \underbrace{\prod_{i=1}^{N} \prod_{t=1}^{T} f(X_{i,t} | \tau_i^*)}_{(b)} \end{aligned} \tag{4.3}$$

where the term (a) captures the failure propagation model and term (b) captures the measurement model.

Proof. According to model (Equation 4.2), given change-points τ^*, $X_{i,t}$ are independent. i.e.

$$f(X_{i,t}, i = 1, \ldots, N, t = 1, \ldots, T | \tau^*) = \prod_{i=1}^{N} \prod_{t=1}^{T} f(X_{i,t} | \tau_i^*) \tag{4.4}$$

According to [45], the likelihood of a failure nodes is the product of the survival probability up to the failure time and the hazard rate at the failure time, i.e. for $\tau_i^* < T$

$$f(\tau_i^* | \{\tau_1^*, \ldots, \tau_N^*\} \setminus \tau_i^*, \tau_i^* < T) = \lambda_i(\tau_i^*) \exp\Big(- \int_0^{\tau_i^*} \lambda_i(t) dt \Big). \tag{4.5}$$

For a node which has no failure before time T, the likelihood of it is the survival function

up to time T, i.e. for $\tau_i^* \geq T$:

$$f(\tau_i^* | \{\tau_1^*, \ldots, \tau_N^*\} \setminus \tau_i^*, \tau_i^* \geq T) = \exp\Big(- \int_0^T \lambda_i(t) dt \Big). \tag{4.6}$$

According to the definition of hazard rate (Equation 4.1), $\lambda_i(t)$ only depends on the change-points before time t. Therefore, we have

$$f(\tau_1^*, \ldots, \tau_N^*) = \prod_{i=2}^{N} f(\tau_{(i)}^* | \tau_{(1)}^* \ldots, \tau_{(i-1)}^*). \tag{4.7}$$

Combine the above and (Equation 4.4), we have the following:

$$\begin{aligned} & f(\tau_i^*, X_{i,t}, i = 1, \ldots, N, t = 1, \ldots, T) \\ & = f(\tau_1^*, \ldots, \tau_N^*) f(X_{1,1}, \ldots, X_{N,T} | \tau_1^*, \tau_2^*, \ldots, \tau_N^*) \\ & = \prod_{i=2}^{N} f(\tau_{(i)}^* | \tau_{(1)}^* \ldots, \tau_{(i-1)}^*) \prod_{i=1}^{N} \prod_{t=1}^{T} f(X_{i,t} | \tau_i^*). \qquad \square \end{aligned} \tag{4.8}$$

Our method can be extended in a more general setting. For the failure propagation model, one can choose different hazard functions. Three models are provided in [45]. In our study, we choose to use the exponential model. For the measurement model, one can select the different distribution, and the measurement can be a high dimension.

4.3 Detection Procedure

Consider the following sequential hypothesis test for detecting a dynamic change. An alarm is raised when there are at least η change-points. To perform the online detection, at each time instance T we consider the following hypothesis test:

$$H_{0,\eta,T} : \tau_{(\eta)}^* > T, \;\; H_{1,\eta,T} : 0 \leq \tau_{(\eta)}^* \leq T.$$

We consider a Shewhart chart type procedure: at each time, we evaluate a general likelihood ratio (GLR) statistics over a sliding window. The GLR statistic can handle the mean, and post-change distribution variance are unknown. As shown in (Equation 4.3), given the failure time and the measurements, the likelihood can be decoupled into two parts: the likelihood of the failure propagation model and the likelihood of measurement model, respectively.

4.3.1 Log likelihood of failure propagation model

Define $\mathcal{C}(i) = \{j \in \mathcal{V} | (j,i) \in \mathcal{E}\}$ to be the set of the ith node's neighbors. Given τ, the log-likelihood function for $[0,T]$ is shown in Proposition 4.3.1:

Proposition 4.3.1. *Given T, a set of failure times $\tau = (\tau_1, \ldots, \tau_N)$, graph $\mathcal{G}$, and parameters $\alpha_{i,j}$ $\forall i,j \in \mathcal{V}$, the log-likelihood function of the failure propagation model is given by*

$$\begin{aligned}\ell_{1,T} &= \log f(\tau_1, \tau_2, \ldots, \tau_N | \{\alpha_{i,j}\}) \\ &= \sum_{\substack{i:\tau_i \le T, \\ \tau_i \neq \tau_{(1)}}} \Big\{ \log \left(\sum_{j \in \mathcal{C}(i)} \alpha_{j,i} \mathbb{I}(\tau_j < \tau_i) \right) \\ &- \sum_{j \in \mathcal{C}(i)} \alpha_{j,i} (\tau_i - \tau_j)^+ \Big\} - \sum_{i:\tau_i > T} \sum_{j \in \mathcal{C}(i)} \alpha_{j,i} (T - \tau_j)^+,\end{aligned} \tag{4.9}$$

where $(\cdot)^+ = \max(\cdot, 0)$, and $\mathbb{I}(\cdot)$ is the indicator function.

Proof. Combine the model (Equation 4.1), equations (Equation 4.5) and (Equation 4.6), the log likelihood of a failure node is, $\forall \tau_i < T$,

$$\begin{aligned}&\log f(\tau_i | \{\tau_1, \ldots, \tau_N\} \setminus \tau_i) = \log \lambda_i(\tau_i) - \int_0^{\tau_i} \lambda_i(t) dt \\ &= \log \Big(\sum_{j \in \mathcal{C}(i), \tau_j < \tau_i} \alpha_{j,i} \Big) - \sum_{j \in \mathcal{C}(i), \tau_j < \tau_i} \alpha_{j,i} (\tau_i - \tau_j).\end{aligned} \tag{4.10}$$

The log-likelihood function of a node without failure before time T is, $\forall \tau_i > T$, can be

written as

$$\begin{aligned}\log f(\tau_i|\{\tau_1,\ldots,\tau_N\}\setminus\tau_i) &= -\int_0^T \lambda_i(t)dt \\ &= -\sum_{j\in\mathcal{C}(i),\tau_j<T} \alpha_{j,i}(T-\tau_j).\end{aligned} \tag{4.11}$$

Combine with Equation 4.7, Equation 4.10 and Equation 4.11, we can derive the proposition. □

4.3.2 Log likelihood of measurement model

Since we assume that the mean and variance of post-change distribution are unknown, we estimate the μ_is and σ_is by maximum likelihood estimation (MLE): $\hat{\mu}_i$, $\hat{\sigma}_i$. Therefore the log-likelihood function of measurements of the ith node, given τ_i, is:

$$\begin{aligned}\ell_{2,i,T} =& \log f(X_{i,t}, t=1,\ldots,T|\tau_i) \\ =& -\sum_{t=1}^{T\wedge(\tau_i-1)} \frac{X_{i,t}^2}{2} - \sum_{t=\tau_i}^{T} \frac{(X_{i,t}-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2} \\ & - \frac{T}{2}\log(2\pi) - (T-\tau_i+1)^+\log(\hat{\sigma}_i).\end{aligned} \tag{4.12}$$

Since we assume that the distribution of measurements at each node is independent given τ, the log likelihood function of all measurements is the summation of the log likelihood function of each node, i.e., $\ell_{2,T} = \sum_{i=1}^{N} \ell_{2,i,T}$. Therefore, given the failure time τ and measurements $X_{i,t}$s, the log likelihood at time T is:

$$\ell_T(\tau, X_{i,t}\, i=1,\ldots,N, t=1,\ldots,T) = \ell_{1,T} + \ell_{2,T}. \tag{4.13}$$

Notice that if $\tau_i > T$ for all i, Equation (Equation 4.9) equals 0 and ℓ_T is the sum of log likelihoods of standard normal distribution for the measurements on each node, according to Equation (Equation 4.12).

To perform the hypothesis test between $H_{0,\eta,T}$ and $H_{1,\eta,T}$, we need to search for τ such that the log-likelihood in (Equation 4.13) is maximized. Define

$$U(\eta) = \{\tau : \sum_{i=1}^{N} \mathbb{I}(\tau_i \leq T) \geq \eta\},$$

and

$$L(\eta) = \{\tau : \sum_{i=1}^{N} \mathbb{I}(\tau_i \leq T) \leq \eta - 1\}.$$

We consider a Shewhart type of detection procedure, and we evaluate the test statistics with the data over a sliding window $[T - L + 1, T]$, where L is the length of the window. To detect a change for at least η change-points, we apply the following GLR test statistics $\forall \eta = 1, \ldots, N$:

$$S_{\eta,T} = \max_{\tau \in U(\eta)} \ell_T(\tau) - \max_{\tau \in L(\eta)} \ell_T(\tau).$$

The corresponding stopping time is

$$\Gamma = \inf\{T > 0 : S_{\eta,T} > b\},$$

for some preset threshold b.

4.4 Computationally Efficient Algorithm

We would like to implement change detection online and detect the cascading changes as quickly as possible in practice. Thus, we need a low-complexity algorithm and only search for propagation paths with at most m nodes affected by the failure. The computation cost of the maximum likelihood under the alternative hypothesis is high. For instance, for a fully connected graph with N nodes with observations in time horizon T, the computation cost is $\mathcal{O}(T^m N!/((N-m)!m!))$. We aim to develop a computationally efficient algorithm

based on a pruning strategy (similar to the ideas in [48, 49]). Our proposed algorithm is described as in Algorithm algorithm 1, which we describe below in more details.

To reduce the computation cost, we propose a *random sampling strategy* as follows. Since the number of possible propagation paths in a fully connected network grows exponentially as the number of nodes increases. Define $\mathcal{F}$ as the failure set that contains the failed nodes, and

$$\mathcal{R} = \{j \notin \mathcal{F} : \exists i \in \mathcal{F}, \alpha_{i,j} > 0\},$$

as the risk set. Then, we generate the next possible failure points by randomly picking P points in $\mathcal{R}$ without replacement with probability vector $\mathbf{p} = (p_i)_{i\in\mathcal{R}}$, where

$$p_i = \tilde{p}_i(\sum_{j\in\mathcal{R}} \tilde{p}_j)^{-1}, \quad \tilde{p}_i = \sum_{j\in\mathcal{F}} \alpha_{j,i}. \tag{4.14}$$

With this scheme, we reduce the number of paths to $O(NP^m)$.

To further reduce computational complexity, we also combine the above with a `Thinning` algorithm by exploiting the monotonicity of e^{-x} involved in the likelihood function, which is summarize in Algorithm algorithm 3. Consider the likelihood

$$\mathcal{L}_T = e^{\ell_T}, \mathcal{L}_{2,i,T} = e^{\ell_{2,i,T}}, \mathcal{L}_{2,T} = e^{\ell_{2,T}}, \mathcal{L}_{1,T} = e^{\ell_{1,T}}.$$

Given a propagation path, we need to compute the maximum $\mathcal{L}_T(\tau)$, which is the product of $\mathcal{L}_{1,T}$ and $\mathcal{L}_{2,T}$. Define the qth percentile of the ith node to be $l_{2,i,q}$. Also define a lower bound l_1 for $\mathcal{L}_{1,T}$. Instead of maximizing $\mathcal{L}_T(\tau)$ over all the possible choices, we maximize it only in a thinned set

$$\{\tau : \mathcal{L}_{2,i,T}(\tau_i) \geq l_{2,i,q},\ \forall i = 1, \ldots, N\} \cap \{\tau : \mathcal{L}_{1,T}(\tau) \geq l_1\}.$$

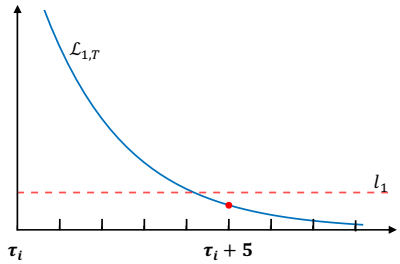

Figure 4.2: Illustration of the searching strategy for τ_j given the previous failure point τ_i. When searching τ_j from $\tau_i + 1$, we have $\mathcal{L}_{1,T}(\tau_i + 4) > l_1$ and $\mathcal{L}_{1,T}(\tau_i + 5) < l_1$; by the monotonicity of e^{-x}, we can stop searching at $\tau_i + 5$.

Specifically, given τ_i, we consider

$$\tau_j \in \{\tau_i + 1, \tau_i + 2 \dots\} \cup \{\tau_j : \mathcal{L}_{2,j,T}(\tau_j) \geq l_{2,j,q}\},$$

from the smallest to the largest until $\mathcal{L}_{1,T} < l_1$ as shown in Figure 4.2. The computation cost of this step is $O(h)$, where h depends on the topology of $\mathcal{G}$, as well as parameters $\alpha_{i,j}$, l_1 and q. Moreover, we can show that $h \leq [L(1-q)]^m$.

With the above strategies, we can reduce the computation cost to $O(NP^m h)$, which is linear to N, the network's size. As shown in the following numerical examples, we can now efficiently compute the test statistics for a 300-bus power system. We combine *random sampling strategy* and *thinning* to reduce the computation cost using a recursion function `genNext` as shown in algorithm 2.

Algorithm 1: Cascading change-point detection

Input: Data $X_{i,T-L+1}, \ldots X_{i,T}, i = 1, \ldots N$

Variables:

- m: the maximal number of change-points
- ℓ: log likelihood given a set of change-points
- τ: a set of change-points
- r: path of change-points (sorted)
- $\ell_{\max}$, $\tau_{\max}$, $r_{\max}$: best log likelihood and parameters
- J : sets of potential change-points of each nodes
- j: current depth of recursion
- q: percentile to threshold the measurement likelihood
- l_1: threshold of failure propagation likelihood

1. $j = 1$

2. $J \leftarrow$ compute potential change-points using `Thinning` (j, m, r, τ, q, l_1, $\{X_{i,t}\}$)

for $x = 1 : N$ **do**

  for t *in* $J(x)$ **do**

    $p_1 = x$, $\tau_1 = t$

    ℓ, r, τ = `genNext`$(j + 1, m, r, \tau, \{X_{i,t}\})$

    if $\ell > \ell_{\max}$ **then**

      $\ell_{\max} = \ell$, $r_{\max} = r$, $\tau_{\max} = \tau$

    end

  end

end

4. **Return** $(\ell_{\max}, r_{\max}, \tau_{\max})$

Algorithm 2: `genNext` function

Input: j, m, r, τ, Data $\{X_{i,t}\}$

Variable K: a set of potential nodes with change-point

if $j > m$ **then**

  $\ell \leftarrow$ compute log-likelihood

  return(ℓ, r, τ)

end

$J \leftarrow$ `Thinning`$(j, m, r, \tau, q, l_1,$ Data$)$

$K \leftarrow$ sample nodes with the probability as (Equation 4.14)

for x *in* K **do**

  for t *in* $J(x)$ **do**

    $r_j = x, \tau_j = t,$

    $\ell, r, \tau =$ `genNext`$(j+1, m, r, \tau, \{X_{i,t}\})$

    if $\ell > \ell_{\max}$ **then**

      $\ell_{\max} = \ell, r_{\max} = r, \tau_{\max} = \tau$

    end

  end

end

Return $(\ell_{\max}, r_{\max}, \tau_{\max})$

Algorithm 3: `Thinning` function

Input: j, m, r, τ, q, l_1, Data $\{X_{i,t}\}$

1. $J = \emptyset$

2. compute $l_{2,x,q}$, $\forall x = 1, \ldots, N$ using given Data.

for $x = 1 : N \setminus r$ **do**

  for $t = \tau_j : T$ **do**

    Given t and τ, compute $\mathcal{L}_{2,x,T}$ and $\mathcal{L}_{1,T}$.

    if $\mathcal{L}_{2,x,T} > l_{2,x,q}$ **then**

      $J(x) = J(x) \cup t$

    end

    if $\mathcal{L}_{1,T} < l_1$ **then**

      exit for

    end

  end

end

4. **Return** (J)

4.5 Numerical Examples

In this section, we perform several numerical examples to demonstrate our proposed method's performance and compare it to existing methods. We consider two commonly used performance metrics in change-point detection: the average run length (ARL) (a large ARL means a low false alarm rate) and the expected detection delay (EDD). More specifically, for a stopping time Γ, we define ARL as $\mathbb{E}_0[\Gamma]$, and we use $\mathbb{E}_1[(\Gamma - \tau^*_{(1)})^+]$ as a measure for EDD (which is a common practice in literature (see, e.g., [50]), where $\mathbb{E}_i$ denotes the expectation with the probability measure under hypothesis H_i. As we increase the threshold, ARL typically increases exponentially, whereas the EDD will increase linearly. A good change-point detection procedure should have a small EDD, given the same ARL.

We consider two case studies: one is to detect the very first change in the network, and the other is to detect the change when there are at least η change-points. In Case I, we compare our methods with generalized likelihood ratio (GLR) and (CuSum), since CuSum is the optimal procedure when the parameter is known [51]) and GLR is a natural generalization of CuSum when the post-change parameter is unknown [52]. In Case II, we compare our method with the state-of-the-art, including the S-CuSum [53] and a traditional method, Generalized Multi-char CuSum, both of which are suitable for Case II. Below the pre-change distribution is $\mathcal{N}(0,1)$ and post-change distribution is $\mathcal{N}(1,1)$.

Case I: Detect the first change-point. In Figure 4.3 (left), we show the results in a 300-bus power system (see MATPOWER [54]). In this relatively large system, we apply our algorithm with $L = 100$, $q = 0.8$, $P = 1$, $l_1 = e^{-5}$, and $m = 5$. Our detection statistic can be computed quite efficiently: the average computation time for each time step is less than 3 seconds. Dashed lines are the results when parameters are known. In this scenario, we compare our proposed method with the exact CuSum and two misspecified CuSum ($\mu = 2, 2.5$). Solid lines are the results when parameters are unknown. In this scenario, we compare our proposed method with GLR. Overall, our method shows the best performance,

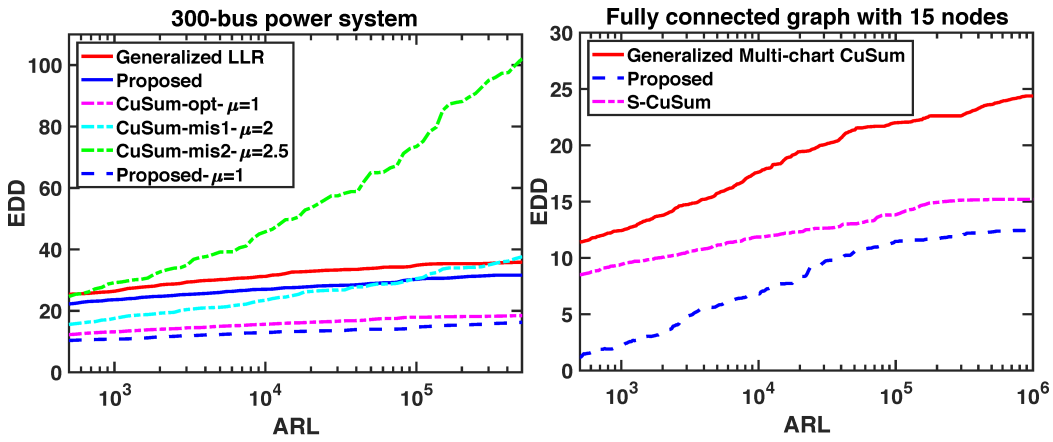

Figure 4.3: (Left) Comparison of CuSum, generalized likelihood ratio, and the proposed method. (Right) Comparison of generalized multi-chart CuSum, S-CuSum, and the proposed method.

which is reasonable because our method is not only based on the likelihood ratio but also considers the likelihood of failure propagation.

Case II: Detect when there are at least η change-points. Here we compare our proposed method with generalized multi-chart CuSum, and S-CuSum[53], because these are the most well-known algorithms for tackling such problems. In this experiment, the graph is fully connected with 15 nodes. The parameters for the algorithm are $L = 100$, $q = 0.8$, $P = 1$, $l_1 = e^{-7}$, and $m = 5$. We set $\eta = 3$. To compute the ARL, we generate data with $\eta - 1$ affected nodes. The result in Figure 4.3 (right) shows that our method outperforms both generalized multi-chart CuSum and S-CuSum.

4.6 Conclusion

In this chapter, we proposed a computationally efficient algorithm to perform the change-point detection by modeling the cascading failure as a temporal diffusion process in a network. Numerical experiments show that our proposed method demonstrates good performance; for an IEEE 300-bus system, which is considered relatively large, our results show that the proposed algorithm can scale up to larger systems.

CHAPTER 5
CHANGE-POINTS DETECTION FOR NETWORK POINT PROCESS VIA SCAN SCORE STATISTICS

5.1 INTRODUCTION

In this chapter, we proposed a change-point detection procedure by scan score statistics in a multivariate Hawkes network. Our scan score statistics are computationally efficient, since we don't need to compute the estimates of the post-change parameters, which is of importance for online detection. We present the theoretical results of our proposed procedure including the analysis of the false alarm rate (FAR) and average run length (ARL) of procedure under null hypothesis. We first compute the variance of our proposed statistics and then use Brownian motion to approximate the distribution of our proposed statistics. Combine the technique in [55], importance sampling can be used for computation of FAR and ARL. We use simulation study to testify our theoretical results and compare our method with an existing change-point detection procedure with generalized likelihood ratio statistics [56]. We also apply our proposed procedure in real-world data such as memetracker and stock market, which show promising results in detecting abrupt change in the network.

The rest of our chapter is organized as follows. Section 5.2 provides the background knowledge of multivariate Hawkes process. Section 5.3 presents the definition of our problem. Section 5.4 proposes our detection procedure and includes the analysis of our scan score statistics. In Section 5.5, there are experiments of simulation study and real-world data application. Finally, Section 5.6 concludes this chapter.

5.2 Background

A multivariate Hawkes process is a self-exitation process over a network. Let M denote the number of nodes in the network and $[M]$ denote $\{1, \dots, M\}$. The data is of the form $\{(u_1, t_1), (u_2, t_2), \dots\}$, where $u_i \in [M]$ denotes the location of ith event and $t_i \in \mathbb{R}^+$ denotes the time of ith event. A multivariate Hawkes Process is actually a special case of spatio-temporal counting process [57]. Let $\mathcal{H}_t$ denote the history before time t, i.e. the σ-algebras of events before time t. $\{\mathcal{H}_t\}_{t\geq 0}$ is a filtration, an increasing sequence of σ-algebras. Let $N_m(t)$ denote the number of events on ith node up to time t, i.e a counting process,

$$N_m(t) = \sum_{t_i \leq t} \mathbb{I}(t_i \leq t, u_i = m).$$

Then, a multivariate Hawkes process can be determined by the following conditional intensity function [58].

$$\lambda_m(t) = \lim_{s \to 0} \frac{\mathbb{P}\{N_m(t+s) > 0 | \mathcal{H}_t\}}{s}. \tag{5.1}$$

For a multivariate Hawkes process, the conditional intensity function takes the form:

$$\lambda_m(t) = \mu_m + \sum_{i \in [M]} \int_0^t g_{i,m}(t-s) N_i(ds). \tag{5.2}$$

Here μ_m is the base intensity and $g_{i,j}(t)$ is the kernel function that characterized the influence of the previous events. Specifically, we assume the commonly used kernel, exponential kernel, i.e.

$$g_{i,j}(t) = \alpha_{i,j} e^{-\beta t}. \tag{5.3}$$

Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ and $\mathbf{A} \in \mathbb{R}^{M \times M}$, of which the (i,j)th entry is $\alpha_{i,j}$. A multivariate Hawkes Process with exponential kernel is parametrized by the base intensity $\boldsymbol{\mu}$, influence matrix $\mathbf{A}$ and decay rate β. Given all the events in time window $[0, T]$. Let K denote the

number of events. The log likelihood function is:

$$\begin{aligned}\ell_T(\mathbf{A}) = &\sum_{k=1}^{K} \log\left(\mu_{u_k} + \sum_{t_i<t_k} \alpha_{u_i,u_k} e^{-\beta(t_k-t_i)}\right) - \sum_{m=1}^{M} \mu_m T \\ &+ \frac{1}{\beta}\sum_{m=1}^{M}\sum_{k=1}^{K} \alpha_{u_k,m}[e^{-\beta(T-t_k)} - 1],\end{aligned} \tag{5.4}$$

Notice that, when $\mathbf{A} = \mathbf{0}$, the process becomes a multivariate Poisson process.

5.3 Problem Setup

In this chapter, we consider a change-point detection problem in a network. In a network with M nodes, there are events on each nodes over time. We say that time $\tau^* > 0$ is a change-point if the following applies. Before time τ^* the events of the network follows a multivariate Hawkes process with parameters $\boldsymbol{\mu}$, $\mathbf{A}_0$ and β. After time τ^* the events of the network follows a multivariate Hawkes of which the influence matrix change from $\mathbf{A}_0$ to $\mathbf{A}_1$. For both the pre-change and post-change multivariate Hawkes process are assumed to be stationary. To detect whether a change-point τ^* exists given data, we consider the following hypothesis test:

$$\begin{aligned}
H_0 :&\lambda_m(t) = \mu_m + \sum_{t_i \le t} \alpha_{u_i,m,0} e^{-\beta(t-t_i)}, \qquad m \in [M], t \ge 0; \\
H_1 :&\lambda_m(t) = \mu_m + \sum_{t_i \le \tau^*} \alpha_{u_i,m,0} e^{-\beta(t-t_i)}, \qquad m \in [M], 0 \le t \le \tau^*; \\
&\lambda_m(t) = \mu_m + \sum_{\tau^* \le t_i \le t} \alpha_{u_i,m,1} e^{-\beta(t-t_i)}, \quad m \in [M], t > \tau^*;
\end{aligned} \tag{5.5}$$

where $\lambda_m(t)$ is the true conditional intensity of node m at time t, $\alpha_{i,j,0}$ and $\alpha_{i,j,1}$ are the (i,j)th entry of $\mathbf{A}_0$ and $\mathbf{A}_1$, respectively.

5.4 Scan Score Statistics Detection Procedure

To perform the hypothesis test (Equation 5.5), we proposed a detection procedure base on scan score statistics. A score statistics is the first derivative of the log likelihood function. In a multivariate Hawkes network, we are interested in the influence between multiple pair of nodes (i.e. the entries in influence matrix $\mathbf{A}$). For each pair, we would have a score statistics. Therefore, we will end up with a high dimensional vector of score statistics. We use the scanning strategy to compute our test statistics as in [7, 59] . Specifically, we divided the whole network into several clusters. At each time t, we compute the interested score statistics in each cluster. Then we get a statistics for each cluster by summing up the standardized score statistics in the corresponding cluster. Finally, we use the maximum over all clusters to be the scan score statistics at time t for the whole network. More details will be discussed in this section.

5.4.1 Score Statistics

Since the change in hypothesis test (Equation 5.5) is caused by the change of influence matrix, we define the following score statistics, given data up to time t, with respect to $\alpha_{p,q}$:

$$S_T^{(p,q)}(\mathbf{A}) \triangleq \frac{\partial \ell_T(\mathbf{A})}{\partial \alpha_{p,q}}. \tag{5.6}$$

Moreover, define $S_T(\mathbf{A})$ is the vector of all elements in $\{S_T^{(p,q)}(\mathbf{A}); p, q \in [M]\}$. According to [57, Theorem 1], we have the following corollary.

Corollary 5.4.1. *Under the assumptions in [57], assume the influence matrix of the multivariate Hawkes Process is* $\mathbf{A}$*. The score function* $S_T(\mathbf{A})$ *satisfies that,* $T^{-\frac{1}{2}} S_T(\mathbf{A}) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\mathbf{A}))$*, where* $\mathcal{I}(\mathbf{A})$ *is the Fisher information.*

Theoretical Result of $\mathcal{I}(\mathbf{0})$

When $\mathbf{A} = \mathbf{0}$, it is possible to compute the theoretical result of $\mathcal{I}(\mathbf{0})$ as shown in the following theorem. For simplicity, let's define $\mathcal{C}(i,t)$ as the set of events at node i before time t, i.e. $\mathcal{C}(i,t) = \{k : t_k < t, u_k = i\}$

Theorem 5.4.2. *Assume the the conditional intensity function has the form as in Equation 5.2 and the kernel function is exponential as in Equation 5.3. According to Equation 5.6,*

$$\begin{aligned} S_T^{(p,q)}(\mathbf{A}) =& \frac{\partial \ell_T(\mathbf{A})}{\partial \alpha_{p,q}} \\ =& \sum_{k \in \mathcal{C}(q,T)} \frac{\sum_{i \in \mathcal{C}(p,t_k)} e^{-\beta(t_k - t_i)}}{\mu_q + \sum_{t_i < t_k} \alpha_{u_i,q} e^{-\beta(t_k - t_i)}} + \frac{1}{\beta} \sum_{k \in \mathcal{C}(p,T)} [e^{-\beta(T - t_k)} - 1] \end{aligned} \tag{5.7}$$

Moreover, when $\mathbf{A} = \mathbf{0}$, as $T \to \infty$, we have the following limits of the (co)variances:

$$\begin{aligned} &\mathrm{Var}[T^{-\frac{1}{2}} S_T^{(q,q)}(\mathbf{0})] \to \frac{1}{2\beta} + \frac{\mu_q}{\beta^2}, \\ &\mathrm{Var}[T^{-\frac{1}{2}} S_T^{(p,q)}(\mathbf{0})] \to \frac{\mu_p}{\mu_q}(\frac{1}{2\beta} + \frac{\mu_p}{\beta^2}), \\ &\mathrm{Cov}[T^{-\frac{1}{2}} S_T^{(p,q)}(\mathbf{0}), T^{-\frac{1}{2}} S_T^{(p',q)}(\mathbf{0})] \to \frac{\mu_p \mu_{p'}}{\mu_q \beta^2}, \end{aligned} \tag{5.8}$$

Remark 5.4.1. *Since we assume the stationary of the multivariate Hawkes process, according to [58], the stationary intensity for multivariate Hawkes process with exponential kernel is:*

$$\tilde{\lambda}(t) = (\mathbf{I} - \mathbf{A}/\beta)^{-1} \boldsymbol{\mu} \tag{5.9}$$

Therefore, given $\mathbf{A}$ and $\boldsymbol{\mu}$, to simplify the computation, we can compute the score statistics with the parameters of stationary distribution, i.e. $\mathbf{A}_{\text{stationary}} = \mathbf{0}$ and $\boldsymbol{\mu}_{\text{stationary}} = \tilde{\lambda}$.

Estimation of $\mathcal{I}(\mathbf{A})$

When $A \neq \mathbf{0}$, it is difficult to compute the variance theoretically. According to [57], we have the following approximation of $\mathcal{I}(\mathbf{A})$.

Theorem 5.4.3. *With the same assumption as in Theorem Theorem 5.4.2. We have the following estimation of fisher information. Let*

$$\hat{\mathcal{I}}_T(\mathbf{A})_{(i,j),(p,q)} = \begin{cases} 0 & \text{if } j \neq q \\ \sum_{k \in \mathcal{C}(q,T)} \frac{(\sum_{k \in \mathcal{C}(i,t)} e^{-\beta(t_k - t_i)})(\sum_{k \in \mathcal{C}(p,t)} e^{-\beta(t_k - t_i)})}{\left(\mu_q + \sum_{t_i < t_k} \alpha_{u_i,q} e^{-\beta(t_k - t_i)}\right)^2} & \text{if } j = q \end{cases} \tag{5.10}$$

We have $\frac{1}{T}\hat{\mathcal{I}}_T(\mathbf{A}) \to \mathcal{I}(\mathbf{A})$*, i.e.* $\forall i, j, p, q$*,*

$$\frac{1}{T}\hat{\mathcal{I}}_T(\mathbf{A})_{(i,j),(p,q)} \to \mathcal{I}(\mathbf{A})_{(i,j),(p,q)}, \tag{5.11}$$

where $\mathcal{I}(\mathbf{A})_{(i,j),(p,q)}$ *is the asymptotic (co)variance of* $T^{-1/2} S_T^{(i,j)}(\mathbf{A})$ *and* $T^{-1/2} S_T^{(p,q)}(\mathbf{A})$*.*

5.4.2 Scan Score Statistics

To combine all the score statistics and complete the detection procedure, we compute the scan statistics based on given clusters. A cluster is a subset of nodes, which is defined as R_i below:

$$R_i = \{v \in V, v \text{ belongs to } i\text{th cluster}\}, i = 1, \ldots, L;$$

where L is the number of clusters.

In practise, to reduce the computation cost, we only compute the score statistics given data in a time window of length w. Specifically, at time t, for ith cluster, we compute all the interested score statistics with data in $[t - w, t]$ and have a vector of score statistics denoted as $S_{t,w}^{(i)}(\mathbf{A})$. Let $\tilde{R}_i$ and $I_{t,w}^{(i)}(\mathbf{A})$ denote the dimension of $S_{t,w}^{(i)}(\mathbf{A})$ and Fisher information of $S_{t,w}^{(i)}(\mathbf{A})$, respectively. Then the test statistics for cluster i at time τ, with window length w

is:

$$\Gamma_{t,w}^{(i)} = \tilde{R}_i^{-1/2}\mathbf{1}^\top I_{t,w}^{(i)}(\mathbf{A})^{-1/2} S_{t,w}^{(i)}(\mathbf{A}) \sim \mathcal{N}(0,1) \tag{5.12}$$

Then at each time t, we compute the scan score statistics for whole network:

$$\Gamma_t = \max_{1\le i\le L} |\Gamma_{t,w}^{(i)}|,$$

Given a threshold b, we stop our procedure and raise an alarm for change-point detection as the following rule:

$$T_b = \inf\{t : \Gamma_t > b\} \tag{5.13}$$

As we discuss in Remark Remark 5.4.1, we can use parameters of stationary distribution to compute our score statistics with less computation cost. In the following we provide the analysis for the case $\mathbf{A} = 0$.

False Alarm Rate of Scan Statistics

According to the Theorem 5.4.2, we can approximate it with Brownian motion as the following:

$$S_t^{p,q}(\mathbf{0}) \approx \sqrt{\frac{\mu_p}{\mu_q}}\big(\frac{1}{\sqrt{2\beta}} a^{(p,q)}(t) + \frac{\sqrt{\mu}_p}{\beta} b^{(q)}(t)\big), \tag{5.14}$$

where $a^{(p,q)}(t)$, $b^{(q)}(t)$ is independent Brownian motion $\forall p, q \in [M]$. With the same manner, for a score statistics for a window length w, we have the following approximation:

$$S_{t,w}^{p,q}(\mathbf{0}) \approx \sqrt{\frac{\mu_p}{\mu_q}}\big(\frac{1}{\sqrt{2\beta}}(a^{(p,q)}(t) - a^{(p,q)}(t-w)) + \frac{\sqrt{\mu}_p}{\beta}(b^{(q)}(t) - b^{(q)}(t-w))\big). \tag{5.15}$$

With above approximation, we have

$$\mathrm{Cov}(S_{t,w}^{p,q}, S_{t+\delta,w}^{i,j}) = (w-\delta)^+ \mathrm{Cov}(S_1^{p,q}, S_1^{i,j}) \tag{5.16}$$

Therefore, for cluster i and j, we have

$$\mathrm{Cov}(\Gamma^{(i)}_{t,w}, \Gamma^{(j)}_{t+\delta,w}) = \frac{(w-\delta)^+}{w}\mathrm{Cov}(\Gamma^{(i)}_{1,1}, \Gamma^{(j)}_{1,1}) \tag{5.17}$$

Suppose there are L cluster, at time t, we want to control the false alarm rate.

$$\begin{aligned}
\mathbb{P}(\Gamma_t > b) &= \mathbb{P}\big(\max_{1\le i\le L} |\Gamma^{(i)}_{t,w}| \ge b\big) = \mathbb{P}\big(\bigcup_{i}^{L} |\Gamma^{(i)}_{t,w}| \ge b\big) \\
&= \mathbb{P}\big(\bigcup_{i}^{L}\{\Gamma^{(i)}_{t,w} \ge b\}\bigcup_{i}^{L}\{\Gamma^{(i)}_{t,w} \le -b\}\big) \\
&= \mathbb{P}\big(\{\max_{1\le i\le L} \Gamma^{(i)}_{t,w} \ge b\}\bigcup\{\min_{1\le i\le L} \Gamma^{(i)}_{t,w} \le -b\}\big) \\
&\le 2\mathbb{P}\big(\max_{1\le i\le L} \Gamma^{(i)}_{t,w} \ge b\big)
\end{aligned} \tag{5.18}$$

Lwt $\Gamma_{t,w}$ denote the vector of $\Gamma^{(i)}_{t,w}$s. To control the upper bound of the false alarm rate, we compute the Equation 5.18 with the technique in [55]:

$$\begin{aligned}
\mathbb{P}\big(\max_{1\le i\le L} \Gamma^{(i)}_{t,w} \ge b\big) &= \mathbb{P}\big(\bigcup_{i=1}^{L}\{\Gamma^{(i)}_{t,w} \ge b, \Gamma^{(i)}_{t,w} \ge \Gamma^{(j)}_{t,w}, j \ne i\}\big) \\
&= \sum_{i=1}^{L}\mathbb{P}\big(\Gamma^{(i)}_{t,w} \ge b, \Gamma^{(i)}_{t,w} \ge \Gamma^{(j)}_{t,w}, j \ne i\}\big) \\
&= \sum_{i=1}^{L}\mathbb{P}\big(AP_i\Gamma_{t,w} \ge \mathbf{b}\big),
\end{aligned} \tag{5.19}$$

where P_i is the permutation matrix interchanging first entry and the ith entry, and

$$A = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \cdots & 0 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} b \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5.20}$$

According to equation (Equation 5.12), $\Gamma_{t,w} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. In [55], they provide an importance sampling algorithm to estimate the equation (Equation 5.19). Σ can be computed with Equation 5.17 according to the network topology and the score statistics in the clusters. Figure 5.1 is an example and the corresponding Σ.

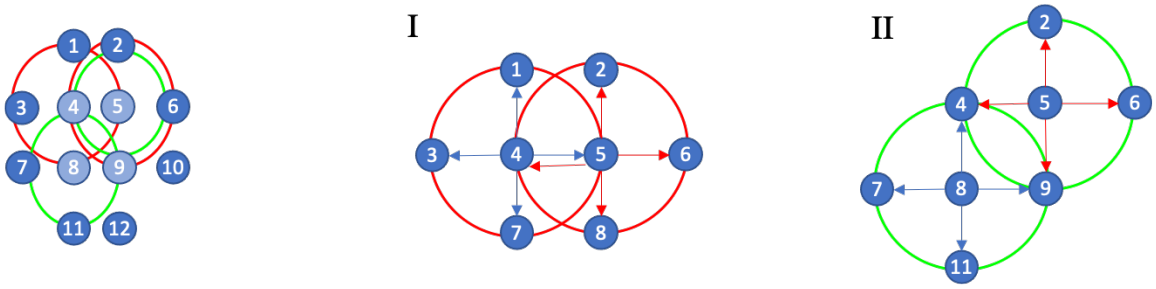

Figure 5.1: In this example, there are 4 clusters, and each cluster includes 5 locations. The light blue nodes are the centers of each cluster and in each cluster we consider the 4 directions from the center to the neighboring nodes as shown in I and II. $S_{t,w}^{(i)} \sim \mathcal{N}(\mathbf{0}_4, w(1/(2\beta) + \mu/\beta^2)\mathbf{I}_4)$ and $\Gamma_{t,w} \sim \mathcal{N}(0, \Sigma)$. For the case I, the Σ_{ij} corresponding to $\Gamma_{t,w}^{(i)}$ and $\Gamma_{t,w}^{(j)}$ equals to 0. For case II, $\Sigma_{ij} = \sigma^2 \triangleq \mu/(\beta + 2\mu)$. Therefore, $\Sigma_{\cdot,1}^\top = (1, 0, 0, \sigma^2)$, $\Sigma_{\cdot,2}^\top = (0, 1, \sigma^2, 0)$, $\Sigma_{\cdot,3}^\top = (0, \sigma^2, 1, 0)$, $\Sigma_{\cdot,4}^\top = (\sigma^2, 0, 0, 1)$.

Remark 5.4.2. *Since our scan statistics are standardized, determining the threshold b with the method in Equation 5.19 does not depend on the window length w. However, with a larger w the approximation in Equation 5.15 would be better. In Table 5.1, we can see, as window length increase, the false alarm rate would be better controlled.*

Table 5.1: Approximation of false alarm rate.

w	b	$\mathbb{P}(\max_i \Gamma_{t,w}^{(i)} > b)$	$\hat{\mathbb{P}}(\Gamma_t^G > b)$
50	3	0.005	0.0174
100	3	0.005	0.0146
200	3	0.005	0.0114
50	2.8	0.01	0.0282
100	2.8	0.01	0.0226
200	2.8	0.01	0.0210

5.4.3 Average Run Length of T_b

Besides false alarm rate, another performance measure of a change-point detection procedure is the *average run length* (ARL) of the stopping time in Equation 5.13, which is denoted as $\mathbb{E}_\infty[T_b]$. To evaluate the ARL, we are going to show that T_b is approximately exponential distribution with some parameter λ_0. The analysis is similar to [60]. Let $f(b) = be^{b^2/2}$. For certain interval $[0, xf(b)]$, we decompose it into k sub-intervals with length m, i.e $xf(b) = km$. For simplicity, we assume k and m are integers. Let indicator X_j denotes $\mathbb{I}\{\max_{t\in((j-1)m,jm]} \Gamma_t > b\}$, and defind $W = \sum_{j=1}^{k} X_j$, then we have $\{W = 0\} = \{T_b > xf(b)\}$. To prove that T_b is approximately exponential, it is same to prove W is approximately Poisson distributed. We herein apply the result from [61]. According to the [61, Theorem I], we establish the following theorem.

Theorem 5.4.4. *Let T_b be the stopping time defined in Equation 5.13, X_j be the indicator defined above and W be the sum of the indicators. With $w \ll m \ll f(b)$, then*

$$\lim_{b\to\infty} |\mathbb{P}(T_b > xf(b)) - e^{-\mathbb{E}W}| = 0 \tag{5.21}$$

According to the construction of W, we have

$$\begin{aligned}\mathbb{E}W &= xf(b)\mathbb{P}(X_j = 1)/m \\ &\leq 2xf(b)\mathbb{P}\{\max_{0<t\leq m, 1\leq i\leq d} \Gamma_{t,w}^{(i)} > b\}/m \leq 2xf(b)\mathbb{P}\{\max_{1\leq i\leq d} \Gamma_{t,w}^{(i)} > b\}.\end{aligned}$$

By Theorem Theorem 5.4.4, $\mathbf{E}_\infty(T) \approx \lambda_0^{-1}$ and

$$\begin{aligned}\lambda_0 &\leq 2\mathbb{P}\{\max_{0<t\leq m, 1\leq i\leq d} \Gamma_{t,w}^{(i)} > b\}/m &(5.22)\\ &\leq 2\mathbb{P}\{\max_{1\leq i\leq d} \Gamma_{t,w}^{(i)} > b\} &(5.23)\end{aligned}$$

5.5 Experiments

5.5.1 Simulated result of ARL and EDD

In this experiment, the network is set up as shown in Figure 5.1. The event in each notes follows a Poisson process with $\mu = 1$, and we set the $\beta = 1$. The window length is set to be 200 and the statistics is computed each $\delta = 10$ time units. In Table 5.2, we show the estimated ARL of simulation with the threshold estimated by (Equation 5.22) and (Equation 5.23) for $\lambda^{-1} \geq 1000$ and $\lambda^{-1} > 2000$, which corresponds to ARL$\geq \lambda^{-1}\delta$. To get the simulated ARL, we generate events in time window $[0, 60000]$ and compute the run length when the statistics exceed the corresponding threshold. Note that this approximation will always underestimate the ARL since we can only generate events in finite time window. We can see the thresholds computed from (Equation 5.22) give us desired results. However (Equation 5.23) tends to over estimate the threshold.

Table 5.2: Verification of approximated ARL in (Equation 5.22) and (Equation 5.23)

	b	theoretic ARL	simulated ARL
Results of (Equation 5.22), $m = 100$	3.3718	10000	9762
Results of (Equation 5.22), $m = 50$	3.3859	10000	9945
Results of (Equation 5.23)	3.6625	10000	22818
Results of (Equation 5.22), $m = 100$	3.5824	20000	18380
Results of (Equation 5.22), $m = 50$	3.5867	20000	18747
Results of (Equation 5.23)	3.8352	20000	32157

Now, let's compare expected detection delay (EDD) of our proposed method with generalized likelihood ratio (GLR) method in [56]. In the experimentsof EDD, the distribution under H_0 is set as mentioned above. The thresholds of our methods are set according to the estimate of Equation 5.22 with $m = 50$, so that our desired ARL are 10000 or 20000 (see detials in Table 5.2). As for the GLR, we compute the log generalized likelihood ratio with frequency 0.1 per time unit, and window length $w = 200$. The maximum likelihood estimates of the $\mathbf{A}_1$ and $\boldsymbol{\mu}_1$ are computed by Newton method. The thresholds of desired ARLs

are estimated with simulation. We compare different settings of post-change distribution, which is shown in Table 5.3. The simulated EDDs are shown in column 4-9 of Table 5.4. We can see our proposed method has better performance in the cases that there are multiple changes in the influence matrix $\mathbf{A}_1$. Even though our method does not consider the change in base intensity $\boldsymbol{\mu}_1$, from case iv, we can see our method still have a better performance when there are significant change in the influence matrix $\mathbf{A}_1$. In case v and vi, we can see, when the main change is the base intensity or the change of influence matrix is weak, GLR has better performance. It is also worth noticing that the computation cost of our proposed methods is much less than the GLR methods since our method does not require to estimate the post-change distribution parameters. The speed of performing our method are about 5 times faster than the GLR method in our experiments.

Table 5.3: Setting of different cases in Table 5.4

	changed parameters in post-change distribution
Case *i*	$\alpha_{4,1} = \alpha_{4,3} = \alpha_{4,5} = \alpha_{4,8} = 0.2$
Case *ii*	$\alpha_{4,1} = \alpha_{4,3} = \alpha_{4,5} = \alpha_{4,8} = 0.5$
Case *iii*	$\alpha_{4,5} = \alpha_{4,8} = \alpha_{9,8} = \alpha_{9,5} = 0.5$
Case *iv*	$\mu_4 = 1.5, \alpha_{4,5} = \alpha_{4,8} = 1$
Case *v*	$\mu_4 = 1.5, \alpha_{4,5} = 0.5$
Case *vi*	$\alpha_{4,5} = 0.5$

Table 5.4: Comparison of EDD

Methods	thresholds	ARLs	Case *i*	Case *ii*	Case *iii*	Case *iv*	Case *v*	Case *vi*
Proposed	3.3859	10000	**101.16**	**45.63**	**46.45**	**26.4**	77.8	152.86
Log GLR	20.599	10000	145.25	62.75	48.95	28.55	**66.6**	**128.45**
Proposed	3.5867	20000	**107.63**	**47.86**	**48.70**	**28.0**	85.4	164.60
Log GLR	21.407	20000	153.55	66.00	51.95	29.85	**69.5**	**134.60**

5.5.2 Real-data

In this section, we apply our scan statistics on a *memetracker data* and *stock data*.

- memetracker data: It tracks texts and phrases, which are called meme, over different websites. This data is used to study the information diffusion via social media and blogs. We use three meme data in [56]. The first data is"Barack Obama was elected as the 44th president of the United State". We use data from the top 40 news website, which includes Yahoo, CNN, Nydaily, The Guardian, etc. We use the data from Nov.01.2008 to Nov.02.2008 as the training data and the data from Nov.03.2008 to Nov.05.2008 as the test data. Our procedure detect a change at the time around 7pm on Nov.03, which is few hours before the votes. The second data is "the summer Olympics game in Beijing". We use data from Aug.01.2008 to Aug.03.2008 as the training data and data from Aug.04.2008 to Aug.15.2008 as the test set.

- Stock data: This data is downloaded from Yahoo Finance. We collect the closing price and trading volume of stock tickers: SPY, QQQ, DIA, EFA, and IWM, which are all index-type stock and can reflect the situation of overall stock market. For each ticker, we construct 3 types of events. High return: the day with return over 90 percentile. Low return: the day with return below 10 percentile. High volume: the day with trading volume over 90 percentile. Therefore, in this data we have a network with 15 nodes. Such extreme trading events are of interest in the study of finance [62]. We use the data from Jan.04.2016 to Dec.31.2018 as the training data and data from Jan.01.2019 to Dec.31.2020 as the test data.

For each data, we apply Newton method to fit the MLE of the parameters for the training set. Then, we use the fitted parameters to compute the scan statistics on the test set. For memetracker data, we construct the cluster by applying community detection methods on the fitted $\hat{A}$. For stock data, each cluster are the events that related to certain ticker. Details are shown in Table 5.5. The change point detected from the "Obama" data, is around 7am on 2008.11.03. The result indicates that the public opinion of Barack Obama changed at around one day before the votes. For the "Olympic" data, our procedure detect a change on Aug.04 which is 3 days before the Olympic game. For the stock data, we detect 3 time

intervals of which the starting dates are 2019.06.21, 2019.08.16 and 2020.03.04. According to the news, the first change-point 2019.06.21 is the date that the S&P 500 hit a new record-high and the three major stock indexes surged with different scale. The second change-point 2019.08.16 is related to the US-China trade war. In August 2019, both US and China made multiple announcement about their tariffs. The last change-point is 2020.03.04 which is 3 business day before the first circuit breaker in 2020. There are a lot change-points after first circuit breaker 2020.03.09, which indicates a long-term change in stock market caused by pandemic and trade-war. The results of real data shows that our proposed scan statistics has a good performance on detecting the real change in different area such as social media, finance markets.

Table 5.5: result of real data

Data	training set	test set	# of cluster	thresholds	change-points
"Obama"	08.11.01-08.11.02	08.11.03-08.11.05	5	4	11.03 7am
"Olympic"	08.08.01-08.08.03	08.08.04-08.08.08	4	4	08.04 6pm
stock data	16.01.04-18.12.31	19.01.01-20.12.31	5	4	19.06.21, 19.08.16, 20.03.04

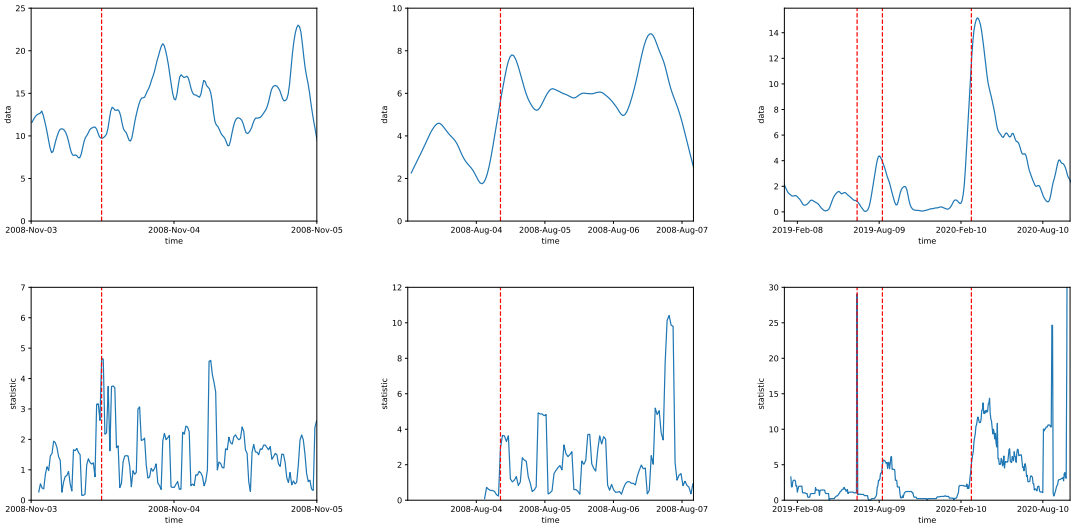

Figure 5.2: Scan Statistics procedure applied on real data. Red line: detected change-points. For the upper plots, the blue line is the smoothed frequency of all events in the network. For the lower plots, the blue line is the scan statistics of proposed procedure. (1,1) & (2,1): data of "Obama". (1,2) & (2,2): data of "Olympic". (1,3) & (2,3): data of stock.

5.6 Conclusion

In this chapter, we proposed a scan score statistics for detecting the change-points of network point processes. We use multivariate Hawkes process to model the sequential event data. Our proposed method is based on score statistics, meaning that we don't need to compute the post-change parameters. Therefore, our method is computational efficient compared to the conventional GLR method, which is of importance in online detection. Moreover, we are able to provide analysis of the false alarm rate and average run length. In experiments, we first use simulated data to verify our theoretical results. We also perform our method in real world data, which shows a promising detection performance.

Appendices

APPENDIX A

APPENDICES OF CHAPTER 2

Proof of Theorem 2.3.2

We argue by a contradiction. Suppose that there is a sequence $\{Y_k\} \subset \mathcal{M}_r$ (with $Y_k \neq \bar{Y}$) converging to $\bar{Y}$ such that $P_\Omega(Y_k) = M$. It follows that $Y_k - \bar{Y} \in \mathbb{V}_{\Omega^c}$. By passing to a subsequence if necessary we can assume that $(Y_k - \bar{Y})/t_k$, where $t_k := \|Y_k - \bar{Y}\|$, converges to some $H \in \mathbb{V}_{\Omega^c}$. Note that $H \neq 0$. Moreover $Y_k = \bar{Y} + t_k H + o(t_k)$, and hence $H \in \mathcal{T}_{\mathcal{M}_r}(\bar{Y})$. That is $H \in \mathbb{V}_{\Omega^c} \cap \mathcal{T}_{\mathcal{M}_r}(\bar{Y})$, and $H \neq 0$ by the construction. This gives the desired contradiction with (Equation 2.17).

Proof of Theorem 2.3.4

Let ϱ be the characteristic rank of mapping $\mathfrak{F}$. Consider $\theta^* \in \Theta$ such that $\varrho = \operatorname{rank}\big(\Delta(\theta^*)\big)$. It follows that matrix $\Delta(\theta^*)$ has an $\varrho \times \varrho$ submatrix whose determinant is not zero. Consider function $\phi : \Theta \to \mathbb{R}$ defined as the determinant of the corresponding $\varrho \times \varrho$ submatrix of $\Delta(\theta)$. We have that $\phi(\cdot)$ is a polynomial function and is not identically zero on Θ since by the construction $\phi(\theta^*) \neq 0$. Since Θ is connected, it follows that the set $\{\theta \in \Theta : \phi(\theta) = 0\}$ is "thin", in particular has Lebesgue measure zero. That is, $\phi(\theta) \neq 0$ and hence $\operatorname{rank}\big(\Delta(\theta)\big) \geq \varrho$ for a.e. $\theta \in \Theta$. Also by the definition of ϱ we have that $\operatorname{rank}\big(\Delta(\theta)\big) \leq \varrho$ for all $\theta \in \Theta$. It follows that $\operatorname{rank}\big(\Delta(\theta)\big) = \varrho$ for a.e. $\theta \in \Theta$. Since rank of $\Delta(V, W, X)$ is the same for all $X \in \mathbb{V}_{\Omega^c}$, this completes the proof of the assertion (i). Since $\operatorname{rank}\big(\Delta(\cdot)\big)$ is a lower semicontinuous function, the assertion (ii) follows.

Now consider a regular point $\bar{\theta} = (\bar{V}, \bar{W}, \bar{X})$ with $\bar{X} = 0$, and the corresponding matrix $\bar{Y} = \bar{V}\bar{W}^\top$. Since $\bar{\theta}$ is regular, we have that rank of $\Delta(\theta)$ is constant (equal ϱ)

for all θ in a neighborhood of $\bar{\theta}$. By the Constant Rank Theorem it follows that there is a neighborhood $\mathcal{V}$ of $\bar{\theta}$ such that the set $\mathcal{S} := \{\mathfrak{F}(\theta) : \theta \in \mathcal{V}\}$ forms a smooth manifold of dimension ϱ in $\mathbb{R}^{n_1 \times n_2}$. The tangent space to this manifold at $\bar{Y}$ is the space $\mathcal{T}_{\mathcal{M}_r}(\bar{Y}) + \mathbb{V}_{\Omega^c}$. Hence if $\varrho = \mathfrak{f}(r, m)$, then

$$\dim \left(\mathcal{T}_{\mathcal{M}_r}(\bar{Y}) + \mathbb{V}_{\Omega^c} \right) = \dim(\mathcal{T}_{\mathcal{M}_r}(\bar{Y})) + \dim(\mathbb{V}_{\Omega^c}).$$

Consequently $\dim \left(\mathcal{T}_{\mathcal{M}_r}(\bar{Y}) \cap \mathbb{V}_{\Omega^c} \right) = 0$, and thus condition (Equation 2.17) follows (compare with Proposition Proposition 2.3.1). On the other hand if $\varrho < \mathfrak{f}(r, m)$, then the manifold $(\mathbb{V}_{\Omega^c} + \bar{Y}) \cap \mathcal{M}_r$, in a neighborhood of $\bar{Y}$, has a positive dimension. Thus in that case the solution of MRMC is not locally unique and condition (Equation 2.17) does not hold. This completes the proof of the assertions (iii) and (iv).

Proof of Theorem 2.3.5

Suppose that Ω is reducible. Then by making permutations of rows and columns if necessary, it can be assumed that M has the block diagonal form as in (Equation 2.23). Let $\bar{Y}$ be a respective minimum rank solution. That is $M_1 = V_1 W_1^\top$, $M_2 = V_2 W_2^\top$ and $\bar{Y} = V W^\top$ with $V = \left(\begin{smallmatrix} V_1 \\ V_2 \end{smallmatrix}\right)$ and $W = \left(\begin{smallmatrix} W_1 \\ W_2 \end{smallmatrix}\right)$ being $n_1 \times r$ and $n_2 \times r$ matrices of rank r. Note that $\bar{Y} = \left(\begin{smallmatrix} M_1 & V_1 W_2^\top \\ V_2 W_1^\top & M_2 \end{smallmatrix}\right)$. By changing V_1 to αV_1 and W_1 to $\alpha^{-1} W_1$ for $\alpha \neq 0$, we change matrix $\bar{Y}$ to matrix $\left(\begin{smallmatrix} M_1 & \alpha V_1 W_2^\top \\ \alpha^{-1} V_2 W_1^\top & M_2 \end{smallmatrix}\right)$. If $V_1 W_2^\top \neq 0$ or $V_2 W_1^\top \neq 0$, we obtain that solution $\bar{Y}$ is not locally unique. On the other hand when both $V_1 W_2^\top = 0$ and $V_2 W_1^\top = 0$, and hence $\bar{Y} = \left(\begin{smallmatrix} M_1 & 0 \\ 0 & M_2 \end{smallmatrix}\right)$, rank r solutions for example are matrices of the form $\bar{Y} = \left(\begin{smallmatrix} M_1 & M_3 \\ 0 & M_2 \end{smallmatrix}\right)$, where columns of matrix M_3 are linear combinations of columns of matrix M_1. If $M_1 = 0$, then we can use matrix $\bar{Y} = \left(\begin{smallmatrix} M_1 & 0 \\ M_3 & M_2 \end{smallmatrix}\right)$ in the similar way. Hence nonuniqueness of rank r solutions follows.

Proof of Theorem 2.3.6

Suppose that Ω is irreducible. Consider a rank one solution $\bar{Y} = vw^\top$ with respective vectors $v = (v_1, ..., v_{n_1})^\top$ and $w = (w_1, ..., w_{n_2})^\top$. We can assume that v_1 is fixed, say $v_1 = 1$. Consider an element M_{1j_1}, $(1, j_1) \in \Omega$, in the first row of matrix M. Since it is assumed that each row has at least one observed entry, such element exists. Since $M_{1j_1} = v_1 w_{j_1}$, it follows that the component w_{j_1} of vector w is uniquely defined. Next consider element M_{i_1,j_1}, $(i_1, j_1) \in \Omega$. Since $M_{i_1 j_1} = v_{i_1} w_{j_1}$, it follows that the component v_{i_1} of vector v is uniquely defined. We proceed now iteratively. Let $\nu \subset \{1, ..., n_1\}$ and $\omega \subset \{1, ..., n_2\}$ be index sets for which the respective components of vectors v and w are already uniquely defined. Let $j \notin \omega$ be such that there is $(i, j') \in \Omega$ with $j' \in \omega$ and hence $w_{j'}$ is already uniquely defined. Since $M_{ij} = v_i w_j$ and $M_{ij'} = v_i w_{j'}$, it follows that w_j is uniquely defined and j can be added to the index set ω. If such column j does not exist, take row $i \notin \nu$ such that there is $(i', j) \in \Omega$ with $i' \in \nu$. Then v_i is uniquely defined and hence i can be added to ν. Since Ω is irreducible, this process can be continued until all components of vectors v and w are uniquely defined.

Proof of Proposition Proposition 2.3.3

Consider function defined in (Equation 2.26). The differential of $f(Y)$ can be written as

$$\mathrm{d}f(Y) = \mathrm{tr}[(P_\Omega(Y) - M)^\top \mathrm{d}Y].$$

Therefore if $Y \in \mathcal{M}_r$ is an optimal solution of the least squares problem (Equation 2.5), then $\nabla f(Y) = P_\Omega(Y) - M$ is orthogonal to the tangent space $T_{\mathcal{M}_r}(Y)$. By (Equation 2.16) this implies optimality conditions (Equation 2.24) .

Proof of Proposition 2.3.4

Consider function ϕ defined in (Equation 2.25), and the problem of minimization of $\phi(Y, \Theta)$ subject to $Y \in \mathcal{M}_r$ with Θ viewed as a parameter. Locally for Y near $\bar{Y} \in \mathcal{M}_r$ the manifold $\mathcal{M}_r$ can be represented by a system of $K = n_1 n_2 - \dim(\mathcal{M}_r)$ equations $g_i(Y) = 0$, $i = 1, ..., K$, for an appropriate smooth mapping $g = (g_1, ..., g_K)$. That is, the above optimization problem can be written as

$$\min \phi(y, \theta) \ \text{ subject to } g_i(y) = 0, \ i = 1, ..., K, \tag{A.1}$$

where with some abuse of the notation we write this in terms of vectors $y = \mathrm{vec}(Y)$ and $\theta = \mathrm{vec}(\Theta)$. Note that the mapping g is such that the gradient vectors $\nabla g_1(\bar{y}), ..., \nabla g_K(\bar{y})$ are linearly independent.

First order optimality conditions for problem (Equation A.1) are

$$\nabla_y L(y, \lambda, \theta) = 0, \ g(y) = 0, \tag{A.2}$$

where $L(y, \lambda, \theta) := f(y, \theta) + \lambda^\top g(y)$ is the corresponding Lagrangian. For $\theta = \theta_0$ this system has solution $\bar{y}$ and the corresponding vector $\bar{\lambda} = 0$ of Lagrange multipliers. We can view (Equation A.2) as a system of (nonlinear) equations in $z = (y, \lambda)$ variables.

We would like now to apply the Implicit Function Theorem to this system of equations to conclude that for all θ near θ_0 it has unique solution near $\bar{z} = (\bar{y}, \bar{\lambda})$. Consider the Jacobian matrix $\left(\begin{smallmatrix} H & G \\ G^\top & 0 \end{smallmatrix}\right)$ of the system (Equation A.2) at $(y, \lambda) = (\bar{y}, \bar{\lambda})$, where $H := \nabla_{yy} \phi(\bar{y}, \theta_0)$ is the Hessian matrix of the objective function and $G := \nabla g(\bar{y}) = [\nabla g_1(\bar{y}), ..., \nabla g_K(\bar{y})]$. We need to verify that this Jacobian matrix is nonsingular. This is implied by condition (Equation 2.17), which is equivalent to condition (Equation 2.27).

Indeed suppose that

$$\begin{bmatrix} H & G \\ G^\top & 0 \end{bmatrix} \begin{bmatrix} v \\ u \end{bmatrix} = 0, \tag{A.3}$$

for some vectors v and u of appropriate dimensions. This means that $Hv + Gu = 0$ and $G^\top v = 0$. It follows that $v^\top Hv = 0$. Condition $G^\top v = 0$ means that v is orthogonal to the tangent space $\mathcal{T}_{\mathcal{M}_r}(\bar{y})$. It follows then by condition (Equation 2.27) that $v = 0$. Then $Gu = 0$ and hence, since G has full column rank, it follows that $u = 0$. Since equations (Equation A.3) have only zero solution, it follows that this Jacobian matrix is nonsingular. Now by implying the Implicit Function Theorem to the system (Equation A.2) we obtain the required result. This completes the proof.

Proof of Proposition 2.4.2

Note that under the specified assumptions, $M_{ij} - Y^*_{ij}$ are of stochastic order $O_p(N^{-1/2})$. We have by Proposition 2.4.1 that an optimal solution of problem (Equation 2.29) converges in probability to Y^*. By the standard theory of least squares (e.g., [63, Lemma 2.2]) we can write the following local approximation near Y^* as (Equation 2.32). It follows that the limiting distribution of $T_N(r)$ is the same as the limiting distribution of N times the first term in the right hand side of (Equation 2.32). Note that $N^{1/2} w_{ij}^{1/2} E_{ij}$ converges in distribution to normal with mean $\sigma_{ij}^{-1}\Delta_{ij}$ and variance one. It follows that the limiting distribution of N times the first term in the right hand side of (Equation 2.32), and hence the limiting distribution of $T_N(r)$, is noncentral chi-square with degrees of freedom $\nu = m - \dim\left(P_\Omega(\mathcal{L})\right)$ and the noncentrality parameter δ_r. Recall that dimension of the linear space $\mathcal{L}$ is equal to the sum of the dimension of its image $P_\Omega\left(\mathcal{L}\right)$ plus the dimension of the kernel $\mathrm{Ker}(P_\Omega)$. It remains to note that condition (Equation 2.17) means that $\mathrm{Ker}(P_\Omega) =$

$\{0\}$ (see Remark 2.3.3), and hence

$$\dim\left(P_{\Omega}(\mathcal{L})\right) = \dim\left(\mathcal{L}\right) = r(n_1 + n_2 - r). \tag{A.4}$$

This completes the proof.

APPENDIX B

APPENDICES OF CHAPTER 3

Proof of Proposition 3.2.1

(i) Since $G(\cdot)$ is twice continuously differentiable, it follows that $J(\cdot)$ is continuous. Thus the function $\mathrm{rank}(J(\cdot))$ is lower semicontinuous, and hence the set $\{\theta \in \Theta : \mathrm{rank}(J(\theta)) \leq \mathfrak{r} - 1\}$ is closed. It follows that its complement set $\{\theta \in \Theta : \mathrm{rank}(J(\theta)) = \mathfrak{r}\}$ is open.

(ii) Let $\theta_0 \in \Theta$ be such that $\mathrm{rank}(J(\theta_0)) = \mathfrak{r}$, such θ_0 exists since the function $\mathrm{rank}(J(\cdot))$ is piecewise constant. Consider an $\mathfrak{r} \times \mathfrak{r}$ submatrix of $J(\theta_0)$ of rank $\mathfrak{r}$, and the associated function $\phi(\theta)$ given by the determinant of this submatrix of $J(\theta)$. Since $G(\cdot)$ is analytic, we have that the function $\phi(\cdot)$ is analytic and is not constantly zero since $\phi(\theta_0) \neq 0$. It follows that the set $\{\theta : \phi(\theta) = 0\}$ has (Lebesgue) measure zero (e.g., [64]). That is, for a.e. θ we have that $\mathrm{rank}(J(\theta)) \geq \mathfrak{r}$. Since by the definition the rank $\mathfrak{r}$ is maximal, it follows that $\mathrm{rank}(J(\theta)) = \mathfrak{r}$ for a.e. $\theta \in \Theta$. This completes the proof.

Proof of Proposition 3.3.1

Since $\mathfrak{M}$ is a smooth manifold near x_0 it can be defined by equations $\phi(x) = 0$ in a neighborhood of x_0 with $\phi : \mathbb{R}^m \to \mathbb{R}^m$ being a smooth near x_0 mapping with nonsingular Jacobian matrix $\nabla\phi(x_0)$. Then optimality condition (Equation 3.9) can be written as: there exists $\lambda \in \mathbb{R}^m$ such that the derivatives of the Lagrangian $L(x, \lambda) := \frac{1}{2}\|\hat{y} - x\|^2 - \lambda^\top \phi(x)$ are zeros at $(\hat{x}, \lambda)$. This can be written as the following system of equations in (x, λ),

$$\nabla_x L(x, \lambda) = 0, \ \phi(x) = 0. \tag{B.1}$$

Note that as $\hat{y}$ and x approach x_0, the corresponding λ tends to 0. The Jacobian matrix of partial derivatives of this system, with respect to (x, λ), at $x = x_0$ and $\lambda = 0$ is $\begin{pmatrix} I_m & \nabla\phi(x_0) \\ \nabla\phi(x_0)^\top & 0 \end{pmatrix}$. This Jacobian matrix is nonsingular. It follows by the Implicit Function Theorem that in a neighborhood $\mathcal{W}$ of x_0 the system (Equation B.1) has unique solution. Moreover by Remark 3.3.1 the neighborhood $\mathcal{W}$ can be such that if $\hat{y} \in \mathcal{W}$, then any optimal solution of the least squares problem is in $\mathcal{W}$. If moreover $\hat{x}$ is in $\mathcal{W}$ and satisfies optimality Equation B.1, then by the uniqueness property $\hat{x}$ should coincide with the corresponding optimal solution. This completes the proof.

Proof of Theorem 3.3.1

Since $\hat{y}$ converges in probability to x_0, the assertion (i) follows from Proposition 3.3.1. Also any minimizer $\hat{x}$ in the right hand side of (Equation 3.8) converges in probability to x_0 (see Remark 3.3.1). Therefore we can perform the asymptotic analysis in a neighborhood of x_0. As in the above proof of Proposition 3.3.1, $\mathfrak{M}$ can be defined by equations $\phi(x) = 0$ in a neighborhood of x_0 with nonsingular Jacobian matrix $\nabla\phi(x_0)$. Let $(\hat{x}, \hat{\lambda})$ be a solution of Equation B.1 in a sufficiently small neighborhood of $(x_0, 0)$. By the Implicit Function Theorem we have that

$$\begin{bmatrix} \hat{x} - x_0 \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} I_m & \nabla\phi(x_0) \\ \nabla\phi(x_0)^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{y} - x_0 \\ 0 \end{bmatrix} + o(\|\hat{y} - x_0\|). \tag{B.2}$$

Also it follows by (Equation 3.7) that $N^{1/2}(\hat{y} - x_0)$ converges in distribution to normal $\mathcal{N}(\gamma, \sigma^2 I_m)$. In particular this implies that $\|\hat{y} - x_0\| = O_p(N^{-1/2})$, and hence

$$\hat{x} - x_0 = P(\hat{y} - x_0) + o_p(N^{-1/2}), \tag{B.3}$$

where

$$P = I_m - \nabla\phi(x_0)\left(\nabla\phi(x_0)^\top \nabla\phi(x_0)\right)^{-1}\nabla\phi(x_0)^\top. \tag{B.4}$$

Note that $\mathcal{T}_{\mathfrak{M}}(x_0) = \{v : \nabla\phi(x_0)^\top v = 0\}$. Therefore matrix P in (Equation B.4) is the orthogonal projection matrix onto the tangent space $\mathcal{T}_{\mathfrak{M}}(x_0)$. Slutsky's theorem together with (Equation B.3) imply that $N^{1/2}(\hat{x} - x_0)$ has the same asymptotic distribution as $P[N^{1/2}(\hat{y} - x_0)]$. Since $N^{1/2}(\hat{y} - x_0)$ converges in distribution to normal $\mathcal{N}(\gamma, \sigma^2 I_m)$, the assertion (iii) follows, and the assertion (iv) follows by similar arguments.

Moreover by (Equation B.3),

$$\hat{y} - \hat{x} = \hat{y} - x_0 - (\hat{x} - x_0) = (I_m - P)(\hat{y} - x_0) + o_p(N^{-1/2}),$$

and since $\|\hat{y} - x_0\| = O_p(N^{-1/2})$ it follows that

$$\|\hat{y} - \hat{x}\|_2^2 = \|(I_m - P)(\hat{y} - x_0)\|_2^2 + o_p(N^{-1}). \tag{B.5}$$

It follows by Slutsky's theorem that the N times right hand side of (Equation B.5) has the same asymptotic distribution as $Z^\top(I_m - P)Z$, where $Z \sim \mathcal{N}(\gamma, \sigma^2 I_m)$. The assertion (ii) follows. This completes the proof.

Theorem 3.3.2 can be proved in a similar way by showing that asymptotically this is equivalent to the linear case.

Proof of Proposition 3.3.2

Let $x = \mathcal{G}(\xi)$ be a well-posed point. Then $\mathcal{T}_{\mathcal{M}}(x) = \{d\mathcal{G}(\xi)h : h \in \mathbb{R}^d\}$, and for any $\zeta \in \mathbb{R}^k$ we have by (Equation 3.14) that dimension of the image of the differential $dG(\xi, \zeta)$

is $\rho + k$. It follows that $\mathfrak{r} \geq \rho + k$. Since $\mathfrak{r} \leq \rho + k$, it follows that $\mathfrak{r} = \rho + k$.

Conversely suppose that $\mathcal{M}$ is a smooth manifold of dimension ρ and $\mathfrak{r} = \rho + k$. Let $\theta \in \Theta$ be such that dimension of the image of $dG(\theta)$ is $\mathfrak{r}$, by Proposition Proposition 3.2.1 we have that a.e. θ is like that. Since $\mathfrak{r} = \rho + k$ and $\mathcal{T}_{\mathcal{M}}(x) = \{d\mathcal{G}(\xi)h : h \in \mathbb{R}^d\}$ we have by (Equation 3.14) that (Equation 3.15) follows. It remains to note that $dG(\theta) = dG(\theta')$ for any points $\theta = (\xi, \zeta)$ and $\theta' = (\xi, \zeta')$ in Θ with the same first component. This completes the proof.

Proof of Proposition 3.4.4

Let ρ be the characteristic rank of mapping

$$\mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r} \ni (A, B, C) \mapsto A \otimes B \otimes C. \tag{B.6}$$

Recall that it always holds that $r(n_1 + n_2 + n_3 - 2) \geq \rho$.

Consider $\xi = (A, B, C)$ such that rank of the Jacobian matrix of mapping (Equation B.6) at (A, B, C) is ρ. For $X = A \otimes B \otimes C$ consider the set

$$\mathcal{G}^{-1}(X) = \left\{(A', B', C') \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r} : \right.$$
$$\left. A' \otimes B' \otimes C' = X\right\}.$$

By the Constant Rank Theorem this set forms a smooth manifold of dimension

$$\dim\left(\mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}\right) - \rho = r(n_1 + n_2 + n_3) - \rho$$

in a neighborhood of the point ξ. If (Equation 3.29) holds, then dimension of this manifold is $2r$, and hence any $(A', B', C') \in \mathcal{G}^{-1}(X)$ in a neighborhood of (A, B, C) can be obtained

by the rescaling. That is, the local identifiability follows.

On the other hand if $r(n_1 + n_2 + n_3) - \rho > 2r$, then this will imply that there exists $(A', B', C') \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}$ near (A, B, C) such that $A' \otimes B' \otimes C' = A \otimes B \otimes C$ and (A', B', C') cannot be obtained from (A, B, C) by the rescaling. That is, the local identifiability does not hold.

Derivation of the Jacobian matrix in section subsection 3.4.6.

For all $k_0 = 1, \ldots, K$, $\forall n, m, n_0 = 1, \ldots, N$ and $f = 0, \ldots, T-1$, the entries of the Jacobian matrix can be derived as follows

$$\begin{aligned}\frac{\partial \mathcal{R}_{n,m,f}}{\partial \rho_{k_0}} =& \sum_{l=1}^{K} \rho_l (\cos(2\pi f(\tau_{n,l} - \tau_{m,k_0})) \\ &+ \cos(2\pi f(\tau_{n,k_0} - \tau_{m,l}))) \\ &\cdot \pi \sqrt{\frac{1}{\alpha_{k_0}\alpha_l}} e^{-\pi^2 f^2 (\frac{1}{\alpha_{k_0}} + \frac{1}{\alpha_l})}.\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{I}_{n,m,f}}{\partial \rho_{k_0}} =& \sum_{l=1}^{K} \rho_l (\sin(2\pi f(\tau_{n,l} - \tau_{m,k_0})) \\ &+ \sin(2\pi f(\tau_{n,k_0} - \tau_{m,l}))) \\ &\cdot \pi \sqrt{\frac{1}{\alpha_{k_0}\alpha_l}} e^{-\pi^2 f^2 (\frac{1}{\alpha_{k_0}} + \frac{1}{\alpha_l})}.\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \mathcal{R}_{n,m,f}}{\partial \alpha_{k_0}} =& -\frac{\pi}{2}\sum_{l=1}^{K}\rho_{k_0}\rho_l(\cos(2\pi f(\tau_{n,l}-\tau_{m,k_0}))\\
&+\cos(2\pi f(\tau_{n,k_0}-\tau_{m,l})))\\
&\cdot \alpha_{k_0}^{-\frac{3}{2}}\alpha_l^{-\frac{1}{2}}e^{-\pi^2 f^2(\frac{1}{\alpha_k}+\frac{1}{\alpha_l})}\\
&+\pi^3 f^2\sum_{l=1}^{K}\rho_{k_0}\rho_l(\cos(2\pi f(\tau_{n,l}-\tau_{m,k_0}))\\
&+\cos(2\pi f(\tau_{n,k_0}-\tau_{m,l})))\alpha_{k_0}^{-\frac{1}{2}}\\
&\cdot \alpha_l^{-\frac{1}{2}}e^{-\pi^2 f^2(\frac{1}{\alpha_{k_0}}+\frac{1}{\alpha_l})}\alpha_{k_0}^{-2}\\
=&\frac{\partial \mathcal{R}_{n,m,f}}{\partial \rho_{k_0}}(-\frac{\rho_{k_0}\alpha_{k_0}^{-1}}{2}+\pi^2 f^2\rho_{k_0}\alpha_{k_0}^{-2}).
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{I}_{n,m,f}}{\partial \alpha_{k_0}} =& -\frac{\pi}{2}\sum_{l=1}^{K}\rho_{k_0}\rho_l(\sin(2\pi f(\tau_{n,l}-\tau_{m,k_0}))\\
&+\sin(2\pi f(\tau_{n,k_0}-\tau_{m,l})))\\
&\cdot \alpha_{k_0}^{-\frac{3}{2}}\alpha_l^{-\frac{1}{2}}e^{-\pi^2 f^2(\frac{1}{\alpha_k}+\frac{1}{\alpha_l})}\\
&\pi^3 f^2\sum_{l=1}^{K}\rho_{k_0}\rho_l(\sin(2\pi f(\tau_{n,l}-\tau_{m,k_0}))\\
&\sin(2\pi f(\tau_{n,k_0}-\tau_{m,l})))\\
&\cdot \alpha_{k_0}^{-\frac{1}{2}}\alpha_l^{-\frac{1}{2}}e^{-\pi^2 f^2(\frac{1}{\alpha_{k_0}}+\frac{1}{\alpha_l})}\alpha_{k_0}^{-2}\\
=&\frac{\partial \mathcal{I}_{n,m,f}}{\partial \rho_{k_0}}(-\frac{\rho_{k_0}\alpha_{k_0}^{-1}}{2}+\pi^2 f^2\rho_{k_0}\alpha_{k_0}^{-2}).
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \mathcal{R}_{n,m,f}}{\partial \tau_{n_0,k_0}} \\
=&\mathbb{I}(n=n_0)\sum_{l=1}^{K}\rho_l\rho_{k_0}(-2\pi f\sin(2\pi f(\tau_{n_0,k_0}-\tau_{m,l})))\\
&\pi\alpha_l^{-\frac{1}{2}}\alpha_{k_0}^{-\frac{1}{2}}e^{-\pi^2f^2(\frac{1}{\alpha_l}+\frac{1}{\alpha_{k_0}})}\\
&+\mathbb{I}(m=n_0)\sum_{l=1}^{K}\rho_l\rho_{k_0}(2\pi\cdot\\
&f\sin(2\pi f(\tau_{n,l}-\tau_{n_0,k_0})))\cdot\pi\alpha_l^{-\frac{1}{2}}\alpha_{k_0}^{-\frac{1}{2}}e^{-\pi^2f^2(\frac{1}{\alpha_l}+\frac{1}{\alpha_{k_0}})}.
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \mathcal{I}_{n,m,f}}{\partial \tau_{n_0,k_0}} \\
=&\mathbb{I}(n=n_0)\sum_{l=1}^{K}\rho_l\rho_{k_0}(2\pi f\cos(2\pi f(\tau_{n_0,k_0}-\tau_{m,l})))\\
&\cdot\pi\alpha_l^{-\frac{1}{2}}\alpha_{k_0}^{-\frac{1}{2}}e^{-\pi^2f^2(\frac{1}{\alpha_l}+\frac{1}{\alpha_{k_0}})}\\
&+\mathbb{I}(m=n_0)\sum_{l=1}^{K}\rho_l\rho_{k_0}(-2\pi f\cos(2\pi f(\tau_{n,l}-\tau_{n_0,k_0})))\\
&\cdot\pi\alpha_l^{-\frac{1}{2}}\alpha_{k_0}^{-\frac{1}{2}}e^{-\pi^2f^2(\frac{1}{\alpha_l}+\frac{1}{\alpha_{k_0}})}.
\end{aligned}
$$

With the above result, we can numerically check the rank of Jacobian matrix $J(\xi)=\frac{\partial\mathcal{G}(\xi)}{\partial\xi}$.

Discussion of estimating the noise variance σ^2.

In the paper, we provide two ways to estimate the variance σ^2 of the noise ε in the model.

1. As it is mentioned in Section Section 3.3, if $N>1$, i.e, we can use sample variance to estimate the σ^2. That is: we have samples $y_{i,j}\ \forall i=1,\ldots,m, j=1,\ldots,N$. Let $\bar{y}_i=(N)^{-1}\sum_{j=1}^{N}y_{i,j}$ and $\hat{\sigma}^2=(mN)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{N}(y_{i,j}-\bar{y}_i)^2$.

2. If $N=1$, let's assume $\varepsilon_i\sim N(0,\sigma^2)$ and $\gamma=0$. Then we can apply Theorem III.2 to construct a consistent estimate of σ^2. Consider $\mathfrak{M}'\subset\mathfrak{M}$ and $\mathfrak{r}'=\dim(\mathfrak{M}')$,

$\mathfrak{r} = \dim(\mathfrak{M})$, let

$$\tilde{T}'_N = \min_{x \in \mathfrak{M}'} \|\hat{y} - x\|_2^2, \; \tilde{T}_N = \min_{x \in \mathfrak{M}} \|\hat{y} - x\|_2^2.$$

Then let,

$$\hat{\sigma}^2 = \frac{\tilde{T}'_N - \tilde{T}_N}{\mathfrak{r} - \mathfrak{r}'}. \tag{B.7}$$

According to Theorem III.2, we know that under the true model $T'_N - T_N$ follows central χ^2 distribution with $\mathfrak{r} - \mathfrak{r}'$ degrees-of-freedom asymptotically. Therefore $\hat{\sigma}^2$ is a consistent estimate of σ^2, i.e. $\hat{\sigma}^2 \to \sigma^2$ as $\mathfrak{r}' - \mathfrak{r} \to \infty$. More specifically, as mentioned in section III.C, we assume that our manifold can be decomposed to be a sum of smooth manifold and linear space. Therefore, for an $x_0 \in \mathfrak{M}' = \mathcal{M} + \mathcal{L}'$, we can construct a linear space $\mathcal{L}$, s.t $L' \subset L$. Then, let $\mathfrak{M} = \mathcal{M} + \mathcal{L}$. we can compute eq.(Equation B.7).

Below, we will show how to use this general strategy to construct the $\mathcal{L}$ in each application mentioned in the paper. The key idea is that we can always leave out some observations to construct the $\mathcal{L}$.

1. Matrix completion: Denote the set of observation indices as Ω_0 manifold: $\mathfrak{M}' = \mathcal{M}_r + \mathcal{L}'$, where $\mathcal{L}' = \{X \in \mathbb{R}^{n_1 \times n_2} : X_{i,j} = 0, \forall (i,j) \in \Omega_0\}$. To estimate the σ^2, we can leave out some observation, i.e. we form a smaller observation set $\Omega_1 \subset \Omega_0$. Then the new manifold is $\mathfrak{M} = \mathcal{M}_r + \mathcal{L}$, where $\mathcal{L} = \{X \in \mathbb{R}^{n_1 \times n_2} : X_{i,j} = 0, \forall (i,j) \in \Omega_1\}$. We can see that $\mathcal{L}' \subset \mathcal{L} \Rightarrow \mathfrak{M}' \subset \mathfrak{M}$. Therefore, according to

eq.(Equation B.7), we can estimate σ^2 as following:

$$\begin{aligned}\tilde{T}'_N &= \min_{X\in\mathcal{M}_r} \sum_{(i,j)\in\Omega_0} (\hat{Y}_{ij} - X_{ij})^2, \\ \tilde{T}_N &= \min_{X\in\mathcal{M}_r} \sum_{(i,j)\in\Omega_1} (\hat{Y}_{ij} - X_{ij})^2, \\ \hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{|\Omega_0| - |\Omega_1|}. \end{aligned} \tag{B.8}$$

2. Complex matrix completion: It is similar to real matrix completion. By leaving out some observations, we have a smaller set of observation indices $\Omega_1 \subset \Omega_0$, and

$$\mathcal{L}' = \{X \in \mathbb{C}^{n_1\times n_2}, X_{ij} = 0,\ \forall (i,j) \in \Omega_0\}$$
$$\mathcal{L} = \{X \in \mathbb{C}^{n_1\times n_2}, X_{ij} = 0,\ \forall (i,j) \in \Omega_1\}$$

Let $\tilde{T}'_N$ be the objective value of eq.(24) in the paper with respect to observation set Ω_0 and $\tilde{T}_N$ be the result with respect to observation set Ω_1. Then, we can estimate the σ^2:

$$\hat{\sigma}^2 = \frac{\tilde{T}'_N - \tilde{T}_N}{|\Omega_0| - |\Omega_1|}$$

3. Rank-r tensor completion: It is similar to matrix completion problem: Denote the manifold of rank-r tensors as $\mathcal{M}_r$, and there is an observation index Ω_0. By leaving out some observations, we have $\Omega_1 \subset \Omega_0$. Let's define,

$$\mathcal{L}' = \{X \in \mathbb{R}^{n_1\times n_2}, X_{ijk} = 0,\ \forall (i,j,k) \in \Omega_0\}$$
$$\mathcal{L} = \{X \in \mathbb{R}^{n_1\times n_2}, X_{ijk} = 0,\ \forall (i,j,k) \in \Omega_1\}$$

and

$$\mathfrak{M}' = \mathcal{M}_r + \mathcal{L}',\ \mathfrak{M} = \mathcal{M}_r + \mathcal{L}.$$

We can see $\mathfrak{M}' \subset \mathfrak{M}$. According to the Theorem III.2, we can construct the $\hat{\sigma}^2$ similar to eq.(Equation B.8),

$$\begin{aligned}
\tilde{T}'_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j,k) \in \Omega_0} (\hat{Y}_{ijk} - X_{ijk})^2, \\
\tilde{T}_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j,k) \in \Omega_1} (\hat{Y}_{ijk} - X_{ijk})^2, \\
\hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{|\Omega_0| - |\Omega_1|}.
\end{aligned}$$

4. Demixing: It can be viewed as a tensor completion problem in our setting. The difference between the demixing problem and rank-r tensor completion problem is the way of parameterizing. In the rank-r tensor completion problem, we parameterize the tensor with rank. In the demixing problem, we parameterize the tensor as the cross-correlation function of the frequency domain signals. However, in estimating σ^2, what matters is the $\mathcal{L}$ part, which is not related to the parameterization of the $\mathcal{M}$ part.

5. Neural networks: Suppose we have m observations, i.e. $y \in \mathbb{R}^m$. Then we say that our set of observation indices are all the indices i.e. $\Omega_0 = \{1, 2, \ldots, m\}$. Then $\mathcal{L}' = \{X \in \mathbb{R}^m : X_i = 0, \forall i \in \Omega_0\} = \{0\}$. By leaving out some observations, we have $\Omega_1 \subset \Omega_0$, $\mathcal{L} = \{X \in \mathbb{R}^m : X_i = 0, \forall i \in \Omega_0\} \supset \mathcal{L}'$, according to the eq.(27) in the paper, σ^2 is estimated as:

$$\begin{aligned}
\tilde{T}'_N &= \min_{U \in \mathbb{R}^{d \times r}} \sum_{i=1}^{m} (y_i - \mathbf{1}^\top q(U^\top x_i))^2, \\
\tilde{T}'_N &= \min_{U \in \mathbb{R}^{d \times r}} \sum_{i \in \Omega_1} (y_i - \mathbf{1}^\top q(U^\top x_i))^2, \\
\hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{m - |\Omega_1|}.
\end{aligned} \tag{B.9}$$

6. Matrix sensing: As mentioned in the paper, matrix sensing is a special case of one-

hidden-layer neural networks with quadratic activation function.

Below we also present two numerical examples to show the performance of the estimate of the sigma:

1. Matrix completion: Table Table B.1 shows a result of estimating σ^2 for each rank r. In this experiment, $n_1 = n_2 = 100$, true rank $r^* = 6$, $|\Omega_0| = 8000$, $\sigma = 10$, $N = 1$. In practise, we may not know the true rank, therefore, we compute the estimate of σ^2 for each rank r ranging from 1 to 8. σ^2 is estimated by $\hat{\sigma}^2$ in eq.(Equation B.8) with $|\Omega_1| = 2000$. When $r < r^*$, $\hat{\sigma}^2$ largely overestimates the σ^2 and decreases hugely as r increases because part of the signal is treated as noise. When $r > r^*$, $\hat{\sigma}^2$ become stable since it is over-fitting the noise. We can also see that when $r = r^*$, our $\hat{\sigma}^2$ is close to σ^2.

   Table B.1: Estimate of σ^2 in matrix completion with true rank $r^* = 6$.

rank	$\hat{\sigma}^2$	rank	$\hat{\sigma}^2$
1	34995.5	5	5050.63
2	26751.3	**6**	**97.7**
3	18719.6	7	96.6
4	11231.8	8	96.7

2. Matrix sensing (One-hidden-layer neural networks with quadratic activation). Table Table B.2 shows a result of estimating σ^2 for each rank r. In this experiment, $d = 50$, true rank $r^* = 3$ (the number of hidden nodes), $m = |\Omega_0| = 500$, $\sigma = 1$, $N = 1$. We compute the estimate of σ^2 for each rank r ranging from 1 to 4. σ^2 is estimated by $\hat{\sigma}^2$ in eq.(Equation B.9) with $|\Omega_1| = 400$. We can see that our estimator $\hat{\sigma}^2$ is close to the true σ^2 when $r = r^*$.

Table B.2: Estimate of σ^2 in matrix sensing ($r^* = 3$).

rank	$\hat{\sigma}^2$	rank	$\hat{\sigma}^2$
1	8952.8	4	1.04
2	1498.8	5	1.12
3	**1.12**	6	0.88

APPENDIX C

APPENDICES OF CHAPTER 5

Proof of Theorem Theorem 5.4.2

Proof.

(i) Since under H_0, $S_T^{(q,q)}(0)$ has the same the distribution as the univariate case.

$$\begin{aligned}
\mathrm{Var}_{H_0}(T^{-\frac{1}{2}}S_T(0)) =& T^{-1}Var(S_T(0)) \\
=& T^{-1}\Big(\frac{T}{2\beta} + \frac{4\mu T - 1}{4\beta^2} \\
& + \frac{e^{-2\beta T}}{4\beta^2} - \frac{3\mu}{2\beta^3} - \frac{\mu e^{-2\beta T}}{2\beta^3} + \frac{2\mu e^{-\beta T}}{\beta^3}\Big) \\
\to& \frac{1}{2\beta} + \frac{\mu}{\beta^2} \text{ as } T \to \infty
\end{aligned}$$

(ii) To prove the variance of $S_T^{(p,q)}$, we use the fact that $\mathrm{Var}_{H_0}[S_T^{(p,q)}(\mathbf{0})] = -\mathbb{E}_{H_0}[\frac{\partial S_T^{(p,q)}(\mathbf{0})}{\partial \alpha_{p,q}}]$

$$\mathbb{E}_{H_0}[-\frac{\partial S_T^{(p,q)}(\mathbf{0})}{\partial \alpha_{p,q}}] \tag{C.1}$$

$$=\mathbb{E}[\frac{1}{\mu_q^2}\sum_{k\in\mathcal{C}(q,T)}(\sum_{i\in\mathcal{C}(p,t_k)}e^{-\beta(t_k-t_i)})^2]$$

$$=\mathbb{E}[\frac{1}{\mu_q^2}\mathbb{E}[\sum_{k\in\mathcal{C}(q,T)}(\sum_{i\in\mathcal{C}(p,t_k)}e^{-\beta(t_k-t_i)})^2|N_q,N_p]] \tag{C.2}$$

$$=\mathbb{E}[\frac{N_q}{\mu_q^2}\mathbb{E}[(\sum_{i=1}^{N_p}Z_i(t_k))^2|N_p,t_k]] \tag{C.3}$$

$$=\mathbb{E}[\frac{N_q}{\mu_q^2}\mathbb{E}[\sum_{i=1}^{N_p}Z_i^2(t_k)+\sum_{i\neq j}^{N_p}Z_i(t)Z_j(t)|N_p,t_k]]$$

$$=\mathbb{E}\Big[\frac{N_q}{\mu_q^2}\Big(N_p\mathbb{E}[Z_i^2(t_k))|t_k]+N_p(N_p-1)\mathbb{E}_{i\neq j}[Z_i(t_k)Z_j(t_k)|t_k]\Big)\Big]$$

$$=\mathbb{E}\Big[\frac{N_q}{\mu_q^2}\Big(\frac{N_p}{2\beta T}(1-e^{-2\beta t_k})+\frac{N_p(N_p-1)}{\beta^2T^2}(1-e^{-\beta t_k})^2\Big)\Big] \tag{C.4}$$

$$=\frac{T}{\mu_q}(\frac{1}{2\beta}+\frac{\mu_q}{\beta^2})+o(T) \tag{C.5}$$

where N_q and N_p are the number of events in $[0,T]$ on nodes q and p respectively. In eq(item C.2), we use the fact that for Poissson process, the arrival times follow $i.i.d.$ uniform distribution when it is conditional on the number of arrivals. With this fact, in eq(item C.3), we define

$$Z_i(t)=\begin{cases}0 & \text{if } t_i\geq t,\\ e^{-\beta(t-t_i)} & \text{if } t_i<t.\end{cases}$$

Since $t_i \overset{i.i.d}{\sim} \mathrm{unif}[0,T]$, then

$$\mathbb{E}Z_i(t)=\frac{1}{T}\int_0^t e^{-\beta(t-u)}du=\frac{1}{\beta T}(1-e^{-\beta t})$$
$$\mathbb{E}Z_i^2(t)=\frac{1}{T}\int_0^t e^{-2\beta(t-u)}du=\frac{1}{2\beta T}(1-e^{-2\beta t}),$$

which proves the eq(item C.4). Since N_p and N_q follow Poisson distribution with mean $T\mu_p$ and $T\mu_q$, respectively. Therefore eq(item C.5) is prooved.

(iii) Follow the similar techniques in (ii), we can prove

$$\mathrm{Cov}_{H_0}[T^{-\frac{1}{2}} S_T^{(p,q)}(\mathbf{0}), T^{-\frac{1}{2}} S_T^{(p',q)}(\mathbf{0})] \to \frac{\mu_p \mu_{p'}}{\mu_q \beta^2}.$$

□

Proof of Theorem 5.4.3

Proof. Follow the definition in [57], let's define the kernel function,

$$g(\mathbf{s}_1, \mathbf{s}_2, t) = \mathbf{s}_2^\top A \mathbf{s}_1 e^{-\beta t},$$

where $\mathbf{s}_i \in \mathbb{R}^M$. Then, we can define the conditional intensity function:

$$\begin{aligned}\Lambda(\mathbf{s}, t) =& \mu(\mathbf{s}) + \int_0^t \int_X g(\mathbf{s}, \mathbf{u}, t-r) N(d\mathbf{u} \times dr) \\ =& \mu(\mathbf{s}) + \sum_{t_i < t} \mathbf{u}_i^\top A \mathbf{s} \cdot e^{-\beta(t-t_i)},\end{aligned}$$

where $\mathbf{u} = e_m$, if $u_i = m$, and e_m is the vector that mth entry is 1 and other entries are 0. Further, define a measure with delta function:

$$\begin{aligned}v(x) =& \sum_{i=1}^M \delta_{e_i}(x) \\ \delta_{e_i}(x) =& \begin{cases} 1 \text{ if } x = e_i \\ 0 \text{ o.w.} \end{cases}\end{aligned}$$

We can write the likelihood function as the following:

$$\ell_T(A) = \int_0^T \int_X \log \Lambda(\mathbf{s}, t; A) N(d\mathbf{s} \times dt) - \int_0^T \int_X \Lambda(\mathbf{s}, t; A) v(d\mathbf{s}) dt$$

We can easily check this define the same multivariate Hawkes process in Equation 5.2, Equation 5.3, Equation 5.4. Define the function Δ as:

$$\Delta_{(i,j),(p,q)} \triangleq \frac{\dot{\Lambda}_{i,j}\dot{\Lambda}_{p,q}}{\Lambda},$$

where $\dot{\Lambda}_{i,j}$ is the partial derivative of Λ with respect to $\alpha_{i,j}$. Therefore, by the result of [57, Equation 4.7], we have:

$$\frac{1}{T}\sum_{k=1}^{K} \frac{\Delta(\mathbf{u}_k, t_k)}{\Lambda(\mathbf{u}_k, t_k)} \to \mathcal{I}(A).$$

By direct computation, we can have the result of Equation 5.10 □

Proof of Theorem 5.4.4

Proof. According to the Theorem I in [61], let's define the "neighbor of dependence" for index j, $J(j) = \{(j-1), j, j+1\}$, with simple modification for $j = 1$ and $j = k$, for $m > w$, X_j and X_i are independent for $i \notin J(j)$. Therefore the dependence of elements not in the neighbor vanished, i.e. b_3 and b_3' equals to 0.

$$\begin{aligned}
b_1 &= \sum_{j=1}^{k} \sum_{i \in J(j)} \mathbb{P}(X_j = 1)\mathbb{P}(X_i = 1) \\
&\leq 3k\mathbb{P}(X_1 = 1)^2 \qquad \text{(C.6)} \\
b_2 &= \sum_{j=1}^{k} \sum_{i \in J(j)\setminus j} \mathbb{P}(X_j = 1, X_i = 1) \\
&\leq 2k\mathbb{P}(X_1 = 1, X_2 = 1) \\
&\leq 2k\mathbb{P}\{T_b \in (0, m-w)\}\mathbb{P}\{T_b \in (m+w, 2m)\} + \\
&\quad 2k\mathbb{P}\{T_b \in (m-w, m+w)\} \qquad \text{(C.7)}
\end{aligned}$$

With the inequality of the tail probability of normal distribution in [65],

$$\begin{aligned}
\mathbb{P}(X_1 = 1) &= \mathbb{P}(T_b \in (0, m]) \\
&= \mathbb{P}\Big\{ \max_{\substack{0<t\leq m,\\ 1\leq i\leq d}} |\Gamma^{(i)}_{t,w}| > b \Big\} \\
&\leq 2\mathbb{P}\Big\{ \max_{\substack{0<t\leq m,\\ 1\leq i\leq d}} \Gamma^{(i)}_{t,w} > b \Big\} \\
&\leq 2md\mathbb{P}(\Gamma^{(i)}_{t,w} > b) \\
&\leq \frac{2md}{b} e^{-\frac{b^2}{2}} \qquad \text{(C.8)}
\end{aligned}$$

With same computation, we can show $\mathbb{P}\{T_b \in (m-w, m+w)\} \leq 4wdb^{-1}e^{b^2/2}$. Therefore with the Theorem 1 in [61], we can show

$$\begin{aligned}
& |\mathbb{P}(T_b > xf(b)) - e^{-\mathbb{E}W}| \qquad \text{(C.9)} \\
= \; & |\mathbb{P}(W = 0) - e^{-\mathbb{E}W}| \qquad \text{(C.10)} \\
< \; & b_1 + b_2 \qquad \text{(C.11)} \\
\leq \; & \frac{12km^2d^2}{b^2e^{b^2}} + \frac{8km^2d^2}{b^2e^{b^2}} + \frac{4kwd}{be^{b^2/2}} \qquad \text{(C.12)} \\
= \; & \frac{12xmd^2}{be^{b^2/2}} + \frac{8xmd^2}{be^{b^2/2}} + \frac{4xwd}{m} \qquad \text{(C.13)}
\end{aligned}$$

□

REFERENCES

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics (FOCS)*, vol. 9, no. 6, pp. 717–772, 2009.

[2] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[3] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Info. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[4] M. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[5] M. Fazel, "Matrix rank minimization with applications," *Ph.D. thesis, Stanford University*, 2002.

[6] R. Zhang, Y. Xie, R. Yao, and F. Qiu, *Online detection of cascading change-points*, 2021. arXiv: 1911.05610 [stat.OT].

[7] J. Chen, S.-H. Kim, and Y. Xie, "S 3 t: A score statistic for spatiotemporal change point detection," *Sequential Analysis*, vol. 39, no. 4, pp. 563–592, 2020.

[8] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PloS one*, vol. 5, no. 9, e12948, 2010.

[9] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *arXiv:1411.8003*, 2014.

[10] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges lin- early for phase retrieval, matrix completion and blind deconvolution.," *arXiv:1711.10467*, 2017.

[11] E. Candés and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, pp. 717–772, 2009.

[12] E. J. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[13] B. Recht, "A simpler approach to matrix completion," *J. Machine Learning Research*, vol. 12, pp. 3414–3430, 2011.

[14] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[15] Y. Chen, "Incoherence-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2909–2923, 2014.

[16] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Coherence matrix completion," *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1881–1889, 2014.

[17] D. Pimentel-Alarcon, N. Boston, and R. D. Nowak, "A characterization of deterministic sampling patterns for low-rank matrix completion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 623–636, 2016.

[18] M. W. Browne, "Statistical inference in factor analysis," in *Topics in Applied Multivariate Analysis*, D. M. Hawkins, Ed., Cambridge University Press, 1982.

[19] A. Shapiro, "Statistical inference of semidefinite programming," *Mathematical Programming*, pp. 1–21, 2018, First Online.

[20] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Practical sketching algorithms for low-rank matrix approximation," *SIAM J. Matrix Anal. Appl.*, vol. 38, no. 4, pp. 1454–1485, Dec. 2017.

[21] G. Pataki, "On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues," *Mathematics of Operations Research*, vol. 23, pp. 339–358, 1998.

[22] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, pp. 805–849, 2012.

[23] A. Shapiro, "Asymptotic theory of overparameterized structural models," *Journal of the American Statistical Association*, vol. 81, pp. 142–149, 1986.

[24] T. W. Anderson and H. Rubin, "Statistical inference in factor analysis," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed., Univ. of California Press, 1956, pp. 111–150.

[25] E. Wilson and J. Worcester, "The resolution of six tests into three general factors," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 25, pp. 73–77, 1939.

[26] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3d. Cambridge: The MIT Press, 2009.

[27] D. McManus, "Who invented local power analysis?" *Econometric Theory*, vol. 7, pp. 265–268, 1991.

[28] J. Steiger, A. Shapiro, and M. Browne, "On the multivariate asymptotic distribution of sequential chi-square statistics," *Psychometrika*, vol. 50, pp. 253–254, 1985.

[29] S. R. Becker, E. J. Candès, and M. C. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical programming computation*, vol. 3, no. 3, p. 165, 2011.

[30] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of machine learning research*, vol. 11, no. Aug, pp. 2287–2322, 2010.

[31] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[32] A. Shapiro, Y. Xie, and R. Zhang, "Matrix completion with deterministic pattern - a geometric perspecve," *IEEE Transactions on Signal Processing*, vol. 67, pp. 1088–1103, 2019.

[33] S. Sternberg, *Lectures on differential geometry*. Englewood Cliff: Prentice Hall, Inc., 1964.

[34] A. Sards, "The measure of critical values of differential maps," *Bull. Amer. Math. Soc*, vol. 48, pp. 883–890, 1942.

[35] H. Federer, *Geometric Measure Theory*. New York: Springer Verlag, 1969.

[36] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*, second, ser. MOS-SIAM Series on Optimization. SIAM, 2014.

[37] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix recovery," *arXiv preprint arXiv:1712.09203*, 2017.

[38] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[39] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoveni, "Effective criteria for specific identifiability of tensors and forms," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, pp. 656–681, 2017.

[40] I. Domanov and L. D. Lathauwer, "Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, pp. 1567–1589, 2015.

[41] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4497–4520, Jul. 2017.

[42] R. Snieder and K. Wapenaar, "Imaging with ambient noise," *Physics Today*, vol. 63, no. 9, pp. 44–49, 2010.

[43] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex demixing, with applications," *Foundations of Computational Mathematics*, vol. 14, no. 3, pp. 503–567, 2014.

[44] L. Xie, Y. Xie, S.-M. Wu, F.-C. Lin, and W. Song, "Communication efficient signal detection for distributed ambient noise imaging," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2018, pp. 1779–1783.

[45] M. Gomez-Rodriguez, E. D. Balduzzi, M. DE, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *International Conference on Machine Learning*, 2011.

[46] P. D. Hines, I. Dobson, E. Cotilla-Sanchez, and M. Eppstein, "Dual graph and random chemistry methods for cascading failure analysis," in *2013 46th Hawaii International Conference on System Sciences*, IEEE, 2013, pp. 2141–2150.

[47] J. Qi, K. Sun, and S. Mei, "An interaction model for simulation and mitigation of cascading failures," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 804–819, 2014.

[48] G. Rigaill, "A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points.," *Journal de la Société Française de Statistique*, vol. 156, no. 4, pp. 180–205, 2015.

[49] O. H. M. Padilla, Y. Yu, D. Wang, and A. Rinaldo, "Optimal nonparametric change point detection and localization," *arXiv preprint arXiv:1905.10019*, 2019.

[50] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection1," *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[51] G. Lorden *et al.*, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[52] T. L. Lai, "Sequential analysis: Some classical problems and new challenges," *Statistica Sinica*, pp. 303–351, 2001.

[53] S. Zou, V. V. Veeravalli, J. Li, and D. Towsley, "Quickest detection of dynamic events in networks," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2280–2295, 2020.

[54] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2010.

[55] Z. I. Botev, M. Mandjes, and A. Ridder, "Tail distribution of the maximum of correlated gaussian random variables," in *2015 Winter Simulation Conference (WSC)*, IEEE, 2015, pp. 633–642.

[56] S. Li, Y. Xie, M. Farajtabar, A. Verma, and L. Song, "Detecting changes in dynamic events over networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 346–359, 2017.

[57] S. L. Rathbun, "Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes," *Journal of Statistical Planning and Inference*, vol. 51, no. 1, pp. 55–74, 1996.

[58] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[59] X. He, Y. Xie, S.-M. Wu, and F.-C. Lin, "Sequential graph scanning statistic for change-point detection," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2018, pp. 1317–1321.

[60] B. Yakir, "Multi-channel change-point detection statistic with applications in dna copy-number variation and sequential monitoring," in *Proceedings of Second International Workshop in Sequential Methodologies*, 2009, pp. 15–17.

[61] R. Arratia, L. Goldstein, L. Gordon, *et al.*, "Two moments suffice for poisson approximations: The chen-stein method," *The Annals of Probability*, vol. 17, no. 1, pp. 9–25, 1989.

[62] P. Embrechts, T. Liniger, and L. Lin, "Multivariate hawkes processes: An application to financial data," *Journal of Applied Probability*, vol. 48, no. A, pp. 367–378, 2011.

[63] A. Shapiro, "Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints," *Biometrika*, vol. 72, pp. 133–144, 1985.

[64] F. M. Fisher, *The identification problem in econometrics*. New York: McGraw-Hill Company, 1966.

[65] W. Feller, "An introduction to probability theory and its applications," *1957*,