Goodness-of-Fit Tests on Manifolds

Alexander Shapiro, Yao Xie, and Rui Zhang

***Abstract*— We develop a general theory for the goodness-of-fit test to 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 noncentral) χ^2 distribution. The parameters of the χ^2 distribution are related to the model order and dimension of the problem. We further present a method to select the model orders sequentially. We demonstrate the broad application of the general theory in machine learning and signal processing, including determining the rank of low-rank (possibly complex-valued) matrices and tensors from noisy, partial, or indirect observations, determining the number of sources in signal demixing, and potential applications in determining the number of hidden nodes in neural networks.**

***Index Terms*— Goodness-of-fit test, manifolds, nested model selection, sequential test.**

I. Introduction

TESTING for goodness-of-fit of a model is a fundamental problem in statistics and signal processing (see, e.g., a survey in [1]). The goal is to describe how well the model fits a set of observations. The model can be represented by a pre-specified distribution, or structured parametric models (such as time series or linear regression models). Commonly seen goodness-of-fit tests include the chi-square and Kolmogorov-Smirnov tests (see, e.g., [2]). The goodness-of-fit test is often used for model diagnosis to determine the appropriate parsimonious models, for instance, selecting the order and type of time series models [3]. For linear regression, a related problem is variable selection [4], which determines a subset of variables that lead to the best overall fit to the data.

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 this paper, 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.

It is worthwhile pointing out that the model goodness-of-fit test here is not the same as the widely known model order selection based on the celebrate AIC and BIC rules, etc.; the related field of is extensive (see, e.g., a recent survey in [1]). The criterion for model order selection therein is the "prediction" or "generalization" error. In contrast, the goodness-of-fit we consider here is to describe how well a model fits a set of observations (thus, we consider "residual" errors). One potential issue with the classic model order selection based on AIC/BIC is that for certain situations, the expected prediction/generalization error may not be easily derived (for linear regression, there are explicit expressions). Such situations happen, for instance, when observations are noisy samples on a manifold. In such cases, the classic AIC and BIC rule may not be easy to carry through and may require significant numerical simulation to estimate the prediction errors. One benefit of the proposed approach is that the distribution of the residual is explicitly characterized. Thus, we can use it conveniently for selecting model orders through a sequential test procedure.

The proposed framework differs from other goodness-of-fit tests, such as the classic Kolmogorov-Smirnov test, which determines whether the empirical distribution is close a "nominal" or "target" distribution, the non-parametric approach based on the Maximum Mean Discrepancy (MMD) divergence [5]–[7], and the Bayesian approach [8]. Our proposed

Manuscript received September 11, 2019; revised September 26, 2020; accepted January 4, 2021. Date of publication January 11, 2021; date of current version March 18, 2021. The work of Alexander Shapiro was supported in part by the National Science Foundation (NSF) under Grant 1633196. The work of Yao Xie was supported in part by NSF under Grant CCF-1442635, Grant DMS-1938106, and Grant DMS-1830210; and in part by the NSF CAREER Award under Grant CCF-1650913. *(Corresponding author: Yao Xie.)*

The authors are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30032 USA (e-mail: ashapiro@isye.gatech.edu; yao.xie@isye.gatech.edu; ruizhang_ray@gatech.edu).

Communicated by S. Boucheron, Associate Editor for Statistical Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3050469.

Digital Object Identifier 10.1109/TIT.2021.3050469

framework also differs from the work on testing the manifold hypothesis [9], i.e., determining whether data lie near a low-dimensional manifold; [9] uses a "worst-case" analysis without assuming prior information about data generation mechanism.

Part of our work is related to low-rank matrix completion from partial and noisy observations. There has been much work done in this field, with notable contributions of [10]–[15]. There are mainly two categories of algorithms, including convex relaxation based on nuclear norm minimization, and non-convex optimization based on alternating minimization. In nuclear minimization (see, e.g., [11]), the rank selection is not explicitly addressed, possibly due to that the focus is on the recovery of the matrix itself. It is not clear how true rank will be recovered using the nuclear norm minimization approach. However, it is known that nuclear norm minimization may be asymptotically biased (see, e.g., [16]). For non-convex optimization-based matrix completion, such as alternating minimization [17], one has to pre-determine the rank of the matrix, and it is typically done empirically by heuristic methods [18].

Our proposed framework also differs from the work on testing the manifold hypothesis in [9]. The approach in [9] is nonasymptotic and, in a sense, nonparametric. It is assumed there that the data is generated from a "true" but unknown distribution. The algorithmic question addressed in [9] is, given a sample of size N, whether it is possible to verify with a high probability existence of a manifold, from a family of d-dimensional $\mathcal{C}^2$-submanifolds, which fits the data with a prescribed accuracy measured in terms of an average squared distance. Unlike our approach, no parametric model is assumed, while a "worst-case" analysis is applied in [9]. On the other hand we consider a nested family of parametrically defined manifolds.

The rest of the paper is organized as follows. Section II presents the background knowledge. Section III contains the main results: the test statistics for model selection on manifolds. Section IV gives several examples to demonstrate the use the general theory in specific settings. Section V presents numerical experiments. Finally, section VI concludes the paper 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 delegated to the Appendix.

II. 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 III.

Consider a nonempty set $\Theta \subseteq \mathbb{R}^d$ and a mapping $G : \Theta \to \mathbb{R}^m$. We assume throughout the paper 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:

$$\min_{\theta\in\Theta} \|\hat{y} - G(\theta)\|_2^2. \tag{1}$$

Define the image of the mapping G,

$$\mathfrak{M} := \{G(\theta) : \theta \in \Theta\}. \tag{2}$$

Then problem (1) can be written as

$$\min_{x\in\mathfrak{M}} \|\hat{y} - x\|_2^2. \tag{3}$$

That is, in problem (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

$$J(\theta) := \partial G(\theta)/\partial\theta, \tag{4}$$

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 II.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 II.1 (Characteristic Rank): We refer to the maximal rank of the Jacobian matrix,

$$\mathfrak{r} := \max_{\theta\in\Theta}\{\mathrm{rank}(J(\theta))\}, \tag{5}$$

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

The following Proposition II.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 [19]; we give its proof in the Appendix.

Proposition II.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., [20]) 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 $\operatorname{rank}(J(\theta_0)) = \mathfrak{r}$ almost surely (with probability one).

Remark II.2: Assuming that the mapping $G(\cdot)$ is $\mathcal{C}^\infty$ smooth, we have by Sard's theorem [21] 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 II.2 (Regularity [19]): 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 II.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., [22, 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 $\mathcal{T}_{\mathfrak{M}}(x_0) = \operatorname{lin}(J(\theta_0))$ 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.,

$$\mathcal{T}_{\mathfrak{M}}(x_0) = \left\{ dG(\theta_0)h : h \in \mathbb{R}^d \right\}. \tag{6}$$

III. Test Statistics on Manifold

We view now the mapping $G(\theta)$ as a considered model of the parameter vector $\theta \in \Theta$, and problem (1) as the least squares estimation (LSE) procedure with $\hat{y}$ being a given data point. More specifically, we assume the following model

$$\hat{y} = x_0 + N^{-1/2}\gamma + \varepsilon, \tag{7}$$

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., [23]).

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

$$T_N := N\hat{\sigma}^{-2} \min_{x \in \mathfrak{M}} \|\hat{y} - x\|_2^2, \tag{8}$$

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

A. Test Statistic on Manifolds

We now consider the general case defined in (7). We will show that for the problem defined on smooth manifolds, similar results in the form of RSS for linear models hold.

Remark III.1: For any $\hat{y} \in \mathbb{R}^m$, the generalized least-square problem (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 (3), then x_k converges to x_0 (e.g., [24, Theorem 7.23]). Under the model (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 (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 (8), then it follows that

$$\hat{y} - \hat{x} \in [\mathcal{T}_{\mathfrak{M}}(\hat{x})]^{\perp}, \tag{9}$$

where $\mathcal{T}_{\mathfrak{M}}(\hat{x})$ is the respective tangent space (see (6)). The following result shows that for $\hat{y}$ sufficiently close to x_0, the necessary optimality condition (9) is also sufficient (cf., [16, Proposition III.4]).

Proposition III.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 (9), then $\hat{x}$ is the unique globally optimal solution of the least squares estimation problem (8).

Since the least-squares problem in (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 (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 (7) we have the following asymptotic results, which are counterparts of the properties when $\mathfrak{M}$ is a linear space (cf., [19]).

Theorem III.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 (8) has unique optimal solution $\hat{x}$,
(ii) The test statistic T_N in (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.

B. 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 (6))

$$\mathcal{T}_{\mathfrak{M}'}(x_0) = \{dG(\theta_0)h : h \in \mathcal{T}_{\Theta'}(\theta_0)\}. \tag{10}$$

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

$$T'_N := N\sigma^{-2} \min_{x \in \mathfrak{M}'} \|\hat{y} - x\|_2^2. \tag{11}$$

We have the following results (cf., [25]).

Theorem III.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.

C. 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 IV. Consider model defined by the following mapping

$$G(\theta) := \mathcal{G}(\xi) + \mathcal{A}(\zeta), \tag{12}$$

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.,

$$\rho := \max_{\xi \in \Xi} \operatorname{rank}(\mathcal{J}(\xi)), \tag{13}$$

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

Consider the corresponding least squares problem (3), the model (7) and the least squares test statistic T_N, defined in (8), for the mapping $G(\theta)$ of the form (12).

Remark III.2: Note that the optimal value of least squares problem (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 III.1 and III.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

$$dG(\theta)(h, z) = d\mathcal{G}(\xi)h + Az,\ h \in \mathbb{R}^d, z \in \mathbb{R}^k. \tag{14}$$

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

Definition III.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

$$\mathcal{T}_{\mathcal{M}}(x) \cap \mathcal{L} = \{0\}. \tag{15}$$

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

$$\mathfrak{r} = \rho + k. \tag{16}$$

For the matrix completion problem the well-posedness condition (at a point) was introduced in [16]. Note that condition (15) means that

$$\dim(\mathcal{T}_{\mathcal{M}}(x) + \mathcal{L}) = \dim(\mathcal{T}_{\mathcal{M}}(x)) + \dim(\mathcal{L}). \tag{17}$$

Of course, a necessary condition for (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 II.2).

Proposition III.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 (12), we have that the point $x_0 \in \mathfrak{M}$ can be represented as

$$x_0 = x^* + v_0 \text{ for some } x^* \in \mathcal{M}, v_0 \in \mathcal{L}. \tag{18}$$

Definition III.2: We say that the model is *identifiable* at x^* (at x_0) if the representation (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 [16, Theorem III.2].

Proposition III.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 (16) can be verified numerically; it is necessary and sufficient for the local identifiability in the generic sense of Proposition III.2. We argue that the well-posedness condition is a minimal property that should be verified for a considered model.

IV. 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 IV.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 II.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 IV-B, IV-C, and IV-F below.

A. 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 [16]; here, we generate a conclusion using the framework of our general theory in this paper.

Consider the noisy matrix completion problem (e.g., [10], [14], [12] 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 (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 (8), associated with the test statistic T_N, can be written as

$$\min_{X \in \mathcal{M}_r} \sum_{(i,j) \in \Omega} (\hat{Y}_{ij} - X_{ij})^2, \tag{19}$$

where $\hat{Y}_{ij}$, $(i,j) \in \Omega$, are observed values of the data matrix. Then the model (7) can be written as

$$\hat{Y}_{ij} = X^*_{ij} + N^{-1/2}\Gamma_{ij} + \varepsilon_{ij},\ (i,j) \in \Omega, \tag{20}$$

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 III.2 and (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

$$\mathfrak{r} \leq r(n_1 + n_2 - r) + n_1 n_2 - |\Omega|, \tag{21}$$

and that the model is well-posed if and only if the equality holds in (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 II, the above assumption (A) holds generically. By Theorem III.1 we have the following result as N tends to infinity (cf., [16]).

Proposition IV.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

$$\delta = \sigma^{-2} \min_{H \in \mathcal{T}_{\mathcal{M}_r(X^*)}} \sum_{(i,j) \in \Omega} (\Gamma_{ij} - H_{ij})^2. \tag{22}$$

Moreover, applying Proposition III.2, we can conclude the following under the assumption:

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

Proposition IV.2: Suppose that Assumption (B) holds. Then: (i) the equality holds in (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 (22), (iii) with probability tending to one, problem (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 III.2 we have the following.

Proposition IV.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 noncentrality 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 IV.2: An application of Theorem III.2 and Proposition IV.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 III.2 and Proposition IV.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 paper and more discussion is provided in the Appendix.

B. 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 $X \in \mathbb{C}^{n_1 \times n_2}$, $V \in \mathbb{C}^{n_1 \times r}$, $W \in \mathbb{C}^{n_2 \times r}$:

$$\min_{V,W} \|\mathrm{X\text{-}V}W^\top\|_2^2 \ \text{s.t.}\ X_{ij} = b_{ij},\ (i,j) \in \Omega. \tag{23}$$

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

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

where $\mathrm{i}^2 = -1$, $V_1 \in \mathbb{R}^{n_1 \times r}$, 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

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

Then

$$VW^\top = (V_1 W_1^\top - V_2\ W_2^\top) + \mathrm{i}(V_1\ W_2^\top + V_2\ W_1^\top).$$

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

$$G_1(\theta) = V_1 W_1^\top - V_2\ W_2^\top + U_1,$$

$$G_2(\theta) = V_1\ W_2^\top + V_2\ W_1^\top + U_2,$$

and $U_1 \in \mathcal{L}_1$ and $U_2 \in \mathcal{L}_1$. Hence we can write the problem (23) in the following form

$$\begin{aligned}
\min_{\theta} \quad & \|X_1 - G_1(\theta)\|_2^2 + \|X_2 - G_2(\theta)\|_2^2 \\
\text{s.t.} \quad & X_{1,ij} = b_{1,ij}, X_{2,ij} = b_{2,ij},\ (i,j) \in \Omega, \\
& X_{1,ij} = X_{2,ij} = 0,\ (i,j) \in \Omega^c.
\end{aligned} \tag{24}$$

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 Section IV-A by a factor of 2.

C. Low-Rank Matrix Sensing

Matrix sensing problems [26] 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 $\langle A, B \rangle := \mathrm{tr}(AB)$ 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.

$$\min_{U \in \mathbb{R}^{d \times r}} f(U) := \sum_{i=1}^{m} \left(y_i - \langle A_i, UU^\top \rangle\right)^2. \tag{25}$$

It is shown in [26] that (25) is the same problem as the problem of fitting one-layer neural networks with quadratic activation in (27), which we discuss next.

D. 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:

$$y_i = \mathbf{1}^\top q(U^{*\top} x_i) + \varepsilon_i, \tag{26}$$

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:

$$\min_{U\in\mathbb{R}^{d\times r}} f(U) := \sum_{i=1}^{m} \left(y_i - \mathbf{1}^\top q(U^\top x_i)\right)^2. \tag{27}$$

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 (27) becomes a least squares problem of the form (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 IV.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.

E. 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 IV-A. 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

$$\Xi := \left\{\xi \in \mathbb{R}^{n_1\times r} \times \mathbb{R}^{n_2\times r} \times \mathbb{R}^{n_3\times r} : \mathcal{G}(\xi) \in \mathcal{M}_r\right\}. \tag{28}$$

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 paper. 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 IV-A. 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 paper 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., [27, section 3.2]).

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

Definition IV.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 (28), has (Lebesgue) measure zero. It is known that for sufficiently small r, the identifiability holds in the generic sense (we refer to [28], [29], 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 IV.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 IV.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,

$$\mathfrak{r} = r(n_1 + n_2 + n_3 - 2). \tag{29}$$

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

Corollary IV.1: If the rank r decomposition is generically identifiable, then formula (29) for the characteristic rank follows.

F. Determining Number of Sources in Blind De-Mixing Problem

De-mixing problem (e.g., [30]) is a fundamental challenge in signal processing, which arises from applications such as ambient noise seismic imaging [31], 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 [32]. 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:

$$x_n(t) = \sum_{k=1}^{K} s_k(t - \tau_{n,k}), \quad n = 1, \ldots, N. \tag{30}$$

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 (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 convolution 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,

$$\begin{aligned} R_{n,m}(f) :=& \mathcal{F}\{r_{n,m}(t)\}(f) \\ =& \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{lk}(f) \cdot e^{2\pi \mathrm{i} f (\tau_{n,l} - \tau_{m,k})}. \end{aligned} \tag{31}$$

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 (30), we can write specifically

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

Now we can write $R_{n,m}(f)$ in (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

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

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 [33] 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

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

TABLE I

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%

TABLE II

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*

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

$$G(\theta) := (\mathcal{R}+M_1, \mathcal{I}+M_2). \tag{32}$$

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 (32) (see Remark IV.1) and the appendix for the derivation of the Jacobian matrix).

V. NUMERICAL EXPERIMENTS

A. Complex Matrix Completion

In this section we consider the complex matrix completion problem (23). To solve the related optimization problem, we use a generalize version the hard thresholding algorithm in [34]. 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

TABLE III

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%

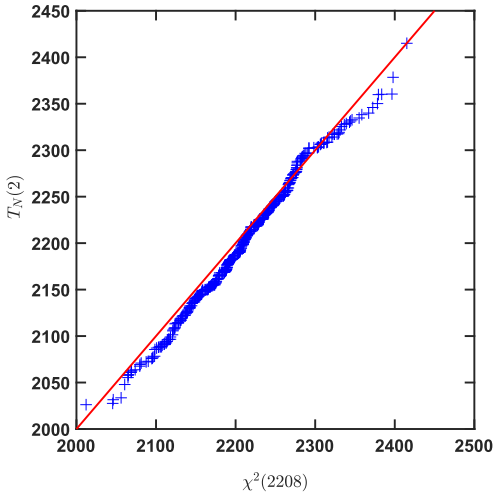

Fig. 1. QQ-plot of test statistics against χ^2 distribution.

all random instances, which is consistent with the results in Section IV-B.

To show the asymptotic distribution of test statistics (Theorem III.1), we generate a rank-2 true matrix $X^* \in \mathbb{C}^{100\times 100}$. The observed entries are contaminated with Gaussian noise:

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

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 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 IV-B). The results in Figure 1 show that the χ^2 distribution fits the test statistics reasonably well. Moreover, we show the result of detecting the rank in table I, 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 I, 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 IV

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 V

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

B. 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

$$\frac{\partial X_{ijl}}{\partial a^k_i} = b^k_j c^k_l,\ \frac{\partial X_{ijl}}{\partial b^k_j} = a^k_i c^k_l,\ \frac{\partial X_{ijl}}{\partial c^k_l} = a^k_i b^k_j.$$

All the other entries in the Jacobian matrix are zero.

Table II shows the rank (evaluated numerically) of the Jacobian matrices for different (n_1, n_2, n_3, r) values. We note

TABLE VI

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%

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 IV.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 II shows some cases in which ranks of the Jacobian matrices are less than $\min\{n_1n_2n_3, r(n_1+n_2+n_3-2)\}$.

C. Determining the Number of Signals in Blind De-Mixing

Consider the ambient noise imaging in a distributed sensor network setting (described in Section IV-F), 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: $\alpha_k \sim \text{Unif}[10,11]$, $\rho_k \sim \text{Unif}[10,11]$, $\tau_{n,k} \sim \text{Unif}[-2.5,2.5]$, $\forall n = 1,\ldots N$ and $k = 1,\ldots,K^*$.

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 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 III 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.

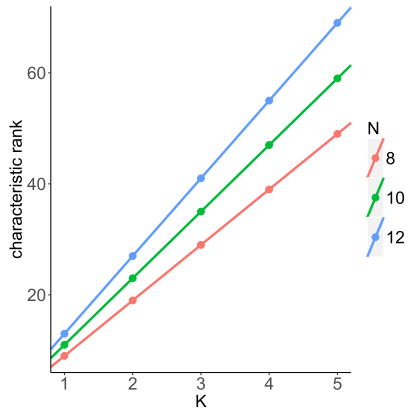

Fig. 2. The characteristic rank of the problem in Section IV-F: K is the number of sources, N is the number of sensors, the points are the rank of the Jacobian matrix of the mapping, and the line is $2K+NK-1$.

D. 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 (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 function to be quadratic activation and sigmoid activation, respectively. Table IV and Table V 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 VI 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 III. 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.

VI. Conclusion

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 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. Providing explicit answers (such as exact values of characteristic ranks) are too complex and beyond the scope of this paper, which we leave for future work.

VII. Appendix

Proof of Proposition II.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., [35]). 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 III.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 (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),

$$\nabla_x L(x, \lambda) = 0,\ \phi(x) = 0. \tag{33}$$

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 (33) has unique solution. Moreover by Remark III.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 equations (33), then by the uniqueness property $\hat{x}$ should coincide with the corresponding optimal solution. This completes the proof.

Proof of Theorem III.1: Since $\hat{y}$ converges in probability to x_0, the assertion (i) follows from Proposition III.1. Also any minimizer $\hat{x}$ in the right hand side of (8) converges in probability to x_0 (see Remark III.1). Therefore we can perform the asymptotic analysis in a neighborhood of x_0. As in the above proof of Proposition III.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 equations (33) in a sufficiently small neighborhood of $(x_0, 0)$. By the Implicit Function Theorem we have that

$$\begin{bmatrix} \hat{x} - x_0 \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} I_m & \nabla\phi(x_0) \\ \nabla\phi(x_0)^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{y} - x_0 \\ 0 \end{bmatrix} + o(\|\hat{y} - x_0\|). \tag{34}$$

Also it follows by (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

$$\hat{x} - x_0 = P(\hat{y} - x_0) + o_p(N^{-1/2}), \tag{35}$$

where

$$P = I_m - \nabla\phi(x_0)\left(\nabla\phi(x_0)^\top \nabla\phi(x_0)\right)^{-1} \nabla\phi(x_0)^\top. \tag{36}$$

Note that $\mathcal{T}_{\mathfrak{M}}(x_0) = \{v : \nabla\phi(x_0)^\top v = 0\}$. Therefore matrix P in (36) is the orthogonal projection matrix onto the tangent space $\mathcal{T}_{\mathfrak{M}}(x_0)$. Slutsky's theorem together with (35) 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 (35),

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

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

$$\|\hat{y} - \hat{x}\|_2^2 = \|(I_m - P)(\hat{y} - x_0)\|_2^2 + o_p(N^{-1}). \tag{37}$$

It follows by Slutsky's theorem that the N times right hand side of (37) 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 III.2 can be proved in a similar way by showing that asymptotically this is equivalent to the linear case.

Proof of Proposition III.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 (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 II.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 (14) that (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 IV.4: Let ρ be the characteristic rank of mapping

$$\mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r} \ni (A, B, C) \mapsto A \otimes B \otimes C. \quad (38)$$

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 (38) at (A, B, C) is ρ. For $X = A \otimes B \otimes C$ consider the set

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

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 (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 IV-F.

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})}. \\
\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})}. \\
\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}). \\
\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}$$

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

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

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 III, 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,

$$\hat{\sigma}^2 = \frac{\tilde{T}'_N - \tilde{T}_N}{\mathfrak{r} - \mathfrak{r}'}. \tag{39}$$

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.(39).

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.(39), we can estimate σ^2 as following:

$$\begin{aligned}
\tilde{T}'_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j) \in \Omega_0} (\hat{Y}_{ij} - X_{ij})^2, \\
\tilde{T}_N &= \min_{X \in \mathcal{M}_r} \sum_{(i,j) \in \Omega_1} (\hat{Y}_{ij} - X_{ij})^2, \\
\hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{|\Omega_0| - |\Omega_1|}.
\end{aligned} \tag{40}$$

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

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

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,

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

and

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

TABLE VII

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

We can see $\mathfrak{M}' \subset \mathfrak{M}$. According to the Theorem III.2, we can construct the $\hat{\sigma}^2$ similar to eq.(40),

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

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:

$$\begin{aligned}
\tilde{T}'_N &= \min_{U \in \mathbb{R}^{d \times r}} \sum_{i=1}^{m} (y_i - \mathbf{1}^\top q(U^\top x_i))^2, \\
\tilde{T}'_N &= \min_{U \in \mathbb{R}^{d \times r}} \sum_{i \in \Omega_1} (y_i - \mathbf{1}^\top q(U^\top x_i))^2, \\
\hat{\sigma}^2 &= \frac{\tilde{T}'_N - \tilde{T}_N}{m - |\Omega_1|}.
\end{aligned} \tag{41}$$

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 VII 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.(40) 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

TABLE VIII
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

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.

2) Matrix sensing (One-hidden-layer neural networks with quadratic activation). Table VIII 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.(41) with $|\Omega_1| = 400$. We can see that our estimator $\hat{\sigma}^2$ is close to the true σ^2 when $r = r^*$.

REFERENCES

[1] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, Nov. 2018.

[2] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*. Cham, Switzerland: Springer, 2005.

[3] P. Brockwell and R. Davis, *Introduction to Time Series and Forecasting*. Cham, Switzerland: Springer, 2010.

[4] J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, Mar. 2005.

[5] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *Proc. JMLR, Workshop Conf. Process.*, 2016, pp. 1–10.

[6] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 276–284.

[7] W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton, "A linear-time kernel goodness-of-fit test," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 262–271.

[8] I. Verdinelli and L. Wasserman, "Bayesian goodness-of-fit testing using infinite-dimensional exponential families," *Ann. Statist.*, vol. 26, no. 4, pp. 1215–1241, Aug. 1998.

[9] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *J. Amer. Math. Soc.*, vol. 29, no. 4, pp. 983–1049, Feb. 2016.

[10] 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, May 2010.

[11] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[12] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Jan. 2010.

[13] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, Feb. 2014.

[14] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 608–622, Jun. 2016.

[15] D. L. Pimentel-Alarcon, N. Boston, and R. D. Nowak, "A characterization of deterministic sampling patterns for low-rank matrix completion," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 623–636, Jun. 2016.

[16] A. Shapiro, Y. Xie, and R. Zhang, "Matrix completion with deterministic pattern—A geometric perspecve," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 1088–1103, Dec. 2018.

[17] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares," *J. Mach. Learn. Res.*, vol. 16, pp. 3367–3402, Dec. 2015.

[18] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Program. Comput.*, vol. 5, no. 2, pp. 201–226, Jun. 2013.

[19] A. Shapiro, "Asymptotic theory of overparameterized structural models," *J. Amer. Stat. Assoc.*, vol. 81, no. 393, pp. 142–149, Mar. 1986.

[20] S. Sternberg, *Lectures on Differential Geometry*. Englewood Cliff, NJ, USA: Prentice-Hall, 1964.

[21] A. Sards, "The measure of critical values of differential maps," *Bull. Amer. Math. Soc*, vol. 48, pp. 883–890, Dec. 1942.

[22] H. Federer, *Geometric Measure Theory*. New York, NY, USA: Springer Verlag, 1969.

[23] D. A. Mcmanus, "Who invented local power analysis?" *Econ. Theory*, vol. 7, no. 2, pp. 265–268, Jun. 1991.

[24] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming* (MOS-SIAM Series on Optimization), 2nd ed. Philadelphia, PA, USA: SIAM, 2014.

[25] J. Steiger, A. Shapiro, and M. Browne, "On the multivariate asymptotic distribution of sequential chi-square statistics," *Psychometrika*, vol. 50, pp. 253–254, May 1985.

[26] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," 2017, *arXiv:1712.09203*. [Online]. Available: http://arxiv.org/abs/1712.09203

[27] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[28] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven, "Effective criteria for specific identifiability of tensors and forms," *SIAM J. Matrix Anal. Appl.*, vol. 38, no. 2, pp. 656–681, Jan. 2017.

[29] I. Domanov and L. D. Lathauwer, "Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 4, pp. 1567–1589, Jan. 2015.

[30] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4497–4520, Jul. 2017.

[31] R. Snieder and K. Wapenaar, "Imaging with ambient noise," *Phys. Today*, vol. 63, no. 9, pp. 44–49, 2010.

[32] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex demixing, with applications," *Found. Comput. Math.*, vol. 14, no. 3, pp. 503–567, Jun. 2014.

[33] L. Xie, Y. Xie, S.-M. Wu, F.-C. Lin, and W. Song, "Communication efficient signal detection for distributed ambient noise imaging," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 1779–1783.

[34] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Jan. 2010.

[35] F. M. Fisher, *The Identification Problem in Econometrics*. New York, NY, USA: McGraw-Hill, 1966.

Alexander Shapiro is currently an A. Russell Chandler III Chair and Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. He has published more than 140 research articles in peer-reviewed journals and is a coauthor of several books. In 2013, he was a recipient of the Khachiyan Prize for Life-time Accomplishments in Optimization awarded by the INFORMS Optimization Society, and of the Dantzig Prize awarded by the Mathematical Optimization Society and Society for Industrial and Applied Mathematics in 2018. He served on editorial board for a number of professional journals. He was the Editor-in-Chief of the *Mathematical Programming* (Series A) journal. In 2020, he was elected to the National Academy of Engineering.

Yao Xie received the Ph.D. degree in electrical engineering (minor in mathematics) from Stanford University. She was a Research Scientist with Duke University. She is currently an Associate Professor and Harold R. and Mary Anne Nash Early Career Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, and the Associate Director of the Machine Learning Center. Her research areas include statistics (in particular sequential analysis and sequential change-point detection), machine learning, and signal processing. She received the National Science Foundation (NSF) CAREER Award in 2017. She is currently an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, *Sequential Analysis: Design Methods and Applications*, and *INFORMS Journal on Data Science*, and serves on the Editorial Board of *Journal of Machine Learning Research*.

Rui Zhang received the B.S. degree in statistics from Sun Yat-sen University in 2015, and the M.S. degree in statistics from the University of Michigan in 2017. He is currently pursuing the Ph.D. degree with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. His research interests include change-point detection, machine learning, and statistics.