Neurocomputing 142 (2014) 307–317

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Determining parameter identifiability from the optimization theory framework: A Kullback–Leibler divergence approach

Zhi-Yong Ran*, Bao-Gang Hu

NLPR&LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form
21 January 2014
Accepted 11 March 2014
Communicated by Ran He
Available online 28 May 2014

Keywords:
Identifiability
Optimization theory
Kullback–Leibler divergence
Hessian matrix
Jacobian matrix

ABSTRACT

This paper reports an extension of the existing investigations on determining identifiability of *statistical* parameter models. By making use of the *Kullback–Leibler divergence (KLD)* in information theory, we cast the identifiability problem into the *optimization theory* framework. This is the first work that studies the identifiability problem from the optimization theory perspective which leads to connections in many areas of scientific research, e.g., identifiability theory, information theory and optimization theory. Within this new framework, we derive identifiability criteria according to the types of models. First, by formulating the identifiability problem of unconstrained parameter models as an unconstrained optimization problem, we derive identifiability criteria by checking the rank of the Hessian matrix of KLD. The resulting theorems extend the existing approaches and work in arbitrary statistical models. Second, by formulating the identifiability problem of parameter-constrained models as a constrained optimization problem, we derive a novel criterion which has a clear algebraic and geometric interpretation. Further, we discuss the pros/cons of the new framework from both theoretical and application viewpoints. Several model examples from the literature are presented to examine their identifiability property.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Identifiability is an essential requirement in system modeling when the parameters to be estimated have a physically interpretable meaning [1–3]. An important part of system modeling involves the derivation of conditions under which a given model structure will be identifiable [1–5]. The identifiability property describes whether there is a *theoretical* possibility for the *unique* determination of model parameters from perfect model specification and noise-free input–output measurements or not [1,3,4]. This property is an important aspect to reflect an *interpretability* and *transparency* degree of the model and hence "*determining identifiability of the models should be addressed before any implementation of estimation*" [2,3]. Moreover, identifiability is closely related to the convergence of a class of estimators including the *maximum likelihood estimator (MLE)* [6,7]. Lack of identifiability gives no guarantee of convergence to the true value of parameter and therefore usually results in severe ill-posed estimation problems [2,6], which is a critical issue if decisions are to be taken on the basis of their numerical values [4]. Besides the ability to detect deficient models in advance, the analysis of identifiability can also bring practical benefits, such as insightful revealing of the relations among inputs, outputs and parameters, which can be very useful for model structure selection and design [2,8]. Therefore, once a model structure has been chosen, one should test identifiability so as to rule out prior unidentifiable models to avoid potential defects.

Most of previous work on system modeling has emphasized the special features of particular model structure, the identifiability issue is often neglected by many researchers, who start from the experimental data and then fit a model structure to the available data to estimate unknown parameter values. This tends to obscure the fact that the problem of identifiability is a general and fundamental one arising in many fields of scientific study. Generally, if a model has hierarchical structures [6,9], unobservable state variables [1,4], latent factors [10], nuisance parameters [11] or coupled submodels [12,13], the model may be unidentifiable. To summarize up, in the areas of machine learning, system identification and pattern recognition, the utility and importance of identifiability can be recognized in at least the following four aspects:

1. Statistical learning theory. Identifiability is a primary assumption in all classical statistical models [14]. However, such an assumption may be violated in a large variety of statistical learning machines. Theoretically, the concept of identifiability is closely related to *singular learning theory* [7]. A statistical

* Corresponding author. Tel.: +86 15201064550.
E-mail addresses: zyran@nlpr.ia.ac.cn (Z.-Y. Ran), hubg@nlpr.ia.ac.cn (B.-G. Hu).

http://dx.doi.org/10.1016/j.neucom.2014.03.055
0925-2312/© 2014 Elsevier B.V. All rights reserved.

learning machine is called *singular* if its Fisher information matrix (FIM) degenerates, or equivalently, if the model is not locally identifiable. In a singular learning machine, the standard statistical paradigm of the *Cramér-Rao bound (CRB)* does not hold [6], the MLE and the Bayesian posterior distribution are no longer subject to Gaussian distribution even in an asymptotical sense [6,7]. Therefore, it is imperative to check identifiability for statistical learning theory.

2. Knowledge-based modeling. Within the context of nonlinear system identification, a common practice is to build a "*black-box*" model in order to achieve accurate prediction or control. However, the fully nonlinear black-box model may be too generic for some situations where there are evidences to include a knowledge-based submodel in the complete model. The practical rule of "*do not estimate what you already know*" would require us to define an ad-hoc model structure if we know that the real system contains a prior known part. Thus, some or all parameters in the complete model have physically interpretable meaning [1,3,10,12,13]; and to identify the true values of such parameters is of practical importance because nonuniqueness of such parameters not only means nonunique description of the physical process but also results in completely erroneous or misleading results. One would not select a model if its parameters cannot be uniquely determined. Therefore, testing identifiability is an essential prerequisite in such models.
3. Model structure learning and model selection. In some application scenarios, one needs to implement joint model and parameter learning, thus requiring one to learn an identifiable model structure [10]; otherwise, the interpretability of the learned model will be severely limited and the model selection criteria such as AIC, BIC and MDL cannot be applied properly [6,8]. For instance, [10] considered a sparse and identifiable linear latent factor and linear Bayesian network model for parsimonious analysis of multivariate data, and showed that the identifiability is a necessary prerequisite for capturing the correlations between the latent factors.
4. Learning algorithm and learning dynamics. In unidentifiable parametric models, the trajectories of dynamics of learning generated by standard gradient descent algorithm are strongly affected by the nonidentifiability, causing plateaus or slow manifolds [6]. It has been shown that once parameters are attracted to unidentifiable points, the learning trajectory is very slow to move away from them. To overcome such slow convergence phenomenon, Amari [9] proposed a *natural gradient descent (NGD)* algorithm, showing that the NGD method works efficiently in such unidentifiable models.

Despite extensive literature exists on the identifiability problem and a number of identifiability criteria for various specific models, the identifiability issue has not been resolved completely. In this paper, we report an extension of the existing investigations on determining identifiability of *statistical* parameter models in two directions. First, in our previous studies, Yang et al. [2,12] considered the identifiability problem in *generalized-constraint neural network (GCNN)* models, and derived identifiability theorems for *Single-input Single-output (SISO)* and *Multiple-input Single-output (MISO)* models. However, their theorems cannot deal with *Multiple-input Multiple-output (MIMO)* models. In [15], Qu et al. studied a kind of GCNN model consisting of RBF neural network and a set of *linear priors* (linear constraints), but the identifiability issue has not been justified. Recently, Ran et al. [16] generalized the concept of GCNN model and proposed a more general *generalized-constraint (GC)* model. Further, the authors [16] derived some new results for MIMO parameter learning machines. For a detailed description of GCNN and GC models, one can see [2,12,15,16] and the references therein. Hence, this paper is an extension of [2,12,15,16] and we further expect to derive identifiability criteria for *parameter-constrained* models whose parameters are constrained by a set of nonlinear equality constraints. Second, based on *Kullback–Leibler divergence (KLD)* in information theory [17], we extend the conventional KLD equation method [18,19] and provide an *optimization theory* [20,21] treatment for identifiability analysis. To the best of our knowledge, this is the first work that studies the problem of identifiability from the optimization theory perspective. The resulting theorems are workable for a large variety of models wherein other methods fail (see Sections 3 and 5 for more details). The main contribution of this paper is given from the following two aspects:

1. From a theoretical viewpoint, we develop a novel perspective of processing identifiability problems based on optimization theory framework. Within the new framework, we formulate the identifiability problem of unconstrained and parameter-constrained models as unconstrained and constrained optimization problems, respectively. The benefit gained will be twofold. First, when information theory through KLD is the link, the interplay between identifiability theory and optimization theory (Fig. 1) is derived theoretically. Second, one is able to achieve a geometrically perceivable insight into the identifiability analysis.
2. From an application viewpoint, we derive several novel theorems for determining identifiability. Based on the theorems, one is able to deal with both unconstrained and parameter-constrained models. Compared with existing techniques, such as KLD equation method [18,19] and orthogonal complement method [22], the main advantage of the new results is that one is able to determine identifiability by calculating the rank of a numerical matrix, thus avoiding the usual bottleneck of seeking for the roots from a set of nonlinear equations. The benefit gained is that the new results lead to a reduction of complexity from NP-complete to $\mathcal{O}(k^3)$, where k is the dimensionality of parameter vector.

The remainder of this paper is organized as follows. Section 2 introduces the basic concepts and gives a concise overview of the literature. Section 3 provides some identifiability theorems for unconstrained parameter models. In Section 4, we provide some identifiability theorems for parameter-constrained models. Section 5 presents some examples to verify the validity of the proposed theorems. Section 6 concludes this paper with a brief summary.

2. Identifiability: basic concepts and existing methods

The identifiability analysis in statistical models is concerned with the possibility of drawing inferences from observed samples to an underlying theoretical distribution. Consider a statistical space $\{\mathbb{R}^n, \sigma(\mathbb{R}^n), P_{\boldsymbol{\theta}}\}$, where $\mathbb{R}^n$ is the sample space, $\sigma(\mathbb{R}^n)$ is the σ-algebra of $\mathbb{R}^n$ and $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^k\}$ is the parametric distribution family in $\sigma(\mathbb{R}^n)$. The identifiability problem is defined in terms of the mapping $\boldsymbol{\theta} \to P_{\boldsymbol{\theta}}$ being one-to-one. Following [14,16,19,23], we give the following definitions.

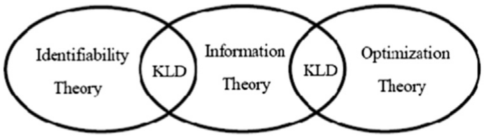

Fig. 1. Schematic diagram of the relationship of identifiability theory, information theory and optimization theory, from which information theory builds a bridge between identifiability theory and optimization theory.

Definition 1. A model $\{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \mathbb{R}^k\}$ is *globally identifiable* if $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2} \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2,\ \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^k$. A model is *locally identifiable* if for every $\boldsymbol{\theta} \in \mathbb{R}^k$, there is an open neighborhood $N(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ such that $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2} \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2,\ \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in N(\boldsymbol{\theta})$.

If a parameter point $\boldsymbol{\alpha} \in \mathbb{R}^k$ is of particular interest, for example, $\boldsymbol{\alpha}$ is assumed to be the real value of the model parameter, we give the following definition.

Definition 2. A parameter point $\boldsymbol{\alpha} \in \mathbb{R}^k$ is *globally identifiable* if $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\alpha}}, \boldsymbol{\theta} \in \mathbb{R}^k \Rightarrow \boldsymbol{\theta} = \boldsymbol{\alpha}$. A parameter point $\boldsymbol{\alpha} \in \mathbb{R}^k$ is said to be *locally identifiable* if there is an open neighborhood $N(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$ such that $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\alpha}},\ \boldsymbol{\theta} \in N(\boldsymbol{\alpha}) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\alpha}$.

If two distinct parameter points $\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2$ define the same model, we say $\boldsymbol{\theta}_1$ is *observationally equivalent* to $\boldsymbol{\theta}_2$ [23], and denote $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$ [14,16,19]. That is, $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2 \Leftrightarrow P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$. Note that the relation "$\sim$" is a proper *equivalent relation* (reflectivity, symmetry and transitivity) [14,16,19]. For $\boldsymbol{\theta} \in \mathbb{R}^k$, we denote the *equivalence class* of $\boldsymbol{\theta}$ by $[\boldsymbol{\theta}] = \{\boldsymbol{\theta}' \in \mathbb{R}^k : \boldsymbol{\theta}' \sim \boldsymbol{\theta}\}$.

Remark 1. From Definitions 1 and 2, one can see that identifiability is a *structural* or *intrinsic* property of the model. In other words, the presence or absence of identifiability is a feature of the model structure, and so, is *independent of the experiment data and the inferential procedure* [3,16,24].

Because of the theoretical and practical importance of identifiability analysis, a large amount of investigations have been devoted to this study, applying various methods and techniques. In [23], Rothenberg proved that the local identifiability of a statistical model $P_{\boldsymbol{\theta}}$ is equivalent to regularity of its FIM $\mathbf{F}(\boldsymbol{\theta}) = (F_{ij}(\boldsymbol{\theta}))_{k\times k}$,

$$F_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial \log\ p(\mathbf{x}, \boldsymbol{\theta})}{\theta_i}\frac{\partial \log\ p(\mathbf{x}, \boldsymbol{\theta})}{\theta_j}\right], \tag{1}$$

where $p(\mathbf{x}, \boldsymbol{\theta})$ is the probability density function (PDF) of $P_{\boldsymbol{\theta}}$, and $\mathbb{E}_{\boldsymbol{\theta}}$ is the expectation operation evaluated at $\boldsymbol{\theta}$. As a special case, [25] studied the connection between identifiability and information regularity in Gaussian family based on holomorphic functions. In [18], the author proposed a KLD equation method which needs to solve a set of nonlinear equations, making it hard to implement in most cases (see Theorem 1 in Section 3). In [14], Dasgupta et al. established an analytical method for constructing new parameters under which an unidentifiable model will be locally identifiable. For parameter-constrained models, the first local identifiability result was proposed in [23]. In [22], the author provided an identifiability criterion which needs to compute the orthogonal complement of a functional matrix, making it a hard task to perform. In [26], Yao et al. studied the regularity and identifiability of blind source separation (BSS) problem with constant modulus (CM) constraints on the sources. Unfortunately, it is very difficult to obtain a global result in more generic settings. In [23], Rothenberg established an FIM-based criterion to test global identifiability for exponential family. Outside the exponential family it does not seem possible to get a necessary and sufficient condition for global identifiability using only the FIM. In [19], the authors proved that global identifiability is a necessary condition for the existence of an unbiased estimator. In [27], Martin et al. extended this result to asymptotically unbiased estimators. They further proved that global identifiability is a necessary condition for the existence of a consistent estimator. The most obvious cause of nonidentifiability is *parameter redundancy*, in the sense that the model can be written in terms of a smaller set of parameters [28,14,2,16]. In [28], Catchpole et al. introduced the concept of parameter redundancy in the context of exponential family. They showed that whether or not a model is parameter redundant can be determined by checking the symbolic rank of a *derivative matrix (DM)*, but their DM-based method only works in the exponential family. In order to detect parameter redundancy in more general models, an exhaustive summary method was presented [29] which generalized the results in [28] to a wider spectrum of models. In [2], Yang et al. proposed a *derivative functional vector (DFV)* method to examine parameter redundancy for the GCNN models, but the DFV method is not applicable in time-variant models. In [16], Ran et al. proposed a regular summary method which can deal with time-variant models including a range of *ordinary differential equation (ODE)* dynamical models. However, their method cannot work in more complicit *partial differential equation (PDE)* models. Recently, Hu [30] analyzed redundancy of Bayesian classifiers whose parameters are given in a form of functionals, not functions. The author further proved that for *M*-class classification problem with reject option, the number of independent parameters in cost matrix is *M*. Although the identifiability problem has been extensively studied in the literature, identifiability issues have not been resolved fully. In Table 1, we list the commonly used methods for checking identifiability together with their associated parametric models.

3. Identifiability criteria for unconstrained parameter models

Essentially, nonidentifiability is the consequence of the lack of enough "information" to discriminate among admissible parameter values in the model. Hence, it is natural to test identifiability with the help of KLD, which is defined as [17]

$$KL(p,q) = \mathbb{E}_p\left(\log\ \frac{p(\mathbf{x})}{q(\mathbf{x})}\right) = \int p(\mathbf{x})\log\ \frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x}, \tag{2}$$

where $p(\mathbf{x})$ and $q(\mathbf{x})$ are two PDFs on $\mathbb{R}^n$. In information theory, the KLD is used to measure the dissimilarity between two PDFs $p(\mathbf{x})$ and $q(\mathbf{x})$ [17]. While in classical statistics, the KLD arises as an

Table 1
General methods for testing identifiability of statistical parameter models.

Parameter space	Framework	Model	Method
Unconstrained	Analytical	Gaussian	Holomorphic function method [25]
	Algebra	Linear model	Rank test method [31,19]
		Exponential family	Derivative matrix (DM) method [28]
		General distribution	Exhaustive summary method [29]
			Derivative functional vector (DFV) method [2]
	Statistics	General distribution	Fisher information matrix (FIM) method [23]
			Statistic method [19,27]
	Information theory	General distribution	KLD equation method [18,19]
			Regular summary method [16]
Constrained	Algebra	Linear model	Rank test method [31]
		+Linear constraints	
	Statistical	General distribution	Reparameterization method [23,32]
		+Nonlinear constraints	KLD equation method [18,19]
			Orthogonal complement method [22]

expected logarithm of the likelihood ratio, and is a measure of the inefficiency of assuming that the distribution is $p(\mathbf{x})$ while the true distribution is $q(\mathbf{x})$ [24]. Compared with the FIM, the KLD is claimed to be a more suitable measure for identifiability analysis, as the KLD is a function of two arguments which makes it can simultaneously deal with global and local identifiability problems [18,19], while FIM is the function of a single variable which makes it can only deal with local identifiability problem [23]. Therefore, the KLD will be a promising tool in identifiability analysis.

In this section, we focus our attention on identifiability issue for unconstrained parameter models whose admissible parameter space is $\mathbb{R}^k$. To proceed, a common existing criterion for testing identifiability is stated as follows [18,19].

Theorem 1. *In a statistical model $p(\mathbf{x},\boldsymbol{\theta})$, a parameter point $\boldsymbol{\alpha}\in\mathbb{R}^k$ is globally (locally) identifiable if and only if $\boldsymbol{\alpha}$ is the unique solution of the equation $KL(\boldsymbol{\alpha},\boldsymbol{\theta})=0$ in $\mathbb{R}^k$ (an open neighborhood of $\boldsymbol{\alpha}$), where*

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{\alpha}}\left(\log\frac{p(\mathbf{x},\boldsymbol{\alpha})}{p(\mathbf{x},\boldsymbol{\theta})}\right)=\int p(\mathbf{x},\boldsymbol{\alpha})\log\frac{p(\mathbf{x},\boldsymbol{\alpha})}{p(\mathbf{x},\boldsymbol{\theta})}\mathrm{d}\mathbf{x}. \tag{3}$$

The proof can be easily verified by the facts that $KL(\boldsymbol{\alpha},\boldsymbol{\theta})\geq 0$ for all $\boldsymbol{\theta}\in\mathbb{R}^k$, $KL(\boldsymbol{\alpha},\boldsymbol{\alpha})=0$, and $KL(\boldsymbol{\alpha},\boldsymbol{\theta})=0\Leftrightarrow p(\mathbf{x},\boldsymbol{\alpha})=p(\mathbf{x},\boldsymbol{\theta})$ for *almost everywhere (a.e.)* $\mathbf{x}\in\mathbb{R}^n$ [17]. However, for many PDFs it is not an easy task to determine all the solutions of the equation $KL(\boldsymbol{\alpha},\boldsymbol{\theta})=0$ in a direct way [19]. To give an example, consider the m-dimensional Gaussian family

$$p(\mathbf{x},\boldsymbol{\theta})=\frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_{\boldsymbol{\theta}}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\boldsymbol{\theta}})\right\}, \tag{4}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is the mean vector and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is the covariance matrix. The KLD can be calculated as [17]:

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta})=\tilde{KL}(\boldsymbol{\alpha},\boldsymbol{\theta})+\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\alpha}}) \tag{5}$$

with

$$\tilde{KL}(\boldsymbol{\alpha},\boldsymbol{\theta})=\frac{1}{2}\left\{\log\frac{|\Sigma_{\boldsymbol{\alpha}}|}{|\Sigma_{\boldsymbol{\theta}}|}+\mathrm{Trace}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}))\right\}. \tag{6}$$

It is easy to see that [17]

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta})=0\Leftrightarrow\boldsymbol{\mu}_{\boldsymbol{\theta}}=\boldsymbol{\mu}_{\boldsymbol{\alpha}},\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}}=\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}. \tag{7}$$

Checking the identifiability of $\boldsymbol{\alpha}$ requires us to solve a system of $m+m(m+1)/2$ nonlinear equations which makes the task an NP-complete problem [4]. Therefore, it is desirable to investigate some effective and efficient approaches to attack this problem. For this purpose, we cast the identifiability problem into the optimization theory framework [20,21].

Definition 3. [20]. A point $\boldsymbol{\alpha}\in\mathbb{R}^k$ is said to be a *local minimum point* of $f(\boldsymbol{\theta})$ if there is a neighbor $N(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$ such that $f(\boldsymbol{\theta})\geq f(\boldsymbol{\alpha})$ for all $\boldsymbol{\theta}\in N(\boldsymbol{\alpha})$. If $f(\boldsymbol{\theta})>f(\boldsymbol{\alpha})$ for all $\boldsymbol{\theta}\in N(\boldsymbol{\alpha}),\boldsymbol{\theta}\neq\boldsymbol{\alpha}$, then $\boldsymbol{\alpha}$ is said to be a *strict local minimum point*.

Definition 4. [20]. A point $\boldsymbol{\alpha}\in\mathbb{R}^k$ is said to be a *global minimum point* of $f(\boldsymbol{\theta})$ if $f(\boldsymbol{\theta})\geq f(\boldsymbol{\alpha})$ for all $\boldsymbol{\theta}\in\mathbb{R}^k$. If $f(\boldsymbol{\theta})>f(\boldsymbol{\alpha})$ for all $\boldsymbol{\theta}\in\mathbb{R}^k$, $\boldsymbol{\theta}\neq\boldsymbol{\alpha}$, then $\boldsymbol{\alpha}$ is said to be a *strict global minimum point*.

With the optimization theory, we can equivalently rewrite Theorem 1 as the following theorem.

Theorem 2. *In a statistical model $p(\mathbf{x},\boldsymbol{\theta})$, a parameter point $\boldsymbol{\alpha}\in\mathbb{R}^k$ is globally (locally) identifiable if and only if $\boldsymbol{\alpha}$ is the strict global (local) minimum point of the unconstrained optimization problem*

$$\mathrm{Minimize}\{KL(\boldsymbol{\alpha},\boldsymbol{\theta}),\boldsymbol{\theta}\in\mathbb{R}^k\}\ . \tag{8}$$

In Fig. 2, we visually illustrate the equivalence of Theorems 1 and 2 with several simple cases. As a matter of fact, the statement of Theorem 2 can be regarded as a dual interpretation of Theorem 1. Specifically, Theorem 1 formulates the identifiability problem as a nonlinear equation system problem, while Theorem 2 formulates the identifiability problem as an unconstrained optimization problem.

Therefore, with the help of the KLD, we transform the identifiability problem of unconstrained parameter models into an unconstrained optimization problem. We now present an identifiability criterion for unconstrained parameter models. *For mathematical simplicity, it should be noted that the interchanges of integral, limitation and derivative are permissible throughout the paper.*

Theorem 3. *Suppose that $p(\mathbf{x},\boldsymbol{\theta})$ is a statistical model, $\boldsymbol{\alpha}\in\mathbb{R}^k$, and that the Hessian matrix*

$$\mathbf{H}(\boldsymbol{\theta})=\left(\frac{\partial^2 KL(\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)_{k\times k} \tag{9}$$

of $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ has a constant rank in a neighbor $N(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$, then $\boldsymbol{\alpha}\in\mathbb{R}^k$ is locally identifiable if and only if $\mathbf{H}(\boldsymbol{\theta})$ is strictly positive definite at $\boldsymbol{\alpha}$.

Proof. For sufficiency. Since

$$\boldsymbol{\alpha}=\mathrm{argmin}\{KL(\boldsymbol{\alpha},\boldsymbol{\theta}),\boldsymbol{\theta}\in\mathbb{R}^k\}, \tag{10}$$

by the first order necessary condition of the local minimum point [20,21], the gradient vector $\nabla KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ of $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ vanishes at $\boldsymbol{\alpha}$. That is,

$$\nabla KL(\boldsymbol{\alpha},\boldsymbol{\alpha}):\ =\nabla KL(\boldsymbol{\alpha},\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\alpha}=0. \tag{11}$$

Apply *Taylor's formula* to $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$, we have

$$\begin{aligned}KL(\boldsymbol{\alpha},\boldsymbol{\theta})&=KL(\boldsymbol{\alpha},\boldsymbol{\alpha})+(\boldsymbol{\theta}-\boldsymbol{\alpha})^{\mathrm{T}}(\nabla KL(\boldsymbol{\alpha},\boldsymbol{\alpha}))\\&\quad+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\alpha})^{\mathrm{T}}\,\mathbf{H}(\boldsymbol{\alpha})(\boldsymbol{\theta}-\boldsymbol{\alpha})+o(\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|^2)\\&=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\alpha})^{\mathrm{T}}\,\mathbf{H}(\boldsymbol{\alpha})(\boldsymbol{\theta}-\boldsymbol{\alpha})+o(\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|^2)\end{aligned} \tag{12}$$

where $o(\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|^2)$ is the higher order infinitesimal of $\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|^2$. Since $\mathbf{H}(\boldsymbol{\alpha})$ is strictly positive definite, $KL(\boldsymbol{\alpha},\boldsymbol{\theta})>0$ for all $\boldsymbol{\theta}\in N(\boldsymbol{\alpha}),\boldsymbol{\theta}\neq\boldsymbol{\alpha}$. Hence, $\boldsymbol{\alpha}$ is a strict local minimum point of the optimization problem (8). By Theorem 2, $\boldsymbol{\alpha}$ is locally identifiable.

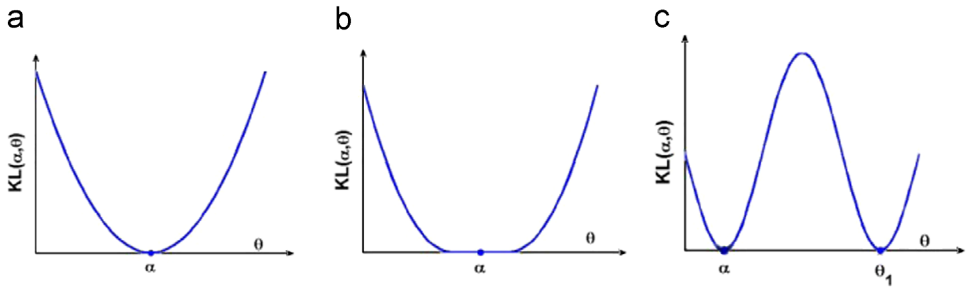

Fig. 2. Schematic illustration of the equivalence of Theorems 1 and 2. In (a), $\boldsymbol{\alpha}$ is globally identifiable. In (b), $\boldsymbol{\alpha}$ is unidentifiable. In (c), $\boldsymbol{\alpha}$ is locally identifiable, but is not globally identifiable.

For necessity. Since $\int p(\mathbf{x},\boldsymbol{\theta})\mathrm{d}\mathbf{x}=1,\ \forall\boldsymbol{\theta}\in\mathbb{R}^k$, by interchange of integral and derivative, we have

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{1}{p(\mathbf{x},\boldsymbol{\theta})}\frac{\partial^2 p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)&=\int\frac{\partial^2 p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\mathrm{d}\mathbf{x}\\ &=\frac{\partial^2\int p(\mathbf{x},\boldsymbol{\theta})\mathrm{d}\mathbf{x}}{\partial\theta_i\partial\theta_j}=\frac{\partial^2 1}{\partial\theta_i\partial\theta_j}=0.\end{aligned}\tag{13}$$

By simple calculation, we obtain

$$\frac{\partial^2\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}=-\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_j}+\frac{1}{p(\mathbf{x},\boldsymbol{\theta})}\frac{\partial^2 p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}.\tag{14}$$

Hence,

$$\mathbb{E}_\theta\left(\frac{\partial^2\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)=-\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_j}\right).\tag{15}$$

By interchange of integral, limitation and derivative, we have

$$\begin{aligned}\mathbb{E}_\alpha\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)&=\int p(\mathbf{x},\boldsymbol{\alpha})\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\mathrm{d}\mathbf{x}\\ &=\int\left(\frac{\partial p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\mathrm{d}\mathbf{x}\\ &=\frac{\partial\int p(\mathbf{x},\boldsymbol{\theta})\mathrm{d}\mathbf{x}}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\\ &=\frac{\partial 1}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}=0\end{aligned}\tag{16}$$

and

$$\begin{aligned}\mathbf{H}(\boldsymbol{\alpha})&=\frac{\partial^2 KL(\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\\ &=-\frac{\partial^2\int p(\mathbf{x},\boldsymbol{\alpha})\log\ p(\mathbf{x},\boldsymbol{\theta})\mathrm{d}\mathbf{x}}{\partial\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\\ &=-\left(\int\frac{\partial^2(p(\mathbf{x},\boldsymbol{\alpha})\log\ p(\mathbf{x},\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^2}\mathrm{d}\mathbf{x}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\\ &=-\int p(\mathbf{x},\boldsymbol{\alpha})\left(\frac{\partial^2\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}\mathrm{d}\mathbf{x}\\ &=-\mathbb{E}_\alpha\left(\frac{\partial^2\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right)\\ &=\mathbb{E}_\alpha\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right).\end{aligned}\tag{17}$$

Hence,

$$\mathbf{H}(\boldsymbol{\alpha})=\mathrm{Cov}_{\boldsymbol{\alpha}}\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right),\tag{18}$$

where $\mathrm{Cov}_{\boldsymbol{\alpha}}$ is the covariance operation evaluated at $\boldsymbol{\alpha}$.

Suppose $\mathbf{H}(\boldsymbol{\alpha})$ is not strictly positive definite, there is a nontrivial vector $\mathbf{v}(\boldsymbol{\theta})\neq 0,\ \ \boldsymbol{\theta}\in N(\boldsymbol{\alpha})$, such that

$$\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})\mathbf{v}(\boldsymbol{\theta})=0,\boldsymbol{\theta}\in N(\boldsymbol{\alpha}).\tag{19}$$

Define a differentiable curve $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma}=\left\{\boldsymbol{\theta}(s)\in N(\boldsymbol{\alpha}):\frac{\mathrm{d}\boldsymbol{\theta}(s)}{\mathrm{d}s}=\mathbf{v}(\boldsymbol{\theta}(s)),\boldsymbol{\theta}(0)=\boldsymbol{\alpha},-1<s<1\right\}.\tag{20}$$

Following the same calculation as Eq.(16), we have

$$\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)=0.\tag{21}$$

Thus, the expectation and variance of $\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)$ are as follows:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\left(\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\right)=\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\mathbb{E}_\theta\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)=0,\\ &\mathrm{Var}_{\boldsymbol{\theta}}\left(\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\right)=\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\left(\mathrm{Cov}_{\boldsymbol{\theta}}\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\right)\mathbf{v}(\boldsymbol{\theta})\\ &=\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})\mathbf{v}(\boldsymbol{\theta})=0,\ \boldsymbol{\theta}\in N(\boldsymbol{\alpha}).\end{aligned}\tag{22}$$

This implies that

$$\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta})\left(\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)=0\quad\text{for}\quad\text{a.e. }\mathbf{x}\in\mathbb{R}^n.\tag{23}$$

Now, for $\boldsymbol{\theta}(s)\in\boldsymbol{\Gamma}$, the following equation

$$\begin{aligned}\frac{\mathrm{d}\log p(\mathbf{x},\boldsymbol{\theta}(s))}{\mathrm{d}s}&=\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\frac{\mathrm{d}\boldsymbol{\theta}(s)}{\mathrm{d}s}\\ &=\mathbf{v}^{\mathrm{T}}(\boldsymbol{\theta}(s))\frac{\partial\log\ p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\\ &=0\end{aligned}\tag{24}$$

holds for all $-1<s<1$. That is,

$$p(\mathbf{x},\boldsymbol{\theta}(s))=const,\ -1<s<1.\tag{25}$$

Thus $p(\mathbf{x},\boldsymbol{\theta}(s))$ is unchanged for all points $\boldsymbol{\theta}(s)\in\boldsymbol{\Gamma},\ -1<s<1$. This means all the parameters $\boldsymbol{\theta}(s)$ in $\boldsymbol{\Gamma}$ are local minimum points of the optimization problem (8). Therefore, by Theorem 2, $\boldsymbol{\alpha}$ is not locally identifiable. □

We present a geometrical interpretation for the identifiability condition in Theorem 3 For a matrix $\mathbf{A}$, we denote $\ker\mathbf{A}=\{\mathbf{u}:\mathbf{Au}=0\}$ to be its kernel space. From Eq.(22), we have $\mathbf{v}(\boldsymbol{\theta})\in\ker\mathbf{H}(\boldsymbol{\theta})$. That is, the set $\ker\mathbf{H}(\boldsymbol{\theta})$ consists of the smooth curves along which $KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s)),\quad -1<s<1$ has completely flat ridge in $\mathbb{R}^k$. However, the conventional KLD equation method [18,19] cannot provide us such geometrical sight.

Most of previous studies on identifiability problem concerned mainly with local identifiability. Up to now, few investigations have been reported on how to examine global identifiability. However, in some cases, we are more interested in global identifiability rather than simply local identifiability [2,16]. Unfortunately, it is very difficult to obtain global results in generic statistical settings. Based on Theorem 3, we propose a *global* result as follows. Compared with the global results in [23,28], our result is valid for any statistical model without restricting to exponential family.

Theorem 4. *Suppose that the Hessian matrix* $\mathbf{H}(\boldsymbol{\theta})$ *of* $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ *is strictly positive definite for all* $\boldsymbol{\theta}\in\mathbb{R}^k,\ \ \boldsymbol{\theta}\neq\boldsymbol{\alpha}$, *then* $\boldsymbol{\alpha}$ *is globally identifiable.*

Proof. Apply *Taylor's formula* to $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ and from Eq.(12), we have

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta})=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\alpha}),\tag{26}$$

where

$$\mathbf{H}(\boldsymbol{\theta}^*)=\frac{\partial^2 KL(\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*},\boldsymbol{\theta}^*=(1-t)\boldsymbol{\alpha}+t\boldsymbol{\theta},\quad 0<t<1\ .\tag{27}$$

Since $\mathbf{H}(\boldsymbol{\theta})$ is strictly positive definite for each $\boldsymbol{\theta}\in\mathbb{R}^k,\ \boldsymbol{\theta}\neq\boldsymbol{\alpha}$,

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta})>0\quad\text{for any }\boldsymbol{\theta}\neq\boldsymbol{\alpha}.\tag{28}$$

That is, $\boldsymbol{\alpha}$ is the strict global minimum point of the optimization problem (8). By Theorem 2, $\boldsymbol{\alpha}$ is globally identifiable. □

4. Identifiability criteria for parameter-constrained models

If the parameter in the unconstrained model is unidentifiable, we can change the nature of modeling approach to make it identifiable. Traditionally, there are two approaches to achieve this purpose. The first approach is to introduce a priori distribution on the unknown parameter to be estimated, and to cast the estimation problem into a *Bayesian* framework [24,33]. The second approach is to impose some *deterministic* constraints (e.g. functional constraints [22,23,32], sparsity constraints [10], monotonicity constraints [34,15], order constraints [15], etc.) on the unknown parameter, and to result in a parameter estimation problem with reduced dimensions [23,32].

In this section, we focus on the identifiability problem for statistical models with nonlinear equality constraints. Formally, we suppose that the admissible parameter space is restricted to

$$\mathbf{S}=\{\boldsymbol{\theta}\in\mathbb{R}^k:\boldsymbol{\Phi}(\boldsymbol{\theta})=0\}, \tag{29}$$

where $\boldsymbol{\Phi}(\boldsymbol{\theta})=(\varphi_1(\boldsymbol{\theta}),\cdots,\varphi_q(\boldsymbol{\theta}))$ and $\varphi_i(\boldsymbol{\theta}),\ 1\le i\le q$ possess continuous partial derivatives. We also suppose that the rank of the Jacobian matrix

$$\mathbf{J}(\boldsymbol{\theta})=\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\theta}}=\left(\frac{\partial\varphi_i}{\partial\theta_j}\right)_{q\times k} \tag{30}$$

of $\boldsymbol{\Phi}(\boldsymbol{\theta})$ is q for all $\boldsymbol{\theta}$. This assumption means that the constraints are non-redundant; otherwise, certain constraints are redundant and can be removed.

Following the same line as the unconstrained case, we formulate the identifiability problem as the following constrained optimization problem.

Theorem 5. *In a statistical model $p(\mathbf{x},\boldsymbol{\theta})$ with constrained parameter space $\mathbf{S}=\{\boldsymbol{\theta}\in\mathbb{R}^k:\boldsymbol{\Phi}(\boldsymbol{\theta})=0\}$, a parameter point $\boldsymbol{\alpha}\in\mathbf{S}$ is globally (locally) identifiable if and only if $\boldsymbol{\alpha}$ is the strict global (local) minimum point of the following constrained optimization problem*

$$\begin{aligned}&\text{Minimize}\quad KL(\boldsymbol{\alpha},\boldsymbol{\theta})\\&\text{Subject to}\quad \boldsymbol{\Phi}(\boldsymbol{\theta})=0.\end{aligned} \tag{31}$$

After transforming the identifiability problem into a constrained optimization problem, we present the following theorem.

Theorem 6. *Suppose that the parameter space of the statistical model $p(\mathbf{x},\boldsymbol{\theta})$ is restricted to $\mathbf{S}=\{\boldsymbol{\theta}\in\mathbb{R}^k:\Phi(\boldsymbol{\theta})=0\}$, $\boldsymbol{\alpha}\in\mathbf{S}$, $\mathbf{M}(\boldsymbol{\theta})$ is a block matrix of the form*

$$\mathbf{M}(\boldsymbol{\theta})=\begin{pmatrix}\mathbf{H}(\boldsymbol{\theta})\\ \mathbf{J}(\boldsymbol{\theta})\end{pmatrix}, \tag{32}$$

where $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian matrix of $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ is the Jacobian matrix of $\boldsymbol{\Phi}(\boldsymbol{\theta})$, if $\mathbf{M}(\boldsymbol{\theta})$ has constant rank in an open neighbor $N(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$, then the following three conditions are equivalent:

(a) $\boldsymbol{\alpha}\in\mathbf{S}$ is not locally identifiable.
(b) The block matrix $\mathbf{M}(\boldsymbol{\alpha})$ is column rank-deficient.
(c) The matrix $\mathbf{H}(\boldsymbol{\alpha})+\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha})$ is rank-deficient.

Proof. (a) ⇒ (b). Since rank $\mathbf{J}(\boldsymbol{\alpha})=q$, by the implicit function theorem [35], we can see that the constrained parameter space $\mathbf{S}$ is, in a neighbor $N(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$, a manifold of $k-q$ dimensionalities. Thus $\mathbf{S}$ is locally homeomorphic to an open set of $\mathbb{R}^{k-q}$. Suppose $\boldsymbol{\alpha}\in\mathbf{S}$ is not locally identifiable, there exists a differentiable nontrivial curve Γ in $\mathbf{S}$,

$$\boldsymbol{\Gamma}=\{\boldsymbol{\theta}(s)\in\mathbf{S}:\boldsymbol{\theta}(s)\in[\boldsymbol{\alpha}],\ \boldsymbol{\theta}(0)=\boldsymbol{\alpha},\quad -1<s<1\} \tag{33}$$

along which all $\boldsymbol{\theta}(s),\quad -1<s<1$ are local minimum points of the optimization problem (31), i.e.,

$$KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s))=0,\quad -1<s<1. \tag{34}$$

Taking second derivative with respect to s for Eq.(34), we have

$$\left(\frac{\partial KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\frac{\mathrm{d}^2\boldsymbol{\theta}}{\mathrm{d}s^2}+\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}\right)^{\mathrm{T}}\frac{\partial^2 KL(\alpha,\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}^2}\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}\right)=0. \tag{35}$$

Moreover, since the curve $\boldsymbol{\Gamma}$ lies on the surface $\mathbf{S}$, we obtain

$$\boldsymbol{\Phi}(\boldsymbol{\theta}(s))=0,\quad -1<s<1. \tag{36}$$

Taking derivative with respect to s for Eq.(36), we have

$$\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\theta}}\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}=0. \tag{37}$$

Combining Eq.(35) with (37), we obtain

$$\begin{cases}\left(\frac{\partial KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\frac{\mathrm{d}^2\boldsymbol{\theta}}{\mathrm{d}s^2}+\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}\right)^{\mathrm{T}}\frac{\partial^2 KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}^2}\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}\right)=0\\ \frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\theta}}\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}s}=0\end{cases}. \tag{38}$$

From Eq.(11), we have

$$\left.\frac{\partial KL(\boldsymbol{\alpha},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right|_{s=0}=\left.\frac{\partial KL(\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\alpha}}=0. \tag{39}$$

Evaluating Eq.(38) at $s=0$, we obtain

$$\begin{cases}\mathbf{u}^{\mathrm{T}}\mathbf{H}(\boldsymbol{\alpha})\mathbf{u}=0\\ \mathbf{J}(\boldsymbol{\alpha})\mathbf{u}=0\end{cases} \tag{40}$$

by letting $\mathbf{u}=(\mathrm{d}\boldsymbol{\theta}/\mathrm{d}s)|_{s=0}$ which is nonzero as $\boldsymbol{\Gamma}$ is non-trivial. From Eq.(10) and the second order necessary condition of local minimum point [20,21], the Hessian matrix $\mathbf{H}(\boldsymbol{\alpha})$ is positive semidefinite. Hence, Eq. (40) can be written as

$$\mathbf{M}(\boldsymbol{\alpha})\mathbf{u}=0. \tag{41}$$

Therefore, the block matrix $\mathbf{M}(\boldsymbol{\alpha})$ is column rank-deficient.

(b) ⇒ (c). Immediate.

(c) ⇒ (a). Suppose that $\mathbf{H}(\boldsymbol{\alpha})+\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha})$ is rank-deficient, there exists a non-trivial vector $\mathbf{u}$ such that

$$(\mathbf{H}(\boldsymbol{\alpha})+\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}))\mathbf{u}=0. \tag{42}$$

Since $\mathbf{H}(\boldsymbol{\alpha})$ is positive semidefinite and $\ker(\mathbf{J}(\boldsymbol{\alpha}))=\ker(\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}))$ [36], we have

$$\mathbf{H}(\boldsymbol{\alpha})\mathbf{u}=\mathbf{J}(\boldsymbol{\alpha})\mathbf{u}=0. \tag{43}$$

This means

$$\mathbf{M}(\boldsymbol{\alpha})\mathbf{u}=0. \tag{44}$$

Since $\mathbf{M}(\boldsymbol{\alpha})$ has constant rank in a neighbor $N(\boldsymbol{\alpha})$ of α, then $\mathbf{M}(\boldsymbol{\theta})$ is rank-deficient for all $\boldsymbol{\theta}\in N(\boldsymbol{\alpha})$. Let $\boldsymbol{\Gamma}$ be the smooth curve in $N(\boldsymbol{\alpha})$ as follows

$$\Gamma=\{\boldsymbol{\theta}(s)\in N(\boldsymbol{\alpha}):\mathbf{M}(\boldsymbol{\theta}(s))\mathbf{u}(s)=0,\boldsymbol{\theta}(0)=\boldsymbol{\alpha},\mathbf{u}(0)=\mathbf{u},\ -1<s<1\}. \tag{45}$$

We then have

$$\mathbf{J}(\boldsymbol{\theta}(s))\mathbf{u}(s)\equiv 0,\quad -1<s<1. \tag{46}$$

Since rank$\mathbf{J}(\boldsymbol{\theta})=q$ for all $\boldsymbol{\theta}$, $\mathbf{u}(s)$ is on the tangent plane of the surface $\mathbf{S}$ at $\boldsymbol{\theta}(s)$ [35]. Define a differentiable curve $\boldsymbol{\Upsilon}$ on $\mathbf{S}$ as follows

$$\boldsymbol{\Upsilon}=\left\{\sigma(s)\in\mathbf{S}:\frac{\mathrm{d}\sigma(s)}{\mathrm{d}s}=\mathbf{u}(s),\sigma(0)=\alpha,\mathbf{u}(0)=\mathbf{u},\quad -1<s<1\right\}. \tag{47}$$

That is, the curve $\boldsymbol{\Upsilon}$ is on the constrained surface $\mathbf{S}$ passing through $\boldsymbol{\alpha}$ with tangent vector $\mathbf{u}(s)$ at $\boldsymbol{\theta}(s)$. Further, from Eq.(45), we have

$$\mathbf{H}(\boldsymbol{\theta}(s))\mathbf{u}(s)=0,\quad -1<s<1. \tag{48}$$

Hence,

$$\mathbf{u}^{\mathrm{T}}(s)\mathbf{H}(\boldsymbol{\theta}(s))\mathbf{u}(s)=0,\quad -1<s<1. \tag{49}$$

Following the same calculation as Eqs.(16) and (17), we have

$$\mathbb{E}_{\boldsymbol{\theta}(s)}\left(\mathbf{u}^{\mathrm{T}}(s)\frac{\partial\log p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right)=\mathbf{u}^{\mathrm{T}}(s)\mathbb{E}_{\boldsymbol{\theta}(s)}\left(\frac{\partial\log p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right)=0, \tag{50}$$

$$\mathrm{Var}_{\boldsymbol{\theta}(s)}\left(\mathbf{u}^{\mathrm{T}}(s)\frac{\partial\log p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}\right)=\mathbf{u}^{\mathrm{T}}(s)\mathbf{H}(\boldsymbol{\theta}(s))\mathbf{u}(s)=0. \tag{51}$$

Hence, we obtain

$$\mathbf{u}^{\mathrm{T}}(s)\frac{\partial\log p(\mathbf{x},\boldsymbol{\theta}(s))}{\partial\boldsymbol{\theta}}=0\quad\text{for a.e. }\mathbf{x}\in\mathbb{R}^n. \tag{52}$$

Taking derivative with respect to s for $KL(\boldsymbol{\alpha},\boldsymbol{\sigma}(s))$, we have

$$\frac{\mathrm{d}KL(\boldsymbol{\alpha},\boldsymbol{\sigma}(s))}{\mathrm{d}s}=\left(\frac{\mathrm{d}\boldsymbol{\sigma}(s)}{\mathrm{d}s}\right)^{\mathrm{T}}\left(\frac{\partial KL(\boldsymbol{\alpha},\boldsymbol{\sigma}(s))}{\partial\boldsymbol{\theta}}\right)$$

$$= -\int \mathbf{u}^{\mathrm{T}}(s)\left(\frac{\partial \log\ p(\mathbf{x}, \boldsymbol{\sigma}(s))}{\partial \boldsymbol{\theta}}\right)p(\mathbf{x}, \boldsymbol{\alpha})\mathrm{d}\mathbf{x}$$
$$= 0. \tag{53}$$

The equation above implies that all the points $\boldsymbol{\sigma}(s)$, $-1 < s < 1$ on $\mathbf{S}$ are local minimum points of the optimization problem (31). Therefore, by Theorem 5, $\boldsymbol{\alpha} \in \mathbf{S}$ is not locally identifiable. □

The identifiability result in Theorem 6 can be viewed in two complementary ways. They are, in a geometrical viewpoint, if $\boldsymbol{\alpha} \in \mathbf{S}$ is not locally identifiable, then $KL(\boldsymbol{\alpha}, \boldsymbol{\sigma}(s))$, $-1 < s < 1$ has complete flat ridge on the constrained parameter space $\mathbf{S}$. From an algebraic viewpoint, $\boldsymbol{\alpha} \in \mathbf{S}$ is locally identifiable if the deficiency in $\mathbf{H}(\boldsymbol{\alpha})$ can be compensated by the rank of $\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha})$.

In some practical applications, it is of interest to study the problem of how many constraints are needed to guarantee identifiability. The following theorem can be applied as a guideline to quantitative experiment design.

Theorem 7. *Suppose* rank$\mathbf{H}(\boldsymbol{\alpha}) = r$, $r \le k$, *the minimum number of constraints needed to achieve local identifiability is* $k - r$.

Proof. From the identifiability condition in Theorem 6 and the fact rank$(\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}))$ = rank$(\mathbf{J}(\boldsymbol{\alpha}))$ [36], we have

$$k = \mathrm{rank}(\mathbf{H}(\boldsymbol{\alpha}) + \mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}))$$
$$\le \mathrm{rank}(\mathbf{H}(\boldsymbol{\alpha})) + \mathrm{rank}(\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}))$$
$$= r + q. \tag{54}$$

Hence,

$$q \ge k - r. \tag{55}$$

That is, the number of constraints needed to guarantee local identifiability is at least $k - r$. Next we prove that the equality is attainable. Consider the spectral decomposition [36] of the symmetric positive semidefinite matrix $\mathbf{H}(\boldsymbol{\alpha})$:

$$\mathbf{H}(\boldsymbol{\alpha}) = (\mathbf{U}_1(\boldsymbol{\alpha}), \mathbf{U}_2(\boldsymbol{\alpha}))\begin{pmatrix} \Sigma(\boldsymbol{\alpha}) & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{U}_1^{\mathrm{T}}(\boldsymbol{\alpha}) \\ \mathbf{U}_2^{\mathrm{T}}(\boldsymbol{\alpha}) \end{pmatrix}, \tag{56}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\alpha}) = \mathrm{diag}\{\lambda_1(\boldsymbol{\alpha}), \ldots, \lambda_r(\boldsymbol{\alpha})\}$ with $\lambda_i(\boldsymbol{\alpha}) > 0$, $i = 1, \ldots, r$. We choose a set of constraints as

$$\boldsymbol{\Phi}(\boldsymbol{\theta}) = \mathbf{U}_2^{\mathrm{T}}(\boldsymbol{\alpha})\boldsymbol{\theta} + \mathbf{b} = 0. \tag{57}$$

Since the Jacobian matrix $\mathbf{J}(\boldsymbol{\theta})$ of $\boldsymbol{\Phi}(\boldsymbol{\theta})$ is $\mathbf{U}_2^{\mathrm{T}}(\boldsymbol{\alpha})$, it is clear to see that

$$\mathbf{H}(\boldsymbol{\alpha}) + \mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha}) = (\mathbf{U}_1(\boldsymbol{\alpha}), \mathbf{U}_2(\boldsymbol{\alpha}))\begin{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{\alpha}) & 0 \\ 0 & \mathbf{I}_{k-r} \end{pmatrix}\begin{pmatrix} \mathbf{U}_1^{\mathrm{T}}(\boldsymbol{\alpha}) \\ \mathbf{U}_2^{\mathrm{T}}(\boldsymbol{\alpha}) \end{pmatrix} \tag{58}$$

where $\mathbf{I}_{k-r}$ is the identity matrix of size $k - r$. Hence, we have

$$\left|\mathbf{H}(\boldsymbol{\alpha}) + \mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha})\right| = \lambda_1(\boldsymbol{\alpha})\ldots\lambda_r(\boldsymbol{\alpha}) > 0. \tag{59}$$

From Theorem 6, $\boldsymbol{\alpha}$ is locally identifiable. Since the number of constraints in Eq.(57) is $k - r$, this means the lower bound $k - r$ is tight. □

In the final part of this section, we briefly discuss the global identifiability problem for parameter-constrained models. A direct result from the *convex optimization theory* [37] is that, if the objective function in Eq.(31) is strictly convex with respect to $\boldsymbol{\theta}$ and the constraint function $\boldsymbol{\Phi}(\boldsymbol{\theta})$ is convex, then the identifiability result in Theorem 6 becomes a global one. However, the $KL(\boldsymbol{\alpha}, \boldsymbol{\theta})$ is not generally convex with respect to $\boldsymbol{\theta}$ although $KL(p, q)$ is convex with respect to the second argument q [38], as q is nonlinear in $\boldsymbol{\theta}$. Thus, we cannot cast the identifiability problem into the convex optimization theory framework, making one difficult to derive a global criterion. Therefore, the global identifiability problem remains a challenging subject in identifiability theory.

5. Applications

In this section, we first present several simple examples from literature to illustrate the validity of the proposed identifiability criteria. Specific examples considered include learning machine, Gaussian linear model with linear constraints, nonlinear regression, and RBF neural network. Further, we present three practical models to study their identifiability property. They are GCNN model, *partially linear support vector machine* (*PL-SVM*) model and signal estimation with power constraints.

Example 1. Consider a statistical learning machine [39]

$$p(y|x, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(y - ax - b)^2\right], \tag{60}$$

where $\boldsymbol{\theta} = (a, b)^{\mathrm{T}}$. The admissible parameter space is $\mathbb{R}^2$. Then each $\boldsymbol{\theta} \in \mathbb{R}^2$ defines a PDF $p(x, y, \boldsymbol{\theta}) = p(x)p(y|x, \boldsymbol{\theta})$ in $\mathbb{R}^2$. Suppose that the true model is

$$p(x, y, \boldsymbol{\alpha}) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}(x^2 + y^2)\right]. \tag{61}$$

The KLD can be calculated as [39]

$$KL(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \int p(x, y, \boldsymbol{\alpha})\log\frac{p(x, y, \boldsymbol{\alpha})}{p(x, y, \boldsymbol{\theta})}\mathrm{d}x\mathrm{d}y$$
$$= \int p(x, y, \boldsymbol{\alpha})\log\frac{p(y|x, \boldsymbol{\alpha})}{p(y|x, \boldsymbol{\theta})}\mathrm{d}x\mathrm{d}y$$
$$= \frac{1}{2}(a^2 + b^2). \tag{62}$$

It is easy to see that $\boldsymbol{\alpha}$ is locally identifiable since $\boldsymbol{\theta} = 0$ is the unique solution of the equation $KL(\boldsymbol{\alpha}, \boldsymbol{\theta}) = 0$. α is locally identifiable by conventional KLD equation method [18,19]. We then use Theorem 3 to test the identifiability of $\boldsymbol{\alpha}$. It is obvious that the Hessian matrix $\mathbf{H}(\boldsymbol{\alpha}) = \mathbf{I}_2$ is strictly positive definite, $\boldsymbol{\alpha}$ is therefore locally identifiable. The two approaches give the same result. In this example, we can see that, compared with the FIM method [23], the main advantage of the KLD-based method is that we do not need the explicit PDF $p(x)$ since it can be eliminated from the KLD calculation, while FIM method cannot deal with this problem if the exact form of $p(x)$ is not available.

Example 2. (Gaussian linear model [19,40]) Consider a simple Gaussian linear model

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{63}$$

where $\mathbf{X}$ is an $n \times k$ design matrix with rank$\mathbf{X} < n < k$, the additive error vector $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma)$. It is well known that local identifiability and global identifiability are synonyms in those linear models [19]. As the distribution of $\mathbf{y}$ depends on $\boldsymbol{\theta}$ through $X\boldsymbol{\theta}$ and $\mathbf{X}$ is not of column full rank, there exist several distinct values of $\boldsymbol{\theta}$ compatible with the same distribution of $\mathbf{y}$, the model is therefore unidentifiable. Alternatively, from Eq.(5), it is easy to see that the KLD is given by

$$KL(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\alpha}\|^2. \tag{64}$$

The Hessian matrix is $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ for any $\boldsymbol{\theta}$, by Theorem 3, the model is unidentifiable since $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is rank deficient. Next we suppose that a linear constraint $A\boldsymbol{\theta} + \mathbf{b} = 0$ is imposed on model (63), where $\mathbf{A}$ is a known row full-rank matrix, $\mathbf{b}$ is a known vector and $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A}^{\mathrm{T}}\mathbf{A}$ is of full rank. We first directly show that this parameter-constrained model is identifiable. Otherwise, there exist two distinct parameters $\boldsymbol{\theta}_1 \ne \boldsymbol{\theta}_2$ such that

$$\mathbf{X}\boldsymbol{\theta}_1 = \mathbf{X}\boldsymbol{\theta}_2 \ \text{ and } \mathbf{A}\boldsymbol{\theta}_1 + \mathbf{b} = 0,\ \mathbf{A}\boldsymbol{\theta}_2 + \mathbf{b} = 0. \tag{65}$$

This leads to

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A}^{\mathrm{T}}\mathbf{A})(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) = 0. \tag{66}$$

This is contradictory to the fact that $\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is of full rank. The parameter-constrained model is therefore identifiable. Then, by using the statement (c) of Theorem 6, the model is identifiable since $\mathbf{H}(\boldsymbol{\theta})+\mathbf{J}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})=\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is of full rank. The two approaches therefore give the same result.

Example 3. (Nonlinear regression [41]). Consider the MIMO nonlinear regression model

$$\mathbf{y}=\mathbf{f}(\mathbf{x},\boldsymbol{\theta})+\boldsymbol{\epsilon}, \tag{67}$$

where $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{y}\in\mathbb{R}^m$ are the input and output vectors, $\theta\in\mathbb{R}^k$, $\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$ is a vector-valued mapping

$$\mathbf{f}(\mathbf{x},\boldsymbol{\theta})=(f_1(\mathbf{x},\boldsymbol{\theta}),\ \ldots,f_m(\mathbf{x},\boldsymbol{\theta}))^{\mathrm{T}}. \tag{68}$$

Suppose that the PDF $p(\mathbf{x})$ is positive for a.e. $\mathbf{x}\in\mathbb{R}^n$ and the noise vector $\boldsymbol{\epsilon}\sim\mathcal{N}(0,\Sigma)$. The joint PDF of $\mathbf{x}$ and $\mathbf{y}$ is

$$p(\mathbf{x},\mathbf{y},\boldsymbol{\theta})=\frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{f}(\mathbf{x},\boldsymbol{\theta}))^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}-\mathbf{f}(\mathbf{x},\boldsymbol{\theta}))\right\}p(\mathbf{x}). \tag{69}$$

Denote $\mathbf{H}(\boldsymbol{\alpha})=(h_{ab}(\boldsymbol{\alpha}))$. From Eq.(18) we have

$$\begin{aligned}h_{ab}(\boldsymbol{\alpha})&=\int\frac{\partial\log\ p(\mathbf{x},\mathbf{y},\boldsymbol{\alpha})}{\partial\theta_a}\frac{\partial\log\ p(\mathbf{x},\mathbf{y},\boldsymbol{\alpha})}{\partial\theta_b}p(\mathbf{x},\mathbf{y},\boldsymbol{\alpha})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\&=\int\left(\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_a}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_b}\right)p(\mathbf{x})\mathrm{d}\mathbf{x},\end{aligned} \tag{70}$$

where

$$\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i}=\left(\frac{\partial f_1(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i},\ldots,\frac{\partial f_m(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i}\right)^{\mathrm{T}},1\le i\le k. \tag{71}$$

From Theorem 3, we can see that $\boldsymbol{\alpha}$ is not locally identifiable if and only if $\mathbf{H}(\boldsymbol{\alpha})$ is not strictly positive definite, i.e., there is a non-zero vector $\mathbf{v}=(v_1,\ldots,v_k)^{\mathrm{T}}$ such that

$$\mathbf{v}^{\mathrm{T}}\mathbf{H}(\boldsymbol{\alpha})\mathbf{v}=\sum_{a,b}v_a h_{ab}(\boldsymbol{\alpha})v_b=0. \tag{72}$$

By Eq.(70), we have

$$\mathbf{v}^{\mathrm{T}}\mathbf{H}(\boldsymbol{\alpha})\mathbf{v}=\int\left(\sum_i v_i\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\sum_i v_i\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i}\right)p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{73}$$

Since $p(\mathbf{x})$ is positive for a.e. $\mathbf{x}\in\mathbb{R}^n$, $\boldsymbol{\alpha}$ is not locally identifiable if and only if

$$\sum_i v_i\frac{\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_i}=0\quad\text{for}\quad\text{a.e. }\mathbf{x}\in\mathbb{R}^n, \tag{74}$$

i.e., the vectors $\partial\mathbf{f}(\mathbf{x},\boldsymbol{\alpha})/\partial\theta_i$, $i=1,\ldots,k$ are linearly dependent. The validity of this identifiability condition is consistent with our intuition.

For a geometric interpretation of the identifiability condition, we equivalently rewrite Eq.(74) as the following m equations

$$\nabla f_i^{\mathrm{T}}\mathbf{v}=0,\quad i=1,\ldots,m, \tag{75}$$

where

$$\nabla f_i=\left(\frac{\partial f_i(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_1},\ \ldots,\frac{\partial f_i(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_k}\right)^{\mathrm{T}}, \tag{76}$$

is the gradient vector of f_i with respect to $\boldsymbol{\theta}$. Each f_i is unchanged along the vector field $\mathbf{v}$ since ∇f_i is orthogonal to $\mathbf{v}$ in the parameter space. In other words, each f_i has completely flat ridge along this vector filed $\mathbf{v}$. It is worthwhile noting that Eqs.(74) and (75) provide a dual interpretation for the identifiability condition. Specifically, Eq. (74) says that the following *partial derivative matrix* (*PDM*)

$$\mathrm{PDM}=\left(\frac{\partial f_i(\mathbf{x},\boldsymbol{\alpha})}{\partial\theta_j}\right)_{m\times k} \tag{77}$$

is *column* linearly dependent. While Eq.(75) says that all the *row* vectors of the PDM are orthogonal to the vector $\mathbf{v}$. Moreover, the identifiability condition is independent of the PDF $p(\mathbf{x})$ and the covariance matrix $\boldsymbol{\Sigma}$. That is to say, even if we do not know the explicit expressions of $p(\mathbf{x})$ and Σ, we can still derive the identifiability condition. In this example, we can see that the FIM [23] and the KLD equation method [18,19] is not applicable since the close-form FIM and KLD cannot be obtained.

Next, we restrict the admissible parameter space to $\mathbf{S}$, and further study the identifiability condition of the parameter-constrained models. From Theorem 6, $\boldsymbol{\alpha}$ is locally identifiable if and only if the matrix $\mathbf{M}(\boldsymbol{\alpha})$ defined in Eq.(32) is column full-rank. To verify the validity of this assertion, from Eq.(19), we can see that $\mathbf{v}(\boldsymbol{\theta})\in\ker\mathbf{H}(\boldsymbol{\theta})$. That is, the set $\ker\mathbf{H}(\boldsymbol{\theta})$ consists of all the directions along which $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ has completely flat ridges, while the set $\ker\mathbf{J}(\boldsymbol{\theta})$ consists of all the feasible directions. If $\mathbf{M}(\boldsymbol{\alpha})$ is of column full-rank, then the set $\ker\mathbf{M}(\boldsymbol{\alpha})$ is trivial. That is, there exists no non-trivial feasible direction such that $KL(\boldsymbol{\alpha},\boldsymbol{\theta})$ has completely flat ridge. Hence, $\boldsymbol{\alpha}$ is locally identifiable since $\boldsymbol{\alpha}$ is the unique local optimum point of the optimization problem (31).

Example 4. (RBF neural network [42]) Consider the RBF neural network

$$y=\theta_1\psi_1(\mathbf{x})+\cdots+\theta_k\psi_k(\mathbf{x})+\epsilon, \tag{78}$$

where $\psi_i(\mathbf{x})=\psi(\|\mathbf{x}-\boldsymbol{\mu}_i\|)$ is a Gaussian RBF with center $\boldsymbol{\mu}_i$ and common covariance matrix $\boldsymbol{\Sigma}$. The unknown parameter $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)^{\mathrm{T}}\in\mathbb{R}^k$. $\epsilon\sim\mathcal{N}(0,\sigma^2)$. It has been shown that $\boldsymbol{\theta}$ is identifiable if and only if $\boldsymbol{\mu}_i$, $1\le i\le k$ are distinct [42]. Or alternatively, it is clear that the PDM of model (78) is $\mathrm{PDM}=(\psi_1(\mathbf{x}),\ \ldots,\psi_k(\mathbf{x}))$. From Example 3, the model is identifiable if and only if $\psi_i(\mathbf{x})$, $1\le i\le k$ are functionally independent. This means that the interpolation equation $\Psi\boldsymbol{\theta}=0$ has a trivial solution, where $\Psi=(\psi_{ij})_{k\times k}$, $\psi_{ij}=\psi(\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|)$. This condition is equivalent to the fact that μ_i, $1\le i\le k$ are distinct [43]. This verifies the validity of Theorem 6.

Example 5. (GCNN model). In [12], a GCNN model given by

$$\begin{aligned}y&=f(x,\boldsymbol{\theta})+\epsilon=g(x,\alpha)\times h(x,\mathbf{w},\mathbf{c})+\epsilon\\&=\mathrm{e}^{-\alpha x}\sum_{i=1}^{n}w_i\mathrm{e}^{-(x-c_i)^2}+\epsilon\end{aligned} \tag{79}$$

is applied to a nonlinear regression problem, where $\mathbf{w}=(w_1,\ldots,w_n)^{\mathrm{T}}$, $\mathbf{c}=(c_1,\ldots,c_n)^{\mathrm{T}}$ and α is a positive real number. The GCNN model $f(x,\boldsymbol{\theta})$ basically consists of two submodels, namely the knowledge-driven submodel $g(x,\alpha)=\mathrm{e}^{-\alpha x}$ which represents the available domain knowledge and the data-driven submodel $h(x,\mathbf{w},\mathbf{c})=\sum_{i=1}^{n}w_i\mathrm{e}^{-(x-c_i)^2}$ which fits the experimental data by making use of the RBF neural network. The two submodels are coupled by a multiplication operation. The unknown parameter $\boldsymbol{\theta}=(\alpha,\mathbf{w},\mathbf{c})^{\mathrm{T}}$ and the additive noise $\epsilon\sim\mathcal{N}(0,\sigma^2)$. The parameter α has a physically interpretable meaning (a dampen coefficient) which is of practical interest since its value reflects the level of the energy dissipation in the real system. All parameters including the physically based parameter α were learned simultaneously from observation data. Although higher generalization capability was obtained in comparison with other methods due to the introduction of domain knowledge, it was also observed through numerical simulations, that it is not possible to obtain a reasonable estimation for this practically important parameter α. In this example, we will rigorously prove that the model is actually unidentifiable, revealing that it is just the nonidentifiability that leads to ambiguity in parameter estimation. For clarity, we consider the following simplified model

$$f(x,\boldsymbol{\theta})=g(x,\alpha)\times h(x,w,c)=w\mathrm{e}^{-c^2}\mathrm{e}^{-x^2+(2c-\alpha)x}, \tag{80}$$

where $\boldsymbol{\theta}=(\alpha,w,c)^{\mathrm{T}}$, since the extension to general form Eq. (79) is rather straightforward. Intuitively, the model is unidentifiable due to the presence of terms $w\mathrm{e}^{-c^2}$ and $2c-\alpha$. From Example 3, we only need to check column dependence of the PDM of $f(x, \boldsymbol{\theta})$. This is easily verified by the following algebraic equation

$$2\frac{\partial f}{\partial \alpha}+2wc\frac{\partial f}{\partial w}+\frac{\partial f}{\partial c}=0. \tag{81}$$

This implies that there is a nontrivial linear dependence among the columns of the PDM of $f(x, \theta)$. Or equivalently,

$$\nabla f^{\mathrm{T}}\mathbf{v}=0, \tag{82}$$

where $\mathbf{v} \propto (2, 2wc, 1)^{\mathrm{T}}$. This implies that f has completely flat ridge along the vector field $\mathbf{v}$ in parameter space. Therefore, the parameter α is unidentifiable due to the coupling effect in the model. The practical implication of nonidentifiability suggests that, in order to identify the physically interpretable parameter α, the current model structure should be reformulated or an additional parameter constraint should be imposed on the unconstrained GCNN model. Compared with the result in [2], the superiority of the proposed method lies that, on the one hand, it provides a dual interpretation of the identifiability condition which is algebraically reasonable and geometrically comprehensible, while the result in [2] is simply an algebraic one. On the other hand, our method explicitly gives the observationally equivalent parameter vector $\mathbf{v}$, while [2] can only detect the redundancy status of the model.

Example 6. (PL-SVM model, [13]). The objective of nonlinear system identification is to establish a relation between input $u(t)$ and output $y(t)$ generated by an unknown target dynamical system. Let $\mathbf{z}(t)=[y(t-1), \ldots, y(t-a), u(t), u(t-1), \ldots, u(t-b)]$ be the regression vector corresponding to the output $y(t)$ in a *nonlinear autoregressive exogenous (NARX)* model of order (a, b). The task is then to estimate a nonlinear function g such that $y(t, \boldsymbol{\theta})=g(\mathbf{z}(t), \theta)+\epsilon(t),\ t=1, \ldots, N$. In [13], the authors studied the case where there is evidence that some of the regressors in the model $g(\mathbf{z}(t), \boldsymbol{\theta})$ is linear. In other words, the nonlinearity of $g(\mathbf{z}(t), \boldsymbol{\theta})$ does not apply over all the components of $\mathbf{z}(t)$, rather a subset of it, leading to the identification of the following partially linear model

$$y(t, \boldsymbol{\theta})=g(\mathbf{b}(t))+\mathbf{a}^{\mathrm{T}}(t)\boldsymbol{\beta}+b+\epsilon(t), \tag{83}$$

where $\mathbf{a}(t)$ and $\mathbf{b}(t)$ represent the subvectors of $\mathbf{z}(t)$ that enter linearly and nonlinearly into the model, respectively. In [13], the following PL-SVM model

$$y(t, \boldsymbol{\theta})=\boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{b}(t))\mathbf{w}+\mathbf{a}^{\mathrm{T}}(t)\boldsymbol{\beta}+b+\epsilon(t) \tag{84}$$

is applied to approximate the unknown target system, where $\boldsymbol{\varphi}$ is the nonlinear feature mapping from input space to *infinite-dimensional* feature space and satisfies $\kappa(\mathbf{u}, \mathbf{v})=\boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{u})\boldsymbol{\varphi}(\mathbf{v})$, where $\kappa(\mathbf{u}, \mathbf{v})$ is the kernel function. The unknown parameter $\boldsymbol{\theta}=(\mathbf{w}, \beta, b)^{\mathrm{T}}$. We suppose that the additive noise $\epsilon(t) \sim \mathcal{N}(0, 1)$. The problem addressed in this example is to determine whether or not the linear part $\mathbf{a}^{\mathrm{T}}(t)\beta+b$ can be fully recovered from the model structure since this part is of practical importance to control or prediction purpose. The numerical experiments in [13] demonstrated the advantages of this structured model in, e.g., better performance results, improved generalization ability, and reduction of effective parameters. However, the identifiability issue is not theoretically verified in [13]. It is clear that for any finite observation data, (β, b) cannot be uniquely determined from the unconstrained model since $\mathbf{w}$ contains infinitely many parameters. To the best of our knowledge, none of the existing methods can deal with this infinite-dimensional parameter case. The goal in this example is to determine the precise conditions under which the parameter vector (β, b) will be identifiable. Formally, we suppose that a constant modulus constraint $\|\mathbf{w}\|=const$ is imposed on the original unconstrained model. From Eq.(5) we can see that the KLD criterion in this example is equivalent to the least squares criterion. Hence, by Theorem 5, the model is identifiable if and only if the following equivalent optimization problem has unique minimum point.

$$\begin{aligned}&\text{Minimize } \tfrac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}+\tfrac{1}{2}\mu\mathbf{w}^{\mathrm{T}}\mathbf{w}\\ &\text{Subject to } \mathbf{y}=\mathbf{B}\mathbf{w}+\mathbf{A}\beta+b\mathbb{I}_N+\boldsymbol{\epsilon}.\end{aligned} \tag{85}$$

Here $\mathbf{y}=(y(1), \ldots, y(N))^{\mathrm{T}}$, $\boldsymbol{\epsilon}=(\epsilon(1), \ldots, \epsilon(N))^{\mathrm{T}}$, $\mathbf{A}$ and $\mathbf{B}$ are matrices with $\mathbf{a}^{\mathrm{T}}(i)$ and $\mathbf{b}^{\mathrm{T}}(i)$ as their rows, respectively, $\mathbb{I}_N$ is a column vector with all its elements equal to 1, and μ is a positive regularized coefficient. This leads to the following Lagrangian

$$\mathscr{L}(\mathbf{w}, b, \boldsymbol{\epsilon}, \beta, \lambda)=\frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}+\frac{1}{2}\mu\mathbf{w}^{\mathrm{T}}\mathbf{w}+\lambda^{\mathrm{T}}(\mathbf{y}-\mathbf{B}\mathbf{w}-\mathbf{A}\beta-b\mathbb{I}_N-\boldsymbol{\epsilon}), \tag{86}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. The resulting *Karush-Kuhn-Tucker (K.K.T.)* conditions [20,21] are obtained as follows:

$$\begin{cases}\frac{\partial \mathscr{L}}{\partial \mathbf{w}}=0 \Rightarrow \mathbf{w}=\mu^{-1}\mathbf{B}^{\mathrm{T}}\lambda\\ \frac{\partial \mathscr{L}}{\partial b}=0 \Rightarrow \mathbb{I}_N^{\mathrm{T}}\lambda=0\\ \frac{\partial \mathscr{L}}{\partial \boldsymbol{\epsilon}}=0 \Rightarrow \lambda=\boldsymbol{\epsilon}\\ \frac{\partial \mathscr{L}}{\partial \beta}=0 \Rightarrow \mathbf{A}^{\mathrm{T}}\lambda=0\\ \frac{\partial \mathscr{L}}{\partial \lambda}=0 \Rightarrow \mathbf{y}=\mathbf{B}\mathbf{w}+\mathbf{A}\beta+b\mathbb{I}_N+\boldsymbol{\epsilon}\end{cases} \tag{87}$$

After elimination of $\mathbf{w}$ and $\boldsymbol{\epsilon}$, we obtain the following system

$$\begin{pmatrix}(\mu^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}) & \mathbf{A} & \mathbb{I}_N\\ \mathbf{A}^{\mathrm{T}} & 0 & 0\\ \mathbb{I}_N^{\mathrm{T}} & 0 & 0\end{pmatrix}\begin{pmatrix}\lambda\\ \beta\\ b\end{pmatrix}=\begin{pmatrix}\mathbf{y}\\ 0\\ 0\end{pmatrix}. \tag{88}$$

Note that $\mu^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}$ is always strictly positive definite, the coefficient matrix of the above system is congruent to the following block matrix

$$\operatorname{diag}\{\mu^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I},\ (\mathbf{A}, \mathbb{I}_N)^{\mathrm{T}}(\mu^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I})^{-1}(\mathbf{A}, \mathbb{I}_N)\}. \tag{89}$$

Hence, a unique solution exists for (β, b) if and only if the matrix $(\mathbf{A}, \mathbb{I}_N)^{\mathrm{T}}(\mu^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I})^{-1}(\mathbf{A}, \mathbb{I}_N)$ is invertible. From elementary linear algebra, this requires that $(\mathbf{A}, \mathbb{I}_N)$ is of full column rank. The practical implication is that the linear part can be fully recovered from model Eq.(84) unless the matrix $(\mathbf{A}, \mathbb{I}_N)$ is of full column rank. The interesting point in this example is that the infinitely dimensional parameter $\mathbf{w}$ appears implicitly as an intermediate step, and be eliminated in the final expression, thus avoiding the direct operations in the infinitely dimensional feature space. This of course attributes to the interplay of the optimization theory and the kernel trick.

Example 7. (Signal estimation with power constraints, [40]) Consider the problem of estimating the discrete-time signal waveform $\boldsymbol{\theta}=(\theta_1,\ \theta_2, \theta_3)^{\mathrm{T}}$, subject to constraints on the squared-modulus of the discrete Fourier transformation (DFT) of θ. We suppose that the sum of the squared moduli on the first frequency interval is to be a known constant. Denote $\mathbf{W}=(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ be the 3×3 unitary matrix of orthonormal DFT columns:

$$\mathbf{w}_i=\frac{1}{\sqrt{3}}\left(1, \mathrm{e}^{-\boldsymbol{j}\frac{2\pi i}{3}}, \mathrm{e}^{-\boldsymbol{j}\frac{4\pi i}{3}}\right)^{\mathrm{T}}, \tag{90}$$

where $\boldsymbol{j}=\sqrt{-1}$. We can write the constraint as $(W\boldsymbol{\theta})_1=const$, where $(W\boldsymbol{\theta})_1$ is the first entry of $W\boldsymbol{\theta}$. The constraint can be equivalently written as $\theta^{\mathrm{T}}\mathbb{I}\mathbb{I}^{\mathrm{T}}\theta=const$. We now specialize to the linear observation model:

$$x_i=\theta_i+\epsilon_i, \quad i=1, 2, \tag{91}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. It is obvious that the unconstrained model (91) is unidentifiable since it is under-determinant. With the introduction of the power constraint, we will prove that the model is locally identifiable. The Hessian matrix of KLD is $\mathbf{H}(\boldsymbol{\alpha})=\operatorname{diag}\{1, 1, 0\}$ and the Jacobian matrix of the constraint is $\mathbf{J}(\boldsymbol{\theta})=\boldsymbol{\theta}^{\mathrm{T}}\mathbb{I}\mathbb{I}^{\mathrm{T}}$. After some algebra operations, we have

$\left|\mathbf{H}(\boldsymbol{\alpha})+\mathbf{J}^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{J}(\boldsymbol{\alpha})\right| \neq 0$. From Theorem 6, the constrained model is locally identifiable.

As demonstrated before, in addition to the deep theoretical insight, the formulation of identifiability problem within the optimization theory framework brings several practical advantages compared with existing methods. First, one can derive identifiability criteria in the case of lost information (e.g. the $p(x)$ in Example 1, the $p(\mathbf{x})$ and Σ in Example 3) while other methods fail. Second, one is able to determine identifiability by calculating the rank of a numerical matrix, thus avoiding the usual bottleneck of seeking for the roots from a set of nonlinear equations. The benefit gained is that the new results lead to a reduction of computational complexity from NP-complete to $\mathcal{O}(k^3)$. Third, for processing identifiability problem with infinitely dimensional unknown parameter (see Example 6), up to now, there exists no theoretical or methodological treatments in this aspect. As far as the authors concerned, the proposed optimization theory framework is perhaps the only suitable tool for dealing with this case. We attribute the derivation to the interplay of the optimization theory and the kernel trick. Nevertheless, identifiability analysis of nonlinear models is still difficult to implement since, whatever the method being used, the complexity increases very fast with the number of parameters, the dimensionality of input/output spaces, the nonlinear degree of models. Especially, in real problems with large dimensionality and high nonlinearity, the Hessian matrix itself is difficult to obtain. Moreover, the adoption of numerical approximation can also result in errors. This is a common difficulty for all existing methods. Therefore, this challenging problem is left for the future research.

6. Conclusion

In this paper, by making use of the KLD in information theory, we cast the identifiability problem into the optimization theory framework. Several novel identifiability criteria are derived for unconstrained and parameter-constrained models. The results partially answered the problem proposed by Yang et al. [2], i.e., the problem of how many, and what types of constraints are required to produce a unique estimation. The pros/cons of the proposed framework are detailed discussed from both theoretical and application viewpoints. Finally, we outline two directions below for future work:

1. One of the major objectives in identifiability theory is to obtain a set of identifying functions and then use them to *reparameterize* the model [14]. In almost all cases, such a set of functions cannot be easily obtained by visual inspection or analytic verification. In the present study, we propose some criteria to test parameter identifiability, but it tells nothing about reparameterization when parameter redundancy is detected. However, it would be highly desirable to seek for generic reparameterization methods. It is still an open problem which is one of the directions of research into the identifiability theory [14,29].
2. In real application scenarios, a vast variety of parameter systems are described by time-variant *ODE* or *PDE* dynamical models [4,5,44]. In spite of what a large literature on model identifiability, we found rare discussions on identifiability of parameter-constrained time-variant models. Therefore, it is worthwhile considering the identifiability issue in those models.

Acknowledgment

This work is supported in part by NSFC No. 61273196.

References

[1] S. Audoly, L. D'Angio, M.P. Saccomani, C. Cobelli, Global identifiability of biokinetic models of linear compartment models, a computer algebra algorithm, IEEE Trans. Biomed. Eng. 45 (1) (1998) 36–47.

[2] S.H. Yang, B.-G. Hu, P.H. Cournède, Structural identifiability of generalized-constraint neural network models for nonlinear regression, Neurocomputing 72 (2008) 392–400.

[3] D. Csercsik, K.M. Hangos, G. Szederkenyi, Identifiability analysis and parameter estimation of a single Hodgkin-Huxley type voltage dependent ion channel under voltage step measurement conditions, Neurocomputing 77 (1) (2012) 178–188.

[4] L. Wang, H. Garnier, System Estimation, Environmental Modeling and Control System Design, Springer, London, 2012.

[5] L. Ljung, System Identification: Theory for the User, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[6] S.I. Amari, H. Park, T. Ozeki, Singularities affect dynamics of learning in neuromanifolds, Neural Comput. 18 (2006) 1007–1065.

[7] S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, J. Mach. Learn Res. 11 (2010) 3571–3594.

[8] N. Murata, S. Yoshizawa, S.I. Amari, Network information criterion-determining the number of hidden units for an artificial neural network model, IEEE Trans. Neural Netw. 5 (6) (1994) 865–872.

[9] S.I. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (1998) 251–276.

[10] R. Henao, O. Winther, Sparse linear identifiable multivariate modeling, J. Mach. Learn Res. 12 (2011) 863–905.

[11] S. Fortunati, F. Gini, M.S. Greco, A. Farina, A. Graziano, S. Giompapa, On the identifiability problem in the presence of random nuisance parameters, Signal Process. 92 (2012) 2545–2551.

[12] B.-G. Hu, H.B. Qu, S.H. Yang, A generalized-constraint neural network model: associating partially known relationships for nonlinear regressions, Inf. Sci. 179 (2009) 1929–1943.

[13] M. Espinoza, J.A.K. Suykens, B.D. Moor, Kernel based partially linear models and nonlinear identification, IEEE Trans. Autom. Control 50 (10) (2005) 1602–1606.

[14] A. Dasgupta, S.G. Self, S.D. Gupta, Nonidentifiable parametric probability models and reparameterization, J. Stat. Plann. Inference 137 (2007) 3380–3393.

[15] Y.J. Qu, B.-G. Hu, Generalized constraint neural network regression model subject to linear priors, IEEE Trans. Neural Netw. 22 (12) (2011) 2447–2459.

[16] Z.Y. Ran, B.-G. Hu, Determining structural identifiability of parameter learning machines, Neurocomputing 127 (2014) 88–97.

[17] T.M. Cover, J.A. Thomas, Element of Information Theory, Wiley, Chichester, 1991.

[18] R. Bowden, The theory of parametric identification, Econometrica 41 (6) (1973) 1069–1074.

[19] C.D.M. Paulino, C.A.D.B. Pereira, On identifiability of parametric statistical models, J. Ital. Stat. Soc. 3 (1994) 125–151.

[20] D.G. Luenberger, Linear and Nonlinear Programming, Addison-Wesley, 1984.

[21] R.K. Sundaram, A. First Course in Optimization Theory, Cambridge University Press, Cambridge, 1996.

[22] P. Stoica, B.C. Ng, On the Cramér-Rao bound under parametric constraints, IEEE Signal Process. Lett. 5 (6) (1998) 177–179.

[23] T.J. Rothenberg, Identification in parametric models, Econometrica 39 (3) (1971) 577–591.

[24] G. Casella, R.L. Berger, Statistical Inference, Duxbury, 2002.

[25] B. Hochwald, A. Nehorai, On identifiability and information-regularity in parameterized normal distributions, Circuit Syst. Signal Process. 16 (1) (1997) 83–89.

[26] Y.W. Yao, G. Giannakis, On regularity and identifiability of blind source separation under constant modulus constraints, IEEE Trans. Signal Process. 53 (4) (2005) 1272–1281.

[27] E.M. Martin, F. Quintana, Consistency and identifiability revisited, Braz. J. Probab. Stat. 16 (2012) 99–106.

[28] E.A. Catchpole, B.J.T. Morgan, Detecting parameter redundancy, Biometrika 84 (1) (1997) 187–196.

[29] D.J. Cole, B.J.T. Morgan, D.M. Titterington, Determining the parametric structure of models, Math. Biosci. 228 (1) (2010) 16–30.

[30] B.G. Hu, What are the differences between Bayesian classifiers and mutual-information classifiers, IEEE Trans, Neural Networks and Learning Systems 25 (2) (2014) 249–264 ⟨http://arxiv.org/abs/1105.0051v2⟩.

[31] C.R. Rao, Linear Statistical Inference and Its Applications, Wiley, New York, 1973.

[32] T.J. Moore, A theory of Cramér-Rao bounds for constrained parametric models (Ph.D. thesis), University of Maryland, 2010.

[33] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, 1985.

[34] F. Lauer, G. Bloch, Incorporating prior knowledge in support vector machines for classification: a review, Neurocomputing 71 (7–9) (2008) 1578–1594.

[35] J.J. Duistermaat, J.A.C. Kolk, Multidimensional Real Analysis, Cambridge University Press, 2004.

[36] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1985.

[37] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[38] T. van Erven, P. Harremoës, Renyi divergence and Kullback-Leibler divergence, ⟨http://arxiv.org/abs/1206.2459v1⟩, http://dx.doi.org/10.1109/TIT.2014.2320500.
[39] S. Watanabe, Algebraic geometrical methods for hierarchical learning machines, Neural Netw. 14 (8) (2001) 1049–1060.
[40] J.D. Gorman, A.O. Hero, Lower bounds for parametric estimation with constraints, IEEE Trans. Inf. Theory 26 (1990) 1285–1301.
[41] G.A.F. Seber, C.J. Wild, Nonlinear Regression, Wiley, New York, 2003.
[42] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 2005.
[43] C.A. Micchelli, Interpolation of scattered data: distance matrices and conditionally positive definite functions, Constr. Approx. 2 (1986) 11–22.
[44] H. Miao, X. Xia, A.S. Perelson, H. Wu, On identifiability of nonlinear ODE models and applications in viral dynamics, SIAM Rev. 53 (1) (2011) 3–39.

Zhi-Yong Ran received his M.Sc. degree in applied mathematics from the Beijing University of Technology, Beijing, China, in 2007. Currently he is a Ph.D. candidate at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include parameter identifiability theory and machine learning.

Bao-Gang Hu (M'94-SM'99) received the M.Sc. degree from the University of Science and Technology, Beijing, China, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, both in mechanical engineering, in 1983 and 1993, respectively. He was a Research Engineer and Senior Research Engineer at C-CORE, Memorial University of Newfoundland, St. John's, NF, Canada, from 1994 to 1997. From 2000 to 2005, he was the Chinese Director of computer science, control, and applied mathematics with the Chinese-French Joint Laboratory, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor at NLPR. His current research interests include pattern recognition and plate growth modeling.