2018 26th Mediterranean Conference on Control and Automation (MED)
Zadar, Croatia, June 19-22, 2018

Ranking Variables Based on Goodness of Fits in Nonlinear Nonparametric System Identification

Er-wei Bai, Changmin Cheng

***Abstract*— Identification of a high dimensional nonlinear nonparametric system is costly. On the other hand for many real-world problems, they are sparse in the sense that not all variables contribute or contribute significantly. If these variables that do not contribute or contribute little can be identified and removed prior to system identification, the identification problem is of lower dimension. In this paper, methods to rank variables based on the Goodness of Fits are proposed without full scale identification.**

I. Introduction

Our goal is identification of a scalar discrete nonlinear non-parametric system

$$y(k) = f(x(k)) + v(k) = f(x_1(k), x_2(k), ..., x_p(k)) \quad \text{(I.1)}$$
$$+v(k),\ k = 1, 2, ..., N$$

where $y(\cdot)$ is the system output and $v(\cdot)$ is an iid noise sequence of zero mean and finite variance σ^2. $x(k) = (x_1(k), ..., x_p(k))$ is the regressor that consists of possibly contributing input variables. The function $f(\cdot)$ is unknown that makes identification nontrivial. Throughout the paper, the system is assumed to be asymptotically stationary which is a common assumption in nonlinear system identification.

The system (I.1) represents a large class of nonlinear systems. The purpose of nonlinear nonparametric identification is to estimate the unknown f based on the available data set $\{y(k), x(k)\}_{k=1}^N$. Clearly, one of the main difficulties is the lack of the structure of f. A very popular approach in the literature is to assume that the unknown system f can be represented by a linear combination of some possibly nonlinear but known basis functions [Hong (2008)], [Peng (2006)]. Therefore, a nonlinear nonparametric identification problem becomes a linear parametric problem. What remain unknown are the coefficients of the basis functions. This problem is linear and can be solved by a number of well developed linear techniques, e.g., the least squares [Bai (2007)]. A big problem with this approach is that in order to have good basis functions, a priori knowledge about the unknown function $f(\cdot)$ must be available which may or may not be practical.

The other popular approach is to apply nonparametric estimation techniques, e.g., the celebrated kernel, local polynomial and statistic approximation methods [Bai (2007)], [Bai (2010)], [Bai (2014)], [Fan (2005)], [Pillonetto (2011)], [Sjoberg (1995)], [Zhao (2015)]. All these methods are based on local averages that suffer from the curse of dimensionality [Bai (2010)], [Bai (2017)], [Zhao (2015)]. A manifestation of the curse of dimensionality is that the data length N needed has to grow exponentially as a function of the dimension p that is practically impossible. The curse of dimensionality problem is fundamental and makes all local average based methods impractical if the dimension p is high.

For many real-world problems, however, they are sparse in the sense that not all variables x_i's contribute. If these variables that do not contribute can be identified and removed prior to system identification, the identification problem is of lower dimension. Moreover, even if some variables contribute but only contribute marginally, by removing those variables would only have a minimal effect on the output. If those variables can be identified and removed prior to identification, the dimension of the system identification problem can be further lowered, and the problem therefore suffers less from the curse of dimensionality. Here we emphasize the word "prior to identification". One certainly can identify all the variables that do not contribute or contribute marginally if an accurate model $f(x_1, ..., x_p)$ is available. How to estimate f with a finite data length is a key when p is high. Because of the curse of dimensionality, an accurate estimate of f is unlikely to obtain unless N is very long. In other words, it is costly to have a reasonable estimate of f when p is high. By identifying and removing variables that do not contribute or contribute little prior to identification reduces the dimension of the problem that alleviates or eliminates the effect of the curse of dimensionality. The fact is that identifying variables that do not contribute or contribute little is likely a much easier problem than the full dimensional identification of a nonlinear nonparametric function $f(x_1, ..., x_p)$.

Identifying and eliminating variables is the topic of variable selection. System identification is a very active research area over the last a few decades. On the other hand except order estimation, variable selection has only received scattered attention in the nonlinear identification literature [Bai (2014)], [Zhao (2015)], [Peduzzi (1980)]. Unsurprisingly, variable selection problem has been studied in the statistical and other literature [Lind (2008)], [Roll (2005)]. The most common one is ANOVA (the analysis of variance) [Lind (2008)], [Roll (2005)]. The system (I.1) can be rewritten as

$$y(k) = \sum_{i=1}^{p} f_i(x_i(k)) + \sum_{i<j} f_{ij}(x_i(k), x_j(k))$$
$$+ ... + f_{12...,p}(x_1(k), ..., x_p(k)) + v(k) \quad \text{(I.2)}$$

This paper was supported in part by grants NSF CNS-1239509

The authors are with Dept. of Electrical and Computer Engineering University of Iowa, Iowa City, Iowa 52242, er-wei-bai@uiowa.edu, Tel:+3193355949 and Fax: +3193356028, Shanghai Jiaotong University, Shanghai, China

978-1-5386-7891-6/18/$31.00 ©2018 IEEE

where f_i's are 1-factor terms, f_{ij}'s 2-factor terms and so on. For simplicity, consider a case of $p=3$ and let τ, β, γ denote x_1, x_2, x_3 respectively. The effect of the output y can be decomposed into

$$\begin{aligned} y_{ijlk} &= \mu + \tau_i + \beta_j + \gamma_l + (\tau\beta)_{ij} + (\tau\gamma)_{il} \\ &+ (\beta\gamma)_{jl} + (\tau\beta\gamma)_{ijl} + v_{ijlk},\ k=1,...,N \end{aligned} \tag{I.3}$$

where $i=1,...,a, j=1,...,b, l=1,...,c$. In (I.3), μ is the overall mean effect, τ_i is the effect of ith level of the variable x_1, β_j is the effect of jth level of the variable x_2 and γ_l is the effect of lth level of the variable x_3. Similarly, $(\tau\beta)_{ij}$, $(\tau\gamma)_{il}$, $(\beta\gamma)_{jl}$ and $(\tau\beta\gamma)_{ijl}$ are the effects of interactions of x_1x_2 of the ijth level, x_1x_3 of the ilth level, x_2x_3 of the jlth level and $x_1x_2x_3$ of ijlth level respectively. Whether the contribution by τ_i, β_j, γ_l, $(\tau\beta)_{ij}$, $(\tau\gamma)_{il}$, $(\beta\gamma)_{jl}$ or $(\tau\beta\gamma)_{ijl}$ is significant or not can be determined by the hypothesis test based on the F-distribution. ANOVA is a powerful tool in determining the contribution of each term in (I.3) though not perfect. First, it requires Gaussian assumption. Second, for a random and continuous input x_i, it has to be quantized into a discrete set of levels in order to apply ANOVA. More critically, the number of terms in (I.3) grows exponentially as the dimension p increases. Practically, ANOVA is limited to a reasonably sized dimension p.

We will focus on the Goodness of Fits in ranking variables.

II. Ranking based on the Goodness of Fits

To convey the ideas clearly, we consider an additive nonlinear system first that is the most applied and investigated nonlinear system in the literature [Bai (2008)], [Chen (1995)], [Fan (2005)],

$$\begin{aligned} y(k) &= f(x_1(k), ..., x_p(k)) + v(k),\ k=1,...,N \\ &= f_1(x_1(k)) + f_2(x_2(k)) + ... + f_p(x_p(k)) + v(k), \end{aligned} \tag{II.4}$$

where f_i's are unknown to be estimated and $v(\cdot)$ is a random sequence of zero mean and finite variance σ^2. It is assumed in this section that x_i's are statistically independent. This condition will be relaxed later. Also, how to check a nonlinear system is additive or not will be discussed later. Further to avoid unnecessary complications, we assume that

$$Ey = 0,\ E(f_i(x_i)) = 0,\ i=1,...,p$$

where E is the expectation operator. This is not a restriction at all. An arbitrary additive system

$$\bar{y}(k) = \bar{f}_1(x_1(k)) + ... + \bar{f}_p(x_p(k)) + v(k)$$

can always be written as

$$\underbrace{\bar{y}(k) - \sum_{i=1}^{p} E\bar{f}_i(x_i)}_{y(k)} = \sum_{i=1}^{p} \underbrace{(\bar{f}_i(x_i(k)) - E\bar{f}_i(x_i))}_{f_i(x_i(k))} + v(k)$$

which is exactly in the form of (II.4). In implementation, $y(k)$ can be obtained by $\bar{y}(k) - \frac{1}{N}\sum_{i=1}^{N}\bar{y}(i)$ since $\frac{1}{N}\sum_{i=1}^{N}\bar{y}(i) \to \sum_{i=1}^{p} E\bar{f}_i(x_i)$ in probability as $N \to \infty$ by the law of large numbers.

In the system identification, the most important quality measure is the Goodness of Fits (GOF) [Soderstrom (1989)]. Let

$$\hat{y}(k) = \hat{f}_1(x_1(k)) + \hat{f}_2(x_2(k)) + ... + \hat{f}_p(x_p(k))$$

be the predicted output based on the estimates $\hat{f}_i$'s of f_i's. The Goodness of Fits is defined as

$$GOF(\hat{f}_1, ..., \hat{f}_p) = 1 - \sqrt{\frac{E(y(k) - \hat{y}(k))^2}{Var(y)}} \tag{II.5}$$

where $Var(y)$ is the variance of y. In practice for a finite length data N, (II.5) is often calculated as

$$GOF(\hat{f}_1, ..., \hat{f}_p) = 1 - \sqrt{\frac{\frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{y}(k))^2}{\frac{1}{N}\sum_{k=1}^{N}(y(k) - \frac{1}{N}\sum_{i=1}^{N}y(i))^2}}$$

$GOF(\hat{f}_1, ..., \hat{f}_p)$ measures how close the predicted $\hat{y}$ is to the actual y. If $y \equiv \hat{y}$, $GOF(\hat{f}_1, ..., \hat{f}_p) = 1$. If $\hat{y}$ is close to y, the GOF is expected to be close to 1.

For an additive system under the assumption that all x_i's are independent, the contribution of x_i in the absence of all other variables $x_j, j \neq i$, is obviously $f_i(x_i)$. This implies that the $GOF(x_i)$ when only x_i contributes is given by

$$GOF(x_i) = 1 - \sqrt{\frac{E(y(k) - f_i(k))^2}{Var(y)}} \tag{II.6}$$

$GOF(x_i)$ is used to compare the contribution of each variable x_i. In the sense of GOF, we say the contribution of x_i is larger than that of x_j if and only if $GOF(x_i) > GOF(x_j)$. Note in (II.6), the output y is given. Thus, $GOF(x_i)$ is completely determined by the term

$$E(y(k) - f_i(x_i(k)))^2 = \sum_{j \neq i} Ef_j^2(x_j) + \sigma^2$$

Since $E(y(k) - f_i(x_i(k)))^2$ and $GOF(x_i)$ move in an opposite way, a variable x_i has the largest contribution in term of GOF

$$\begin{aligned} & i = arg\max_i GOF(x_i) \\ \Longleftrightarrow\ & i = arg\min_i \sum_{j \neq i} Ef_j^2(x_j) \\ \Longleftrightarrow\ & i = arg\max_i Ef_i^2(x_i) \end{aligned}$$

By the assumption that x_i's are independent and $Ef_i(x_i) = 0$, it follows that

$$f_i(x_i) = E(y \mid x_i)$$

We now define the importance measure $GOFM(x_i)$ based on the Goodness of Fits as

$$GOFM(x_i) = E[E^2(y \mid x_i)] = Ef_i^2(x_i) \tag{II.7}$$

The exact relationship between $GOF(x_i)$ and $GOFM(x_i)$ can be established,

$$GOF(x_i) = 1 - \sqrt{\frac{E(y(k) - f_i(k))^2}{Var(y)}}$$

$$= 1 - \sqrt{\frac{\sum_{j=1}^{p} Ef_j^2(x_j) + \sigma^2 - Ef_i^2(x_i)}{\sum_{j=1}^{p} Ef_j^2(x_j) + \sigma^2}}$$

$$= 1 - \sqrt{1 - \frac{Ef_i^2(x_i)}{Var(y)}}$$

$$= 1 - \sqrt{1 - \frac{GOFM(x_i)}{Var(y)}}$$

where $0 \le GOFM(x_i)/Var(y) \le 1$. When $GOFM(x_i) = Var(y)$, i.e, all contributions are from x_i, $\frac{GOFM(x_i)}{Var(y)} = 1$ and $GOF(x_i) = 1$. If x_i does not contribute, $\frac{GOFM(x_i)}{Var(y)} = 0$ and $GOF(x_i) = 0$. In conclusion, $GOFM(x_i)$ is an simply alternative representation of $GOF(x_i)$.

Now, if we order $GOFM(x_i)$ as

$$GOFM(x_{j_1}) \ge GOFM(x_{j_2}) \ge ... \ge GOFM(x_{j_p})$$

then,the contribution of x_i to y is in the order of

$$x_{j_1}, x_{j_2}, ..., x_{j_p} \tag{II.8}$$

in the sense of GOF. Further, given a $d > 0$,

$$GOF(x_{j_1}, x_{j_2}, ..., x_{j_d}) \ge GOF(x_{i_1}, x_{i_2}, ..., x_{i_d})$$

for any d-subset $(i_1, i_2, ..., i_d) \in (1, 2, ..., p)$ as shown in (II.8). This property is very useful because the best d-subset $j_1, ..., j_d$ can be chosen one at a time by finding the largest $GOFM(x_i)$ or equivalently $GOF(x_{i_j})$ in the remaining set.

As discussed, $GOFM(x_i)$ can be used to determine which variables x_i's contribute, and which variables x_i's do not contribute or contribute a little, and therefore can be eliminated prior to actual nonlinear system identification. It can be done in two ways, individual contribution or accumulative contribution. For individual contribution, let d_1 be the threshold, say $d_1 = 0.03$ or 3%. If $GOFM(x_i)/Var(y) < d_1$, the variable x_i is considered to have no contribution or contribute a little and so can be eliminated. For accumulative contribution, let d_2 be the threshold, say $d_2 = 0.95$ or 95% and arrange the contribution of x_i's in the order of (II.8). Let d be the smallest integer such that

$$\frac{GOFM(x_{j_1})}{Var(y)} + ... + \frac{GOFM(x_{j_d})}{Var(y)} \ge d_2$$

Then, all the variables $x_{j_{d+1}}, ..., x_{j_p}$ are considered to have a little contribution and can be eliminated prior to identification.

Once defined, the next question is how to calculate $GOFM(x_i)$ for each i based on the available data $\{y(k), x(k)\}_{k=1}^{N}$. The problem is that $f_i(x_i(k))$ is not available. However, (II.7) provides a numerical algorithm to calculate $GOFM(x_i)$'s by replacing the statistical averages by the sampled means.

Algorithm to calculate $GOFM(x_i) = E(E^2(y|x_i))$, $i = 1, 2, ..., p$, based on the data

$$\{y(k), x(k)\}_{k=1}^{N}. \tag{II.9}$$

Step 1: Let β be any value satisfying $0 < \beta < 1$. Choose N so that N^β is an integer. For instance, when $\beta = 1/2$, N can be $4, 9, 25, 36, ...$ Collect data $\{y(k), x(k)\}_{k=1}^{N}$.

Step 2: For each $i = 1, 2, ..., p$, divide the range of x_i into H_i non-overlap slices, $I_1(1), ..., I_i(H_i)$. Let the number of $x_i(k) \in I_i(h)$, $h = 1, 2, ..., H_i$ that falls into each slide $I_i(h)$ be $l_i(h) = N^\beta$, $i = 1, ..., p, h = 1, ..., H_i$.

Step 3: Within each slide $I_i(h)$, $h = 1, ..., H_i$, compute the sampled mean of $E^2(y \mid x_i)$ by

$$M_i(h) = (\frac{1}{l_i(h)} \sum_{x_i(k) \in I_i(h)} y(k))^2$$

$$= (\frac{1}{N^\beta} \sum_{x_i(k) \in I_i(h)} y(k))^2$$

Step 4: Calculate the sampled mean of $E(E^2(y \mid x_i)) = Ef_i^2(x_i)$ by

$$\widehat{GOF}M(x_i) = \sum_{h=1}^{H_i} \frac{l_i(h)}{N} M_i(h)$$

$$= \sum_{h=1}^{N^{1-\beta}} \frac{N^\beta}{N} M_i(h)$$

Theorem 2.1: Assume that $E(y \mid x_i)$ is Lipschitz and $E|y| < \infty$. Then, in probability as $N \to \infty$ for each i

$$\widehat{GOF}M(x_i) \to GOFM(x_i) = Ef_i^2(x_i)$$

Proof: Since $N = N^{1-\beta}N^\beta$, $H_i = N^{1-\beta}$ and $l_i(h) = N^\beta$ are integers, and $H_i = N^{1-\beta}$, $l_i(h) = N^\beta \to \infty$, $l_i(h)/N \to 0$ as $N \to \infty$. Further $l_i(h) = N^\beta$ is regular defined by [Walk (2008)]. Thus the convergence of $M_i(h) \to E^2(y|x_i)$ comes from Theorem 1 of [Walk (2008)]. The convergence of $\widehat{GOF}M(x_i) \to GOFM(x_i)$ follows from the facts that $M_i(h) \to E^2(y|x_i)$, $H_i \to \infty$ and the law of large numbers. This completes the proof.

We emphasize again that $\widehat{GOF}M(x_i)$ for each i can be calculated without full scale identification of f or f_i's. In short, the contribution of each variable x_i can be ranked prior to system identification.

III. General Nonlinear Systems

In this section, we extend the idea and calculation to a general nonlinear nonparametric system.

If input variables are correlated or interact even independent, ranking the contribution of each variable x_i in terms of Goodness of Fits is very hard if possible. The contribution of one variable is often coupled with the contribution of other variables. Consider a system with independent x_i's,

$$y(k) = f_1(x_1(k)) + f_2(x_2(k)) + f_{13}(x_1(k), x_3(k))$$

$$= x_1(k) + x_2(k) + x_1(k)x_3(k)$$

Suppose the amplitude of x_1 is so small and negligible compared to that of x_2. However, the amplitude of x_3 is large and the product term x_1x_3 could be dominate. Thus, there is no clear answer to the question how to evaluate the contribution of x_1. On the other hand, if we give up the concept of the contribution of each variable and instead adopt the concept of the contribution of each term, the contribution of each term f_1, f_2, f_{13} can be evaluated. To this end, we consider a general system again with upto 2-factor terms and independent x_i's.

$$y(k) = \sum_{i=1}^{p} f_i(x_i(k)) + \sum_{i<j} f_{ij}(x_i(k), x_j(k)) + v(k)$$

$$= \sum_{i=1}^{M} \phi_i(k) + v(k) \tag{III.10}$$

where $M = \frac{p(p+1)}{2}$.

The reason for considering the system upto 2-factor terms is for simplicity. All results that will be derived can be trivially but cumbersomely extended to any nonlinear systems. Now we assume that all the terms $\phi_i(k)$ in (III.10) satisfy

$$E\phi_i(k) = 0,\ E\phi_i(k)\phi_j(k) = 0$$

for all $1 \le i, j \le p$ and $i \ne j$.

This is not a restriction at all. Any system can be normalized to have this property. To be more precisely, consider an arbitrary system upto 2 factor-terms,

$$\bar{y}(k) = \sum_{i=1}^{p} \bar{f}_i(x_i(k)) + \sum_{i<j} \bar{f}_{ij}(x_i(k), x_j(k)) + v(k) \tag{III.11}$$

Let the conditional expectations for given x_{j_1}, and/or x_{j_2} be respectively,

$$E(y \mid x_{j_1}),\ E(y \mid x_{j_1}, x_{j_2}),$$

$$E(f_{j_1 j_2}(x_{j_1}, x_{j_2}) \mid x_{j_1}),\ and\ E(f_{j_1 j_2}(x_{j_1}, x_{j_2}) \mid x_{j_2}).$$

Now define

$$\begin{array}{l}
f_{j_1 j_2}(x_{j_1}, x_{j_2}) = \bar{f}_{j_1 j_2}(x_{j_1}, x_{j_2}) \\
-E(\bar{f}_{j_1 j_2}(x_{j_1}, x_{j_2}) \mid x_{j_2}) - E(\bar{f}_{j_1 j_2}(x_{j_1}, x_{j_2}) \mid x_{j_1}) \\
+E\bar{f}_{j_1 j_2}(x_{j_1}, x_{j_2}), \quad 1 \le j_1 < j_2 \le p \\
f_1(x_1) = \bar{f}_1(x_1) + \sum_{i=2}^{p} E(\bar{f}_{1i}(x_1, , x_i) \mid x_1) \\
-E\{\bar{f}_1(x_1) + \sum_{i=2}^{p} E(\bar{f}_{1i}(x_1, x_i) \mid x_1)\} \\
f_j(x_j) = \bar{f}_j(x_j) + \sum_{i=j+1}^{p} E(\bar{f}_{ji}(x_j, x_i) \mid x_j) \\
+\sum_{i=1}^{j-1} E(\bar{f}_{ij}(x_i, x_j) \mid x_j) \\
-E\{\bar{f}_j(x_j) + \sum_{i=j+1}^{p} E(\bar{f}_{ji}(x_j, x_i) \mid x_j) \\
+\sum_{i=1}^{j-1} E(\bar{f}_{ij}(x_i, x_j) \mid x_j)\},\ \ j = 2, 3, ..., p-1 \\
f_p(x_p) = \bar{f}_p(x_p) + \sum_{i=1}^{p-1} E(\bar{f}_{ip}(x_i, x_p) \mid x_p) \\
-E\{\bar{f}_p(x_n) + \sum_{i=1}^{p-1} E(\bar{f}_{ip}(x_i, x_p) \mid x_p)\}
\end{array}$$

Now, we are in a position to define data-dependent orthogonal basis functions ϕ_i, $i = 1, ..., M$.

$$\begin{array}{lcl}
y(k) & = & \bar{y}(k) - E\bar{y} \\
\phi_j(x_j) & = & f_j(x_j), \\
 & \Longrightarrow & \phi_1, ..., \phi_p, \\
 & & j = 1, 2, ..., p \\
\phi_{\frac{2p}{2}-1+j}(x_1, x_j) & = & f_{1j}(x_1, x_j), \\
 & \Longrightarrow & \phi_{p+1}, ..., \phi_{2p-1} \\
 & & j = 2, ..., p \\
\phi_{\frac{2p-1}{2}2-2+j}(x_2, x_j) & = & f_{2j}(x_2, x_j), \\
 & \Longrightarrow & \phi_{2p}, ..., \phi_{3p-3}, \\
 & & j = 3, ..., p \\
\phi_{\frac{2p-2}{2}3-3+j}(x_3, x_j) & = & f_{3j}(x_3, x_j), \\
 & \Longrightarrow & \phi_{3p-2}, ..., \phi_{4p-6}, \\
 & & j = 4, ..., p \\
\vdots & \vdots & \vdots
\end{array} \tag{III.12}$$

$$\begin{array}{ll}
\phi_{\frac{2p-(p-3)}{2}(p-2)-(p-2)+j}(x_{p-2}, x_j) & = f_{(p-2)j}(x_{p-2}, x_j), \\
\Longrightarrow & \phi_{\frac{p^2+p}{2}-2}, \phi_{\frac{p^2+p}{2}-1}, \\
 & j = p-1, p \\
\phi_{\frac{2p-(p-2)}{2}(p-1)-(p-1)+j}(x_{p-1}, x_j) & = f_{(p-1)j}(x_{p-1}, x_j), \\
 & \Longrightarrow \phi_{\frac{p^2+p}{2}}, \\
 & j = p
\end{array} \tag{III.13}$$

Clearly, $\phi_j(x_j)$'s, $j = 1, ..., p$, represent the 1-factor terms and $\phi_i(x_{j_1}, x_{j_2})$'s, $i = p+1, ..., M$, are 2-factor terms. When the meaning is clear from the context, with a little abuse of notion, we interchangeably use

$$\begin{array}{ll}
\phi_j[k] & = \phi_j(x_j(k)),\ \ j = 1, ..., p \\
\phi_j[k] & = \phi_j(x_1(k), x_{j-p+1}(k)),\ \ j = p+1, ..., 2p-1 \\
\phi_j[k] & = \phi_j(x_2(k), x_{j-2p+3}(k),\ \ j = 2p, ..., 3p-3 \\
 & \vdots \\
\phi_j[k] & = \phi_j(x_{p-2}(k), x_{j-M+p+1}(k),\ \ j = M-2, M-1 \\
\phi_j[k] & = \phi_j(x_{p-1}(k), x_p(k)),\ \ j = M = p(p+1)/2.
\end{array}$$

In implementation $y(k) = \bar{y}(k) - E\bar{y}$ can be replaced by $y(k) = \bar{y}(k) - \frac{1}{N}\sum_{i=1}^{N} \bar{y}(i)$. We have the following result.

Theorem 3.1: Consider the system (III.11). Then we have:

1) The system (III.11) can be represented by the data driven basis functions ϕ_i's,

$$y[k] = \sum_{i=0}^{M} \phi_i[k] + v[k] \tag{III.14}$$

where $M = p + p(p-1)/2 = p(p+1)/2$.

2) The data driven basis functions ϕ_i's are orthogonal. i.e., for all $1 \le j \le M$ and $0 \le j_1 < j_2 \le M$,

$$E\phi_j[k] = 0, \quad E\phi_{j_1}[k]\phi_{j_2}[k] = 0.$$

3) The unknown ϕ_j's are the expectations or conditional

expectations of the output,

$$\begin{aligned}
&\phi_j(x_j) = E(y[k] \mid x_j),\ j = 1, ..., p,\\
&\phi_{\frac{2n}{2}-1+j}(x_1, x_j)\\
&\quad = E(y[k] \mid x_1, x_j) - \phi_1(x_1) - \phi_j(x_j)\ j = 2, ..., p,\\
&\phi_{\frac{2p-1}{2}2-2+j}(x_2, x_j)\\
&\quad = E(y[k] \mid x_2, x_j) - \phi_2(x_2) - \phi_j(x_j),\ j = 3, ..., p,\\
&\qquad\vdots\\
&\phi_{\frac{2p-(p-3)}{2}(p-2)-(p-2)+j}(x_{p-2}, x_j)\\
&\quad = E(y[k] \mid x_{p-2}, x_j) - \phi_{p-2}(x_{p-2}) - \phi_j(x_j),\\
&\qquad\qquad j = p-1, p,\\
&\phi_{\frac{2p-(p-2)}{2}(p-1)-(p-1)+j}(x_{p-1}, x_j)\\
&= E(y[k] \mid x_{p-1}, x_j) - \phi_{p-1}(x_{p-1}) - \phi_j(x_j),\ j = p
\end{aligned}$$

Proof: The first part is directly from the definition of ϕ_i's. Also from the definition, it is easily verified that $E\phi_j[k] = 0$ for $j = 1, ..., p$. $E\phi_j[k] = 0$, $j = p+1, ..., M$ follows from $Ef_{j_1 j_2}(x_{j_1}, x_{j_2}) = 0$. We now show $E\phi_{j_1}[k]\phi_{j_2}[k] = 0$. For $0 \le j_1 < j_2 \le p$, $E\phi_{j_1}[k]\phi_{j_2}[k] = E\phi_{j_1}[k]E\phi_{j_2}[k] = 0$ because of independence of x_{j_1} and x_{j_2}. The proofs for other j_1 and j_2 follow from the same arguments as

$$E\phi_1[k]\phi_{p+1}[k] = E\phi_1(x_1(k))\phi_{p+1}(x_1(k), x_2(k))$$

$$= E\{\phi_1(x_1(k))E\{\phi_{p+1}(x_1(k), x_2(k) \mid x_1(k)\}\} = 0.$$

To show the third part, observe

$$y[k] = \sum_{j=1}^{p} f_j(x_j(k)) + \sum_{1 \le j_1 < j_2 \le p} f_{j_1 j_2}(x_{j_1}(k), x_{j_2}(k)) + v[k],$$

$$E(y[k] \mid x_j) = f_j(x_j) = \phi_j(x_j),\ j = 1, ..., p$$

$$E(y[k] \mid x_{j_1}, x_{j_2}) = f_{j_1}(x_{j_1}) + f_{j_2}(x_{j_2}) + f_{j_1 j_2}(x_{j_1}, x_{j_2})$$

$$= \phi_{j_1}(x_{j_1}) + \phi_{j_2}(x_{j_2}) + f_{j_1 j_2}(x_{j_1}, x_{j_2}),\ 1 \le j_1 < j_2 \le p$$

Then, the conclusion follows from the definition of ϕ_j's.

By the theorem, we now have

$$Ey^2(k) = \sum_{i=1}^{M} \phi_i^2(k) + \sigma^2$$

The contribution of ϕ_i in the absence of all other $\phi_j, j \ne i$, is obviously ϕ_i. This implies that the $GOF(x_i)$ when only ϕ_i contributes is given by

$$GOF(\phi_i) = 1 - \sqrt{\frac{E(y(k) - \phi_i(k))^2}{Var(y)}} \qquad \text{(III.15)}$$

In the sense of GOF, we say the contribution of ϕ_i is larger than that of ϕ_j if and only if $GOF(\phi_i) > GOF(\phi_j)$. Similarly, we define the importance measure $GOFM(\phi_i)$ based on the Goodness of Fits as

$$GOFM(\phi_i) = E\phi_i^2 \qquad \text{(III.16)}$$

The exact relationship between $GOF(\phi_i)$ and $GOFM(\phi_i)$ can be similarly established,

$$\begin{aligned}
GOF(\phi_i) &= 1 - \sqrt{\frac{E(y(k) - \phi_i(k))^2}{Var(y)}}\\
&= 1 - \sqrt{\frac{\sum_{j=1}^{p} E\phi_j^2(k) + \sigma^2 - E\phi_i^2(k)}{\sum_{j=1}^{p} E\phi_j^2(k) + \sigma^2}}\\
&= 1 - \sqrt{1 - \frac{E\phi_i^2(k)}{Var(y)}}\\
&= 1 - \sqrt{1 - \frac{GOFM(\phi_i)}{Var(y)}}
\end{aligned}$$

Similarly, $GOFM(\phi_i)$ can be used to determine which terms ϕ_i's contribute, and which terms ϕ_i's do not contribute or contribute a little and therefore can be eliminated prior to actual nonlinear system identification. It can be done in two ways, individual contribution or accumulative contribution. For individual contribution, let d_1 be the threshold, say $d_1 = 0.03$ or 3%. If $GOFM(\phi_i)/Var(y) < d_1$, the term ϕ_i is considered to have no contribution or contribute a little and so can be eliminated. For accumulative contribution, let d_2 be the threshold, say $d_2 = 0.95$ or 95% and arrange the contribution of ϕ_i's in the order of

$$GOFM(\phi_{j_1}) \ge GOFM(\phi_{j_2}) \ge ... \ge GOFM(\phi_{j_M})$$

Let d be the smallest integer such that

$$\frac{GOFM(\phi_{j_1})}{Var(y)} + ... + \frac{GOFM(\phi_{j_d})}{Var(y)} \ge d_2$$

Then, all the terms $\phi_{j_{d+1}}, ..., \phi_{j_M}$ are considered to have a little contribution and can be eliminated prior to identification.

Now what left is to find the estimates of $GOFM(\phi_i) = E\phi_i^2$ based on the available data $\{y(k), x(k)\}$. Clearly

$$E\phi_j^2 = E(E^2(y|x_j)),\ j = 1, ..., p$$

and their estimates can be exactly calculated by Algorithm (II.9) with the convergence. For $m > p$, $\phi_m(x_i, x_j)$'s are in the form of

$$\phi_m(x_i, x_j) = E(y|x_i, x_j) - \phi_i(x_i) - \phi_j(x_j)$$

It follows that

$$E\phi_m^2(x_i, x_j) =$$
$$E(E^2(y|x_i, x_j)) - E(E^2(y|x_i)) - E(E^2(y|x_j))$$

The last two terms can be again calculated by Algorithm (II.9). The first term can be calculated in a similar way as in Algorithm (II.9).

Algorithm to calculate $E(E^2(y|x_i, x_j))$.

Step 1: Let β be any value satisfying $0 < \beta < 1$. Choose N so that N^β is an integer.

Step 2: Divide the range of x_i into H_i non-overlap slices, $I_1(1), ..., I_i(H_i)$. Let the number of $x_i(k) \in I_i(h),\ h = 1, 2, ..., H_i$ that falls into each slide $I_i(h)$ be $l_i(h) = N^\beta$, $i = 1, ..., p, h = 1, ..., H_i$.

Step 3: Within each pair of slides $I_i(h_1), I_j(h_2)$, let $I_{[x_i(k) \in I_i(h_1), x_j(k) \in I_j(h_2)]}$ be the indicator function and

$$I_{ij}(h_1, h_2) = \sum_{k=1}^{N} I_{[x_i(k) \in I_i(h_1), x_j(k) \in I_j(h_2)]}$$

	f_1	f_2	f_3	f_4	f_5	f_{45}
$\frac{GOFM}{Var(y)}$	.3581	.1842	.2710	.0064	.0057	.1964

TABLE I
THE GOFM OF EACH TERM

Compute the sampled mean of $E^2(y|x_i, x_j)$ by

$$M_{ij}(h_1,h_2) = \left(\frac{1}{I_{ij}(h_1,h_2)} \cdot \sum_{x_i(k)\in I_i(h_1), x_j(k)\in I_j(h_2)} y(k)\right)^2.$$

Step 4: Calculate the sampled mean of $E(E^2(y|x_i, x_j))$ by

$$\sum_{h_1=1}^{N^{1-\beta}} \sum_{h_2=1}^{N^{1-\beta}} \frac{I_{ij}(h_1,h_2)}{N} M_{ij}(h_1,h_2)$$

This result holds for a general nonlinear nonparametric system with dependent variables.

IV. NUMERICAL SIMULATION

Consider an 5-dimensional system

$$\begin{aligned} y(k) &= f\big(x_1(k), x_2(k), ..., x_5(k)\big) + v(k) \\ &= a_1(x_1(k))^2 + a_2 e^{x_2(k)} + a_3 \cos(x_3(k)) \\ &+ a_4 x_4(k) x_5(k) + v(k), k = 1, ..., 3600 \end{aligned} \quad \text{(IV.17)}$$

In simulation, $a_1 = 1, a_2 = 0.5, a_3 = 2.7, a_4 = 1$.

The contribution of each terms were calculated as in Eq.(IV.17). The results are shown in Table 1. Since $GOFM(f_4(x_4)) \approx 0$ and $GOFM(f_5(x_5)) \approx 0$, terms $f_4(x_4)$ and $f_5(x_5)$ don't exist. Therefore, there are only terms $f_1(x_1)$,$f_2(x_2)$,$f_3(x_3)$ and $f_{4,5}(x_4, x_5)$ in the system.

V. CONCLUDING REMARKS

In this paper, the rankings of variables are studied in a system identification setting. The idea is that with the ranking, variables that contribute significantly rank ahead of those that do not contribute or contribute only marginally. Therefore variable selection can be carried out based on the ranking prior to system identification.

REFERENCES

[Bai (2008)] Bai, E.W. and K Chan (2008) "Identification of an additive *Automatica*, **44**, pp.430-436

[Bai (2007)] Bai, E.W. and Y. Liu (2007) "Recursive direct weight optimization in nonlinear system identification: a minimal probability approach", *IEEE Trans on Automatic Control*, **52**, pp1218-1231

[Bai (2010)] Bai, E.W. (2010) "Non-Parametric Nonlinear System Identification: An Asymptotic Minimum Mean Squared Error Estimator", *IEEE Trans on Automatic Control*, **55**, pp.1615-1626

[Bai (2014)] Bai,E.W., K Li, W Zhao and W Xu (2014) "Kernel based approaches to local nonlinear Nonparametric variable selection", *Automatica*, **50**, pp.100-113

[Bai (2017)] Bai, E.W., C. Cheng and W. Zhao (2017) "Variable Selection of High-Dimensional Non-Parametric Nonlinear Systems by Derivative Averaging to Avoid the Curse of Dimensionality", *IEEE Conf on Decision and Control*

[Chan (2003)] Chan, K, A. Kristoffersen and N.Stenseth (2003) "Burmann expansion and test for additivity", *Biometrika*, **90**, pp.209-222

[Chen (1995)] Chen,R., J. Liu and R. Tsay (1995) "Additivity tests for nonlinear autoregression", *Biometrika*, **82**, pp.369-383

[Fan (2005)] Fan, J.Q. and Q.W. Yao *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag, 2005.

[Hong (2008)] Hong, X, Mitchell, S. Chen, C. Harris, K. Li and G.W Irwin (2008) "Model selection approaches for nonlinear system identification:a review", *Int. J of System Science*, **39**, pp.925-949

[Peng (2006)] Li, K, J. Peng and E.W. Bai (2006) "A two-stage algorithm for identification of nonlinear dynamic systems", *Automatica*, **42**, pp.1187-1196

[Lind (2008)] Lind, I and L. Ljung (2008) "Regressor and structure selection in NARX models using a structure ANOVA approach", *Automatica*, **44**, pp.383-395

[Peduzzi (1980)] Peduzzi, P (1980) "A stepwise variable selection procedure for nonlinear regression methods", *Biometrics* **36** pp.510-516

[Pillonetto (2011)] Pillonetto, G., M. Quang and A. Chiuso (2011) "A new kernel-based approach for nonlinear system identification", *IEEE Transactions on Automatic Control*, **56**, pp.2825-2840.

[Roll (2005)] Roll, J, A Nazin and L. Ljung (2005) "Nonlinear system identification via direct weight optimization", *Automatica*, **41**, pp.475-490

[Sjoberg (1995)] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, and P. Glorennec, H. Hjalmarsson and A. Duditsky (1995) "Nonlinear black-box modeling in system identification: A unified overview", *Automatica*, **31**, pp. 1691-1724

[Soderstrom (1989)] Soderstrom, T and P. Stoica (1989) System Identification, Prentice Hall, New York

[Walk (2008)] Walk, H. (2008) "A universal strong law of large numbers for conditional expectations via nearest neighbors", —bf 99, pp.1035-1050

[Zhao (2015)] Zhao, W., H.F. Chen, E.W. Bai, and K. Li (2015) "Kernel Based Local Order Estimation of Nonlinear Nonparametric Systems", *Automatica*, **51**, pp.243-254