Proceedings of the 6th International Conference of Control, Dynamic Systems, and Robotics (CDSR'19)
Ottawa, Canada – June 6-7, 2019
Paper No. CDSR 137
DOI: 10.1159/cdsr19.137

Overlap Group Lasso for Variable Selection in Nonlinear Nonparametric System Identification

Changming Cheng and Er-wei Bai
Dept. of Electrical and Computer Engineering
University of Iowa, Iowa City, Iowa 52242
er-weibai@uiowa.edu

Abstract - Identification of a nonlinear nonparametric system is not easy. On the other hand, many systems are sparse in the sense that not all variables contribute. If these variables that do not contribute can be detected and removed, the identification problem becomes lower dimensional and is relatively easy to deal with. The goal of the paper is to develop an overlap group Lasso method to detect which variables contribute and which variables do not. The algorithm developed favors sparsity in terms of partial derivatives that is the necessary and sufficient condition for a variable to contribute.

1. Introduction

In modeling of engineering, social or biomedical problems, a parsimony model is often preferred. It is based on the principle that is to explain the system input-output behavior by the simplest model possible that does not violate the data in any significant ways. A parsimony model is essential if the purpose of the model is for prediction. Input-output data usually is composed of two contributions, a replicable contribution that similar outputs are expected if similar inputs are applied and an non-replicable contribution that is due to random uncertainty including noise. A good model should capture the replicable part but very little of the non-replicable part. Overly simplified models that underfit the data are unable to capture the replicable part and provide poor prediction even the same inputs are applied. Overly complex models tend to overfit uncertainty including noise that may or may not be predictable and again provide poor prediction. Thus in practice a parsimony model is often preferred to discriminate system properties from noise.

There are also practical reasons to choose a parsimony model in many nonlinear system identification settings, including the complexity of identification algorithms and the cost of identification. Unlike linear system identification, identification of a nonlinear system is a very hard task [1, 2, 19, 20]. Further, identification of a high dimensional nonlinear system is much more complex than that of a low dimensional one [12, 22]. For instance, the required data length depends on the dimension of the system. In general, the data length required increases exponentially as a function of the number of variables in the systems. When the dimension is high, the curse of dimensionality becomes an issue. In such a case, it is simply unrealistic to collect such a long data sequence. By keeping model parsimony alleviates such difficulties.

We consider in this paper a nonlinear system

$$\begin{gathered} y(k) = f(x(k)) + v(k) \\ = f(x_1(k), x_2(k), \dots, x_q(k)) + v(k) \\ for\ k = 1,2, \dots, N \end{gathered} \tag{1}$$

where $y(k)$ is the system output at time k and $v(k)$ is an iid noise sequence of zero mean and finite variance, and is independent of the input variable $x_i(k)$, $i = 1, \dots, q$. The regressor $x(k) = (x_1(k), \dots, x_q(k))^T$ consists of possible contributing input variables. The structure of the nonlinear function f is unknown. The system (1) represents a large class of nonlinear systems including the well known finite impulse response nonlinear systems [5] and the nonlinear auto-regressive systems with exogenous inputs (NARX)[11,14], by letting

$$\begin{gathered} x(k) = (u(k-1), \dots, u(k-q)), or \\ x(k) = f(y(k-1), \dots, y(k-m), \\ u(k-1), \dots, u(k-m)) \end{gathered}$$

respectively.

The goal of identification is to identify the unknown system $f(\cdot)$ based on the available input-output data set $\{y(k), x(k)\}_{k=1}^{N}$. As discussed, nonlinear system identification is not an easy task specially if the dimension p is high. Fortunately for many practical applications, systems are sparse in the sense that not all variables contribute to the system. If the variables that do not contribute can be identified and removed, the dimension for identification could be smaller, also making the model parsimony.

2. Overlap group Lasso approach

In this section, we propose a way to detect if the variable x_i contributes or not for every i. The objective is that the algorithm developed should not suffer from the curse of dimensionality like the local approach and should have a fixed number of unknown parameters to be estimated that does not increase as the data length increases as in both the local and RKHS approaches. To this end, assume that there are some basis functions $\phi_i(x)$'s, $j = 1,2,\ldots,M$ so that the unknown function $f(\cdot)$ can be written as

$$f(x) = \sum_{j=1}^{M} \beta_j^* \phi_j(x) \tag{2}$$

for some $M > 0$ and unknown coefficients β_j^*'s. Admittedly the assumption (2) is stronger than in RKHS representation $f(x) = \sum_{k=1}^{\infty} \alpha_k K(x(k), x)$ that could have infinitely many terms. Thus, more prior information on the unknown $f(\cdot)$ is needed. However, (2) does have applications for many practical systems. For instance suppose f is unknown but is known that it is a polynomial of the maximum L. Then the assumption (2) is valid and the choice of $\phi_j(x)$ is obvious.

Now, decompose the coefficients set $\{\beta_1^*, \ldots, \beta_M^*\}$ into q subsets $S_1, \ldots, S_q$,

$$S_i = \{\beta_j^* | \phi_j(x)\ involves\ x_i, j = 1,2,\ldots,M\},$$
$$i = 1,2,\ldots,q$$

Clearly $\cup_{i=1}^{q} S_i \subset \{\beta_1^*, \ldots, \beta_M^*\}$, but there could be overlaps between S_i and S_j. If no confusion we may also refer S_i by the indices of $\beta_j^* \in S_i$, i.e., $S_i = \{j | \phi_j(x)\ involves\ x_i\}$.

We now make a simple but important observation:

$$\begin{gathered}\frac{\partial f}{\partial x_i} \equiv 0 \Leftrightarrow \beta_j^* = 0\ \ if\ \ \beta_j^* \in S_i \\ \Leftrightarrow \sum_{j \in S_i} \beta_j^{*2} = 0\end{gathered}$$

We give an example here. Consider a polynomial of x_1, x_2, x_3 of order 2,

$$\begin{gathered}f(x) = \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 + \beta_4^* x_1^2 + \beta_5^* x_2^2 \\ + \beta_6^* x_3^2 + \beta_7^* x_1 x_2 + \beta_8^* x_2 x_3 + \beta_9^* x_2 x_3 \\ = \beta_1^* \phi_1(x) + \beta_2^* \phi_2(x) + \beta_3^* \phi_3(x) + \beta_4^* \phi_4(x) \\ + \beta_5^* \phi_5(x) + \beta_6^* \phi_6(x) + \beta_7^* \phi_7(x) \\ + \beta_8^* \phi_8(x) + \beta_9^* \phi_9(x)\end{gathered} \tag{3}$$

Then,

$$S_1 = \{\beta_1^*, \beta_4^*, \beta_7^*, \beta_8^*\},$$
$$S_2 = \{\beta_2^*, \beta_5^*, \beta_7^*, \beta_9^*\},$$

$$S_3 = \{\beta_3^*, \beta_6^*, \beta_8^*, \beta_9^*\}$$

Now,

$$\frac{\partial f}{\partial x_1} = \beta_1^* \frac{\partial \phi_1(x)}{\partial x_1} + \beta_4^* \frac{\partial \phi_4(x)}{\partial x_1} + \beta_7^* \frac{\partial \phi_7(x)}{\partial x_1} + \beta_8^* \frac{\partial \phi_8(x)}{\partial x_1}$$

It is easily verified that

$$\frac{\partial f}{\partial x_1} \equiv 0 \Leftrightarrow \beta_1^* = \beta_4^* = \beta_7^* = \beta_8^* = 0 \Leftrightarrow \sum_{j \in S_1} \beta_j^{*2} = 0$$

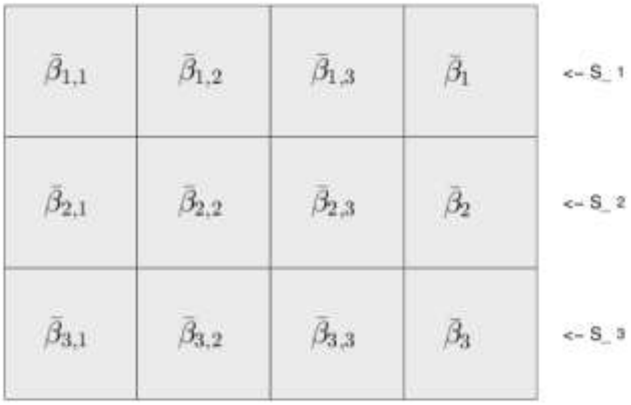
Similarly,

$$\frac{\partial f}{\partial x_2} \equiv 0 \Leftrightarrow \sum_{j \in S_2} \beta_j^{*2} = 0, \frac{\partial f}{\partial x_3} \equiv 0 \Leftrightarrow \sum_{j \in S_3} \beta_j^{*2} = 0$$

For a polynomial of the form (3), the sets S_i's can be visualized by re-parameterizing (3) as

$$f(x) = \sum_{i=1}^{3} \bar{\beta}_i x_i + \sum_{i=1}^{3} \sum_{j=1}^{3} \bar{\beta}_{i,j} x_i x_j$$

where $\bar{\beta}_{i,j} = \bar{\beta}_{j,i}$. Then S_i consists of all the coefficients in the ith row as shown in Figure 1.



$\bar{\beta}_{1,1}$	$\bar{\beta}_{1,2}$	$\bar{\beta}_{1,3}$	$\bar{\beta}_1$	<-- S_ 1
$\bar{\beta}_{2,1}$	$\bar{\beta}_{2,2}$	$\bar{\beta}_{2,3}$	$\bar{\beta}_2$	<-- S_ 2
$\bar{\beta}_{3,1}$	$\bar{\beta}_{3,2}$	$\bar{\beta}_{3,3}$	$\bar{\beta}_3$	<-- S_ 3

Fig. 1: Illustration of S_i's.

Now consider (2). The objective is to estimate the unknown β_j^*'s and determine if $\sum_{j \in S_i} \beta_j^{*2} = 0$ or not for each i. To this end, let

$$
\begin{gathered}
J(\beta) = J(\beta_1, \ldots, \beta_M) = \\
\frac{1}{2}\sum_{k=1}^{N} [y(k) - \sum_{j=1}^{M} \beta_j \phi_j(x(k))]^2 \\
+N \sum_{i=1}^{q} \lambda_{Ni} \sqrt{\sum_{j \in S_i} \beta_j^2}
\end{gathered}
\tag{4}
$$

for some λ_{Ni} and consider the following minimization

$$
\min_{\beta} J(\beta) \tag{5}
$$

The (4) can also be rewritten in a matrix form,

$$
\begin{gathered}
J(\beta) = \frac{1}{2} \parallel Y - \Phi\beta \parallel^2 + N \sum_{i=1}^{q} \lambda_{Ni} \sqrt{\sum_{j \in S_i} \beta_j^2} \\
\text{where, } Y = (y(1), \ldots, y(N))', \ \beta = (\beta_1, \ldots, \beta_M)' \text{ and} \\
\Phi = \begin{pmatrix} \phi_1(x(1)) & \cdots & \phi_M(x(1)) \\ \phi_1(x(2)) & \cdots & \phi_M(x(2)) \\ \vdots & \ddots & \vdots \\ \phi_1(x(N)) & \cdots & \phi_M(x(N)) \end{pmatrix}
\end{gathered}
\tag{6}
$$

Note the cost function of (6) is similar to Lasso [25] but with a different penalty term. This is because each $\frac{\partial f}{\partial x_i}$ is represented by a group of coefficients $\beta_j^* \in S_i$. Lasso minimization amounts to the selection of each coefficient β_j^* rather than a group of coefficients $\beta_j^* \in S_i$ corresponding to one variable x_i. When applied, Lasso tends to make selection based on the strength of individual $\phi_i(x)$ rather than the strength of the group of coefficients $\beta_j^* \in S_i$, often resulting in selecting more variables than necessary. A key difference is that sparsity here is not defined by $\phi_j(x)$'s but in terms of variables x_i's. The natural group tension of (4) favors sparsity in terms of partial derivatives $\frac{\partial f}{\partial x_i}$ and thus improves over Lasso in terms of variable selection.

Theorem 2.1 *Consider the system (2), the cost function (4) and the minimization (5). Assume* $\frac{1}{N}\Phi^T\Phi > 0$. *Denote* $\hat{\beta}$ *the least squares solution of* $min \sum_{k=1}^{N} (y(k) - \sum_{j=1}^{M} \hat{\beta}_j \phi_j(x(k)))^2$. *Let*

$$
\lambda_{Ni} = \frac{N^{-\epsilon}}{\sqrt{\sum_{j \in S_i} \hat{\beta}_j^2}} \tag{7}
$$

for some $1/2 < \epsilon < 1$. Then, we have in probability as $N \to \infty$,

1) $\beta - \beta^* = O(N^{-1/2})$.
2) If $\sqrt{\sum_{j \in S_i} \beta_j^{*2}} = 0$,

$$Prob\{\sqrt{\sum_{j\in S_i}\beta_j^2}=0\}\to 1.$$

Proof: (1) Note that $J(\beta)$ is a strictly convex function if $\frac{1}{N}\Phi^T\Phi>0$. Therefore from [8], it suffices to show that β is $\sqrt{N}$-consistency, if for any $\varepsilon>0$, there exists a large enough $C>0$ such that

$$\liminf_{N\to\infty}P\{\inf_{u\in R^M,\|u\|=C}J(\beta^*+N^{-1/2}u)>J(\beta^*))\}>1-\varepsilon. \tag{8}$$

Note

$$\begin{aligned}
&J(\beta^*+N^{-1/2}u)-J(\beta^*))\\
&=\frac{1}{2}\parallel Y-\Phi(\beta^*+N^{-1/2}u)\parallel^2\\
&-\frac{1}{2}\parallel Y-\Phi\beta^*\parallel^2\\
&+N\sum_{i=1}^{q_0}\lambda_{Ni}\sqrt{\sum_{j\in S_i}(\beta_j^*+N^{-1/2}u_j)^2}\\
&-N\sum_{i=1}^{q_0}\lambda_{Ni}\sqrt{\sum_{j\in S_i}\beta_j^{*2}}\\
&+N\sum_{i=q_0+1}^{q}\lambda_{Ni}\sqrt{\sum_{j\in S_i}(N^{-1/2}u_j)^2}\\
&\geq\frac{1}{2N}u^T\Phi^T\Phi u-\frac{1}{\sqrt{N}}u^T(Y-\Phi\beta^*)\\
&+N\sum_{i=1}^{q_0}\frac{\alpha N^{-\epsilon}}{\sqrt{\sum_{j\in S_i}\beta_j^{*2}}}\cdot\\
&(\sqrt{\sum_{j\in S_i}(\beta_j^*+N^{-1/2}u_j)^2}-\sqrt{\sum_{j\in S_i}\beta_j^{*2}})\\
&=O(\parallel u\parallel^2)+O(u)+O(N^{-(\epsilon+1/2-1)})
\end{aligned}\tag{9}$$

The first term in (9) is quadratic function in u, and the second term is linear in u, and the third term converges to zero. Thus for large enough C, (8) holds, then

$$\beta-\beta^*=O(\frac{1}{\sqrt{N}})$$

(2) Suppose that β_w^* is in $S_{q_0+1},\ldots,S_q$, i.e., β_w^* is in some $\sqrt{\sum_{j\in S_i}\beta_j^{*2}}=0$ or $\beta_w^*=\beta_j^*$ for some $i=q_0+1,\ldots,q$

and $j \in S_i$.

Let Φ_w denote the wth column of Φ. The necessary condition for β_w to be optimal is

$$\begin{aligned}
0 &= \frac{1}{\sqrt{N}} \frac{\partial J}{\partial \beta_w} \\
&= \frac{1}{2\sqrt{N}} \Phi_w^T (Y - \Phi\beta) \\
&+ \sqrt{N} \sum_{\substack{all\ terms\ with\ \beta_w \\ and \sqrt{\sum_{j\in S_i} \beta_j^2} = 0}} \frac{N^{-\epsilon}}{\sqrt{\sum_{j\in S_i} \beta_j^2}} \frac{\beta_w}{\sqrt{\sum_{j\in S_i} \beta_j^2}} \\
&+ \sqrt{N} \sum_{\substack{all\ terms\ with\ \beta_w \\ and \sqrt{\sum_{j\in S_i} \beta_j^2} \neq 0}} \frac{N^{-\epsilon}}{\sqrt{\sum_{j\in S_i} \beta_j^2}} \frac{\beta_w}{\sqrt{\sum_{j\in S_i} \beta_j^2}} \\
&= \frac{1}{2\sqrt{N}} \Phi_w^T (Y - \Phi\beta^*) - \frac{1}{2\sqrt{N}} \Phi_w^T \Phi (\beta - \beta^*) \\
&+ O(N^{1-\epsilon}) + O(N^{-(1-\epsilon)})
\end{aligned} \tag{10}$$

The first term in (10) is of the order $O(1)$, and so is the second term because of $\beta - \beta^* = O(N^{-1/2})$ according to the proof of (1) in Theorem 2.1. The third term goes to ∞ and the fourth term goes to zero. Thus if $\beta_w \neq 0$, the terms in (10) is $O(N^{1-\epsilon}) \to \infty$. This implies that $\beta_w \neq 0 \Rightarrow \frac{1}{\sqrt{N}} \frac{\partial J}{\partial \beta_w} \neq 0$ that contradicts that β_w is optimal. Thus optimal β_q equals zero for N large enough.

We make a few comments.

- From the theorem, if x_i contributes or not can be determined by checking if $\sqrt{\sum_{j\in S_i} \beta_j^2} = 0$.
- The representation of $f(\cdot)$ is global and so the curse of dimensionality is no longer an issue. Also the number of parameters $\beta_j, j = 1,2\ldots, M$, to be estimated is fixed independent of the data length N.
- The minimization (5) is reminiscent of the group Lasso [27] but not the same. In the group Lasso setting, the coefficients are decomposed into disjoint subsets with no overlap. In out setting, overlap is allowed and in fact often present.
- The minimization (5) is also reminiscent of the overlapping groups Lasso as in [13] but again totally different. Though in [13], overlapping is allowed, e.g., $\beta_j^* \in S_{i1}, \ldots, S_{il}$. Then, β_j^* has to be decomposed in a unique way into $\beta_{ji1}^* \in S_{i1}, \ldots, \beta_{jil}^* \in S_{il}$ such that $\beta_j^* = \beta_{ji1}^* + \ldots + \beta_{jil}^*$. In our setting, such equalities usually do not hold. Consequently, the convergence proofs and the algorithms developed in [13] do not apply to our minimization.

The minimization problem (5) can be solved by modifying LARS [6]. However, LARS is known to have some difficulties if $\phi_i(x)$'s are highly correlated. We propose to solve (5) by the coordinate descent algorithm [26] well known in the literature.

Algorithm to solve (5):

Step 0: Set $m = 0$ and any initial estimate $\beta(0)$.

Step 1: At any m, set $l = 1$

Step: 1.1: Fix $\beta_1(m), \ldots, \beta_{l-1}(m), \beta_{l+1}(m), \ldots, \beta_p(m)$ and solve

$$\min_{\beta_l} J(\beta_l)$$

where J is in (4) by fixing $(p-1)$ β_j's and leaving one β_l for minimization. This one dimensional minimization can be solved by the gradient descent method or by the line search method. Set the solution as $\beta_l(m+1)$.

Step: 1.2: If $l < q$, set $l = l + 1$ and go to Step 1.1. If $l = q$, go to Step 2.
Step 2: If $\| \beta(m+1) - \beta(m) \| / \| \beta(m) \|$ is smaller than a prescribed threshold, stop. Otherwise, set $m = m + 1$ and go to Step 1.

The following result is from [26].

Theorem 2.2 *Consider the above algorithm and the minimization (5). Assume* $\frac{1}{N}\Phi^T \Phi > 0$. *Then, the sequence* $\beta(m)$ *generated by the above algorithm converges to the solution of (5).*

Note so far we only consider the case that M is fixed and $N \to \infty$. The result can be extended to the case $N > M$, and $M, N \to \infty$ under some restrictive conditions. To this end, define $H_c = \{S_1, \ldots, S_{q_0}\}$, $H_r = \{S_{q_0+1}, \ldots, S_q\}$ and also Φ_i, Φ_{H_c}, that contain ϕ_i or the corresponding parts of H_c, respectively. H_c is the part of coefficients that $|\beta_i^*| > 0$ and H_r is the part of coefficients that $|\beta_i^*| = 0$.

Assumption 2.1

1) $\alpha_2 I > \frac{1}{N}\Phi^T \Phi > \alpha_1 I$ almost surely as $N \to \infty$.

2) Assume the cardinality of $S_i, i = 1, \ldots, q$ is d_i, and $d_0 = \sum_{i=1}^{q_0} d_i$, $d = \sum_{i=1}^{q} d_i$. Let $\alpha = \min_{i \in H_c} \| \beta_i^* \|_\infty$ and assume

$$\frac{1}{\alpha}[\sqrt{\frac{\log d_0}{N}} + \sqrt{d_0}\lambda_{Ni}] \to 0.$$

3) For some $0 < \epsilon < 1$ and every $i \in H_r$,

$$\| \Phi_i^T \Phi_{H_c} (\Phi_{H_c}^T \Phi_{H_c})^{-1} \|_2 \leq \frac{1-\epsilon}{\sqrt{q_0}}$$

4) $\frac{1}{\lambda_{Ni}}\sqrt{\frac{\log(d-d_0)}{N}} \max_{i \in H_r} \sqrt{d_i} \to 0.$

Then following the proof of Theorem 4.2 [16], we have

Theorem 2.3 *Consider the system (2), the cost function (4) and the minimization (5). Under the assumption 2.1, the probability that* $\| \tilde{\beta}_i \|_2 > 0$ *for all* $i \in H_c$, *and* $\tilde{\beta}_i = 0$ *for all* $i \in H_r$ *converges to 1 as* $N > M, M, N \to \infty$.
Interested readers are referred to [16] for more details.

3. Numerical simulation

Consider an 10-dimensional second order polynomial system with the input $u(k-i) = x_i(k)$ where $u(\cdot)$ is iid Gaussian of zero mean and unity variance.

$$\begin{aligned} y(k) &= f\big(x_1(k), x_2(k), \ldots, x_{10}(k)\big) + v(k) \\ &= u(k-3) + u(k-4) + u(k-3)^2 \\ &\quad +0.5u(k-3)u(k-4) + 0.5u(k-4)^2 \\ &\quad + u(k-6) + u(k-6)^2 + u(k-3)u(k-6) \\ &\quad +u(k-4)u(k-6) + v(k) \\ &= x_3(k) + x_4(k) + x_3(k)^2 + 0.5x_3(k)x_4(k) \\ &\quad +0.5x_4(k)^2 + x_6(k) + x_6(k)^2 + x_3(k)x_6(k) \\ &\quad +x_4(k)x_6(k) + v(k) \\ &= \sum_{i=1}^{10} \bar{\beta}_i x_i(k) + \sum_{i,j=1}^{10} \bar{\beta}_{i,j} x_i(k) x_j(k) + v(k), \end{aligned} \tag{11}$$

with $\bar{\beta}_{i,j} = \bar{\beta}_{j,i}$. $v(\cdot)$ represents iid Gaussian of zero mean and standard deviation 0.5. In simulation, no knowledge of $f(\cdot)$ is available.

Consider the overlap group Lasso method, Similar to Figure 1 for the example (3), the set S_i's are clearly defined in terms of $\bar{\beta}_i$ and $\bar{\beta}_{i,j}$. In fact S_i consists of all the coefficients in the ith row as in Figures 2 and 3. Since only x_3, x_4 and x_6 contribute and $x_1, x_2, x_5, x_7, x_8, x_9$ and x_{10} do not contribute

$$\sqrt{\bar{\beta}_i^2 + \bar{\beta}_{i,1}^2 + \ldots + \bar{\beta}_{i,10}^2} = 0,$$

for $i = 1,2,5,7,8,9,10$ and

$$\sqrt{\bar{\beta}_i^2 + \bar{\beta}_{i,1}^2 + \ldots + \bar{\beta}_{i,10}^2} > 0,$$

for $i = 3,4,6$.

With $N = 80$ and $\lambda_{Ni} = \frac{N^{-0.6}}{\sqrt{\sum_{j \in S_i} \hat{\beta}_j^2}}$ as given by (7) where $\hat{\beta}_j$'s are the least squares estates, the results of the overlap group Lasso are shown in Figure 2. Light gray indicates that the corresponding estimate of $\bar{\beta}_i$ or $\bar{\beta}_{i,j}$ is exactly zero and dark gray indicates that the corresponding estimate of $\bar{\beta}_i$ or $\bar{\beta}_{i,j}$ is non-zero. Figure 1 clearly demonstrates that all the coefficients in $S_i, i = 1,2,5,7,8,9,10$ are exactly zero and some coefficients in $S_i, i = 3,4,6$ are nonzero. Thus, the variables x_3, x_4 and x_6 contribute and $x_1, x_{2,}, x_5, x_7, x_8, x_9, x_{10}$ do not contribute. For comparison, the standard Lasso is also applied for the same example and the same simulation parameters. The results are shown in Figure 3. Obviously, it does not reveal which variable contributes and which one does not. The reason is that Lasso favors sparsity in terms of individual terms but not in terms of variable selection. On the other hand, the overlap group Lasso proposed in this paper favors sparsity in terms of partial derivatives $\frac{\partial f}{\partial x_i}$ and thus improves the variable selection capability.

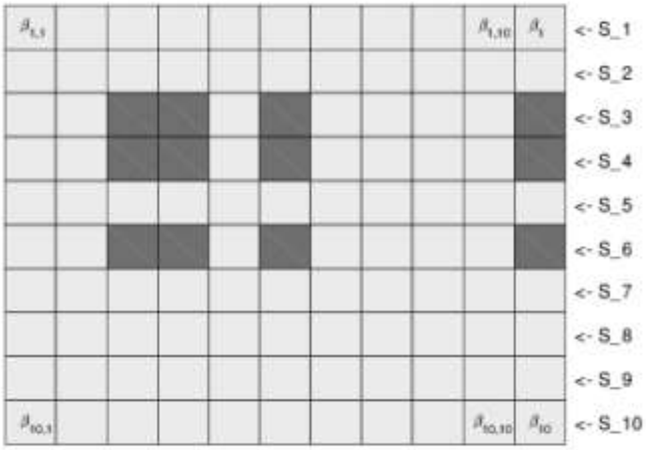


Fig. 2: $\bar{\beta}_i, \bar{\beta}_{i,j} \in S_i, i = 1,\ldots,10$ with overlap group Lasso.

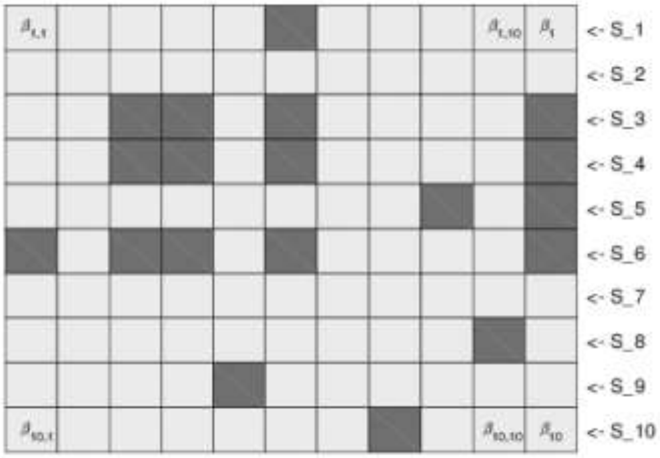


Fig. 3: $\bar{\beta}_i, \bar{\beta}_{i,j} \in S_i, i = 1,\ldots,10$ with Lasso.

4. Concluding remarks

In this paper, an overlap group Lasso method is developed that aims to determine which variables contribute and which ones do not. The algorithm favors sparsity in terms of variables and therefore overperforms the standard Lasso that favors sparsity in terms of individual terms.

There is an important issue that is not discussed in the paper: how to make sense of the ranking if variables are statistically correlated. When variables are correlated, contributions from one input are contaminated by the contributions from other correlated variables. This makes ranking very hard if possible. Also, the ranking depends on the definition of importance. In the papers, it is based on the summation of the coefficients squares. It is not clear at this point how to rank the importance of variables, e.g., in terms of the Goodness of Fit. These directions deserve further studies.

References

[1] E. W. Bai and K. Chan, “Identification of an additive nonlinear system,” *Automatica*, vol. 44, pp.430-436, 2008.

[2] E. W. Bai and Y. Liu, “Recursive direct weight optimization in nonlinear system identification: a minimal probability approach,” *IEEE Trans on Automatic Control*, 52, pp. 1218-1231, 2007.

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

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

[5] C. M. Cheng, Z. K.,Peng, W. M. Zhang and G. Meng,“Volterra-series-based nonlinear system modeling and its engineering applications: A state-of-the-art review,” *Mech. Syst. Signal Process*, vol. 87, pp. 340-364, 2017.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression," *The Annals of statistics*, vol. 32, pp. 407-499, 2004.

[7] J. Fan and I. Gijbels, *Local Polynomial Modeling and Its Applications*. New York: Chapman and Hall/CRC, 1996.

[8] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, pp. 1348-1360, 2001.

[9] K. Fukumuzu and C. Leng “Gradient-Based Kernel Dimension Reduction for Regression,” *J. of the American Statistical Association*, vol. 109, pp. 359-370, 2014.

[10] Helmbold and B. Willamson (Eds), *Computational Learning Theory*. Springer, 2001, pp.416-426.

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

[12] Kennel, M. R. Brown and H. Abarbanel, “Determining embedding dimension for phase-space reconstruction using geometrical construction,” *Physical Review*, vol. 45, pp. 3403-3411, 1992.

[13] L. Jacob, G. Obozinski, and J. P. Vert, “Group lasso with overlap and graph lasso", in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 433-440.

[14] K. Li, J. Peng and E. W. Bai, “A two-stage algorithm for identification of nonlinear dynamic systems,” *Automatica*,

vol. 42, pp. 1187-1196, 2006.

[15] S. Mukherjee and D. X. Zhou, “Learning coordinate covariances via gradients," *Journal of Machine Learning Research*, vol. 7, pp. 519-549, 2006.

[16] Y. Nardi and A. Rinaldo, “On the asymptotic properties of the group lasso estimator for linear models,” *Electronic Journal of Statistics* , vol. 2, pp. 605-633, 2008.

[17] P. Peduzzi, “A stepwise variable selection procedure for nonlinear regression methods,” *Biometrics*, vol. 36, pp. 510-516, 1980.

[18] H. Peng, F. Long, C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1226-1238, 2005.

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

[20] J. Roll, A. Nazin and L. Ljung, “Nonlinear system identification via direct weight optimization,” *Automatica*, vol. 41, pp. 475-490, 2005.

[21] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri, “Nonparametric sparsity and regularization," *The Journal of Machine Learning Research*, vol. 14, pp. 1665-1714, 2013.

[22] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by local linear embedding,” *Science*, vol. 290, pp. 2323-2326, 2000.

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

[24] I. M. Sobol, S. Kucherenko, “Derivative based global sensitivity measures and their link with global sensitivity indices,” *Mathematics and Computers in Simulation*, vol. 79, no. 10, pp. 3009-3017, 2009.

[25] R. Tibshirani, “Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, pp. 267-288, 1996.

[26] T. T. Wu and K. Lange, “Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, pp. 224-244, 2008.

[27] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, pp. 49-67, 2006.

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

[29] D. Zhou, “Derivative reproducing properties for kernel methods in learning theory,” *J. of Computational and Applied Mathematics*, vol. 220, pp. 456-463, 2008.