The following publication J. Wang et al., "Manifold-Regularized Multitask Fuzzy System Modeling With Low-Rank and Sparse Structures in Consequent Parameters," in IEEE Transactions on Fuzzy Systems, vol. 30, no. 5, pp. 1486-1500, May 2022 is available at https://doi.org/10.1109/TFUZZ.2021.3062691.

Manifold-regularized Multitask Fuzzy System Modeling with Low-rank Structure and Sparse Consequent Parameters

Zhuangzhuang Zhao, Jun Wang, Zhaohong Deng, Xiang Pan, Kup-Sze Choi, Lejun Gong, Jun Shi, and Shitong Wang

***Abstract*—Multitask modeling methods for Takagi-Sugeno-Kang (TSK) fuzzy systems exhibit better generalization ability attributed to the utilization of the knowledge of inter-task correlation. However, these methods usually ignore the balance between the sharing of the common knowledge across multiple tasks and the preservation of the task-specific characteristics of each rule. To this end, we propose a novel manifold-regularized multitask modeling method for TSK fuzzy system by introducing low-rank structure and sparse consequent parameters. Specifically, we decompose the consequent parameters into two components – the low-rank structure shared by multiple tasks and the task-specific component that encodes the sparse characteristics of the individual tasks. An efficient Augmented Lagrange Multiplier is developed to solve the optimization problem. The experimental results demonstrate that the proposed model significantly outperforms the existing methods.**

***Index Terms*—TSK fuzzy system, multitask learning, low-rank structure.**

I. Introduction

FUZZY systems are developed based on fuzzy logic and fuzzy inference. They are specialized in describing the uncertainty of knowledge and expression, and are able to approximate uncertain nonlinear systems better than conventional machine learning models [1]. Various models have been developed as a result of recent advances in fuzzy systems. Among them, the Takagi-Sugeno-Kang (TSK) fuzzy system is the most popular one that provides an effective framework to reduce nonlinear systems into multiple local linear structures [2-6].

TSK fuzzy systems include several fuzzy rules, which are expressed in the form of IF-THEN statements. The modeling of TSK fuzzy systems includes two fundamental steps [7, 8]. In the first step, the premise of the fuzzy rules is extracted by partitioning the data into several groups. In data driven modeling, clustering methods can be used to achieve this goal. In the second step, the consequent parameters of the fuzzy rules are learned. From the perspective of machine learning, this step is always considered as a linear regression problem. Many criterions for learning the consequent parameters have been developed. The classical learning criterions include the least squares [9, 10] and its extensions, the maximum interval method [11] and so on.

Like most supervised machine learning models, TSK fuzzy systems require sufficient training data. However, in many real-world applications, training samples are often limited whereas the dimensionality of the consequent parameters is high, thus leading to overfitting problems. By multitask learning, the learning performance of individual tasks can be improved by utilizing the joint information obtained from the related tasks [12-17]. The tasks are assumed to be similar so that the learning of one task can be benefitted from the learning of the others. That is to say, multitask learning learns the shared information of multiple tasks, which can be applied to different but related tasks to improve the generalization ability of each task.

In multitask learning setting, multitask modeling methods have been developed for TSK fuzzy systems. For example, Jiang et al. proposed a multitask TSK fuzzy system by considering the inter-task relation in a shared hidden subspace [18]. This model however always produces highly complex fuzzy model with a large number of consequent parameters. To obtain an accurate and concise fuzzy model, Wang et al. proposed the mtSparseTSK model which jointly learned a compact set of fuzzy rules and consequent parameters through

This work was supported in part by the National Natural Science Foundation of China under Grants 61772239, 61702225, 61772254 and 61872190, the Natural Science Foundation of Jiangsu Province under Grant BK20181339, BK20161268, BK20151299, and BK20151358, Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, NJUPT (No. BDSIP1904), and the Research Grants Council of the HKSAR (PolyU 512006/19E)

Corresponding author: Jun Wang (wangjun_shu@shu.edu.cn), Zhaohong Deng (dengzhaohong@jiangnan.edu.cn).

Z. Zhao, Z. Deng, X. Pan and S. Wang are with the school of artificial intelligence and computer science, Jiangnan University, Wuxi 214122, China.

K. S. Choi is with the Center of Smart Health, School of Nursing, the Hong Kong Polytechnic University, Hong Kong, (e-mail: kschoi@ieee.org).

J. Wang and J. Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; J. Wang is also with the Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China.

L. Gong is with the Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China.

a unified procedure by group sparse learning [19]. Nevertheless, mtSparseTSK only considers the shared consequent parameters of the tasks and ignores the task-shared and task-specific characteristics of the consequent parameters and the intrinsic relations (e.g. feature-feature relation).

In multitask fuzzy modeling, the discriminant information may lie in a low-rank subspace spanning across the consequent parameters of the multiple fuzzy models. In this space, multiple fuzzy models can be correlated by the common information of the consequent parameters of different rules of the multiple tasks. This is essentially a low-rank structure. In addition, there may be significant difference between individual learning tasks that would result in sparse discriminative consequent parameters for each task. Therefore, identifying the low-rank structure and the sparse consequent parameters are critical in multitask fuzzy modeling. To this end, by assuming that the multiple tasks share a low-rank structure and that each task has discriminative sparse consequent parameters, we propose a novel manifold-regularized multitask modeling method called LR-S-mtTSK. Specifically, we decompose the consequent parameters into two components – the low-rank structure shared by multiple tasks and the task-specific structure that encodes the sparse characteristics of the individual tasks. This can be implemented by imposing low-rank constraints on the shared structure of the multiple tasks; and by applying the sparse constraints on the task-specific component to retain the specific information of each task. To enhance the performance of multitask learning, we further devise a new manifold regularization method to reflect the feature-feature relation. That is, if a pair of original features are related to each other, the same or similar relation is expected to be preserved between the corresponding consequent parameters within each rule. We formulate a novel learning criterion and further optimize it by proposing the efficient Augmented Lagrange Multiplier (ALM) procedure.

The rest of the paper is organized as follows: Section 2 introduces the fundamentals of multitask TSK fuzzy model. Section 3 presents a novel modeling method of multitask fuzzy system. Section 4 presents the optimization method of the proposed model. Section 5 presents the experiments and gives a comparative analysis of the results. Section 6 draws the conclusions.

II. Overview of Multitask TSK Fuzzy Systems

Classical single-task TSK fuzzy system utilizes a collection of local linear submodels to approximate a nonlinear model. Multitask fuzzy system is a combination of multiple single-task TSK fuzzy models. In multitask settings, we use $\mathbf{x}_t = (x_t^1, x_t^2, \cdots, x_t^D)' \in \mathbb{R}^D$ to denote a feature vector in the t-th task, $t = 1, \cdots, T$, where x_t^d, $d = 1, \cdots, D$, is the d-th variable of $\mathbf{x}_t$; T, N_t and D denote the number of tasks, samples and features respectively.

The m-th fuzzy rule of the t-th task can be written as follows:

$$\begin{aligned} &\textbf{IF } x_t^1 \textbf{ is } \boldsymbol{A}_t^{m,1} \wedge x_t^2 \textbf{ is } \boldsymbol{A}_t^{m,2} \wedge \ \dots \ \wedge x_t^D \textbf{ is } \boldsymbol{A}_t^{m,D}, \\ &\textbf{THEN} \\ &f_t^m(\mathbf{x}_t) = w_t^{m,0} + w_t^{m,1} x_t^1 + w_t^{m,2} x_t^2 + \cdots + w_t^{m,D} x_t^D \end{aligned} \tag{1}$$

$m = 1, \cdots, M$, $t = 1, \cdots, T$

where $\boldsymbol{A}_t^{m,d}$ denotes the a fuzzy subset of input variable x_t^d for m-th rule in the t-th task, $m = 1, \cdots, M$, $d = 1, \cdots, D$; M is the number of fuzzy rules and $\wedge$ denotes the conjunction operation. In this paper, we use the Gaussian function in Eq. (2) as membership function,

$$\mu_{A_t^{m,d}}(x_t^d) = \exp\left(-\frac{\left(x_t^d - c^{m,d}\right)^2}{2\sigma^{m,d}}\right) \tag{2}$$

where c_d^m and σ_d^m denote the mean and variance of the features that are shared by multiple tasks. They can be computed as follows:

$$c^{m,d} = \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mu_{t,i}^m x_{t,i}^d \Big/ \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mu_{t,i}^m \tag{3}$$

$$\sigma^{m,d} = h \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mu_{t,i}^m (x_{t,i}^d - c^{m,d}) \Big/ \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mu_{t,i}^m \tag{4}$$

where $\mu_{t,i}^m$ is the fuzzy membership of the i-th sample in the t-th task for the m-th cluster, and h is an adjustable parameter. The output of the t-th task is given as follows:

$$\hat{\mathbf{y}}_t = \sum_{m=1}^{M} \varphi^m(\mathbf{x}_t) l^m(\mathbf{x}_t) \tag{5}$$

where $l^m(\mathbf{x}_t) = w_t^{m,0} + w_t^{m,1} x_t^1 + w_t^{m,2} x_t^2 + \cdots + w_t^{m,D} x_t^D$ denotes the consequence of the m-th fuzzy rule in the t-th task, and $\mathbf{w}_t^m = \left(w_t^{m,0}, w_t^{m,1}, \cdots, w_t^{m,D}\right)'$ is the consequent parameters of the m-th fuzzy rule of the t-th task; $\varphi^m(\mathbf{x}_t)$ denotes the firing strength of the t-th fuzzy rule which can be expressed as:

$$\begin{aligned} \varphi^m(\mathbf{x}_t) &= \frac{\mu^m(\mathbf{x}_t)}{\sum_{k=1}^{M} \mu^k(\mathbf{x}_t)}, \\ \mu^m(\mathbf{x}_t) &= \prod_{d=1}^{D} \mu_{A_t^{m,d}}(x_t^d) \end{aligned} \tag{6}$$

Let $\{\mathbf{x}_{t,i}, y_{t,i}\}$ be the i-th sample and the corresponding label in the t-th task, $\mathbf{x}_{t,i} = (x_{t,i}^1, x_{t,i}^2, \cdots, x_{t,i}^D)' \in \mathbb{R}^D$, $i = 1, \cdots, N_t$, $t = 1, \cdots, T$, $\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_{t,1}' \\ \vdots \\ \mathbf{x}_{t,N_t}' \end{pmatrix} \in \mathbb{R}^{N_t \times D}$, $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \cdots, y_{t,N_t})' \in \mathbb{R}^{N_t}$. The output of the fuzzy model is thus given by:

$$\hat{\mathbf{y}}_t = \sum_{m=1}^{M} \boldsymbol{\Phi}_t^m \mathbf{w}_t^m \tag{7}$$

where

$$\boldsymbol{\Phi}_t^m = diag\left(\varphi^m\left(\mathbf{x}_{t,1}\right), \cdots, \varphi^m\left(\mathbf{x}_{t,N_t}\right)\right)(\mathbf{1}, \mathbf{X}_t) \in \mathbb{R}^{N_t \times (D+1)} \tag{8}$$

is the m-th dictionary of the fuzzy model in the t-th task, $\mathbf{1}$ is an all-1 vector, and $\mathbf{w}_t^m$ is a vector containing the consequent parameters for the m-th rule in the t-th task. The modeling of multitask TSK fuzzy systems can be formulated as

$$\hat{\mathbf{y}}_t = \boldsymbol{\Phi}_t \mathbf{w}_t, \tag{9}$$

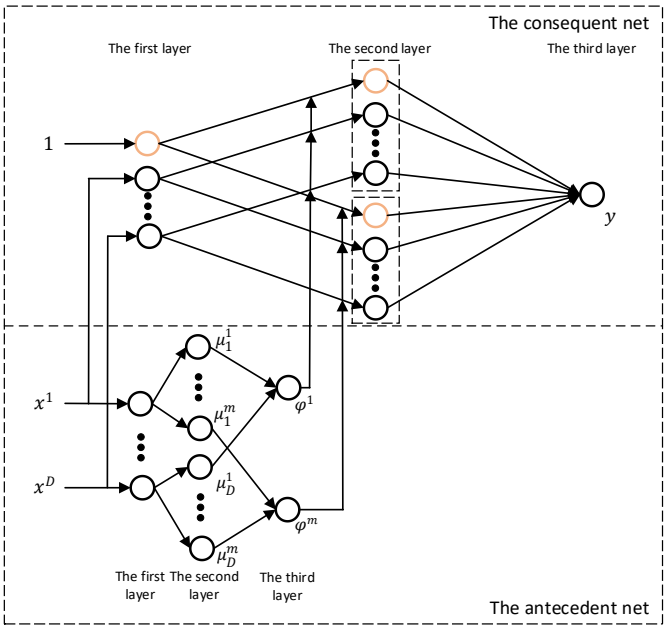

Fig. 1 The antecedent and consequent network of fuzzy system. In the antecedent network, the first layer is the input layer whose nodes are directly connected to the components of the input vector. The second layer contains the membership of each component of the input to the m-th cluster. Each node in the third layer represents a fuzzy rule, which is used to match the antecedents of the fuzzy rules and calculate the applicability of each rule.

where $\mathbf{\Phi}_t \in \mathbb{R}^{N_t \times (D+1)M}$ is the dictionary of the fuzzy model in the t-th task and $\mathbf{\Phi}_t^m$ is the subdictionary of the m-th rule, $m = 1, \cdots, M$. We use $\mathbf{w}_t = ((\mathbf{w}_t^1)', (\mathbf{w}_t^2)', \cdots, (\mathbf{w}_t^M)')' \in \mathbb{R}^{(D+1)M}$ to denote a vector containing all the consequent parameters for the t-th task. For convenience, we define $\mathbf{W}^m = (\mathbf{w}_1^m, \mathbf{w}_2^m, \cdots, \mathbf{w}_T^m) \in \mathbb{R}^{(D+1)\times T}$ which contains the consequent parameters in the m-th rules across all tasks. We further define $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T) = \begin{pmatrix} \mathbf{W}^1 \\ \mathbf{W}^2 \\ \vdots \\ \mathbf{W}^M \end{pmatrix} \in \mathbb{R}^{(D+1)M \times T}$ which contains all the consequent parameters of the multiple tasks.

A. Manifold-regularized multitask fuzzy systems

To learn the consequent parameters of the proposed model, we map the training data into M high dimensional feature space using Eq. (9)-(11). The original problem is therefore transferred into a linear regression problem. Fig. 1 shows the mapping procedure, which is equivalent to a feedforward neural network structure.

The general learning criterion of manifold-regularized multitask fuzzy system is formulated as follows:

$$\min_W \sum_{t=1}^T L(\mathbf{y}_t, \mathbf{\Phi}_t \mathbf{w}_t) + \alpha \cdot R(\mathbf{W})$$

where the first term $\sum_{t=1}^T L(\mathbf{y}_t, \mathbf{\Phi}_t \mathbf{w}_t)$ is a global loss function that considers the difference between the predicted labels and their ground truth in each task; the second term $R(\mathbf{W})$ handles the inter-correlation between multiple tasks in the local model; α is the trade-off parameter of the regularization term.

Let $N = \sum_{t=1}^T N_t$. We define $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \in \mathbb{R}^{N \times D}$ that

$$\mathbf{\Phi}_t = (\mathbf{\Phi}_t^1, \mathbf{\Phi}_t^2, \cdots, \mathbf{\Phi}_t^M) \quad (10)$$

$$\mathbf{w}_t = ((\mathbf{w}_t^1)', (\mathbf{w}_t^2)', \cdots, (\mathbf{w}_t^M)')' \quad (11)$$

contains all the input samples in multiple tasks; $\mathbf{\Phi}^m = \begin{pmatrix} \mathbf{\Phi}_1^m \\ \vdots \\ \mathbf{\Phi}_T^m \end{pmatrix} \in \mathbb{R}^{N \times (D+1)}$, where $\mathbf{\Phi}_t^m$ is computed with Eq. (8). $(\mathbf{\Phi}^m)^i$ and $(\mathbf{\Phi}^m)^j$ denote two column vectors??? that are generated from the i-th and j-th features in $\mathbf{X}$. If the i-th and j-th features of $\mathbf{X}$ are related to each other, their corresponding consequent parameters in the fuzzy rules should also be related. Therefore, the i-th and j-th rows in $\mathbf{W}^m$, i.e. $(\mathbf{W}^m)_i$ and $(\mathbf{W}^m)_j$, should have the same or similar relations with those between $(\mathbf{\Phi}^m)^i$ and $(\mathbf{\Phi}^m)^j$ (the blue area in Fig. 2). To formulate this intrinsic geometric distribution, we define

$$R^m(\mathbf{W}^m) = \frac{1}{2}\sum_{i,j=1}^{D+1} g_{ij} \left\| (\mathbf{W}^m)_i - (\mathbf{W}^m)_j \right\|^2, \quad (12)$$

where g_{ij} is the element of the similarity matrix $\mathbf{G}^m = [g_{ij}^m] \in \mathbb{R}^{(D+1)\times(D+1)}$ that encodes the relation between two columns in $\mathbf{\Phi}^m$. To compute the similarity between two features in the mapped data, we use the Gaussian kernel function as the similarity measure, i.e.,

$$g_{ij}^m = exp\left(-\frac{\left\| (\mathbf{\Phi}^m)^i - (\mathbf{\Phi}^m)^j \right\|_2^2}{2\sigma^2} \right) \quad (13)$$

where σ is the kernel width, $(\mathbf{\Phi}^m)^i$ and $(\mathbf{\Phi}^m)^j$ are the i-th and j-th column in $\mathbf{\Phi}^m$. To construct the similarity matrix $\mathbf{G}^m$, we regard the features as the nodes of an adjacency graph, in which the edge weights denote the similarity of the columns in $\mathbf{\Phi}^m$. The construction of the adjacency graph follows the criteria below. If node j is one of the K nearest neighbors of node i, g_{ij}^m can be derived from Eq. (13). Otherwise, g_{ij}^m is set to zero. The manifold regularization term for feature-feature relation can be formulated as follows:

$$\begin{aligned} R(\mathbf{W}) &= \sum_{m=1}^M R^m(\mathbf{W}^m) \\ &= \frac{1}{2}\sum_{m=1}^M \sum_{i,j}^{D+1} g_{ij}^m \left((\mathbf{W}^m)_i - (\mathbf{W}^m)_j \right)^2 \\ &= \sum_{m=1}^M tr((\mathbf{W}^m)'(\mathbf{L_G})^m(\mathbf{W}^m)) \\ &= \sum_{t=1}^T \sum_{m=1}^M (\mathbf{w}_t^m)'(\mathbf{L_G})^m \mathbf{w}_t^m \end{aligned} \quad (14)$$

where $(\mathbf{L_G})^m = (\mathbf{L_H})^m - \mathbf{G}^m$, $(\mathbf{L_H})^m$ is a diagonal matrix whose diagonal elements are the column-wise sum of the similarity matrices $\mathbf{G}^m$, i.e., $(\mathbf{G}^m)_{ii} = \sum_{j=1}^{D+1} g_{ij}$. The minimization of Eq. (14) can be explained with the mechanism that the greater the similarity between $(\mathbf{\Phi}^m)^i$ and $(\mathbf{\Phi}^m)^j$, the smaller the difference between $(\mathbf{W}^m)_i$ and $(\mathbf{W}^m)_j$.

The learning criterion of the manifold-regularized multitask fuzzy system is therefore formulated as follows:

$$\min_W \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{m=1}^M \mathbf{\Phi}_t^m \mathbf{w}_t^m \right\|_2^2 + \alpha \sum_{t=1}^T \sum_{m=1}^M (\mathbf{w}_t^m)'(\mathbf{L_G})^m \mathbf{w}_t^m \quad (15)$$

The first term in Eq. (15) corresponds to supervised learning

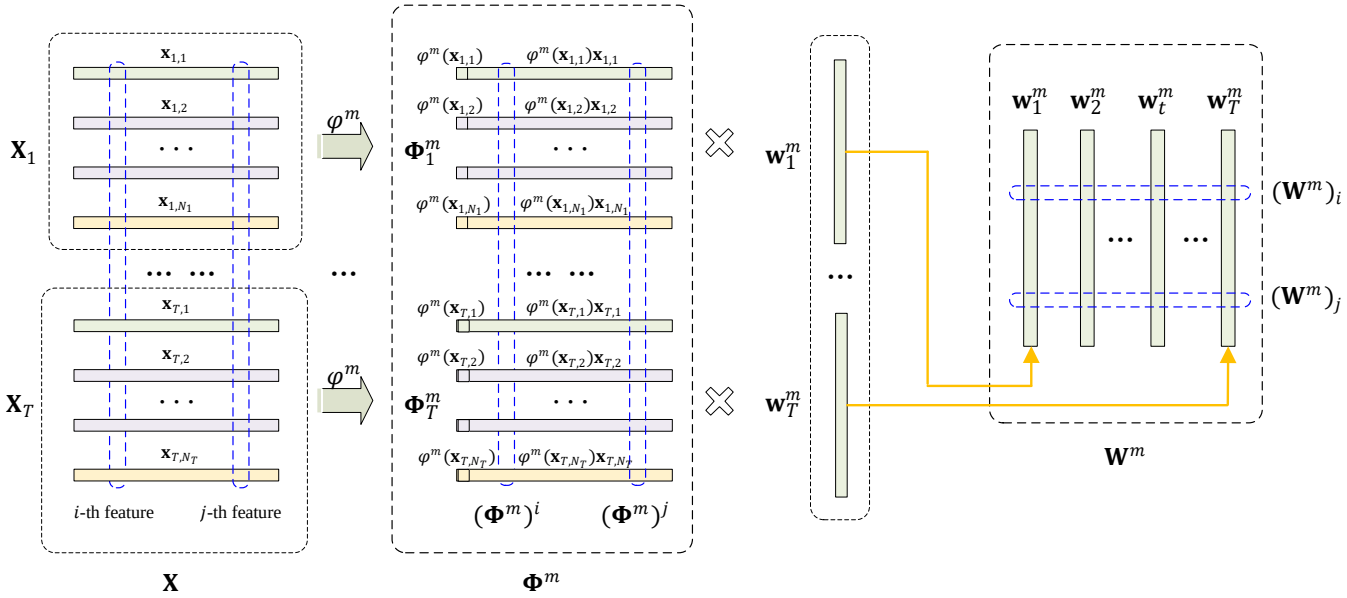

Fig. 2 Mapping the multitask data $\mathbf{X}$ into the $(D+1)$-dimensional feature space by the fuzzy mapping corresponding to the m -th rule. The blue dotted rectangles denote the 'feature-feature' relationship, which yields the manifold regularization in the learning criterion.

that involves labeled data. The second term is the manifold regularization term that corresponds to unsupervised learning on the intrinsic geometric distribution of all tasks.

B. Manifold-regularized multitask fuzzy systems with low-rank structure and sparse constraints

In multitask learning setting, different tasks share a common structure while the original data of the individual tasks have its own characteristics that are different from other tasks. We therefore decompose the joint matrix $\mathbf{W}$ into two components, i.e. $\mathbf{W}=\mathbf{V}+\mathbf{E}$, where $\mathbf{V}$ is the task-shared component, representing that the consequent parameters over multiple tasks have similar structures; $\mathbf{E}$ is the task-specific component which represents the specific characteristics of the tasks [20]. If the tasks are closely related, the task-shared component is dominant in multitask learning and the task-specific component tends to zero. On the contrary, the task-shared component tends to zero. This yields the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{V},\mathbf{E}}(\beta\cdot rank(\mathbf{V})+\lambda\cdot\|\mathbf{E}\|_1)\\&\text{s.t.}\,\mathbf{W}=\mathbf{V}+\mathbf{E}\end{aligned}\tag{16}$$

where β and λ are regularization coefficients. Obviously, it is difficult to solve $rank(\mathbf{V})$, We therefore express it in nuclear norm [21] so that Eq. (16) becomes equivalent to:

$$\begin{aligned}&\min_{\mathbf{V},\mathbf{E}}(\beta\cdot\|\mathbf{V}\|_*+\lambda\cdot\|\mathbf{E}\|_1)\\&\text{s.t.}\,\mathbf{W}=\mathbf{V}+\mathbf{E}\end{aligned}\tag{17}$$

Finally, we define the multitask learning objective function as follows:

$$\begin{aligned}\min_{\mathbf{W}}&\sum_{t=1}^{T}\left\|\mathbf{y}_t-\sum_{m=1}^{M}\boldsymbol{\Phi}_t^m\mathbf{w}_t^m\right\|_2^2\\&+\alpha\sum_{t=1}^{T}\sum_{m=1}^{M}(\mathbf{w}_t^m)'(\mathbf{L}_\mathbf{G})^m\mathbf{w}_t^m\\&+\beta\|\mathbf{V}\|_*+\lambda\|\mathbf{E}\|_1\\&\text{s.t.}\,\mathbf{W}=\mathbf{V}+\mathbf{E}\end{aligned}\tag{18}$$

III. Optimization

The optimization problem in Eq. (18) can be solved using ALM [22] which is defined as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{V},\mathbf{E},\mathbf{Y}}&\sum_{t=1}^{T}\left\|\mathbf{y}_t-\sum_{m=1}^{M}\boldsymbol{\Phi}_t^m\mathbf{w}_t^m\right\|_2^2\\&+\alpha\sum_{t=1}^{T}\sum_{m=1}^{M}(\mathbf{w}_t^m)'(\mathbf{L}_\mathbf{G})^m\mathbf{w}_t^m\\&+\beta\|\mathbf{V}\|_*+\lambda\|\mathbf{E}\|_1\\&+\langle\mathbf{Y},\mathbf{W}-\mathbf{V}-\mathbf{E}\rangle\\&+\frac{\mu}{2}\|\mathbf{W}-\mathbf{V}-\mathbf{E}\|_2^2\end{aligned}\tag{19}$$

where $\mathbf{Y}\in\mathbb{R}^{(D+1)\times T}$ is a Lagrange multiplier matrix, and $\mu>0$ is a penalty parameter; $\langle\cdot,\cdot\rangle$ denotes the inner product of two matrices. By using the LADMAP [23] method, Eq. (19) can be rewritten as,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{V},\mathbf{E},\mathbf{Y}}&\sum_{t=1}^{T}\left\|\mathbf{y}_t-\sum_{m=1}^{M}\boldsymbol{\Phi}_t^m\mathbf{w}_t^m\right\|_2^2\\&+\alpha\sum_{t=1}^{T}\sum_{m=1}^{M}(\mathbf{w}_t^m)'(\mathbf{L}_\mathbf{G})^m\mathbf{w}_t^m\\&+\beta\|\mathbf{V}\|_*+\lambda\|\mathbf{E}\|_1\\&+\frac{\mu}{2}\left\|\mathbf{W}-\mathbf{V}-\mathbf{E}+\frac{\mathbf{Y}}{\mu}\right\|_2^2\\&\text{s.t.}\,\mathbf{W}=\mathbf{V}+\mathbf{E}\end{aligned}\tag{20}$$

Eq. (20) involves four matrices $\mathbf{W}$, $\mathbf{V}$, $\mathbf{E}$ and $\mathbf{Y}$ that makes it difficult to update them simultaneously. Alternatively, we optimize each matrix in turn while fixing the others. Hence, the function can be solved iteratively as follows:

(1) Fixing $\mathbf{V}$, $\mathbf{E}$ and $\mathbf{Y}$, the sub-problem w.r.t. $\mathbf{W}$ can be written as follows:

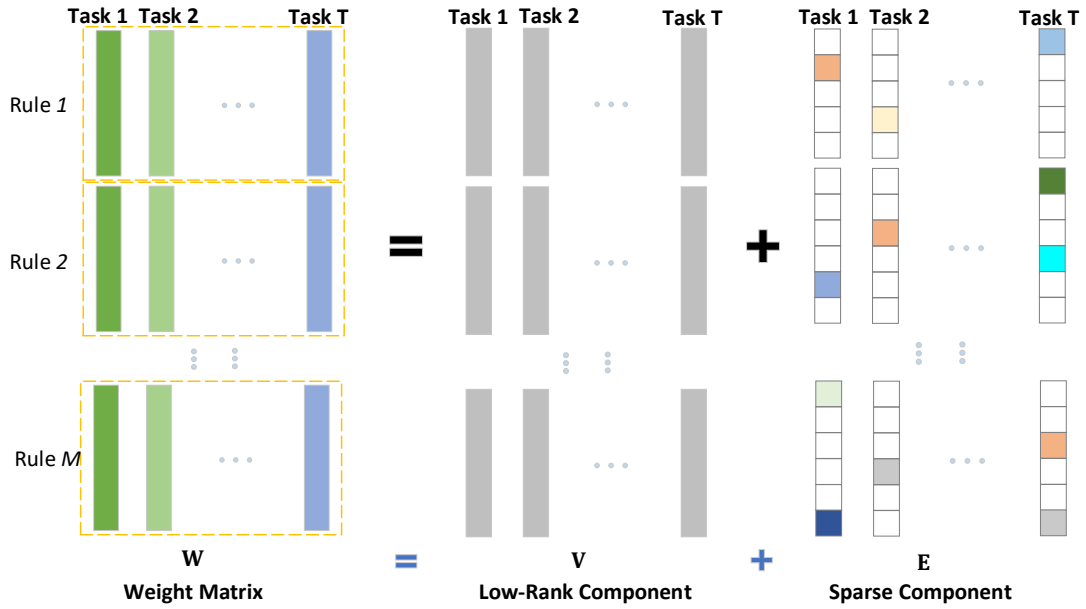

Fig. 3 In multitask learning, the joint matrix **W** can be decomposed into two components: the low-rank matrix **V** that represents the consequent parameters of multiple tasks have similar structures, and the matrix **E** that denotes the task-specific component.

$$\min_{\mathbf{W}} \sum_{t=1}^{T} \left\| \mathbf{y}_t - \sum_{m=1}^{M} \mathbf{\Phi}_t^m \mathbf{w}_t^m \right\|_2^2 + \alpha \sum_{t=1}^{T} \sum_{m=1}^{M} (\mathbf{w}_t^m)' (\mathbf{L_G})^m \mathbf{w}_t^m + \frac{\mu}{2} \left\| \mathbf{W} - \mathbf{V} - \mathbf{E} + \frac{\mathbf{Y}}{\mu} \right\|_2^2 \tag{21}$$

where $\mathbf{w}_t^m$ is a vector of the regression parameters for the m-th fuzzy rule of the t-th task and $\mathbf{w}_t = ((\mathbf{w}_t^1)', (\mathbf{w}_t^2)', \cdots, (\mathbf{w}_t^M)')' \in \mathbb{R}^{(D+1)M}$. Since the column vectors $\mathbf{w}_t$ are independent of each other, we can solve $\mathbf{W}$ by solving each $\mathbf{w}_t$ in turn. By setting the derivative of the objective function in Eq. (20) with respect to $\mathbf{w}_t^m$ to zero, we obtain:

$$-2(\mathbf{\Phi}_t^m)' \left(\mathbf{y_t} - \sum_{i=1}^{M} \mathbf{\Phi}_t^i \mathbf{w}_t^i \right) + 2\alpha (\mathbf{L_G})^m \mathbf{w}_t^m + \mu \left(\mathbf{w}_t^m - \mathbf{v}_t^m - \mathbf{e}_t^m + \frac{\mathbf{Y}_t^m}{\mu} \right) = \mathbf{0} \tag{22}$$

Setting $\mathbf{A}_t^m = 2\alpha(\mathbf{L_G})^m + \mu\mathbf{I}$, $\mathbf{B}_t^m = 2(\mathbf{\Phi}_t^m)'$, $\mathbf{C}_t^m = 2(\mathbf{\Phi}_t^m)'\mathbf{y}_t + \mu\left(\mathbf{v}_t^m + \mathbf{e}_t^m - \frac{\mathbf{Y}_t^m}{\mu}\right)$, Eq. (22) can be rewritten as follows:

$$\mathbf{A}_t^m \mathbf{w}_t^m + \mathbf{B}_t^m \sum_{i=1}^{M} \mathbf{\Phi}_t^i \mathbf{w}_t^i = \mathbf{C}_t^m \tag{23}$$

Hence, $\mathbf{w}_t$ is given by

$$\mathbf{w}_t = (\mathbf{A}_t + \mathbf{B}_t)^{-1} \mathbf{C}_t \tag{24}$$

where $\mathbf{A}_t$, $\mathbf{B}_t$ and $\mathbf{C}_t$ are block matrices which are constructed as follows:

$$\mathbf{A}_t = \begin{bmatrix} \mathbf{A}_t^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{A}_t^M \end{bmatrix}, \quad \mathbf{B}_t = \begin{bmatrix} \mathbf{B}_t^1 \mathbf{\Phi}_t^1 & \cdots & \mathbf{B}_t^1 \mathbf{\Phi}_t^M \\ \vdots & \ddots & \vdots \\ \mathbf{B}_t^M \mathbf{\Phi}_t^1 & \cdots & \mathbf{B}_t^M \mathbf{\Phi}_t^M \end{bmatrix}, \quad \mathbf{C}_t = \begin{bmatrix} \mathbf{C}_t^1 \\ \vdots \\ \mathbf{C}_t^M \end{bmatrix} \tag{25}$$

Finally, we get $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T)$.

(2) Fixing **W**, **E**, **Y**, the sub-problem w.r.t. **V** can be written as follows:

$$\min_{\mathbf{V}} \beta \|\mathbf{V}\|_* + \frac{\mu}{2} \left\| \mathbf{W} - \mathbf{V} - \mathbf{E} + \frac{\mathbf{Y}}{\mu} \right\|_2^2 \tag{26}$$

The above problem could be solved by

$$\mathbf{V} = \mathcal{D}_{\frac{\beta}{\mu}} \left(\mathbf{W} - \mathbf{E} + \frac{\mathbf{Y}}{\mu} \right) \tag{27}$$

where $\mathcal{D}$ is a singular value thresholding operator [24].

(3) Fixing **W**, **V**, **Y**, the sub-problem w.r.t. **E** can be written as follows:

$$\min_{\mathbf{E}} \lambda \|\mathbf{E}\|_1 + \frac{\mu}{2} \left\| \mathbf{W} - \mathbf{V} - \mathbf{E} + \frac{\mathbf{Y}}{\mu} \right\|_2^2 \tag{28}$$

The above problem could be solved by

$$\mathbf{E} = \mathcal{S}_{\frac{\lambda}{\mu}} \left(\mathbf{W} - \mathbf{V} + \frac{\mathbf{Y}}{\mu} \right) \tag{29}$$

(4) Finally, we update the Lagrange multiplier matrix **Y** and the regularization parameter μ using the following equations:

$$\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{W} - \mathbf{V} - \mathbf{E}) \tag{30}$$

$$\mu = \rho\mu \tag{31}$$

where ρ is a positive scalar.

The LR-S-mtTSK algorithm includes two main steps: generating the dictionary of the fuzzy rules and learning the consequent parameters. The time complexity of the first step is

$O(ITNMD)$, where I, T, N, M and D are the number of iterations, tasks, samples, clusters and features respectively. The time complexity of the second step is determined by ALM, whose time complexity is $O(ITMD^2)$. The LR-S-mtTSK algorithm is described as follows.

The LR-S-mtTSK algorithm

Input: Multitask training sets $\mathbf{X}_1,\cdots,\mathbf{X}_T$ and the corresponding labels $\mathbf{y}_1,\cdots,\mathbf{y}_T$; fuzzy rule number M; regularization parameters h, α, β, λ, μ; positive scalar $\rho > 1$;

Training process

1. Generate fuzzy dictionary:
   (i) use FCM to cluster the samples of all tasks and obtain M cluster centers.
   (ii) compute the fuzzy membership of each sample using the Gaussian function in Eq. (2) and generate the dictionary of the TSK fuzzy system for each task.
2. Jointly learn the consequent parameters of the rules across multiple tasks: Optimize $\mathbf{W}$, $\mathbf{V}$ and $\mathbf{E}$ by solving the optimization problem in Eq. (18).
   2.1. Initialization: Set $\mathbf{W}$ to random matrices, set $\mathbf{V} = \mathbf{W}$ and $\mathbf{E} = \mathbf{W} - \mathbf{V}$. Lagrange multiplier matrix $\mathbf{Y} = \frac{\mathbf{W}}{\|\mathbf{W}\|_2}$.
   2.2. while Eq. (20) not converged do
   Updating $\mathbf{W}$ using Eq. (24);
   Updating $\mathbf{V}$ using Eq. (27);
   Updating $\mathbf{E}$ using Eq. (29);
   Updating $\mathbf{Y}$ using Eq. (30);
   Updating μ using Eq. (31);
   end while
   2.3. Get the optimal consequent parameters W
3. Generate fuzzy rules: Generate fuzzy rules for each task based on the dictionary and the optimized consequent parameters.

Output: Multitask fuzzy system LR-S-MTTSFS

IV. Experiments

A. Experimental settings

Experiments were conducted to evaluate the effectiveness of LR-S-mtTSK on both regression and classification tasks using nested 5-folded cross-validation. In the experiments, the whole dataset was partitioned into 5 subsets with equal size. Then, each of the subsets was selected in turn for testing while the remaining four are used training. In the training procedure for each fold, another round of cross-validation was performed to determine the optimal hyper-parameters. This process was repeated 5 times.

For regression tasks, the Relative Root Square Error (RRSE) in Eq. (32) was used as the evaluation measure.

$$J_{RRSE} = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \Big/ \sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \qquad (32)$$

where $\hat{y}_i$ is prediction output of the model and $\bar{y}_i$ is the mean of the actual labels. A model with smaller RRSE indicates that it has better generalization ability. For classification tasks, Accuracy (ACC), Sensitivity (SEN), Specificity (SPE) in Eq. (33)-(35), Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) were used as the evaluation measures,

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \qquad (33)$$

$$Sensitivity = (TP)/(TP + FN) \qquad (34)$$

$$Specificity = (TN)/(FP + TN) \qquad (35)$$

where Accuracy, Sensitivity and Specificity are computed by true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

We compared the performance of LR-S-mtTSK with that of five regression models and seven classification models. The details of these models are presented in Table I. Table II shows the search grids adopted for setting the hyper-parameters in each method. Note that some of the models can be used for both regression and classification tasks. For single-task experiment, we run the single-task models on each task separately to test the performance.

B. Regression

1) Synthetic datasets

We evaluated the regression performance of LR-S-mtTSK on two synthetic datasets to simulate the real-world multitask learning scenes. Each synthetic data set was generated by three functions, which were considered as three similar tasks. The first dataset SIDF (same input different function) was generated by multiple functions with the same input data. To generate the

TABLE I
METHODS UNDER COMPARISON

Methods	Description	Task type
LR-S-mtTSK	The proposed LR-S-mtTSK method.	Regression/ Classification
MT-TSK-FS	Multi-task TSK fuzzy system using inter-task correlation information [18].	Regression/ Classification
Least_L21	Least squares multitask learning method with the L_21 norm regularization (Least_L21 implementation in the MALSAR package [26]).	Regression/ Classification
L2-TSFS	L2-norm penalty-based insensitive TSK fuzzy model [27].	Regression/ Classification
TSFS-SVR-L	TS-fuzzy-system-based support vector regression; linear kernel was adopted in SVR [28].	Regression
TSFS-SVR-G	TS-fuzzy-system-based support vector regression; Gaussian kernel was adopted in SVR [28].	Regression
TSFS-SVM-L	TS-fuzzy-system-based support vector machine; linear kernel was adopted.	Classification
TSFS-SVM-G	TS-fuzzy-system-based support vector machine; Gaussian kernel was adopted.	Classification
SVM-L	Classical single-task support vector machine classifier; linear kernel was adopted**Error! Reference source not found.**.	Classification
SVM-G	Classical single-task support vector machine classifier; linear kernel was adopted.	Classification

dataset, we first generated a group of samples, and then fed the samples into different functions to obtain multiple outputs. It is essentially a multiple input and multiple output (MIMO) system. The second dataset DIDF (different input different function) was generated by multiple functions with different input data. To produce the DIDF dataset, we generated three groups of samples, each corresponding to a function. After that, we fed $(x_1, \cdots, x_5)$ of each sample into the function specific to the group to obtain a multitask dataset containing several multi-input single-output learning tasks. Notice that x_i $(i = 1, \cdots, 5)$ followed uniform distribution and σ followed normal distribution. Table III gives the details of the synthetic datasets in our experiments.

TABLE II
SEARCH GRIDS OF HYPER-PARAMETERS

Methods	Search grids of hyper-parameters
LR-S-mtTSK	Rule number $M \in \{1,2,\cdots,10\}$; regularization parameters: $\alpha \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\beta \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\lambda \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\mu \in \{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$.
LR-S-mtTSK (α=0)	Rule number $M \in \{1,2,\cdots,10\}$; regularization parameters: $\beta \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\lambda \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\mu \in \{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$.
LR-S-mtTSK (β=0)	Rule number $M \in \{1,2,\cdots,10\}$; regularization parameters: $\alpha \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\lambda \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\mu \in \{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$.
LR-S-mtTSK (λ=0)	Rule number $M \in \{1,2,\cdots,10\}$; regularization parameters: $\alpha \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\beta \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}\}$, $\mu \in \{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$.
MT-TSK-FS	Rule number $M \in \{5,10,\cdots,100\}$; $\tau_k \in \{2^{-4}, 2^{-3}, \cdots, 2^{5}\}$; $\lambda \in \{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$.
Least_L21	$\gamma \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$; $\lambda \in \{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$.
L2-TSFS	Rule number $M \in \{5,10,\cdots,100\}$; $\tau \in \{2^{-4}, 2^{-3}, \cdots, 2^{5}\}$; $r \in \{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$.
TSFS-SVR-L	Rule number $M \in \{5,10,\cdots,100\}$; $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$; $r \in \{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$.
TSFS-SVR-G	Rule number $M \in \{5,10,\cdots,100\}$; $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$; $r \in \{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$;
TSFS-SVM-L	Rule number: $M \in \{5,10,\cdots,100\}$; $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$; $r \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$;
TSFS-SVM-G	Rule number: $M \in \{5,10,\cdots,100\}$; Gaussian kernel bandwidth $\sigma^2 \in \{10^{-3}, 10^{-4}, \cdots, 10^{3}\}$; $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$; $r \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$;
SVM-L	$\gamma \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$; $C \in \{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$;
SVM-G	Gaussian kernel bandwidth $\sigma^2 \in \{10^{-3}, 10^{-4}, \cdots, 10^{3}\}$; $\gamma \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$; $C \in \{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$;

2) Real-world datasets

We further evaluated the performance of LR-S-mtTSK on four real-world datasets. Details of the datasets are given as follows:

(i) The Glutamic Acid Fermentation Process dataset was generated from an MIMO system. The input variables included fermentation time h, glucose concentration $S(h)$, thalli concentration $X(h)$, glutamic acid concentration $P(h)$, stirring speed $R(h)$, and ventilation $Q(h)$, in which $h = 0,2,\ldots,28$ was the time point. The output variables included glucose concentration S(h+2), thalli concentration $X(h+2)$, and glutamic acid concentration $P(h+2)$ at a future time point $h+2$. Each output variable was regarded as a task.

(ii) The Slump dataset was used to predict concrete slump flow. It included 103 samples with 7 input variables. Slump flow was dependent on variables including cement, slag, fly ash and so on. Each output variable was regarded as a task.

(iii) The Communities and Crime dataset, used to predict Per Capita Violent Crimes, was obtained from UCI Machine Learning Repository. Each sample was a community subject with many features related to crime, such as the percentage of urban population, median household income, the number of police per capita and so on. We considered the regression for each county as a task. Five counties with similar size were selected for multitask learning.

(iv) The Housing dataset was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) and included 505 samples. Each sample in the dataset described the profile of a Boston suburb or town. Each sample had 13 feature variables. The dataset was divided into three groups by different intervals of variable "RAD" and the regression on each group was considered as a learning task.

3) Results of comparison

The comparative results of the LR-S-mtTSK and the other methods on synthetic and real-world datasets are shown in Table IV and Table V, from which the following conclusions can be drawn. First, when an appropriate algorithms is used, accurate regression model can be trained effectively under the framework of multitask learning. This implies that integrating inter-task relation into the multitask model can improve the generalization performance. On both synthetic and real-world datasets, the performance of multitask models is significantly better than that of the other single-task methods in most cases. Second, the better performance of the multitask models on the Slump dataset and the Communities and Crime dataset indicates that reasonably dividing single-task datasets into multi task datasets is helpful to achieve better prediction results. Third, LR-S-MTTSFS showed better average performance than the other methods, verifying that the integration of low-rank structure and sparse constrains into multitask fuzzy modeling is an effective approach.

TABLE III
THREE-TASK SYNTHETIC DATASETS CREATED FOR REGRESSION EXPERIMENTS

Scenes	Tasks	Functions	Inputs		σ	Size of data set
			x_1,x_2,x_3, x_5	x_4		
SIDF	Task 1	$f_1(\mathbf{x}) = 15x_1\cos(\pi x_2) + 10x_3(x_4-1)^2 + 5x_5 + \sigma$	[-1,1]	[-3,3]	[-1,1]	50
	Task 2	$f_2(\mathbf{x}) = 5x_1\cos(\pi x_2) + 15x_3(x_4-2)^2 + 10x_5 + \sigma$				50
	Task 3	$f_3(\mathbf{x}) = 10x_1\cos(\pi x_2) + 5x_3(x_4-3)^2 + 15x_5 + \sigma$				50
DIDF	Task 1	$f_1(\mathbf{x}) = 15x_1\cos(\pi x_2) + 10x_3(x_4-1)^2 + 5x_5 + \sigma$	[-1,1]	[-3,3]	[-1.1]	50
	Task 2	$f_2(\mathbf{x}) = 5x_1\cos(\pi x_2) + 15x_3(x_4-2)^2 + 10x_5 + \sigma$	[-1,1]	[-3,3]	[-1,1]	50
	Task 3	$f_3(\mathbf{x}) = 10x_1\cos(\pi x_2) + 5x_3(x_4-3)^2 + 15x_5 + \sigma$	[-1,1]	[-3,3]	[-1,1]	50

TABLE IV
REGRESSION PERFORMANCE ON SYNTHETIC MULTITASK DATASETS IN TERMS OF RRSE

Datasets	Methods / Task	LR-S-mtTSK	LR-S-mtTSK (α=0)	LR-S-mtTSK (β=0)	LR-S-mtTSK (λ=0)	MT-TSK-FS	Least_L21	L2-TSFS	TSFS-SVR-L	TSFS-SVR-G
SIDF	Task 1	0.5085 ±0.0795	**0.4678 ±0.0962**	0.5893 ±0.1386	0.6150 ±0.1390	0.5174 ±0.1541	1.0448 ±0.0445	0.7963 ±0.0397	0.4894 ±0.0568	0.6204 ±0.0767
	Task 2	**0.4307 ±0.0845**	0.4780 ±0.0853	0.5155 ±0.1098	0.5617 ±0.1645	0.5331 ±0.0958	0.9953 ±0.0446	0.8085 ±0.0411	0.4968 ±0.0855	0.4596 ±0.0724
	Task 3	**0.4218 ±0.0809**	0.4529 ±0.0600	0.6089 ±0.1404	0.5554 ±0.1148	0.4294 ±0.0673	0.7812 ±0.0446	0.6784 ±0.0353	0.4731 ±0.0467	0.4394 ±0.0474
	Average	**0.4506 ±0.0681**	0.4719 ±0.0630	0.5506 ±0.0886	0.5766 ±0.1399	0.5122 ±0.0961	0.9831 ±0.0305	0.7921 ±0.0344	0.4947 ±0.0546	0.5098 ±0.0539
DIDF	Task 1	0.4772 ±0.1082	0.5934 ±0.2470	0.7015 ±0.0655	0.5506 ±0.1505	**0.4523 ±0.0850**	1.0065 ±0.0789	0.7081 ±0.0378	0.5967 ±0.1135	0.4829 ±0.0902
	Task 2	**0.2691 ±0.0483**	0.4622 ±0.2107	0.5859 ±0.1420	0.4259 ±0.2468	0.3843 ±0.0712	1.0200 ±0.0278	0.7668 ±0.0595	0.3175 ±0.0796	0.5761 ±0.1966
	Task 3	**0.2886 ±0.0352**	0.4490 ±0.1932	0.6216 ±0.1385	0.4384 ±0.1956	0.3329 ±0.0761	0.8343 ±0.0308	0.7358 ±0.0561	0.3418 ±0.0575	0.3634 ±0.0614
	Average	**0.2979 ±0.0383**	0.4637 ±0.1788	0.6021 ±0.1105	0.4435 ±0.2215	0.3770 ±0.0682	0.9739 ±0.0167	0.7418 ±0.0388	0.3660 ±0.0419	0.5320 ±0.1332

TABLE V
REGRESSION PERFORMANCE ON REAL-WORLD DATASETS IN TERMS OF RRSE

Datasets	Method / Task	LR-S-mtTSK	LR-S-mtTSK (α=0)	LR-S-mtTSK (β=0)	LR-S-mtTSK (λ=0)	MT-TSK-FS	Least_L21	L2-TSFS	TSFS-SVR-L	TSFS-SVR-G
Fermentation	Task 1	**0.0928 ±0.0005**	0.0936 ±0.0023	0.0951 ±0.0016	0.0935 ±0.0023	0.1025 ±0.0024	0.3124 ±0.0020	0.1436 ±0.0014	0.1007 ±0.0029	0.1218 ±0.0095
	Task 2	0.3211 ±0.0167	0.3192 ±0.0141	0.3111 ±0.0101	**0.3031 ±0.0046**	0.3208 ±0.0176	1.6925 ±0.0144	0.3187 ±0.0016	0.5232 ±0.0153	0.6002 ±0.0251
	Task 3	**0.0773 ±0.0003**	0.0806 ±0.0021	0.0801 ±0.0025	0.0781 ±0.0036	0.0821 ±0.0032	0.2418 ±0.0018	0.1298 ±0.0010	0.0861 ±0.0029	0.1521 ±0.0237
	Average*	**0.0666 ±0.0002**	0.0678 ±0.0012	0.0685 ±0.0008	0.0671 ±0.0019	0.0728 ±0.0019	0.2201 ±0.0011	0.1051 ±0.0007	0.0731 ±0.0013	0.1010 ±0.0093
Slump	Task 1	1.1313 ±0.0605	1.2336 ±0.1053	1.1582 ±0.0591	1.2146 ±0.0516	1.4959 ±0.0361	1.5586 ±0.0193	**0.9526 ±0.0109**	1.2992 ±0.1970	1.6684 ±0.1855
	Task 2	**0.9482 ±0.0116**	0.9681 ±0.0562	0.9776 ±0.0389	1.0139 ±0.0199	1.0123 ±0.0462	1.0839 ±0.0161	1.0964 ±0.0241	1.0902 ±0.0863	1.1458 ±0.1165
	Task 3	0.3341 ±0.0192	0.3160 ±0.0171	0.3104 ±0.0269	**0.3084 ±0.0406**	0.3687 ±0.0251	0.3955 ±0.0090	0.7526 ±0.0376	0.4688 ±0.0826	0.3312 ±0.0440
	Average*	**0.5380 ±0.0081**	0.5514 ±0.0198	0.5638 ±0.0136	0.5684 ±0.0189	0.5716 ±0.0192	0.5781 ±0.0078	0.7064 ±0.0104	0.6187 ±0.0254	0.5912 ±0.0239
Communities and Crime	Task 1	**0.3236 ±0.0431**	0.3363 ±0.0413	0.5404 ±0.0884	0.5280 ±0.0129	0.3548 ±0.0400	0.6366 ±0.0106	0.6089 ±0.0772	0.5128 ±0.2057	0.3990 ±0.1536
	Task 2	**0.6801 ±0.0175**	0.7070 ±0.0306	0.7141 ±0.0330	0.7272 ±0.0303	0.7019 ±0.0485	0.6837 ±0.0138	0.9119 ±0.0320	0.7491 ±0.0639	0.7452 ±0.1101
	Task 3	**0.7614 ±0.0321**	0.7700 ±0.0271	0.9867 ±0.0609	1.0264 ±0.0508	0.7726 ±0.1310	0.9012 ±0.0081	1.1225 ±0.1384	1.1469 ±0.3363	1.0428 ±0.2660
	Task 4	0.7582 ±0.0501	0.8146 ±0.0509	0.8113 ±0.0540	0.7902 ±0.0678	0.7365 ±0.0383	0.8624 ±0.0997	1.0992 ±0.1032	0.8122 ±0.0850	**0.7241 ±0.0866**
	Average*	**0.5684 ±0.0064**	0.5868 ±0.0135	0.6492 ±0.0221	0.6543 ±0.0180	0.5778 ±0.0033.	0.6137 ±0.0137	0.8117 ±0.0315	0.7080 ±0.0532	0.6504 ±0.0785
Housing	Task 1	**0.3299 ±0.0131**	0.3517 ±0.0286	0.3418 ±0.0349	0.3361 ±0.0477	0.4038 ±0.0396	0.8399 ±0.0067	0.4441 ±0.0125	0.3643 ±0.0273	0.5401 ±0.0418
	Task 2	**0.4287 ±0.0212**	0.4378 ±0.0753	0.4810 ±0.0788	0.4763 ±0.0619	0.4314 ±0.0268	0.8589 ±0.0122	0.4753 ±0.0342	0.6106 ±0.1170	0.7614 ±0.0801
	Task 3	**0.6750 ±0.0295**	0.7130 ±0.0484	0.6965 ±0.0454	0.7249 ±0.0403	0.8390 ±0.1166	0.9903 ±0.0255	0.7632 ±0.0342	0.7774 ±0.0607	0.9851 ±0.0815
	Average*	**0.4220 ±0.0119**	0.4530 ±0.0486	0.4686 ±0.0355	0.4598 ±0.0340	0.4806 ±0.0248	0.7739 ±0.0087	0.4997 ±0.0152	0.5453 ±0.0670	0.6770 ±0.0478

*Average means the average performance of the multi-task model for the multiple tasks.

TABLE VI
THREE-TASK SYNTHETIC DATASETS CREATED FOR CLASSIFICATION EXPERIMENTS

Tasks (Center)	Number of subjects (Patients/NCs)	Sex (Male/Female)
NYU	183(105/78)	146/37
UCLA_1	72(32/40)	62/10
UM_1	96(54/42)	72/24
YALE	55(28/27)	40/15

TABLE VII
PERFORMANCE IN THE CLASSIFICATION OF ASD AND NC

Centers \ Models		LR-S-mtTSK	LR-S-mtTSK (α=0)	LR-S-mtTSK (β=0)	LR-S-mtTSK (λ=0)	MT-TSK-FS	Least_L21	L2_TSFS	TSFS-SVM-L	TSFS-SVM-G	SVM-L	SVM-G
NYU	ACC	**0.7158 ±0.0125**	0.6967 ±0.0173	0.6989 ±0.0151	0.6951 ±0.0119	0.7033 ±0.0217	0.6699 ±0.0221	0.6661 ±0.0121	0.6907 ±0.0132	0.6842 ±0.0160	0.6940 ±0.0232	0.6863 ±0.0318
	SEN	0.5769 ±0.1200	0.4455 ±0.1089	0.3179 ±0.0695	0.5256 ±0.1269	**0.8162 ±0.0795**	0.3449 ±0.0699	0.7590 ±0.0803	0.5487 ±0.0150	0.5641 ±0.0181	0.5564 ±0.0410	0.5897 ±0.0353
	SPE	0.8190 ±0.0834	**0.8833 ±0.0842**	0.8585 ±0.0477	0.8210 ±0.0962	0.5513 ±0.1364	0.9114 ±0.1224	0.5410 ±0.1193	0.7962 ±0.0205	0.7733 ±0.0304	0.7962 ±0.0299	0.7581 ±0.0461
	AUC	**0.7636 ±0.0128**	0.7255 ±0.0191	0.7207 ±0.0267	0.7287 ±0.0111	0.7265 ±0.0286	0.6736 ±0.0279	0.7015 ±0.0146	0.7383 ±0.0136	0.7383 ±0.0136	0.6763 ±0.0241	0.7224 ±0.0410
UCLA_1	ACC	**0.7118 ±0.0216**	0.6719 ±0.0241	0.6542 ±0.0281	0.6639 ±0.0231	0.6000 ±0.0269	0.6972 ±0.0313	0.7069 ±0.0181	0.6778 ±0.0204	0.6000 ±0.0555	0.6309 ±0.0509	0.6611 ±0.0188
	SEN	**0.7188 ±0.0728**	0.6938 ±0.1117	0.6650 ±0.1266	0.6225 ±0.1247	0.3531 ±0.2699	0.7525 ±0.1589	0.6688 ±0.1519	0.7800 ±0.0187	0.6250 ±0.1304	0.6350 ±0.0604	0.6800 ±0.0322
	SPE	0.7031 ±0.0925	0.6445 ±0.1380	0.6786 ±0.1631	0.7156 ±0.1527	**0.7975 ±0.1769**	0.6281 ±0.1107	0.7375 ±0.1195	0.5500 ±0.0375	0.5688 ±0.0996	0.6250 ±0.0988	0.6375 ±0.0468
	AUC	0.7115 ±0.0300	0.6771 ±0.0341	0.6578 ±0.0346	0.6545 ±0.0358	0.5541 ±0.0468	0.7143 ±0.0373	**0.7268 ±0.0210**	0.6709 ±0.0187	0.6709 ±0.0187	0.6359 ±0.0686	0.6553 ±0.0134
UM_1	ACC	**0.7552 ±0.0224**	0.7383 ±0.0156	0.7052 ±0.0326	0.7333 ±0.0330	0.7490 ±0.0257	0.7354 ±0.0168	0.6542 ±0.0153	0.6896 ±0.0290	0.7021 ±0.0083	0.7042 ±0.0346	0.6958 ±0.0212
	SEN	0.6429 ±0.0826	0.6399 ±0.0916	0.4762 ±0.0952	0.5905 ±0.1321	0.7704 ±0.0934	0.7333 ±0.0768	**0.8352 ±0.0818**	0.6143 ±0.0610	0.6524 ±0.0190	0.6190 ±0.0621	0.6429 ±0.0151
	SPE	**0.8426 ±0.0387**	0.8148 ±0.0857	0.8254 ±0.0764	0.8414 ±0.0764	0.7214 ±0.1049	0.7370 ±0.1325	0.4214 ±0.1038	0.7481 ±0.0251	0.7407 ±0.0166	0.7704 ±0.0343	0.7370 ±0.0296
	AUC	**0.8070 ±0.0198**	0.7718 ±0.0179	0.7503 ±0.0436	0.7617 ±0.0324	0.7761 ±0.0351	0.7774 ±0.0177	0.6145 ±0.0193	0.7352 ±0.0246	0.7352 ±0.0246	0.7651 ±0.0145	0.7487 ±0.0317
YALE	ACC	**0.7273 ±0.0493**	0.6886 ±0.0338	0.7209 ±0.0493	0.7073 ±0.0349	0.6073 ±0.0392	0.6855 ±0.0390	0.5818 ±0.0250	0.5491 ±0.0556	0.5164 ±0.0392	0.6000 ±0.0445	0.5891 ±0.0272
	SEN	**0.7037 ±0.1296**	06620 ±0.1394	0.6333 ±0.1277	0.6556 ±0.0845	0.6214 ±0.2096	0.6852 ±0.1327	0.444 ±0.099	0.5333 ±0.0798	0.4222 ±0.1493	0.5037 ±0.0687	0.5778 ±0.1231
	SPE	**0.7500 ±0.1174**	0.7143 ±0.1393	0.7245 ±0.1321	0.7571 ±0.0746	0.5926 ±0.1874	0.6857 ±0.1325	0.7143 ±0.0368	0.5643 ±0.0915	0.6071 ±0.1917	0.6929 ±0.0286	0.6000 ±0.1020
	AUC	**0.7232 ±0.0377**	0.6762 ±0.0310	0.7204 ±0.0653	0.7001 ±0.0429	0.5480 ±0.0549	0.6724 ±0.0453	0.5365 ±0.0238	0.5786 ±0.0440	0.5786 ±0.0440	0.6263 ±0.0572	0.6054 ±0.0330

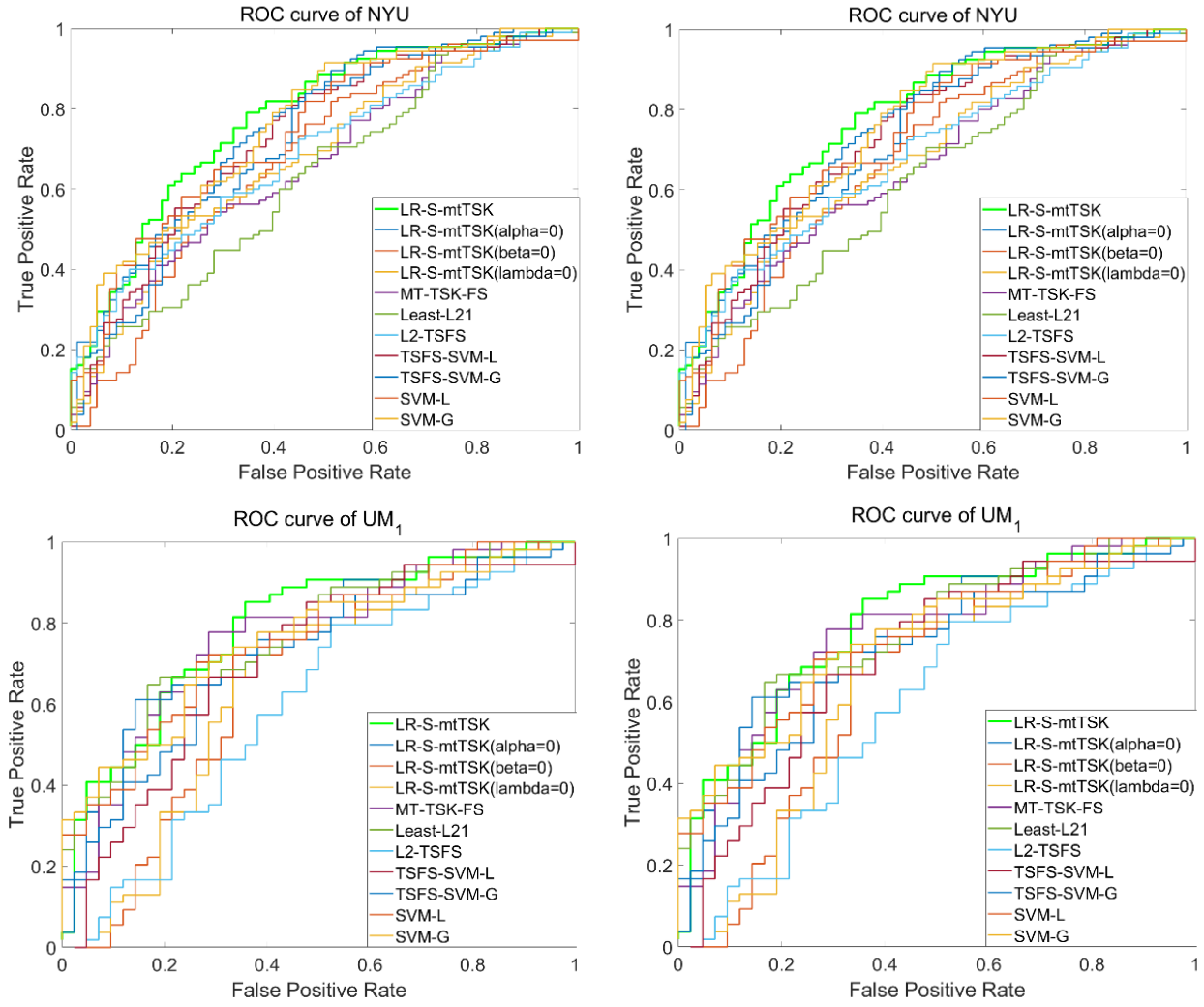

Fig. 4 ROC curves on ASD classification.

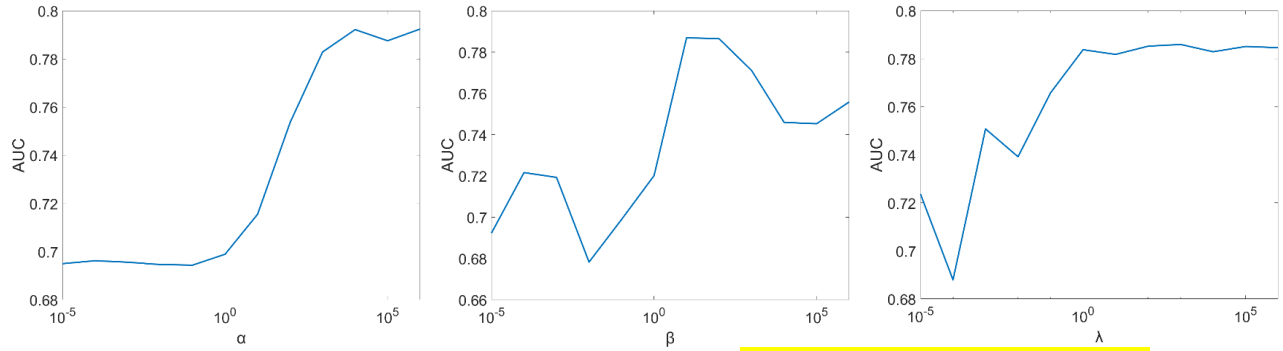

Fig. 5 Classification performance when α, β, λ take different values. (the first sub-figure misses the lower border???)

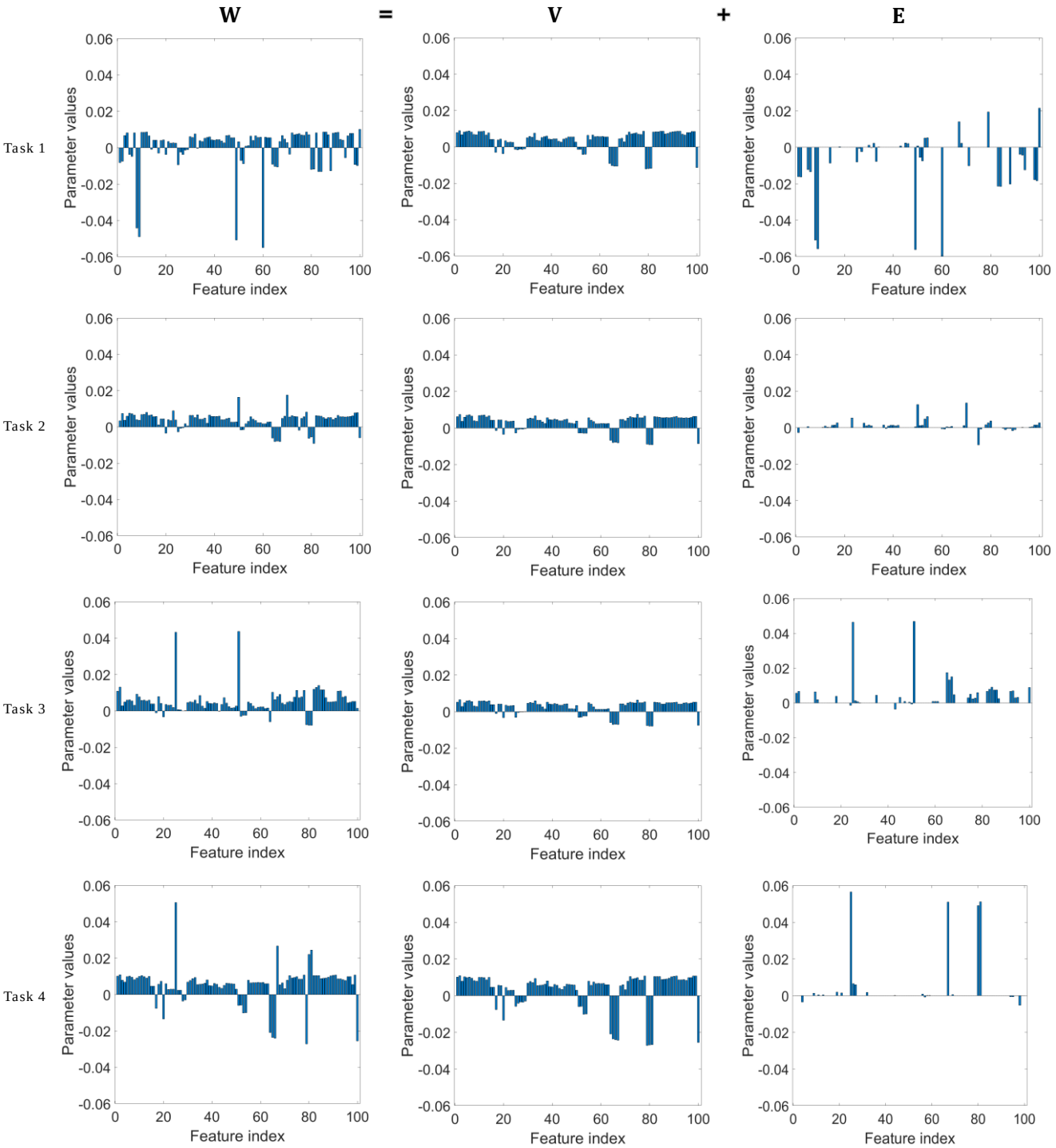

Fig. 6 In multitask learning, the joint matrix **W** is decomposed into two components: one is a low-rank matrix **V** which indicates that the consequent parameters of rules across multiple tasks have similar structures; the other is **E** that denotes the task-specific component.

C. Classification

To evaluate the performance of the proposed algorithm on classification tasks, we employed the ABIDE dataset for multi-center autism spectrum disorder (ASD) diagnosis [29], where the functional magnetic resonance imaging (fMRI) scans from four imaging centers, denoted by NYU, UCLA_1, UM_1 and YALE, were used. Details of the data were shown in Table VI. In each imaging center, the numbers of ASD patients and normal controls (NCs) were comparable. We considered ASD classification in each imaging center as a learning task. In our experiments, the fMRI scans were preprocessed using the method in [29] and a multitask dataset with 400-dimensional features were thus generated. This is a typical multitask learning problem that each task only includes a limited number of samples with high dimensional features. The proposed LR-S-mtTSK model and the other methods were evaluated using the ASD classification dataset. The results are shown in Table VII and Fig. 4. It is obvious that the LR-S-mtTSK method

has higher classification accuracy than the methods under comparison. This verify that incorporating low-rank structure and sparse consequent parameters in multitask TSK fuzzy modeling is effective in dealing with the problem of insufficient and low-quality training samples.

D. Effect of the low-rank component in $\boldsymbol{W}$

Experiments were conducted to evaluate the effect of each regularization term on ASD classification. In the experiments, the optimal hyperparameters were first determined by 5 cross-validation. Then, we fixed the other hyperparameters and varied the values of α, β, λ within $\{10^{-5}, 10^{-3}, \cdots, 10^{6}\}$. The classification performance with respect to each coefficient is shown in Fig. 5. The results show that, with reasonable hyperparameters, the proposed regularization term is beneficial to multitask modeling. We further selected 100 columns from the matrices $\mathbf{W}$, $\mathbf{V}$ and $\mathbf{E}$, and visualized their optimized values in Fig. 6. It can be seen that that $\mathbf{W}$ was decomposed into the low-rank components $\mathbf{V}$ and the sparse components $\mathbf{E}$.

V. Conclusions

While multitask modeling for TSK fuzzy systems have better generalization ability than single task modeling methods, the existing multitask modeling methods ignore the balance between the sharing of the common knowledge and the preservation of discriminative ability of the consequent parameters of the rules. To this end, we proposed a novel manifold-regularized multitask TSK fuzzy modeling method with low-rank structure and sparse consequent parameters. The method uses fuzzy clustering to obtain the fuzzy dictionaries for each task, and learns the low-rank structure of the tasks and the sparse consequent parameters of each task for the task-specific information. Furthermore, the method considers 'feature-feature' relation in each rule by introducing manifold regularizations, where ALM was used to solve the optimization problem.

The experimental results indicate that the proposed method is superior to other comparison methods and that by considering the balance between the sharing of the common knowledge and the preservation of discriminative ability, the performance of multitask fuzzy systems can be effectively improved. However, there are still issues that require further investigation. For example, we can assume that, if two samples in different tasks are similar, the corresponding outputs should also be similar. Therefore, the intrinsic 'sample-sample' relation can also be considered for multitask modeling for TSK fuzzy systems. On the other hand, the proposed method ignores the inter-task correlation in generating the premise of the fuzzy rules across multiple tasks. It is worthwhile to study the effect of introducing inter-task information into clustering on the modeling accuracy of multitask fuzzy systems.

References

[1] T. Dam, and A. Deb, "Block Sparse Representations in Modified Fuzzy C-Regression Model Clustering Algorithm for TS Fuzzy Model Identification," in 2015 IEEE Symposium Series on Computational Intelligence, 2015, pp. 1687-1694.

[2] M. Ghalehnoie, M. R. Akbarzadeh-Tootoonchi, and N. Pariz, "Fuzzy control design for nonlinear impulsive switched systems using a nonlinear Takagi-Sugeno fuzzy model," *Transactions of the Institute of Measurement and Control*, Jan 15, 2020.

[3] Y. Z. Jiang *et al.*, "Realizing Two-View TSK Fuzzy Classification System by Using Collaborative Learning," *IEEE Transactions on Systems Man Cybernetics-Systems,* vol. 47, no. 1, pp. 145-160, Jan, 2017.

[4] C. S. Li *et al.*, "TSK Fuzzy Model Identification Based on a Novel Hyperplane-Shaped Membership Function," *IEEE Transactions on Fuzzy Systems,* vol. 25, no. 5, pp. 1364-1370, Oct, 2017.

[5] E. Lughofer, and S. Kindermann, "SparseFIS: Data-Driven Learning of Fuzzy Systems With Sparsity Constraints," *IEEE Transactions on Fuzzy Systems,* vol. 18, no. 2, pp. 396-411, Apr, 2010.

[6] B. Rezaee, and M. H. F. Zarandi, "Data-driven fuzzy modeling for Takagi-Sugeno-Kang fuzzy system," *Information Sciences,* vol. 180, no. 2, pp. 241-255, Jan 15, 2010.

[7] P. Xu *et al.*, "Concise Fuzzy System Modeling Integrating Soft Subspace Clustering and Sparse Learning," *IEEE Transactions on Fuzzy Systems,* vol. 27, no. 11, pp. 2176-2189, Nov, 2019.

[8] J. Yen, L. Wang, and C. W. Gillespie, "Improving the interpretability of TSK fuzzy models by combining global learning and local learning," *IEEE Transactions on Fuzzy Systems,* vol. 6, no. 4, pp. 530-537, Nov, 1998.

[9] K. Y. Seng, I. Nestorov, and P. Vicini, "Fuzzy least squares for identification of individual pharmacokinetic parameters," *IEEE Trans Biomed Eng,* vol. 56, no. 12, pp. 2796-805, Dec, 2009.

[10] K. A. Toh, and H. L. Eng, "Between classification-error approximation and weighted leasTSKquares learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 4, pp. 658-669, Apr, 2008.

[11] T. Jebara, "Multitask Sparsity via Maximum Entropy Discrimination," *Journal of Machine Learning Research,* vol. 12, pp. 75-110, Jan, 2011.

[12] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning,* vol. 73, no. 3, pp. 243-272, Dec, 2008.

[13] P. H. Gong, J. P. Ye, and C. S. Zhang, "Multi-Stage Multi-Task Feature Learning," *Journal of Machine Learning Research,* vol. 14, pp. 2979-3010, Oct, 2013.

[14] B. Kang, W. P. Zhu, and D. Liang, "Robust multi-feature visual tracking via multi-task kernel-based sparse learning," *Iet Image Processing,* vol. 11, no. 12, pp. 1172-1178, Dec, 2017.

[15] A. Maurer, M. Pontil, and B. Romera-Paredes, "The Benefit of Multitask Representation Learning," *Journal of Machine Learning Research,* vol. 17, 2016.

[16] J. Wang *et al.*, "Multi-Task Diagnosis for Autism Spectrum Disorders Using Multi-Modality Features: A Multi-Center Study," *Human Brain Mapping,* vol. 38, no. 6, pp. 3081-3097, Jun, 2017.

[17] Y. Xue *et al.*, "Multi-task learning for classification with Dirichlet process priors," *Journal of Machine Learning Research,* vol. 8, pp. 35-63, Jan, 2007.

[18] Y. Z. Jiang *et al.*, "Multi-task TSK fuzzy system modeling using inter-task correlation information," *Information Sciences,* vol. 298, pp. 512-533, Mar 20, 2015.

[19] J. Wang *et al.*, "Multitask TSK Fuzzy System Modeling by Jointly Reducing Rules and Consequent Parameters," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1-13, 2019.

[20] Y. Zhang *et al.*, "Low-Rank-Sparse Subspace Representation for Robust Regression," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2972-2981.

[21] G. C. Liu *et al.*, "Robust Recovery of Subspace Structures by Low-Rank Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 171-184, Jan, 2013.

[22] F. Meng, X. M. Yang, and C. H. Zhou, "The Augmented Lagrange Multipliers Method for Matrix Completion from Corrupted Samplings with Application to Mixed Gaussian-Impulse Noise Removal," *Plos One,* vol. 9, no. 9, Sep 23, 2014.

[23] X. Ren, and Z. C. Lin, "Linearized Alternating Direction Method with Adaptive Penalty and Warm Starts for Fast Solving Transform Invariant Low-Rank Textures," *International Journal of Computer Vision,* vol. 104, no. 1, pp. 1-14, Aug, 2013.

[24] J. F. Cai, E. J. Candes, and Z. W. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *Siam Journal on Optimization,* vol. 20, no. 4, pp. 1956-1982, 2010.

[25] T. Zhang, and Z. M. Tang, “Improved Algorithm Based on Non-negative Low Rank and Sparse Graph for Semi-supervised Learning,” *Journal of Electronics & Information Technology,* vol. 39, no. 4, pp. 915-921, Apr, 2017.

[26] J. C. Jiayu Zhou, Jieping Ye. "Malsar: Multi-task learning via structural regularization—User’s manual version 1.1," https://github.com/jiayuzhou/MALSAR

[27] Z. H. Deng *et al.*, “Scalable TSK Fuzzy Modeling for Very Large Datasets Using Minimal-Enclosing-Ball Approximation,” *IEEE Transactions on Fuzzy Systems,* vol. 19, no. 2, pp. 210-226, Apr, 2011.

[28] C. F. Juang, S. H. Chiu, and S. J. Shiu, “Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation,” *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans,* vol. 37, no. 6, pp. 1077-1087, Nov, 2007.

[29] R. C. C. P. Bellec. "ABIDE Preprocessed," http://fcon_1000.projects.nitrc.org/indi/abide/.