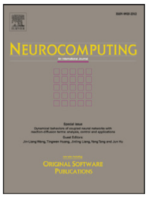
Neurocomputing 355 (2019) 134–142

Contents lists available at ScienceDirect

Neurocomputing

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



# Seemingly unrelated extreme learning machine

Li Zhao[a,b], Xizhao Wang[a,b,*]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

*Article history:*
Received 4 March 2019
Revised 12 April 2019
Accepted 18 April 2019
Available online 10 May 2019

Communicated by Dr. Nianyin Zeng

*Keywords:*
Seemingly unrelated regression (SUR) model
Extreme learning machine (ELM)
Significant correlation
Prediction error

ABSTRACT

A seemingly unrelated regression (SUR) system simultaneously studies the groups of samples that are related with each other through the covariance of decision attributes. Each group of samples is studied by a regression equation, and the error terms of the regression equations are correlated in SUR system. Extreme Learning Machine (ELM) is a training method for single-hidden layer feedforward neural network, it is widely used in many machine learning domains due to the good generalization capability and fast training speed. Since a single ELM ignores the correlated information among different groups of samples, it fails to solve the SUR problem effectively. This ineffectiveness becomes more obvious with the correlation among equations going up. In order to overcome this problem, an extended ELM model is proposed in this paper, described as Seemingly Unrelated ELM (SUELM). SUELM simultaneously learns multiple ELMs by sufficiently using the correlated information among different groups of samples, thus it can solve the SUR problem effectively. In comparison with a single ELM, SUELM significantly improves the performance. Simulation results show that SUELM performs better than the single ELM with respect to mean square error and generalization ability, especially when significant correlations exist among different groups of samples. This paper provides an effective way for solving the SUR problem by adopting ELM as the learning model.

## 1. Introduction

Seemingly unrelated regression (SUR) system, proposed by Zellner [1], studies the samples that are related with each other through the covariance of decision attributes. Since sample correlations widely exist in many real applications, SUR system has been studied in a variety of fields including geography [2], economics [3], biological sciences [4] and so on. Specifically, Hubert M [5] illustrated the influence of SUR model on studying the relationship between the foreign direct investment by multinational corporations and several macroeconomic variables. Fiebig D G [6] studied gasoline demand using a sample of data comprise 18 countries each of which had 19 annual data points. They demonstrated that SUR system could result in improvements in inferences if the procedures were applied to the t-ratios rather than to the standard errors.

Algebraically, the SUR model is

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \qquad (1)$$
$$E(\boldsymbol{\epsilon}_i) = 0,\ E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j') = \sigma_{ij} \mathbf{I}_T,\ i, j = 1, \ldots,\ N,$$

where $\mathbf{X}_i$ is a $T \times L_i$-matrix with $rank(\mathbf{X}_i) = L_i$, $\beta_i$ is a $L_i$-vector of unknown coefficients, $\boldsymbol{\epsilon}_i$ is a $T$-vector of random errors, $\mathbf{Y}_i$ is a $T$-vector and $\mathbf{I}_T$ is a $T \times T$ identity matrix.

Many works have been proposed and devoted to the development of SUR models. Zellner provided the pioneer work in this area [7], Srivastava and Giles reviewed the early literature in their book [8], and Fiebig gave a survey on this topic [9]. More specifically, a feasible ridge estimator was studied by Roozbeh et al. [10–12] to build up semiparametric SUR models; the highly accurate likelihood method was used by Fraser et al. [13] to analyze the SUR model; a direct Monte Carlo approach was derived by Zellner and Ando [14] by using Bayesian analyses method; and the best equivariant estimator was obtained by Kurata and Matsuura [15] with a symmetric error. However, when $N$ is larger than $T$, the above estimation for regression coefficients are not available since the covariance matrix is no longer positive definite. With the advent of the era of big data, more and more information concerning a certain domain can be acquired easily, and therefore, a company can use broader information to make decision. It implies that SUR model with large N will be common in practical. Zhao et al. [16] proposed an improved two-stage conditional expectation estimator, which does not need to compute the inverse of the covariance matrix. The improved two-stage conditional expectation estimator can be used in the situation of big $N$ but it suffers

[*] Corresponding author at: College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.
*E-mail addresses:* xizhaowang@ieee.org, xzwang@szu.edu.cn (X. Wang).

https://doi.org/10.1016/j.neucom.2019.04.067
0925-2312/© 2019 Elsevier B.V. All rights reserved.

from high complex computation. In order to decrease the times of iterations, Zhao and Xu [17] further proposed a generalized canonical correlation variable improved estimator which only uses the elements of the covariance matrix. It performs well when the correlation between the equations is not too small.

It is worth noting that although the SUR problem has been studied already in many works, all of them focus on the correlation coefficient analysis in statistics. Nowdays The study which connects together uncertainty and learning from data has aroused wide attention [18]. To the best of our knowledge, using SUR model for machine learning has not been investigated yet. Since sample correlation can also exist in machine learning data, how to combine SUR with machine learning models is a new problem that has high research and practical values. The main purpose of combining SUR with machine learning model is to enhance the prediction performance of a learning system through mining relationships among error terms. It is essentially a regression problem that aims to minimize the regression errors and maximize the generalization capability based on correlated samples.

On the other hand, neural network is a powerful supervised learning technique that has shown great performance on various regression problems with strong computing capability. In traditional feedforward neural networks, all parameters are adjusted iteratively by back propagation (BP) algorithm based on the gradient descent technique [19]. In recent years some kinds of learning algorithms has been proposed [20–22]. ELM is a non-iterative training method for single-hidden layer feedforward neural networks. The weight parameters connecting the hidden and input layers are randomly chosen, and the weight parameters connecting the output and hidden layers are analytically solved [23]. Many scholars have made substantial efforts to further develop the ELM model [24,25]. Among the works, different theories and techniques have been combined with ELM to improve its performance, such as Haptic recognition [26] feature selection [27] naive Bayesian [28] active learning [29] short-term load forecasting [30].

Traditional ELM gets the regression model by directly solving a linear system. It does not take into account the correlated information among samples, and therefore, cannot solve the SUR problem effectively. When strong correlation exists in the data, this ineffectiveness becomes more obvious. In order to solve this problem, this paper tries to connect SUR with ELM and proposes an ELM-based SUR model, which is named Seemingly Unrelated ELM (SUELM). It simultaneously trains a group of ELMs by incorporating the covariance information of the data, which will finally be used for the prediction.

Since the SUELM model makes a sufficient use of the correlated information in the data, it significantly improves the prediction accuracy in comparison with a single ELM, especially when one group of samples is correlated with another group. The improvement is more significant if the correlation is stronger. It is worthy of noting that the SUELM model will degenerate back to traditional ELM if the groups of samples are independent.

The rest of this paper is structured as follows. The basic definitions of SUR model and ELM will be introduced in Section 2. The framework of the SUELM model will be proposed and some of its characters will be studied in Section 3. Experimental results and an application to air quality index prediction between cities are presented in Section 4. Finally, conclusions and remarks are presented in Section 5.

## 2. Basic knowledge of SUR and ELM

The basic knowledge of SUR and ELM will be introduced in this section.

### 2.1. Seemingly unrelated regression model

SUR system consists of several individual equations. There is no explicit connection such as one equation's observation is another equation's response, but there exists an implicit relation represented by correlated disturbances of response variables.

The model (1) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \mathbf{X}_2 & \ldots & \mathbf{O} \\ \vdots & \vdots & & \vdots \\ \mathbf{O} & \mathbf{O} & \ldots & \mathbf{X}_N \end{bmatrix},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}.$$

The dimension of $\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\beta}_i, \boldsymbol{\epsilon}_i$ is the same as that in Eq. (1). Hence $\mathbf{X}$ is an $NT \times L$-matrix, both $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $NT$-vectors, $\mathbf{O}$ represents zero matrix with corresponding dimensions and $\boldsymbol{\beta}$ is $L$-vector, where $L = \sum_{i=1}^{N} L_i$. The expectation of the error term $\boldsymbol{\epsilon}$ is vector $\mathbf{0}$ and the covariance matrix of it is

$$\mathrm{COV}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_T,$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})_{N\times N}$ and the symbol $\otimes$ denotes Kronecker product of two matrices.

Different methods have been used to estimate the coefficients of the regression equations. The least squares estimator of $\beta$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = (\hat{\boldsymbol{\beta}}_{1OLS}', \ldots, \hat{\boldsymbol{\beta}}_{NOLS}')',$$

where

$$\hat{\boldsymbol{\beta}}_{iOLS} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_i, \ \ i = 1, 2, \ldots, N.$$

The residuals are

$$\hat{\boldsymbol{\epsilon}}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{iOLS} \triangleq \mathbf{N}_i\mathbf{Y}_i,$$

where $\mathbf{N}_i = \mathbf{I}_T - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i, \ i = 1, 2, \ldots, N.$

When $\boldsymbol{\Sigma}$ is known, the generalized least squares estimator of $\beta$ is

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T)\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T)\mathbf{Y}. \tag{3}$$

Commonly, $\boldsymbol{\Sigma}$ is unknown, Zellner replaced $\boldsymbol{\Sigma}$ with its consistent estimator $\hat{\boldsymbol{\Sigma}}$ and got the Zellner's two-stage estimator, that is,

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_T)\mathbf{X})^{-1}\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_T)\mathbf{Y},$$

where

$$\hat{\boldsymbol{\Sigma}} = (\sigma_{ij})_{(N\times N)}, \quad \sigma_{ij} = \frac{1}{T}\hat{\boldsymbol{\epsilon}}_i'\hat{\boldsymbol{\epsilon}}_j = \frac{1}{T}\mathbf{Y}_i'\mathbf{N}_i\mathbf{N}_j\mathbf{Y}_j. \tag{4}$$

### 2.2. Extreme learning machine

ELM is a generalized single-hidden layer feedforward neural network with random weights. Due to the non-iterative mechanism, it has a much faster training speed than traditional BP methods. Therefore, ELM has been widely used in many regression problems [31,32]. The basic framework of ELM is introduced as follows.

Suppose there is a training set that contains $N$ random samples

$$\mathcal{X} = \{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^{N} \subset \mathcal{R}^n \times \mathcal{R}^m,$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$ is the input feature vector that consists of the conditional attributes, $\mathbf{t}_i = [t_{i1}, t_{i2}, \ldots, t_{im}]$ is the output vector that consists of the decision attributes. The number of conditional attributes is $n$ and the number of decision attributes is $m$. The mathematical model of standard ELM with $\tilde{N}$ hidden nodes and activation function $g(x)$ is

$$\sum_{j=1}^{\tilde{N}} \boldsymbol{\beta}_j g(\mathbf{w}_j \cdot \mathbf{x}_i + b_j) = \mathbf{t}_i, i = 1, \ldots, N, \tag{5}$$

where $\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jn}]'$ is a weight vector connecting the input nodes and the *jth* hidden node, $\boldsymbol{\beta}_j = [\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm}]'$ is a weight vector connecting the output nodes and the $j$th hidden node, and $b_j$ is the bias of the *jth* hidden node $(j = 1, \ldots, \tilde{N})$, $\mathbf{w}_j \cdot \mathbf{x}_i$ means the inner product of $\mathbf{w}_j$ and $\mathbf{x}_i$, and sigmoid function

$$g(x) = \frac{1}{1 + exp(-x)}$$

is chosen as the activation function.

Huang et al. in [33] rewritten (5) as

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T},$$

where $\mathbf{H}$ is the hidden layer output matrix,

$$\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}}, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \begin{pmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \ldots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \ldots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}) \end{pmatrix}_{N \times \tilde{N}},$$

$\mathbf{T}$ is the decision attributes matrix denoted as

$$\mathbf{T} = \begin{pmatrix} \mathbf{t}_1' \\ \vdots \\ \mathbf{t}_N' \end{pmatrix}_{N \times m},$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_{\tilde{N}}' \end{pmatrix}_{\tilde{N} \times m}.$$

Conventionally, for the sake of training a single-hidden layer feedforward neural network, we expect to find specific $\hat{\mathbf{w}}_i, \hat{b}_i, \hat{\boldsymbol{\beta}}$ $(i = 1, \ldots, \tilde{N})$ such that

$$\|\mathbf{H}((\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{\tilde{N}}, \hat{b}_1, \ldots, \hat{b}_{\tilde{N}})\hat{\boldsymbol{\beta}} - \mathbf{T}\| = \min_{\mathbf{w}_i, b_i, \boldsymbol{\beta}} \|\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}})\boldsymbol{\beta} - \mathbf{T}\|$$

which is amount to minimizing the cost function

$$E = \sum_{i=1}^{N} \left( \sum_{j=1}^{\tilde{N}} \boldsymbol{\beta}_j g(\mathbf{w}_j \cdot \mathbf{x}_i + b_j) - \mathbf{t}_i \right)^2. \tag{6}$$

The input weights $\mathbf{w}_i$ and hidden layer biases $b_i$ are chosen randomly in the ELM model. For given $\mathbf{w}_i$ and $b_i$, $1 \leq i \leq \tilde{N}$, training the ELM is amount to finding a least square estimation of linear regression equation $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$, i.e.,

$$\|\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}},\ b_1, \ldots, b_{\tilde{N}})\hat{\boldsymbol{\beta}} - \mathbf{T}\| = \min_{\boldsymbol{\beta}} \|\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}},\ b_1, \ldots, b_{\tilde{N}})\boldsymbol{\beta} - \mathbf{T}\|. \tag{7}$$

From the literature [34], the least-squares solution of (7) is $\boldsymbol{\beta} = \mathbf{H}^\dagger\mathbf{T}$, where $\mathbf{H}^\dagger$ is the Moore-Penrose generalized inverse matrix of $\mathbf{H}$.

Since the parameters in ELM need not to be adjusted iteratively, the training speed of it is much faster than the conventional gradient-based learning algorithms. Huang et al. [35] prove that when the number of hidden nodes is infinitely approximate to the number of training samples, ELM can be infinitely close to any given function. However, when significant correlations exist among the groups of samples, a single ELM learned in isolation may have bad generalization capability and serious overfitting problem. Hence, it is expected to propose an improved ELM method by making use of the correlated information in data.

## 3. Seemingly unrelated extreme learning machine

The SUELM model will be proposed in this section, followed by some discussions on its framework, mathematical formulation, and learning algorithm.

### 3.1. The framework of SUELM

Seemingly unrelated samples refer to $K$ groups of samples that are correlated with each other through the decision attributes. The conditional attributes of each group can be the same or different. For example, group $p$ can have conditional attributes

$$\mathbf{x}^p = [x_1, x_2, x_3, x_4],$$

while group $q$ can have conditional attributes

$$\mathbf{x}^q = [x_2, x_4, x_5, x_6].$$

It is highlighted that the relation among the $K$ groups is reflected by the covariance of decision attributes, not by the conditional attributes. It is also the reason why we call it seemingly unrelated samples.

Suppose we have $K$ groups of seemingly unrelated samples. Each group has $N$ independent observations, i.e., we have $N \times K$ distinct samples

$$\{(\mathbf{x}_i^p, t_i^p),\ \ i = 1, 2, \ldots, N,\ \ p = 1, 2, \ldots, K\},$$

where $\mathbf{x}_i^p = [x_{i1}^p, x_{i2}^p, \ldots, x_{in_p}^p] \in \mathcal{R}^{n_p}$, $t_i^p \in \mathcal{R}^1$, superscript $p$ indicates which group the sample belongs to, and

$$cov(t^p, t^q) = \sigma_{pq}, \quad p = 1, \ldots, K,\ \ q = 1, \ldots, K,$$

there exist $p \neq q$ such that $\sigma_{pq} \neq 0$. When using ELM to solve this problem, traditional method is to train $K$ independent ELMs respectively based on the $K$ groups of samples, without considering the covariance information among different groups. In order to improve the performance, we propose the SUELM model. It trains the $K$ ELMs at the same time, then optimizes the model by combining the outputs of the $K$ ELMs based on some correlation analyses. The structure of SUELM is shown in Fig. 1. Similar to traditional model, there are three layers in the *pth* ELM in SUELM. The first layer is the input layer. The number of input nodes equals to the number of conditional attributes used to describe samples. We suppose that the input layer has $n_p$ nodes, and each node represents the real input value of a conditional attribute. The second layer is the hidden layer which contains $\tilde{N}_p$ nodes, and the last layer is the output layer that has only one node.

The *pth* ELM in SUELM with activation function $g(x)$ and $\tilde{N}_p$ hidden nodes is given as

$$\sum_{j=1}^{\tilde{N}_p} \boldsymbol{\beta}_j^p g(\mathbf{w}_j^p \cdot \mathbf{x}_i^p + b_j^p) = t_i^p, i = 1, \ldots, N, \tag{8}$$

where $\mathbf{w}_j^p = [w_{j1}^p, w_{j2}^p, \ldots, w_{jn_p}^p]$ is the weight vector connecting the *jth* hidden node and the input nodes; $\boldsymbol{\beta}_j^p$ is the output weight connecting the *jth* hidden node and the output node; and $b_j^p$ is the bias of the *jth* hidden node, where $1 \leq j \leq \tilde{N}_p$.

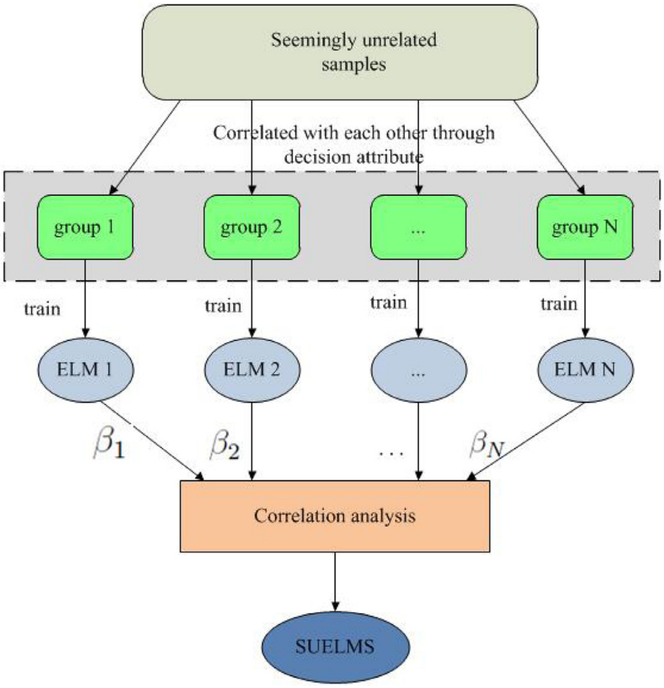


**Fig. 1.** The SUELM.

### 3.2. The mathematical model of SUELM

Once the values of $\mathbf{w}_i$ and $b_i$ are fixed, traditional method is to estimate weights $\beta^p$ using least-squares estimation. However, it neglects the correlations of decision attributes among the seemingly unrelated samples, thus the estimator for the individual equation is usually sub-optimal. In the proposed SUELM, the least squares estimator for the $p$th ELM is

$$\hat{\boldsymbol{\beta}}^p_{ols} = (\mathbf{H}'_p\mathbf{H}_p)^{-1}\mathbf{H}'_p\mathbf{T}_p$$

where

$$\mathbf{H}_p = \begin{pmatrix} g(\mathbf{w}^p_1 \cdot \mathbf{x}^p_1 + b^p_1) & \dots & g(\mathbf{w}^p_{\tilde{N}} \cdot \mathbf{x}^p_1 + b^p_{\tilde{N}}) \\ \vdots & \vdots & \vdots \\ g(\mathbf{w}^p_1 \cdot \mathbf{x}^p_N + b^p_1) & \dots & g(\mathbf{w}^p_{\tilde{N}} \cdot \mathbf{x}^p_N + b^p_{\tilde{N}}) \end{pmatrix}_{N\times\tilde{N}}, \qquad (9)$$

and

$$\mathbf{T}_p = \begin{pmatrix} t^p_1 \\ \vdots \\ t^p_N \end{pmatrix}_{N\times 1}.$$

It is worth noting that $\hat{\boldsymbol{\beta}}^p_{ols}$ minimizes the cost function

$$E = \sum_{i=1}^{N}\sum_{j=1}^{\tilde{N}}(\boldsymbol{\beta}^p_j g_j(\mathbf{w}^p_j \cdot \mathbf{x}^p_i + b^p_j) - t^p_i)^2. \qquad (10)$$

Obviously, the least squares estimator guarantees to minimize the training error, but the testing error cannot be minimized. In SUELM, we consider the correlation between the decision attributes, and train the $K$ ELMs at the same time. Let

$$\mathbf{H} = diag(\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_K)_{NK\times NK} \qquad (11)$$

and

$$\mathbf{T} = [\mathbf{T}'_1, \mathbf{T}'_2, \dots, \mathbf{T}'_K]'_{NK\times 1}$$

where

$$Cov(\mathbf{T}) = \boldsymbol{\Sigma^{-1}} \otimes \mathbf{I}_N,$$

and $\boldsymbol{\Sigma} = (\sigma_{pq})_{K\times K}$.

Based on Eq. (3), we can get

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{H}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{H})^{-1}\mathbf{H}'(\boldsymbol{\Sigma^{-1}} \otimes \mathbf{I}_N)\mathbf{T}.$$

In general, $\boldsymbol{\Sigma}$ is unknown, replace $\boldsymbol{\Sigma}$ with its consistent estimator $\hat{\boldsymbol{\Sigma}}$, we have

$$\hat{\boldsymbol{\beta}}_{FG} = (\mathbf{H}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\mathbf{H})^{-1}\mathbf{H}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\mathbf{T},$$

where $\hat{\boldsymbol{\Sigma}}$ is based on the residuals of least squares solution $\hat{\boldsymbol{\beta}}^p_{ols}$, i.e.,

$$\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{pq})_{(K\times K)},$$

$$\hat{\sigma}_{pq} = \frac{1}{N}(\mathbf{T}_p - \mathbf{H}_p\hat{\boldsymbol{\beta}}^p_{ols})'(\mathbf{T}_q - \mathbf{H}_q\hat{\boldsymbol{\beta}}^q_{ols}).$$

We adjust the weight parameters $(\hat{\boldsymbol{\beta}}^{1'}_{ols}, \hat{\boldsymbol{\beta}}^{2'}_{ols}, \dots, \hat{\boldsymbol{\beta}}^{K'}_{ols})'$ to $\hat{\boldsymbol{\beta}}_{FG}$.

Finally, the training and testing processes of SUELM are described in Algorithms 1 and 2, respectively. In comparison with

**Algorithm 1:** The SUELM trainning algorithm.

```
Input:
    Training set with K groups of samples,
    $\mathcal{X} = \{(\mathbf{x}^p_i, t^p_i)|\mathbf{x}^p_i \in \mathcal{R}^{n_p}, t \in \mathcal{R}, i = 1, \dots, N,\ p = 1, \dots, K\}$
  where $N$ is the number of training samples in each group, $\mathcal{R}$ is aset of real numbers;
    Activation function $g(x)$;
    Number of hidden neorons $\tilde{N}_p,\ p = 1, \dots, K$.
Output:
    Parameters $\mathbf{w}$, $\mathbf{b}$ and $\hat{\boldsymbol{\beta}}_{FG}$.
1 : Randomly choose input weight $\mathbf{w}^p_i$ and bias $b^p_i$ for each ELM, $i = 1, \dots, \tilde{N}_p,\ p = 1, \dots, K$.
2 : Compute the hidden layer output matrix $\mathbf{H}_p,\ p = 1, \dots, K$ and $\mathbf{H}$ according to Eqs. (9) and (11).
3 : Calculate the output weights beta for each ELM based on least squares method
        $\hat{\boldsymbol{\beta}}^p_{ols} = (\mathbf{H}'_p\mathbf{H}_p)^{-1}\mathbf{H}'_p\mathbf{T}_p$, where $p = 1, 2, \dots, K$.
4 : Compute the correlation between each group
      $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{pq})_{(K\times K)}$,
      $\hat{\sigma}_{pq} = \frac{1}{N}(\mathbf{T}_p - \mathbf{H}_p\hat{\boldsymbol{\beta}}^p_{ols})'(\mathbf{T}_q - \mathbf{H}_q\hat{\boldsymbol{\beta}}^p_{ols})$.        (12)
5 : Compute the general feasible output weight
        $\hat{\boldsymbol{\beta}}_{FG} = (\mathbf{H}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\mathbf{H})^{-1}\mathbf{H}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\mathbf{T}$.
```

**Algorithm 2:** The SUELM testing algorithm.

```
Input:
    Testing set with K groups of samples,
    $\mathcal{X} = \{(\hat{\mathbf{x}}^p_i)|\hat{\mathbf{x}}^p_i \in \mathcal{R}^{n_p}, i = 1, \dots, N',\ p = 1, \dots, K\}$ where $N'$ is
  the number of testing samples in each group;
    Activation function $g(x)$;
    Number of hidden neorons $\tilde{N}_p,\ p = 1, \dots, K$;
    Parameters $\mathbf{w}$, $\mathbf{b}$ and $\hat{\boldsymbol{\beta}}_{FG}$.
Output:
    Prediction result $\hat{\mathbf{T}}$.
1 : Compute the hidden layer output matrix $\mathbf{H}_p,\ p = 1, \dots, K$ and $\mathbf{H}$ according to Eqs. (9) and (11).
2 : Compute the output $\hat{\mathbf{T}} = \mathbf{H}\hat{\boldsymbol{\beta}}_{FG}$.
```

the original ELM, SUELM additionally needs to calculate the correlation, which has a very small amount of computation load. Therefore our proposed model has a computational cost almost as same as the standard ELM has. It mainly is to compute a generalized inverse of matrix.

### *3.3. The characteristic of SUELM algorithm*

In this section, we investigate the main characteristic of SUELM, and analyze its advantage in solving SUR problem. Considering the weight parameters between hidden and output layers in ELM as a random vector, in the following, we give a theoretical analysis on the prediction error of SUELM.

**Theorem 1.** *For seemingly unrelated samples, in the class of linear unbiased estimators of* $\mathbf{w}_j$, *if the covariance of the decision attribute* $\boldsymbol{\Sigma}$ *is known, the SUELM can yield the minimal mean square prediction error.*

**Proof.** Suppose $\mathbf{X}_i^p$, $p = 1 \ldots K$ is an $n_p$-dimensional vector of conditional attributes in the *pth* group of samples, and $t_i^p$ is the corresponding decision attribute. The hidden layer output matrix of the *pth* SUELM is $\mathbf{H}_o^p$. And the output of it is $\hat{t}_o^p$. Then the output of the SUELM is

$$\hat{\mathbf{T}}_o = (\hat{t}_o^1, \hat{t}_o^2, \ldots \hat{t}_o^K).$$

Let $\mathbf{H}_o = diag(\mathbf{H}_o^1, \mathbf{H}_o^2, \ldots \mathbf{H}_o^k)$. Consider the unbiased linear estimator of weight say $\hat{\boldsymbol{\beta}} = \mathbf{AT}$.

Since $\mathbf{AT}$ is an unbiased estimator of $\boldsymbol{\beta}$, we know that

$$E(\mathbf{AT}) = \boldsymbol{\beta}$$

According to $\mathbf{T} = \mathbf{H}\boldsymbol{\beta}$, it follows that $\mathbf{AH}\boldsymbol{\beta} = \boldsymbol{\beta}$. It leads to

$$\mathbf{AH} = \mathbf{I}, \tag{13}$$

and the expectation of $\hat{\mathbf{T}}_o$ is $E(\hat{\mathbf{T}}_o) = E(\mathbf{H}_o\mathbf{AT}) = E(\mathbf{H}_o\mathbf{AH}\boldsymbol{\beta}) = E(\mathbf{H}_o\boldsymbol{\beta}) = E(\mathbf{T}_o)$. The prediction error can be written as

$$\begin{aligned}\mathbf{e}_o = \hat{\mathbf{T}}_o - \mathbf{T}_o = \hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o) + E(\mathbf{T}_o) - (\mathbf{T}_o) = (\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o))\\ - (\mathbf{T}_o - E(\mathbf{T}_o)) = (\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o)) - \boldsymbol{\epsilon_o}.\end{aligned}$$

where $\boldsymbol{\epsilon}_o = \mathbf{T}_o - E(\mathbf{T}_o)$.

The mean square prediction error is expressed as

$$\begin{aligned}E(\mathbf{e}_o'\mathbf{e}_o) =& E[(\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o)) - \boldsymbol{\epsilon_o}]'[(\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o)) - \boldsymbol{\epsilon_o}]\\ =& E(\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o))'(\hat{\mathbf{T}}_o - E(\hat{\mathbf{T}}_o)) + E\boldsymbol{\epsilon_o}'\boldsymbol{\epsilon_o}\\ =& E(\mathbf{H}_o\mathbf{AT} - \mathbf{H}_o\boldsymbol{\beta})'(\mathbf{H}_o\mathbf{AT} - \mathbf{H}_o\boldsymbol{\beta}) + E\boldsymbol{\epsilon_o}'\boldsymbol{\epsilon_o}\\ =& E(\mathbf{H}_o\mathbf{AT} - \mathbf{H}_o\mathbf{AH}\boldsymbol{\beta})'(\mathbf{H}_o\mathbf{AT} - \mathbf{H}_o\mathbf{AH}\boldsymbol{\beta}) + E\boldsymbol{\epsilon_o}'\boldsymbol{\epsilon_o}\\ =& Etr(\boldsymbol{\epsilon}'\mathbf{A}'\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\epsilon}) + E(\boldsymbol{\epsilon_o}'\boldsymbol{\epsilon_o})\\ =& Etr(\mathbf{A}'\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\epsilon\epsilon') + E(\boldsymbol{\epsilon_o}'\boldsymbol{\epsilon_o})\\ =& trA'\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\Sigma} + tr\boldsymbol{\Sigma}_c.\end{aligned}$$

In order to find the minimum of $E(\mathbf{e}_o'\mathbf{e}_o)$ under the given constraint conditions (13), we let

$$L(\mathbf{A}, \mathbf{M}) = tr\mathbf{A}'\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\Sigma} + tr\boldsymbol{\Sigma}_c + tr\mathbf{M}'(\mathbf{AH} - \mathbf{I}), \tag{14}$$

where $\mathbf{M}$ is a matrix of Lagrange multipliers.

Taking the derivative of (14) with respect to $\mathbf{A}$, we get

$$\frac{\partial L(\mathbf{A}, \mathbf{M})}{\partial(\mathbf{A})} = 2\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\Sigma} - \mathbf{MH}'.$$

Let

$$2\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\Sigma} - \mathbf{MH}' = 0, \tag{15}$$

and by multiplying from the right with $\boldsymbol{\Sigma}^{-1}\mathbf{H}$, we obtain that

$$2\mathbf{H}_o'\mathbf{H}_o\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{H} - \mathbf{MH}'\boldsymbol{\Sigma}^{-1}\mathbf{H} = 0.$$

According to condition (13), we have that

$$\mathbf{M} = 2\mathbf{H}_o'\mathbf{H}_o(\mathbf{H}'\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}. \tag{16}$$

Substituting Eq. (16) into Eq. (15), we have that

$$\mathbf{A} = (\mathbf{H}'\boldsymbol{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}'\boldsymbol{\Sigma}^{-1}.$$

Furthermore

$$\mathbf{AT} = \hat{\boldsymbol{\beta}}_{GLS}.$$

Thus we prove that the $\hat{\boldsymbol{\beta}}_{GLS}$ minimizes the mean-squared error of forcast. □

From Theorem 1, we can see that in the groups of seemingly unrelated samples, when the covariance of the decision attributes is known, the SUELM yields the minimal mean square prediction error. However, in practice, the covariance is hard to know, we have to estimate the unknown covariance based on the residuals. In this case, SUELM is more efficient than ELM only when the samples are highly correlated.

## 4. Performance evaluation

The experimental comparisons between SUELM and ELM will be given in this section.

### *4.1. Simulation data*

We first describe the data sets used for our simulations. Suppose that our data sets are denoted as

$$\{(\mathbf{x}_i^p, \mathbf{t}_i^p)\}_{i=1}^N \subset \mathcal{R}^{n_p} \times \mathcal{R},\ p = 1, 2,$$

where $p = 1, 2$ represents 2 groups and $i = 1, 2, \ldots, N$ represents $N$ observations for each group. For given $i$ and $p$, the conditional attribute $\mathbf{x}_i^p$ is a $K$-dimensional vector which is sampled from a $K$-variate normal distribution denoted as $N(0, I)$, where the mean is a $K-$dimensional zero vector and the variance is a $K \times K$ identity matrix. We suppose that $\mathbf{x}_i^p$ $(i = 1, 2, \ldots, N)$ are independently and identically distributed (i.i.d.) for $p = 1$ and 2. It is noted that, for given $i$ and $p$, the decision attribute $\mathbf{t}_i^p$ is a real value. Furthermore, we assume that the vector

$$(\mathbf{t}_1^1, \mathbf{t}_2^1, \ldots, \mathbf{t}_N^1)'$$

is correlated with

$$(\mathbf{t}_1^2, \mathbf{t}_2^2, \ldots, \mathbf{t}_N^2)',$$

i.e., the two output variables are related to each other. Then we can generate the two vectors, i.e., the matrix $(\mathbf{t}_i^1, \mathbf{t}_i^2)_{N\times 2}$ based on the normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where $\rho$ is a value in [0,1], representing the correlation between $\mathbf{t}^1$ and $\mathbf{t}^2$. In summary, for give $\rho \in [0, 1]$ we generate two real-valued matrices with $N$ rows and $(K+1)$ columns i.e., $(\mathbf{x}_i^1, \mathbf{t}_i^1)_{N\times(K+1)}$ and $(\mathbf{x}_i^2, \mathbf{t}_i^2)_{N\times(K+1)}$.

### *4.2. Evaluation criterion*

The mean squared error (MSE) is used to evaluate the performance of the SUELM and the single ELM. The MSE of a predictor is given as follows:

$$MSE = \sum_{i=1}^{N}(\hat{\mathbf{t}}_i - \mathbf{t}_i)^2/N,$$

$$MSE = \sum_{i=1}^{N} \frac{((\hat{\mathbf{y}}_i - \mathbf{y}_i) \cdot (\hat{\mathbf{y}}_i - \mathbf{y}_i))}{N},$$

where $\hat{\mathbf{t}}_i$ is the $n$-dimensional vector predicted by the model and $\mathbf{t}_i$ is the ground truth vector.

MSE corresponds to the expected value of the squared error loss. Statistically, MSE of a predictor measures the average of the squares of the errors, that is, the difference between the predicted value and its ground truth values. MSE is the second moment of the error which is considered as a random variable. MSE combines the variance of the predictor with its bias. When the estimator is unbiased, MSE is the variance of the predictor. It is an easily computable quantity. The same as the variance, MSE has the same measurement units as the square of the quantity being estimated.

The root-mean-square error (RMSE) of a predictor is obtained by taking the square root of MSE [36], which is given as:

$$RMSE = \sqrt{\sum_{i=1}^{N} (\hat{\mathbf{t}}_i - \mathbf{t}_i)^2 \Big/ N}.$$

A similar evaluation criterion is the mean absolute error (MAE) of the predictor [37], which is given as:

$$MAE = \frac{\sum_{i=1}^{N} |\hat{\mathbf{t}}_i - \mathbf{t}_i|}{N}.$$

### 4.3. Experimental process

According to Section 4.1, we have two real-valued matrices with $N$ rows and $(K+1)$ columns for given $\rho$. We then select 50% samples from each group to form 2 training sets, and the remaining 50% samples are taken as 2 testing sets. From each training set we train 2 individual ELM models. For convenience, we set the number of nodes in the two ELMs in the same way. The input weights and hidden layer biases are chosen randomly. The number of input nodes equals to the number of conditional attributes. The number of output nodes equals to the number of decision attributes. We try different number of hidden nodes, and find that, if the number of hidden nodes is small, the training accuracy will be low. If the number of hidden nodes is large, the model will be overfitting. We list the training and testing accuracy for different numbers of hidden nodes, and then choose the optimal one. No other parameters to tune. Then by using the correlated information we simultaneously train the two ELMs which form the SUELM. The degree of correlation is represented by $\rho$.

First, we evaluate the relationship between data correlation and the extent of improvement of SUELM to ELM. We set the number of input nodes as 8, the number of hidden nodes as 15, and the sigmoid activation function is employed. To illustrate that the higher correlation the two groups have, the better the prediction accuracy will be, $\rho$ is chosen to be $\frac{i}{20}$, where $i = 0, 1, \ldots, 19$. The improvement is evaluated by the relative mean square error (rMSE)

$$rMSE = \frac{MSE_{ELM} - MSE_{SUELM}}{MSE_{ELM}}.$$

Fig. 2 shows the simulation results.

Then, we evaluate the effect of the hidden nodes number on the performance of the trained models. The number of the input nodes in ELM is set as 8, and the correlation is set as 0.9. We choose different number of hidden nodes, as $\tilde{N} = 2, 3, \ldots, 19$ and $20, 30, \ldots, 200$. The training error and prediction error are displayed in Table 1 and Fig. 3.

### 4.4. Result and discussion

In this section we conduct some statistical analyses of experimental data. Fig. 2 describes the relation between the improvement of testing accuracy and the correlation among the groups of samples. By equally dividing the interval between the maximum and minimum absolute values of correlation coefficients, twenty relevance levels are formed. For the sake of clarity, we only plot the results with fixed number of input nodes and number of hidden nodes. As a notation, the same conclusion can be obtained for different numbers of input nodes and hidden nodes, i.e., when certain correlation exists among different groups of samples, SUELM

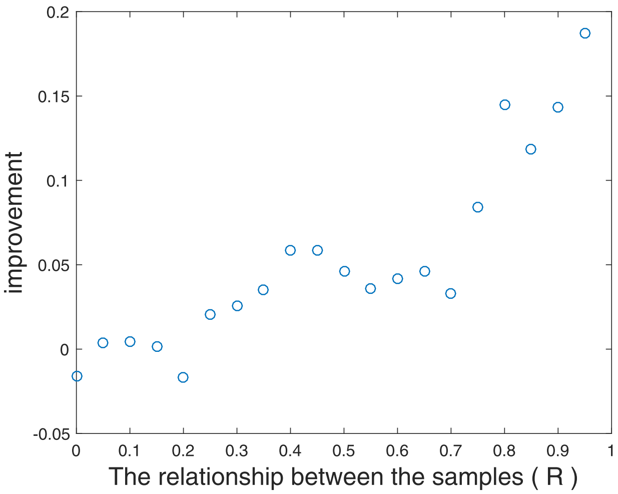
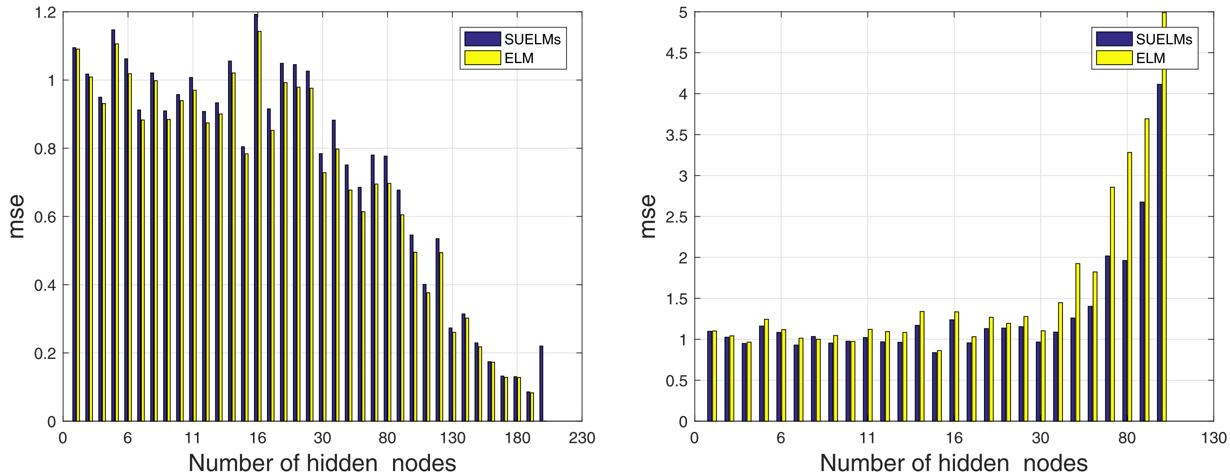


**Fig. 2.** The relation between the improvement of testing accuracy and the correlation among the groups of samples.

**Table 1**
Comparison of the SUELM and the single ELM with different number of hidden nodes.

| $\tilde{N}$ | Training MSE | | Testing MSE | |
|---|---|---|---|---|
| | SUELM | ELM | SUELM | ELM |
| 2 | 1.0948 | 1.0906 | 1.0975 | 1.1011 |
| 3 | 1.0174 | 1.0087 | 1.0257 | 1.0416 |
| 4 | 0.9496 | 0.9312 | 0.9493 | 0.9653 |
| 5 | 1.1471 | 1.1060 | 1.1618 | 1.2444 |
| 6 | 1.0621 | 1.0184 | 1.0843 | 1.1181 |
| 7 | 0.9123 | 0.8828 | 0.9299 | 1.0141 |
| 8 | 1.0207 | 0.9979 | 1.0342 | 1.0007 |
| 9 | 0.9095 | 0.8845 | 0.9542 | 1.0458 |
| 10 | 0.9574 | 0.9395 | 0.9770 | 0.9734 |
| 11 | 1.0073 | 0.9701 | 1.0218 | 1.1217 |
| 12 | 0.9078 | 0.8742 | 0.9688 | 1.0941 |
| 13 | 0.9332 | 0.9004 | 0.9641 | 1.0837 |
| 14 | 1.0555 | 1.0207 | 1.1706 | 1.3395 |
| 15 | 0.8043 | 0.7835 | 0.8377 | 0.8622 |
| 16 | 1.1921 | 1.1425 | 1.2381 | 1.3346 |
| 17 | 0.9155 | 0.8524 | 0.9564 | 1.0315 |
| 18 | 1.0489 | 0.9923 | 1.1308 | 1.2686 |
| 19 | 1.0453 | 0.9791 | 1.1368 | 1.1954 |
| 20 | 1.0260 | 0.9759 | 1.1542 | 1.2779 |
| 30 | 0.7841 | 0.7282 | 0.9665 | 1.1038 |
| 40 | 0.8823 | 0.7977 | 1.0878 | 1.4477 |
| 50 | 0.7511 | 0.6775 | 1.2601 | 1.9249 |
| 60 | 0.6854 | 0.6140 | 1.4021 | 1.8225 |
| 70 | 0.7801 | 0.6951 | 2.0176 | 2.8577 |
| 80 | 0.7769 | 0.6971 | 1.9623 | 3.2821 |
| 90 | 0.6774 | 0.6048 | 2.6764 | 3.6927 |
| 100 | 0.5457 | 0.4951 | 4.1142 | 4.9893 |
| 110 | 0.4010 | 0.3764 | 4.6369 | 6.2599 |
| 120 | 0.5349 | 0.4941 | 4.2913 | 5.5862 |
| 130 | 0.2732 | 0.2603 | 8.2401 | 9.3262 |
| 140 | 0.3143 | 0.3022 | 9.9592 | 11.3634 |
| 150 | 0.2294 | 0.2178 | 8.3755 | 8.4308 |
| 160 | 0.1745 | 0.1729 | 10.6489 | 10.8133 |
| 170 | 0.1323 | 0.1285 | 27.9650 | 28.7373 |
| 180 | 0.1303 | 0.1280 | 21.5275 | 22.3757 |
| 190 | 0.0859 | 0.0831 | 61.7638 | 60.2691 |



(a) training MSE

(b) testing MSE

**Fig. 3.** Comparison between the SUELM and the single ELM with different number of hidden nodes.

significantly improves the prediction accuracy in comparison with a single ELM. The conclusion is consistent with Theorem 1.

Theorem 1 shows that SUELM performs very well when the covariance matrix is known. However, the covariance matrix is scarce knowledge in practice. Unknown $\Sigma$ means efficiency loss for SUELM, especially when $\rho$ is small. Therefore, from Fig. 2 we can see that when $\rho$ is small the SUELM is not as good as ELM. Zhao et al. [16] has proved that in SUR model

$$Cov(\hat{\boldsymbol{\beta}}_{FG}) = a\sigma_{11}(\mathbf{X}_1'\mathbf{X}_1)^{-1}(1-b\rho^2), \tag{17}$$

where $a > 0$ and $b > 0$ are constants related to training samples. Eq. (17) is a monotonically decreasing function of the square of correlation coefficient $\rho$. The larger the correlation between the samples the smaller the covarance of $\hat{\boldsymbol{\beta}}_{FG}$. The covariance of $\hat{\boldsymbol{\beta}}_{FG}$ reaches to its maximum at $\rho = 0$, and reaches to its minimum at $\rho^2 = 1$. The covarance of $\hat{\mathbf{T}}_{SUELM}$ has the same changing rule as the covarance of $\hat{\boldsymbol{\beta}}_{FG}$. As both of $\hat{\mathbf{t}}_{ELM}$ and $\hat{\mathbf{T}}_{SUELM}$ are unbiased, the smaller the covariance the higher the improvement of prediction precision. Fig. 2 shows that the improvement is more significant if the correlation is higher. The simulation results are in accord with the theoretical analysis, which demonstrates that the SUELM is applicable for samples with high correlations.

From Table 1 and Fig. 3 we can see that ELM is better than SUELM in the training set, but SUELM is better than ELM in the testing set. In fact, the performance of a model to predict unseen data, which can be measured by generalization error [38], is more important. Let $C = \{(\mathbf{x}, \mathbf{t})\}$ be a finite space of samples. We wish to learn a function

$$h(\mathbf{x}) : \mathbf{x} \to \mathbf{t}.$$

Suppose there is a joint probability distribution $P(\mathbf{x}, \mathbf{t})$ on $\mathbf{x}$ and $\mathbf{t}$, and the training set contains $m$ i.i.d. instances $S = \{(\mathbf{x}_1, \mathbf{t}_1), (\mathbf{x}_2, \mathbf{t}_2) \dots, (\mathbf{x}_m, \mathbf{t}_m)\}$ sampled from $P(\mathbf{x}, \mathbf{t})$. The hypothesis of a joint probability distribution enables us to model the uncertainty in predictions. A non-negative real-valued loss function $L(\hat{\mathbf{t}}, \mathbf{t})$ is also used to measure how different the prediction $\hat{\mathbf{t}}$ of a hypothesis is from the true outcome $\mathbf{t}$. The risk associated with hypothesis $h(\mathbf{x})$ is then defined as the expectation of the loss function:

$$R(h) = \mathbf{E}[L(h(\mathbf{x}), \mathbf{t})] = \int L(h(\mathbf{x}), \mathbf{t})\, dP(\mathbf{x}, \mathbf{t}).$$

Ordinarily, the risk function $R(h)$ can not be calculated since the joint distribution $P(\mathbf{x}, \mathbf{t})$ is unknown to the learning algorithm. However, by averaging the loss function on the training set we can get the empirical risk as an approximation:

$$R_{\text{emp}}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(\mathbf{x}_i), \mathbf{t}_i).$$

SUELM has a generalization ability better than single ELM. The reason is that the SUELM focuses on the general risk which is the expectation of the loss function. It is a global concept that represents the predictive power of the model for all samples. However the optimization goal of the single ELM is to minimize the training mean squared error which is referred to as empirical risk.

Table 1 and Fig. 3 display the changing trend of the training and testing MSE along with the hidden nodes number. The training MSE goes down with the increase of the hidden nodes number, while the testing MSE goes up with the increase of the hidden nodes number.

For fixed $N$, training error decreases with the increasing amount of $\tilde{N}$. This conclusion is corresponding with Huang et al. [23]. However, the prediction error on testing set increases with the increasing amount of $\tilde{N}$, which demonstrates an over-fitting problem. The concepts of over-fitting and generalization error are closely connected. Over-fitting occurs when the learned function $h(\mathbf{x})$ becomes sensitive to the noise in the testing samples. Therefore, the function can have high training accuracy but the performance will be bad on unseen data from the joint probability distribution of $\mathbf{x}$ and $\mathbf{t}$. In general, the more overfitting occurs, the larger the generalization error will be.

### 4.5. A Real application to air quality index prediction between cities

In this section we illustrate that how the proposed SUELM can be applied to air quality index prediction between cities. The prediction accuracy can be improved by using correlated information.

The air quality monitoring data and the related meteorological data from February 1, 2017 to December 31, 2017 are collected from three monitoring sites, i.e., Beijing, Tianjin, and Guangzhou, respectively. All the air quality data are collected from Data center of the Ministry of Ecology and Environment of China.[1] And all the meteorological data are collected from China Meteorological Data Service Center.[2]

[1] http://datacenter.mep.gov.cn/.

[2] http://data.cma.cn.

**Table 2**
Air quality and meteorological day data.

| City | Data | TEMP | PRESS | HUMD | WS | AQI |
|---|---|---|---|---|---|---|
| Beijing | 2017/02/01 | −7 | 10,329 | 20 | 23 | 55 |
| | 2017/02/02 | −31 | 10,270 | 34 | 12 | 108 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 2017/12/30 | 4 | 10,254 | 76 | 15 | 145 |
| | 2017/12/31 | −19 | 10,222 | 61 | 15 | 64 |
| Tianjin | 2017/02/01 | −26 | 10,367 | 20 | 19 | 52 |
| | 2017/02/02 | −25 | 10,310 | 34 | 17 | 109 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 2017/12/30 | −4 | 10,291 | 76 | 17 | 242 |
| | 2017/12/31 | −24 | 10,261 | 61 | 10 | 203 |
| Guangzhou | 2017/02/01 | 163 | 10,136 | 72 | 27 | 32 |
| | 2017/02/02 | 154 | 10,144 | 72 | 23 | 43 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 2017/12/30 | 169 | 10,144 | 81 | 40 | 83 |
| | 2017/12/31 | 150 | 10,146 | 65 | 41 | 64 |

**Table 3**
Comparison of the SUELM and the single ELM with the data from Beijing and Tianjin monitoring sites.

| | Beijing | | | |
|---|---|---|---|---|
| | Training set | | Test set | |
| | SUELM | ELM | SUELM | ELM |
| Average MSE | 0.0084 | 0.0077 | 0.0087 | 0.0104 |
| SD of MSE | 0.0005 | 0.0004 | 0.0032 | 0.0053 |
| | Tianjin | | | |
| | Training set | | Test set | |
| | SUELM | ELM | SUELM | ELM |
| Average MSE | 0.0086 | 0.0081 | 0.0090 | 0.0104 |
| SD of MSE | 0.0007 | 0.0007 | 0.0039 | 0.0059 |

The air quality monitoring data consists of the air quality index abbreviated as AQI, the meteorological data consists of temperature abbreviated as TEMP, atmosphere pressure abbreviated as PRESS, humidity abbreviated as HUMD and wind speed abbreviated as WS. All data are collected day by day as show in Table 2. Through the Pearson correlation analysis, it was found that the air quality index of Beijing is highly correlated with that of Tianjin, where the Pearson correlation coefficent is 0.877, while the correlation of air quality indices in Beijing and Guangzhou are very low, where the Pearson correlation coefficent is 0.03.

The data sets from Beijing-Tianjin, and Beijing-Guangzhou are used to train both the SUELM and ELM model. In practice, we normalize the input data into interval [0, 1] by Eq. (18), and the output data are anti-normalized by Eq. (19), i.e.,

$$x'_i = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}, \tag{18}$$

$$\mathbf{t} = \mathbf{y}'(y_{\max} - y_{\min}) + y_{\min}, \tag{19}$$

where $x_i$ is the $i$th component value of the input vector $\mathbf{x}$, $x_{imin}$ is the minimum value of the input vector component in training sample space, $x_{imax}$ is the maximum value of the input vector component in training sample space, and $x'_i$ is the component value of normalization of $x_i$. $\mathbf{y}'$ is the component value of normalization of $\mathbf{y}$, $y_{min}$ is the minimum component value of the output vector $\mathbf{y}$ in training sample space, and $y_{max}$ is the maximum value of the output vector $\mathbf{y}$ in training sample space.

It is noteworthy that owing to the random assignment mechanism of the input weights, the results of each run may be different. We carried out 1000 experimental trials for each data set. In each trial, 70% data are randomly picked as training set, and the remaining 30% data are used as the testing set. The results of each trial are different, and we compare the mean and standard deviation of the MSE based on the 1000 trial results. Table 3 illustrates that the average MSE and the standard deviation of SUELM are smaller than those of ELM on testing set based on the data collected from Beijing and Tianjin. The reason can be easily found from the Pearson correlation analysis, i.e., the data from Beijing and Tianjin are highly correlated. Table 4 illustrates that the performance of SUELM is almost as the same as ELM when the correlation between the two sets of data is small.

**Table 4**
Comparison of the SUELM and the single ELM with the data from Beijing and Guangzhou monitoring sites.

| | Beijing | | | |
|---|---|---|---|---|
| | Training set | | Testing set | |
| | SUELM | ELM | SUELM | ELM |
| Average MSE | 0.0086 | 0.0086 | 0.0087 | 0.0087 |
| SD of MSE | 0.0003 | 0.0003 | 0.0016 | 0.0016 |
| | Guangzhou | | | |
| | training set | | testing set | |
| | SUELM | ELM | SUELM | ELM |
| Average MSE | 0.0125 | 0.0125 | 0.0135 | 0.0135 |
| SD of MSE | 0.0007 | 0.0007 | 0.0011 | 0.0011 |

## 5. Concluding remarks

This paper aims at developing a new approach to efficient learning in random-weight neural networks. Using the correlation information we connect SUR with ELM together to propose an SUELM. It simultaneously trains a group of ELMs by incorporating the covariance information of the data. When the samples are uncorrelated, the SUELM is equivlent to ELM. Theoretically, it has been proved that the SUELM can yield a minimal mean squared error when the correlation is known. And simulation results indicate that SUELM has a generalization capability much better than ELM, especially when the two groups of samples are highly correlated to each other.

There are two limitations of the proposed model. The first is that the model can be used only when the data in one group is coming from the same population, which seriously limits the models applicability. The second is that we have not yet a mathematical formulation to describe the models generalization capability. Overcoming the two limitations and then proposing an improved model is our further study on this topic.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants 61772344, 61732011, 71371063, and 61811530324) and in part by Basic Research Project of Knowledge Innovation Program in ShenZhen (JCYJ20180305125850156).

## References

[1] A. Zellner, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, J. Am. Stat. Assoc. 57 (298) (1962) 348–368.
[2] L. Wang, H. Lian, R.S. Singh, On efficient estimators of two seemingly unrelated regressions, Stat. Probab. Lett. 81 (5) (2011) 563–570.
[3] H. Wang, Sparse seemingly unrelated regression modelling: Applications in finance and econometrics, Comput. Stat. Data Anal. 54 (11) (2010) 2866–2877.
[4] P. Foschi, E.J. Kontoghiorghes, A computationally efficient method for solving SUR models with orthogonal regressors, Linear Algebra Appl. 388 (2004) 193–200.
[5] M. Hubert, T. Verdonck, Yorulmaz, Fast robust SUR with economical and actuarial applications, Stat. Anal. Data Min. Asa Data Sci. J. 10 (2) (2016) 77–88.
[6] F. Denzil G, J.H. Kim, Estimation and inference in SUR models when the number of equations is large, Econ. Rev. 19 (1) (2000) 105–130.

[7] A. Zellner, Estimators for seemingly unrelated regression equations : Some exact finite sample results, J. Am. Stat. Assoc. 58 (304) (1963) 977–992.
[8] V.K. Srivastava, D.E.A. Giles, Seemingly Unrelated Regression Equations Models, Marcel Dekker, Inc., 1987. 1987
[9] D.G. Fiebig, Seemingly Unrelated Regression. A Companion to Theoretical Econometrics, Springer, Berlin Heidelberg, 2007.
[10] M. Roozbeh, M. Arashi, M. Gasparini, Seemingly unrelated ridge regression in semiparametric models, Commun. Stat. Theory Methods 41 (8) (2012) 1364–1386.
[11] M. Roozbeh, M. Arashi, Feasible ridge estimator in seemingly unrelated semiparametric models, Commun. Stat. Simulat. Comput. 43 (10) (2014) 2593–2613.
[12] Z. Zeebari, G. Shukur, B.M.G. Kibria, Modified ridge parameters for seemingly unrelated regression model, Commun. Stat. Theory Methods 41 (9) (2012) 1675–1691.
[13] D. Fraser, M. Rekkas, A. Wong, Highly accurate likelihood analysis for the seemingly unrelated regression problem, J. Econ. 127 (1) (2005) 17–33.
[14] A. Zellner, T. Ando, A direct monte carlo approach for bayesian analysis of the seemingly unrelated regression model, J. Econ. 159 (1) (2010) 33–45.
[15] H. Kurata, S. Matsuura, Best equivariant estimator of regression coefficients in a seemingly unrelated regression model with known correlation matrix, Ann. Inst. Stat. Math. (2015) 1–19.
[16] L. Zhao, L. Yan, X. Xu, High correlated residuals improved estimation in the high dimensional SUR model, Commun. Stat. Simulat. Comput. (2017). In press
[17] L. Zhao, X. Xu, Generalized canonical correlation variables improved estimation in high dimensional seemingly unrelated regression models, Stat. Probab. Lett. 2017 (126) (2017) 119–126.
[18] R. Wang, D. Chen, S. Kwong, Fuzzy rough set based active learning, IEEE Trans. Fuzzy Syst. 22 (6) (2014) 1699–1704.
[19] Y. Chauvin, D.E. Rumelhart, Back-Propagation: Theory, Architecture, and Applications, L. Erlbaum Associates Inc., 1995.
[20] R. Wang, X. Wang, S. Kwong, C. Xu, Incorporating diversity and informativeness in multiple-instance active learning, IEEE Trans. Fuzzy Syst. 25 (6) (2017) 1460–1475.
[21] N. Zeng, Z. Wang, B. Zineddin, et al., Image-based quantitative analysis of gold immunochroma to graphic strip via cellular neural network approach[j], IEEE Trans. Med. Imaging 33 (5) (2014) 1129–1136.
[22] N. Zeng, Z. Wang, Y. Li, et al., A hybrid EKF and switching PSO algorithm for joint state and parameter estimation of lateral flow immunoassay models[j], IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (2) (2012) 321–329.
[23] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: Theory and applications designs, Neurocomputing 70 (2006) (2006) 489–501.
[24] J. Cao, K. Zhang, M. Luo, C. Yin, X. Lai, Extreme learning machine and adaptive sparse representation for image classification, Neural Netw. Offic. J. Int. Neural Netw. Soc. 81 (2016) 91–102.
[25] X. Zhao, W. Cao, H. Zhu, Z. Ming, R.A.R. Ashfaq, An initial study on the rank of input matrix for extreme learning machine, Int. J. Mach. Learn. Cybern. 9 (5) (2018) 867–879.
[26] F. Li, H. Liu, X. Xu, F. Sun, Haptic recognition using hierarchical extreme learning machine with local-receptive-field, Int. J. Mach. Learn. Cybern. 10 (3) (2019) 541–547.
[27] H. Zhao, X. Guo, M. Wang, T. Li, C. Pang, D. Georgakopoulos, Analyze eeg signals with extreme learning machine based on PMIS feature selection, Int. J. Mach. Learn. Cybern. 9 (2) (2018) 243–249.
[28] J. Wang, L. Zhang, J.J. Cao, D. Han, NBWELM: naive bayesian based weighted extreme learning machine, Int. J. Mach. Learn. Cybern. 9 (1) (2018) 21–35.
[29] R. Wang, C.-Y. Chow, S. Kwong, Ambiguity based multiclass active learning, IEEE Trans. Fuzzy Syst. 24 (1) (2016) 242–248.
[30] N. Zeng, H. Zhang, W. Liu, et al., A switching delayed PSO optimized extreme learning machine for short-term load forecasting[j], Neurocomputing 240 (2017) 175–182.
[31] J.M. Martnez-Martnez, P. Escandell-Montero, E. Soria-Olivas, J.D. Martn-Guerrero, R. Magdalena-Benedito, J. Gmez-Sanchis, Regularized extreme learning machine for regression problems, Neurocomputing 74 (17) (2011) 3716–3721.
[32] X. Luo, X. Yang, C. Jiang, X. Ban, Timeliness online regularized extreme learning machine, Int. J. Mach. Learn. Cybern. 9 (3) (2018) 465–476.
[33] G. Huang, Q. Zhu, C. Siew, Real-time learning capability of neural networks, IEEE Trans. Neural Netw. 17 (4) (2006) 863–878.
[34] N. Liang, P. Saratchandran, G. Huang, N. Sundararajan, Classification of mental tasks from EEG signals using extreme learning machine, Int. J. Neural Syst. 16 (1) (2006) 29–38.
[35] G.B. Huang, L. Chen, C.K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw. 17 (4) (2006) 879–892.
[36] X.Z. Wang, H.J. Xing, Y. Li, Q. Hua, C.R. Dong, W. Pedrycz, A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning, IEEE Trans. Fuzzy Syst. 23 (5) (2015) 1638–1654.
[37] J.H. Lin, T.M. Sellke, E.J. Coyle, Adaptive stack filtering under the mean absolute error criterion[j], IEEE Trans. Acoust. Speech Signal Process. 38 (6) (1990) 938–954.
[38] X.Z. Wang, R. Wang, C. Xu, Discovering the relationship between generalization and uncertainty by incorporating complexity of classification, IEEE Trans. Cybern. 8 (2) (2018) 703–715.

**Li Zhao** received the bachelors degree in information and computation science from the College of mathematics and computer, Hebei University, Hebei, China, in 2007, and the Ph.D. degree from the School of Mathematics and Statistics, at Beijing Institute of Technology, Beijing, in 2017. She is currently a Postdoctoral Researcher at the College of Computer Science and Software Engineering, Shenzhen University, China. Her current research interests include pattern recognition, machine learning, relative data analysis, and high dimensional statistical inference.

**Professor Xizhao Wang** received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998. He is currently a Professor with the Big Data Institute, Shenzhen University, Shenzhen, China. His current research interests include uncertainty modeling and machine learning for big data. He has edited more than ten special issues and published three monographs, two textbooks, and more than 200 peer-reviewed research papers. By the Google scholar, the total number of citations is over 5000. He is on the list of Elsevier 2015/2016 most cited Chinese authors. He is the Chair of the IEEE SMC Technical Committee on Computational Intelligence, the Editor-in-Chief of Machine Learning and Cybernetics Journal, and Associate Editor for a couple of journals in the related areas. He was a recipient of the IEEE SMCS Outstanding Contribution Award in 2004 and a recipient of the IEEE SMCS Best Associate Editor Award in 2006