International Journal of
Molecular Sciences

MDPI

Article

Descriptor Selection via Log-Sum Regularization for the Biological Activities of Chemical Structure

Liang-Yong Xia [1], Yu-Wei Wang [1], De-Yu Meng [2], Xiao-Jun Yao [1], Hua Chai [1] and Yong Liang [1,*]

[1] State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau 999078, China; xia2yin1234@gmail.com (L.-Y.X.); wangyw09@gmail.com (Y.-W.W.); xjyao@must.edu.mo (X.-J.Y.); chch890113@gmail.com (H.C.)

[2] Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China; dymeng@mail.xjtu.edu.cn

[*] Correspondence: yliang@must.edu.mo

Received: 17 November 2017; Accepted: 21 December 2017; Published: 22 December 2017

Abstract: The quantitative structure-activity relationship (QSAR) model searches for a reliable relationship between the chemical structure and biological activities in the field of drug design and discovery. (1) Background: In the study of QSAR, the chemical structures of compounds are encoded by a substantial number of descriptors. Some redundant, noisy and irrelevant descriptors result in a side-effect for the QSAR model. Meanwhile, too many descriptors can result in overfitting or low correlation between chemical structure and biological bioactivity. (2) Methods: We use novel log-sum regularization to select quite a few descriptors that are relevant to biological activities. In addition, a coordinate descent algorithm, which uses novel univariate log-sum thresholding for updating the estimated coefficients, has been developed for the QSAR model. (3) Results: Experimental results on artificial and four QSAR datasets demonstrate that our proposed log-sum method has good performance among state-of-the-art methods. (4) Conclusions: Our proposed multiple linear regression with log-sum penalty is an effective technique for both descriptor selection and prediction of biological activity.

Keywords: QSAR; biological activity; descriptor selection; regularization; log-sum

1. Introduction

The quantitative structure-activity relationship (QSAR) model searches for a reliable relationship between chemical the structure and biological activities in the field of drug design and discovery [1]. In the study of QSAR, the chemical structure is encoded by a substantial number of descriptors, such as thermodynamic, shape descriptors, etc. Generally, only a few descriptors that are relevant to biological activities are of interest to the QSAR model. Descriptor selection aims to eliminate redundant, noisy and irrelevant descriptors [2]. The flow diagram shows the process of QSAR modeling in Figure 1.

Generally, descriptor selection techniques can be categorized into four groups in the study of QSAR: classical methods, artificial intelligence-based methods, miscellaneous methods and regularization methods.

The classical methods have been proposed in the study of QSAR; as an example, forward selection adds the most significant descriptors until none improves the model to a statistically-significant extent. Backward elimination starts with all candidate descriptors, subsequently deleting descriptors without any statistical significance. Generally, stepwise regression builds a model by adding or removing predictor variables based on a series of F-tests or *t*-tests. The variable selection and modeling method based on the prediction [3] uses leave-one-out cross-validation (Q^2), predicted to select meaningful and important descriptors. Leaps-and-bounds regression [4] selects a subset of descriptors based on the residual sum of squares (RSS).

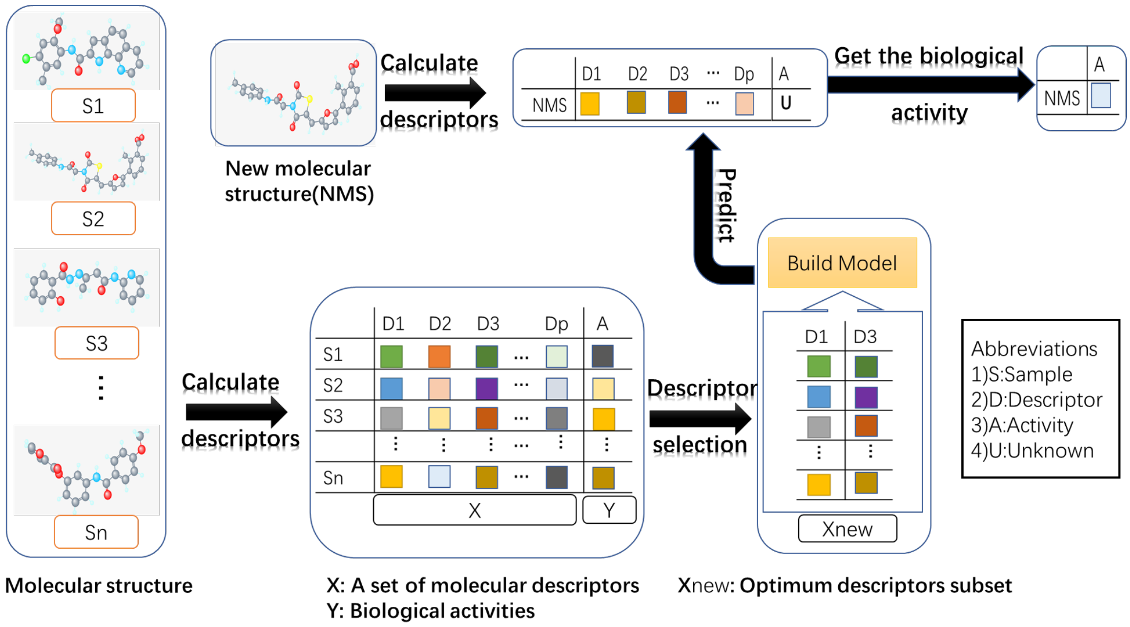

Figure 1. The flow diagram shows the process of QSAR modeling. (1) Collecting molecular structures and their activities; (2) calculating molecular descriptors, which can produce thousands of parameters for each molecular structure; (3) removing redundant or irrelevant descriptors via descriptor selection; (4) building the model with the optimum descriptor subset; (5) predicting the biological activity of a new molecular structure using the established model. Different color blocks represent different values.

Recently, artificial intelligence-based methods have been designed for descriptor selection, such as the genetic algorithm [5], which uses the code, selection, exchange and mutation operations to select the important descriptors. Particle swarm optimization [6] has a series of initial random particles and then selects the descriptors by updating the velocity and positions. Artificial neural networks [7] are composed of many artificial neurons that are linked together according to a specific network architecture and select input nodes (descriptors) to predict the output node (biological activity). Simulated annealing [8] can be performed with the Metropolis algorithm based on Monte Carlo techniques, which performs descriptor selection. Frank et al. [9] used Bayesian regularized artificial neural networks with automatic relevance determination (ARD) in the study of QSAR. ARD has the capacity to allow the network to estimate the importance of each input, neglects irrelevant or highly correlated indices in the modeling and uses the most important variables for modeling the activity data. The ant colony system [10], inspired by real ants, searches a path, which is connected to a number of selected descriptors, between the colony and a source of food.

The miscellaneous methods used for descriptor selection in the development of QSAR include *K* nearest neighbor (KNN) [11], the replacement method (RM) [12], the successive projections algorithm (SPA) [13] and uninformative variable elimination-partial least squares (UVE-PLS) [14], just to name a few. KNN uses a similarity measure (Euler distance) to select the descriptor and predict the biological activity. RM has the capacity to find an optimal subset of the descriptors via the standard deviation. SPA is a simple operation to eliminate collinearity to reduce the descriptors. UVE-PLS has been proposed to increase the predictive ability of the standard PLS method via eliminating the variables that cannot contribute to the model and to make a comparison between experimental variables and added noise variables with respect to the degree of contribution to the model.

The regularization is an effective technique in descriptor selection and has been used in QSRR [15], QSPR [16] and QSTR [17] in the field of chemometrics. However, some individuals have poured their interest and attention into the study of QSAR. For example, LASSO (L_1) (least absolute shrinkage and selection operator) [18] has the capacity to perform descriptor selection. Algamal et al. proposed the L_1-norm to select the significant and meaningful descriptors for anti-hepatitis C virus activity of thiourea derivatives in the QSAR classification model [19]. Xu et al. proposed $L_{1/2}$ [20] regularization, which has more sparsity. Algamal et al. proposed a penalized linear regression model with the $L_{1/2}$-norm to select the significant and meaningful descriptors [21]. Theoretically, the L_0 regularization produces better solutions with more sparsity [22], but it is an NP problem. Therefore, Candes et al. proposed the log-sum penalty [23], which approximates the L_0 regularization much better.

In this paper, we utilized the log-sum penalty, which is non-convex in Figure 2. A coordinate descent algorithm, which uses novel univariate log-sum thresholding for updating the estimated coefficients, has been developed for the QSAR model. Experimental results on artificial and four QSAR datasets demonstrate that our proposed log-sum method has good performance among state-of-the-art methods. The structure of this paper is organized as follows: Section 2 introduces a coordinate descent algorithm, which uses novel univariate log-sum thresholding for updating the estimated coefficients and gives a detailed description of the datasets. In Section 3, we discuss the experimental results on simulated data and four QSRA datasets. Finally, we give some conclusions in Section 4.

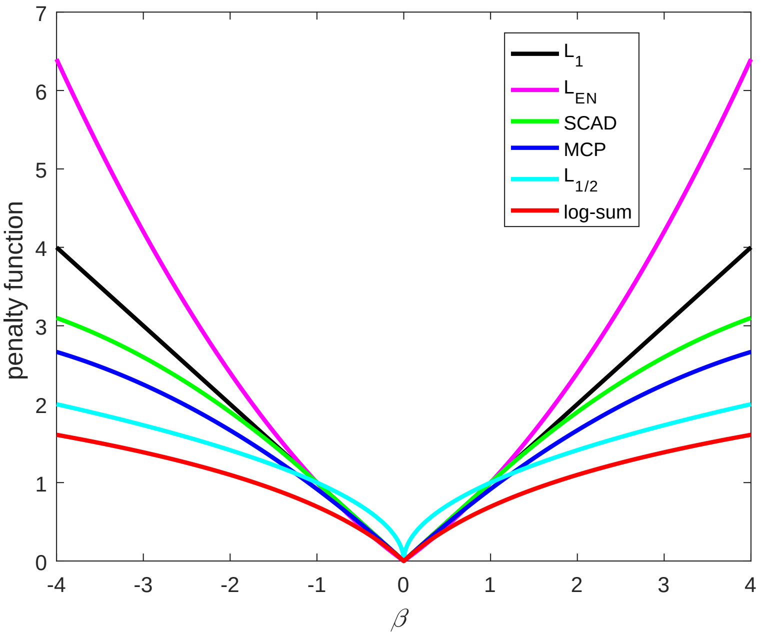

Figure 2. L_1 and L_{EN} are convex, and SCAD, MCP, $L_{1/2}$ and log-sum are non-convex. The log-sum approximates to L_0.

2. Methods

In this paper, there exists a predictor X and a response y, which represent the chemical structure and corresponding biological activities, respectively. Suppose we have n samples, $D = (X_1, y_1), (X_2, y_2), ..., (X_n, y_n)$, where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is the i-th input pattern with dimensionality p, which means X_i has p descriptors, and x_{ij} denotes the value of descriptor j for the i-th sample. The multiple linear regression is expressed as:

$$y_i = x_{i1}\beta_1 + ... + x_{ip}\beta_p + \beta_0 \tag{1}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ are the coefficients.

Given X and y, $\beta_0, \beta_1, ..., \beta_p$ are estimated based on an objective function. The linear regression of the objective function can be formulated:

$$min\{\frac{1}{2n}\|y - X\beta\|^2\} \tag{2}$$

where $y = (y_1, ..., y_n)^T$ is the vector of n response variables, $X = \{X_1, X_2, ..., X_n\}$ is $n \times p$ matrix with $X_i = (x_{i1}, ..., x_{ip})$ and $||.||$ denotes the L_2-norm. When the number of variables is larger than the number of samples ($p \gg n$), this can result in over-fitting. Here, we introduced a penalty function in the objective function to estimate the coefficient. We have rewritten Equation (2):

$$min\{\frac{1}{2n}\|y - X\beta\|^2 + P_\lambda(\beta)\} \tag{3}$$

where $P_\lambda()$ is a penalty function indexed by the regularized parameter $\lambda > 0$.

2.1. Coordinate Decent Algorithm for Different Thresholding Operators

In this paper, we used the coordinate descent algorithm to implement different penalized multiple linear regression. The algorithm is a "one-at-a-time" algorithm and solves β_j, and other $\beta_{k\neq j}$ (representing the parameters remaining after the j-th element is removed) are fixed [22]. Equation (3) can be rewritten as:

$$R(\beta) = argmin\{\frac{1}{2n}(y_i - (\sum_{k\neq j} x_{ik}\beta_k + x_{ij}\beta_j))^2 + \lambda\sum_{k\neq j} P(\beta_k) + P(\beta_j)\} \tag{4}$$

where k represents other variables except the j-th variable.

Take the derivative with respect to β_j :

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}(-x_{ij}(y_j - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)) + \lambda P(\beta_j) = 0 \tag{5}$$

Denote $\widetilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k$, $\widetilde{r}_i^{(j)} = y_i - \widetilde{y}_i^{(j)}$, $w_j = \sum_{i=1}^{n} x_{ij}\widetilde{r}_i^{(j)}$ where $\widetilde{r}_i^{(j)}$ represents the partial residuals with respect to the j-th covariate. To take into account the correlation of descriptors, Zhou et al. have proposed elastic net (L_{EN}) [24], which emphasizes a grouping effect. The L_{EN} penalty function is given as follows:

$$P(\beta) = (1 - a)\frac{1}{2}\|\beta\|_{L_2}^2 + a\|\beta\|_{L_1} \tag{6}$$

The penalty function of L_{EN} is a combination of the L_1 penalty ($a = 1$) and the ridge penalty ($a = 0$). Therefore, Equation (5) is rewritten as follows:

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}(-x_{ij}(y_j - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)) + \lambda(1 - a)\beta_j + \lambda a = 0 \tag{7}$$

Donoho et al. proposed the univariate solution [25] for a L_{EN}-penalized regression coefficient as follows:

$$\beta_j = f_{L_{EN}}(w_j, \lambda, a) = \frac{S(w_j, \lambda a)}{1 + \lambda(1 - a)} \tag{8}$$

where $S(w_j, \lambda a)$ is the soft thresholding operator for the L_1 if a is equal to one; Formula (8) can be rewritten as follows:

$$\beta_j = Soft(w_j, \lambda) = \begin{cases} w_j + \lambda & \text{if } w_j < -\lambda \\ w_j - \lambda & \text{if } w_j > \lambda \\ 0 & \text{if } -\lambda \leq w_j \leq \lambda \end{cases} \tag{9}$$

Fan et al. have proposed the smoothly clipped absolute deviation (SCAD) [26], which can produce a sparse set of solutions and approximately unbiased coefficients for large coefficients. The penalty function is shown as follows:

$$p_{\lambda,a}(\beta) = \begin{cases} \lambda\beta & \text{if } \beta \neq \lambda \\ \frac{a\lambda\beta - \frac{1}{2}(\beta^2+\lambda^2)}{a-1} & \text{if } \lambda < \beta < a\lambda \\ \frac{\lambda(a^2-1)}{2(a-1)} & \text{if } \beta > a\lambda \end{cases} \tag{10}$$

Additionally, the SCAD thresholding operator is given as follows:

$$\beta_j = f_{\text{SCAD}}(w_j, \lambda, a) = \begin{cases} S(w_j, \lambda) & \text{if } |w_j| < 2\lambda \\ \frac{S(w_j, a\lambda/(a-1))}{1-1/(a-1)} & \text{if } 2\lambda < |w_j| \leq a\lambda \\ w_j & \text{if } |w_j| > a\lambda \end{cases} \tag{11}$$

Similar to the SCAD penalty, Zhang et al. have proposed the maximum concave penalty (MCP) [27]. The formula of the penalty function is shown as:

$$p_{\lambda,a}(\beta) = \begin{cases} \lambda\beta & \text{if } \beta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \beta > \gamma\lambda \end{cases} \tag{12}$$

Additionally, the MCP thresholding operator is given as follows:

$$\beta_j = f_{\text{MCP}}(w_j, \lambda, \gamma) = \begin{cases} \frac{S(w_j,\lambda)}{1-1/\gamma} & \text{if } |w_j| \leq \gamma\lambda \\ w_j & \text{if } |w_j| > \gamma\lambda \end{cases} \tag{13}$$

where γ is the experience parameter.

Xu et al. proposed $L_{1/2}$ regularization [20]. Formula (3) can be rewritten:

$$min\{\frac{1}{2n}\|y - X\beta\|^2 + \lambda\sum_{j}^{p}|\beta_j|^{\frac{1}{2}}\} \tag{14}$$

and the univariate half thresholding operator for a $L_{1/2}$-penalized linear regression coefficient is as follows:

$$\beta_j = Half(w_j, \lambda) = \begin{cases} \frac{2}{3}w_j(1 + \cos\frac{2(\pi - \phi_\lambda(w_j))}{3}) & \text{if } |w_j| > \frac{3}{4}(\lambda)^{\frac{2}{3}} \\ 0 & otherwise \end{cases} \tag{15}$$

where $\phi_\lambda(w) = \frac{\lambda}{8}(\frac{|w|}{3})^{-\frac{3}{2}}$.

In this paper, we applied the log-sum penalty to the linear regression model. We could rewrite Formula (3) as follows:

$$min\{\frac{1}{2n}\|y - X\beta\|^2 + \lambda\sum_{j}^{p}log(|\beta_j| + \varepsilon)\} \tag{16}$$

where $\varepsilon > 0$ should be set arbitrarily small, to make the log-sum penalty closely resemble the L_0-norm. Equation (16) has a local minimal. The proof is given in the Appendix A:

$$\beta_j = f_{log-sum}(w_j, \lambda, \varepsilon) = D(w_j, \lambda, \varepsilon) = \begin{cases} sign(w_j)\frac{c_1+\sqrt{c_2}}{2} & \text{if } c_2 > 0 \\ 0 & \text{if } c_2 \leq 0 \end{cases} \tag{17}$$

where $\lambda > 0, 0 < \varepsilon < \sqrt{\lambda}, c_1 = \omega_j - \varepsilon$ and $c_2 = c_1^2 - 4(\lambda - w_j\varepsilon)$.

According to different thresholding operators, we can define three properties for to satisfy the coefficient estimator, unbiasedness, sparsity and continuity, in Figure 3.

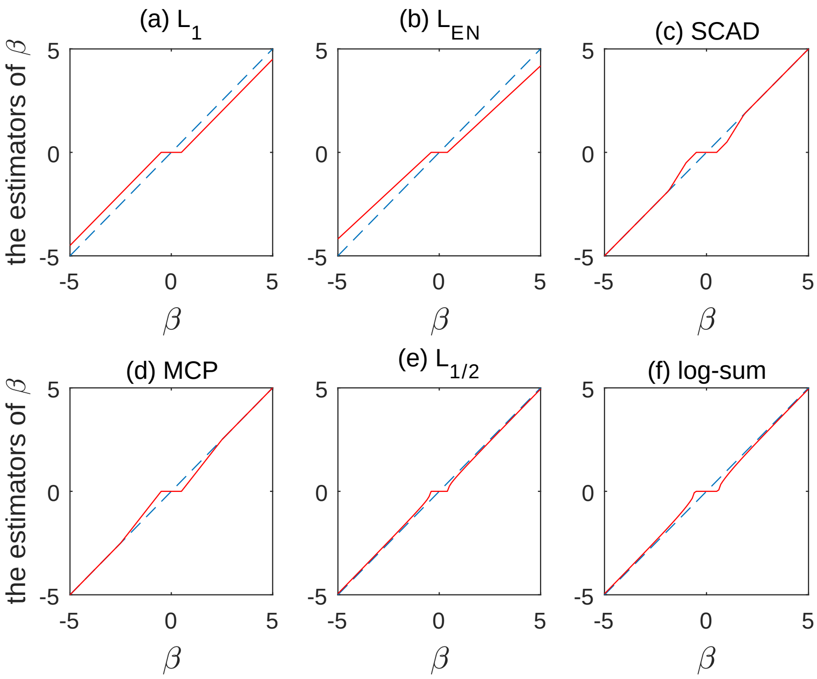

Figure 3. Plot of thresholding functions for: (**a**) L_1; (**b**) L_{EN}; (**c**) SCAD; (**d**) MCP; (**e**) $L_{1/2}$; and (**f**) log-sum.

2.2. Dataset

2.2.1. Simulated Data

In this work, we constructed the simulation. The process of the construction was given as follows:

Step I: The simulated dataset was generated from multiple linear regression using the normal distribution to produce X. Here, the number of row is sample n and the number of column is variable p.

$$y = X\beta + \overbrace{\sigma\epsilon}^{intercept}, \epsilon \sim N(0,1) \tag{18}$$

where $y = (y_1, ..., y_n)^T$ is the vector of n response variables, X = {X_1, X_2, ..., X_n} is the generated matrix with $X_i = (x_{i1}, ..., x_{ip})$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is the random error and σ controls the signal to noise.

Step II: Add a different correlation parameter ρ to the simulation data.

$$x_{ij} = \rho \times x_{11} + (1-\rho)x_{ij}, i \sim (1, ..., n), j \sim (2,3,4,5,6) \tag{19}$$

Step III: In order to get a high quality model and variable selection, the coefficients (20) are set in advance from 1–20.

$$\beta = \overbrace{\underbrace{2, -2, -1, 1.5, 3, 2.5, 3, 2, ..., 2}_{20}, \underbrace{0, 0, 0, ..., 0}_{1980}}^{2000} \tag{20}$$

where β is the coefficient.

Step IV: We can get y from Equations (18)–(20).

In the simulation study, we firstly generated 100 groups of data with different sample sizes $n = 100$ and $n = 200$. Secondly, the correlation coefficient $\rho = 0.2, 0.4$ and the noise control parameter $\sigma = 0.3, 0.9$, were considered in the model. Thirdly, the coefficients (20) are set in advance.

Fourthly, the multiple linear regression with different penalties to select variables and build the model, including our proposed method, was used. Finally, due to the generation of 100 groups of data, the results obtained by different methods need to be averaged.

2.2.2. Real Data

We could obtain four public QSAR datasets, including the global half-life index [28], endocrine disruptor chemical (EDC) estrogen receptor (ER)-binding [29], (Benzo-)Triazoles toxicity in Daphnia magna [30] and apoptosis regulator Bcl-2 [31]. A brief description of these datasets is shown in Table 1. We utilized random sampling to divide datasets into training datasets and test datasets (80% for the training set and 20% for the test set [32]). Six commonly-used parameters in regression problems are employed to evaluate the model performance, including the square correlation coefficients of the leave-one-out cross-validation (Q^2_{LOO}), the root mean squared error of cross-validation ($RMSE_{CV}$), the square correlation coefficients of fitting for the training set (R^2_{train}), the root mean squared error for the training set ($RMSE_{train}$), the square correlation coefficients of fitting for the test set (R^2_{test}) and the root mean squared error for the test set ($RMSE_{test}$). According to existing literature [33], we have learned that the value of Q^2_{LOO} is not the best measure for QSAR model evaluation. Therefore, we poured more interest and attention into (R^2_{test}) and ($RMSE_{test}$).

Table 1. A brief description of four public datasets used in the experiments.

Dataset Name	No. of Samples	No. of Descriptors	No. of Samples (Training)	No. of Samples (Test)
BTAZD	97	1083	78	19
EDCER	129	1089	104	25
GHLI	250	1120	200	50
BCL2	508	1562	407	101

Algorithm: A coordinate descent algorithm for log-sum penalized multiple linear regression.

```
Step 1: Initialize all $\beta_j(m) = 0 (j = 1, 2, 3, ..., p), \lambda, \varepsilon$,set $m = 0$;
Step 2: Calculate the function (16) based on $\beta(m)$
Step 3: Update each $\beta_j(m)$ and cycle $j = 1, 2, 3, ..., p$
        Step 3.1: $\tilde{r}_i^{(j)}(m) = y_i(m) - \tilde{y}_i^{(j)}(m) = y_i(m) - \sum_{k \neq j} x_{ik}\beta_k(m)$
                  and $w_j(m) = x_{ij}(r_i(m) - \tilde{r}_i^{(j)}(m))$
        Step 3.2: Update $\beta_j(m) = D(w_j, \lambda, \varepsilon)$
Step 4: Let $m \leftarrow (m + 1)$,$\beta(m + 1) \leftarrow \beta(m)$
Step 5: Repeat Steps 2 and 3 until $\beta(m)$ converges
```

3. Results

In this work, five methods are compared to our proposed method, including multiple linear regression with L_{EN}, L_1, SCAD, MCP and $L_{1/2}$ penalties, respectively.

3.1. Analyses of Simulated Data

Tables 2 and 3 describe the number of variables that are selected (non-zero coefficient) by different methods within 2000 variables and within pre-set variables (20), respectively. For example, when $n = 200, \rho = 0.4$ and $\sigma = 0.9$, the average number of variables selected is 23.73 within 2000 variables by the log-sum in Table 2. In pre-set variables (20), we got 19.95 variables by the log-sum in Table 3. Therefore, we could calculate the average accuracy ($19.95 \div 23.73 \times 100\% = 84.07\%$) for the simulation datasets obtained by log-sum in Table 4. From Tables 2–4, for example, when the correlation parameter ρ and the noise control parameter σ decrease, the average accuracy of log-sum improves. When $n = 100$ and $\sigma = 0.9$, the average accuracy of log-sum is from 83.77–98.7%, where the correlation parameter ρ

is from 0.4–0.2. When $n = 200$ and $\rho = 0.4$, the results obtained by log-sum are 84.07% and 86.39% with the noise control parameter σ =0.9, 0.3. In addition, compared to other methods, the average accuracy obtained by our proposed log-sum method is better, for example when $n = 200$, $\rho = 0.4$ and $\sigma = 0.9$, the result of the log-sum is 84.07% higher than 3.19%, 20.20%, 49.20%, 83.22% and 81.74% of the L_{EN}, L_1, SCAD, MCP and $L_{1/2}$. In other words, our proposed log-sum method has the capacity to obtain good performance in the simulation dataset.

Table 2. The average number of variables selected in total by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum. In bold, the best performance is shown.

	Sample Size	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
$\rho = 0.2, \sigma = 0.3$	$n = 100$	381.60	92.92	19.09	23.36	19.13	**19.00**
	$n = 200$	498.81	34.18	19.03	**19.00**	19.09	**19.00**
$\rho = 0.2, \sigma = 0.9$	$n = 100$	382.24	93.26	27.74	25.79	21.77	**21.54**
	$n = 200$	499.49	95.83	36.48	23.65	23.83	**23.15**
$\rho = 0.4, \sigma = 0.3$	$n = 100$	378.96	93.98	19.26	24.67	19.98	**19.11**
	$n = 200$	495.66	97.51	40.87	24.04	24.42	**23.79**
$\rho = 0.4, \sigma = 0.9$	$n = 100$	379.35	93.46	29.22	26.08	22.48	**22.04**
	$n = 200$	495.64	98.97	40.61	23.95	24.43	**23.73**

Table 3. The average number of variables selected with a pre-set value (20) obtained by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum.

	Sample Size	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
$\rho = 0.2, \sigma = 0.3$	$n = 100$	12.23	14.45	19.09	18.81	19.13	19.00
	$n = 200$	16.22	20.00	19.03	19.00	19.09	19.00
$\rho = 0.2, \sigma = 0.9$	$n = 100$	12.24	14.30	19.93	19.42	19.74	19.81
	$n = 200$	16.26	20.00	20.00	20.00	20.00	20.00
$\rho = 0.4, \sigma = 0.3$	$n = 100$	11.84	13.57	18.88	18.40	18.65	18.88
	$n = 200$	15.79	19.99	19.97	19.93	19.96	19.93
$\rho = 0.4, \sigma = 0.9$	$n = 100$	11.88	13.55	19.48	18.81	19.14	19.00
	$n = 200$	15.80	19.99	19.98	19.93	19.97	19.95

Table 4. The average accuracy (%) for the simulation data sets obtained by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum. In bold, the best performance is shown.

	Sample Size	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
$\rho = 0.2, \sigma = 0.3$	$n = 100$	3.20%	15.55%	**100.00%**	80.52%	**100.00%**	**100.00%**
	$n = 200$	3.25%	58.51%	**100.00%**	**100.00%**	**100.00%**	**100.00%**
$\rho = 0.2, \sigma = 0.9$	$n = 100$	3.12%	14.44%	98.03%	74.58%	93.34%	**98.80%**
	$n = 200$	3.19%	20.50%	48.86%	82.90%	81.74%	**83.77%**
$\rho = 0.4, \sigma = 0.3$	$n = 100$	3.20%	15.33%	71.85%	75.30%	90.68%	**91.97%**
	$n = 200$	3.26%	20.87%	54.87%	84.57%	83.93%	**86.39%**
$\rho = 0.4, \sigma = 0.9$	$n = 100$	3.19%	20.50%	48.86%	82.90%	81.74%	**83.77%**
	$n = 200$	3.19%	20.20%	49.20%	83.22%	81.74%	**84.07%**

3.2. Analyses of Real Data

As shown in Table 5 and Figures 4 and 5, the R^2_{train} and $RMSE_{train}$ of the L_1, $L_{1/2}$ and MCP are 0.87, 0.87, 0.88 and 0.64, 0.62, 0.27, better than the values of 0.85, 0.86, 0.88 and 0.69, 0.63, 0.28 of the log-sum for the GHLI, EDCER and BATZD datasets, respectively. However, our proposed log-sum method is the best in terms of Q^2 and $RMSE_{CV}$. In the BATZD dataset, the $RMSE_{CV}$ obtained by log-sum is 0.23, lower than the values of 0.30, 0.30, 0.30, 0.28 and 0.26 of other methods. In the BCL2 dataset, the Q^2 obtained by log-sum is 0.75, higher than the 0.51, 0.57, 0.73, 0.73 and 0.67 of other methods. Moreover, a small subset of descriptors was selected by our proposed method; for example, for the EDCER dataset, the result of log-sum is 10, lower than the 47, 36, 17, 11 and 12 of L_{EN}, L_1, SCAD, MCP and $L_{1/2}$. Furthermore, for R^2_{Test} and $RMSE_{test}$, for the GHLI dataset, the best method is log-sum (0.75 and 0.88); L_{EN} and L_1 are second (0.74 and 0.90); MCP is third (0.73 and 0.91); $L_{1/2}$ is fourth (0.72 and 0.92); and the last is SCAD (0.72 and 0.93). Therefore, our proposed method is better than the other methods. In addition, we gave the experimental and predicted values for the four datasets.

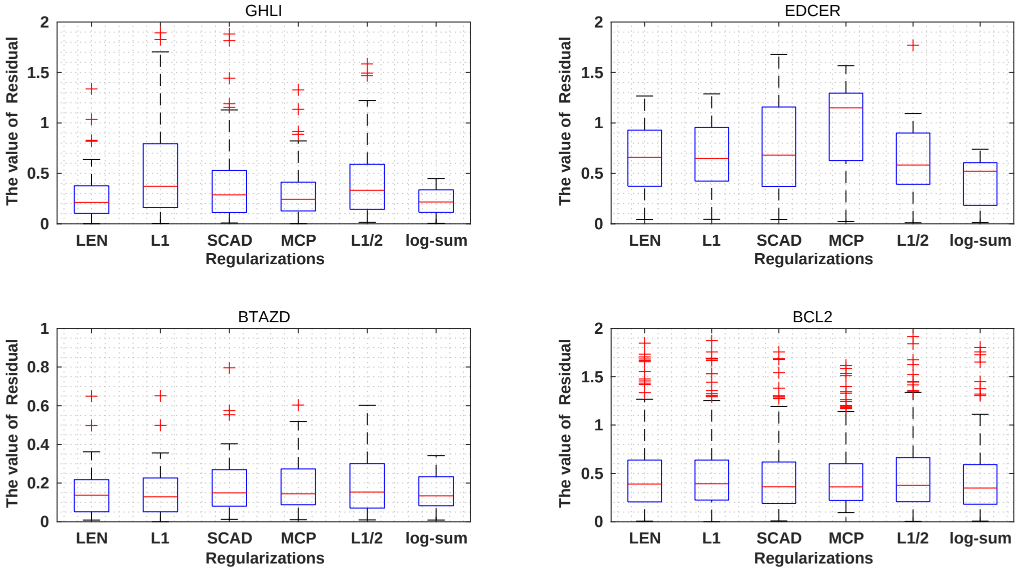

Figure 4. The value of residual ($|y - y^{pred}|$) on different datasets.

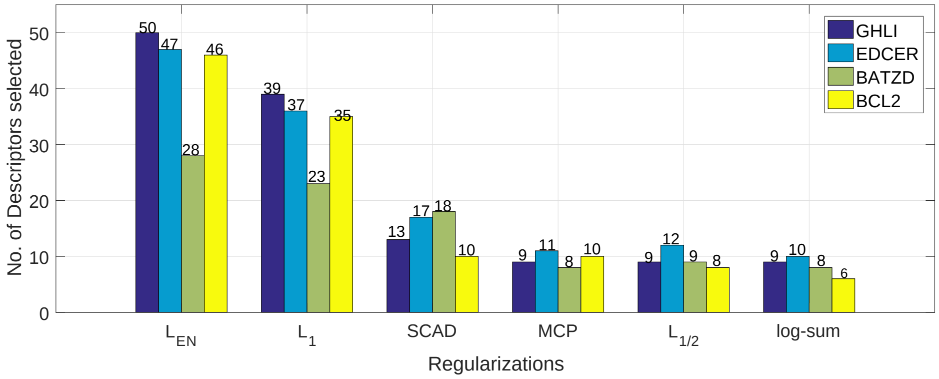

Figure 5. The number of descriptors obtained by the multiple linear regression with the different penalties on different datasets(different colors represent different datasets).

Table 5. Experimental results on the four datasets (the results are emphasized by our proposed method in bold and italic).

Datasets	Methods	R^2_{train}	RMSE$_{train}$	Q^2_{LOO}	RMSE$_{cv}$	R^2_{test}	RMSE$_{test}$
GHLI	L_{EN}	0.87	0.65	0.74	0.68	0.74	0.90
	L_1	0.87	0.64	0.75	0.67	0.74	0.90
	SCAD	0.84	0.71	0.82	0.62	0.72	0.93
	MCP	0.85	0.68	0.80	0.65	0.73	0.91
	$L_{1/2}$	0.82	0.75	0.81	0.62	0.72	0.92
	log-sum	***0.85***	***0.69***	***0.84***	***0.57***	***0.75***	***0.88***
EDCER	L_{EN}	0.81	0.74	0.70	0.70	0.64	1.23
	L_1	0.82	0.73	0.73	0.68	0.63	1.25
	SCAD	0.86	0.63	0.74	0.69	0.70	1.12
	MCP	0.83	0.70	0.74	0.69	0.65	1.21
	$L_{1/2}$	0.87	0.62	0.75	0.65	0.64	1.24
	log-sum	***0.86***	***0.63***	***0.79***	***0.62***	***0.70***	***1.12***
BATZD	L_{EN}	0.87	0.28	0.73	0.30	0.60	0.52
	L_1	0.88	0.28	0.74	0.30	0.60	0.52
	SCAD	0.86	0.30	0.77	0.30	0.62	0.51
	MCP	0.88	0.27	0.83	0.29	0.64	0.50
	$L_{1/2}$	0.86	0.29	0.84	0.26	0.64	0.50
	log-sum	***0.88***	***0.28***	***0.88***	***0.23***	***0.68***	***0.47***
BCL2	L_{EN}	0.75	0.57	0.51	0.53	0.61	0.67
	L_1	0.74	0.58	0.58	0.51	0.61	0.67
	SCAD	0.72	0.59	0.73	0.45	0.59	0.69
	MCP	0.74	0.57	0.73	0.46	0.58	0.70
	$L_{1/2}$	0.73	0.60	0.68	0.48	0.57	0.70
	log-sum	***0.68***	***0.64***	***0.75***	***0.43***	***0.65***	***0.63***

First of all, in Tables 6–9, the number of top-ranked informative descriptors identified by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum is 9, 10, 8 and 6 based on the value of the coefficients. Secondly, the common descriptors are emphasized in bold. Thirdly, as shown in Table 10, the number of descriptors is from the class of 2D. Then, the majority of descriptors are belong to the atom-type electrotopological state and autocorrelation of descriptors types. Finally, the name of the descriptors obtained by the log-sum method is exhibited in Table 11.

Table 6. The 9 top-ranked descriptors identified by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum from the GHLI dataset (the common descriptors are emphasized in bold).

Rank	**GHLI**					
	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
1	***JGI7***	***JGI7***	Mp	***JGI7***	minsCl	***ATSC4c***
2	***ETA_Eta_B_RC***	***ETA_Eta_B_RC***	MDEC-44	***ATSC4c***	ATSC1e	***GATS1e***
3	***BCUTc-1l***	***BCUTc-1l***	***GATS1e***	***GATS1e***	minaaN	***ATSC1p***
4	***Mv***	***Mv***	***ATSC1p***	AATS0e	WPOL	***MATS8m***
5	***ATSC4c***	***MDEN-23***	GGI9	meanI	***nHdsCH***	***maxwHBa***
6	***MDEN-23***	***ATSC4c***	***maxHBa***	***nHdsCH***	ALogP	***maxHBa***
7	***GATS1e***	***GATS1e***	***maxwHBa***	***maxHBa***	nFG12Ring	***ATSC7s***
8	ETA_Epsilon_3	***ETA_Epsilon_4***	***MATS8m***	***ATSC7s***	AATS6i	AATS0v
9	***ETA_Epsilon_4***	minHCsatu	SIC1	ATS4v	AATSC8m	ATS4p

Table 7. The 10 top-ranked descriptors identified by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum from the EDCER dataset (the common descriptors are emphasized in bold).

Rank	**EDCER**					
	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
1	***JGI10***	***JGI10***	***JGI10***	***JGI10***	***JGI10***	***JGI10***
2	***VE2_Dt***	***VE2_Dt***	MATS1i	***JGI6***	GATS1c	MATS1c
3	***JGI7***	***JGI6***	***AATSC2s***	***AATSC2s***	***GATS2s***	***hmax***
4	***AATSC8p***	***AATSC8p***	***hmax***	***AATSC8p***	***hmax***	***nssO***
5	***JGI6***	***JGI7***	***JGI6***	***hmax***	***GATS5v***	piPC6
6	***hmax***	***hmax***	nBase	nHBint2	nTG12Ring	***nFG12HeteroRing***
7	***SpMin4_Bhm***	***SpMin4_Bhm***	GATS8p	nHBd	***nssO***	***maxaaCH***
8	***GATS5v***	***GATS5v***	***nFG12HeteroRing***	***maxaaCH***	***maxaaCH***	SHBint2
9	***GATS2s***	***GATS2s***	MATS5v	C3SP2	ETA_Beta_ns_d	TIC1
10	SpMin5_Bhs	nAcid	***maxaaCH***	SHBint8	MDEC-24	AATSC8m

Table 8. The 8 top-ranked descriptors identified by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum from the BATZD dataset (the common descriptors are emphasized in bold).

Rank	**BATZD**					
	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
1	***JGI4***	***JGI4***	***VE2_Dze***	***SpMax1_Bhi***	***SpMax1_Bhi***	***SpMax1_Bhi***
2	***VE2_Dze***	***VE2_Dze***	JGI3	MATS5m	GATS1p	GATS1v
3	***MATS5v***	***ndS***	***ndS***	***GATS3s***	***ndS***	***GATS3s***
4	SdS	***MATS5v***	***CrippenLogP***	C4SP3	GATS3m	GATS8c
5	***CrippenLogP***	***CrippenLogP***	***nHother***	***CrippenLogP***	***GATS3s***	naaS
6	mindS	***MDEO-22***	minddssS	ALogP	***LipoaffinityIndex***	AATSC4i
7	***MDEO-22***	***nF9Ring***	GATS4m	***nHother***	nHsOH	***LipoaffinityIndex***
8	maxdS	ETA_Epsilon_4	***nF9Ring***	***ATSC8i***	***ATSC8i***	SpDiam_Dzp

Table 9. The 6 top-ranked descriptors identified by L_{EN}, L_1, SCAD, MCP, $L_{1/2}$ and log-sum from the BCL2 dataset (the common descriptors are emphasized in bold).

Rank	**BCL2**					
	L_{EN}	L_1	**SCAD**	**MCP**	$L_{1/2}$	**Log-Sum**
1	***JGI7***	***AATSC8p***	AATSC4s	***JGI7***	***MATS4s***	***AATSC8p***
2	VE2_D	***MATS4s***	***IC2***	***MATS4s***	***IC2***	***IC2***
3	***AATSC8p***	***MATS5m***	***MDEN-13***	***IC2***	***E3m***	GATS4s
4	***MATS5m***	***IC2***	minHsNH2	***E3m***	***MDEN-13***	***maxHBint2***
5	***MATS4s***	***MDEN-13***	***maxHBint2***	GATS8p	***maxHBint2***	***minsOH***
6	***IC2***	SpMax1_Bhi	nT8Ring	***MDEN-13***	***minsOH***	SwHBa

Table 10. The detailed information of the descriptors obtained by the log-sum method.

Descriptor Type	**Class**	**Descriptor**
Autocorrelation	2D	AATS0v; AATSC4i; AATSC8m; ATS4p; ATSC1p; ATSC4c; ATSC7s; GATS1e; GATS1v; GATS3s; GATS8c; MATS1c; MATS8m; AATSC8p; GATS4s
Atom-type electrotopological state	2D	Hmax; LipoaffinityIndex; maxaaCH; maxHBa; maxwHBa; naaS; nssO; SHBint2; maxHBint2; minsOH; SwHBa
Barysz matrix	2D	SpDiam_Dzp
Burden modified eigenvalues	2D	SpMax1_Bhi
Information content	2D	TIC1
Path counts	2D	piPC6
Ring count	2D	nFG12HeteroRing
Topological charge	2D	JGI10
Information content	2D	IC2

Table 11. The name of the descriptors obtained by the log-sum method.

Descriptor	Name
AATS0v	Average Broto–Moreau autocorrelation-lag 0/weighted by van der Waals volumes
AATSC4i	Average centered Broto–Moreau autocorrelation-lag 4/weighted by first ionization potential
AATSC8m	Average centered Broto–Moreau autocorrelation-lag 8/weighted by mass
ATS4p	Average centered Broto–Moreau autocorrelation-lag 1/weighted by polarizabilities
ATSC1p	Centered Broto–Moreau autocorrelation-lag 1/weighted by polarizabilities
ATSC4c	Average centered Broto–Moreau autocorrelation-lag 4/weighted by charges
ATSC7s	Average centered Broto–Moreau autocorrelation-lag 7/weighted by I-state
GATS1e	Geary autocorrelation-lag 1/weighted by Sanderson electronegativities
GATS1v	Geary autocorrelation-lag 1/weighted by van der Waals volumes
GATS3s	Geary autocorrelation-lag 3/weighted by I-state
GATS8c	Geary autocorrelation-lag 8/weighted by charges
hmax	Maximum H E-state
JGI10	Mean topological charge index of order 10
LipoaffinityIndex	Lipoaffinity index
MATS1c	Moran autocorrelation-lag 1/weighted by charges
MATS8m	Moran autocorrelation-lag 8/weighted by mass
maxaaCH	Maximum atom-type E-state: :CH:
maxHBa	Maximum E-states for (strong) hydrogen bond acceptors
maxwHBa	Maximum E-states for weak hydrogen bond acceptors
naaS	Count of atom-type E-state::C:-
nFG12HeteroRing	Number of >12-membered fused rings containing heteroatoms (N, O, P, S or halogens)
nssO	Count of atom-type E-state: -O-
piPC6	Conventional bond order ID number of order 6 (ln(1 + x)
SHBint2	Sum of E-state descriptors of strength for potential hydrogen bonds of path length 2
SpDiam_Dzp	Spectral diameter from Barysz matrix/weighted by polarizabilities
SpMax1_Bhi	Largest absolute eigenvalue of Burden-modified matrix - n 1/weighted by the relative first ionization potential
TIC1	Total information content index (neighborhood symmetry of 1-order)
SwHBa	Sum of E-states for weak hydrogen bond acceptors
AATSC8p	Average centered Broto–Moreau autocorrelation-lag 8/weighted by polarizabilities
IC2	Information content index (neighborhood symmetry of 2-order)
GATS4s	Geary autocorrelation-lag 4/weighted by I-state
maxHBint2	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 2
minsOH	Minimum atom-type E-state: -OH

4. Conclusions

In the field of drug design and discovery, only a few descriptors are of interest to the QSAR model. Therefore, descriptor selection plays an important role in the study of QSAR. In this paper, we proposed univariate log-sum thresholding for updating the estimated coefficients and developed a coordinate descent algorithm for log-sum penalized multiple linear regression.

Both experimental results on artificial and four QSAR datasets demonstrate that our proposed multiple linear regression with log-sum penalty is still better than L_1, L_{EN}, SCAD, MCP and $L_{1/2}$. Therefore, our proposed log-sum method is the effective technique in both descriptor selection and prediction of biological activity.

In this paper, we introduced random sampling, which is easy to use, for QSAR data preprocessing. However, this method does not take into account additional knowledge. Therefore, we plan to integrate a self-paced learning mechanism, which learns easy samples first and then gradually takes into consideration complex samples, making the model more and more mature, with our proposed method in future work.

Acknowledgments: This work was supported by the Macau Science and Technology Development Funds $GrantNo.$003/2016/AFJ from the Macau Special Administrative Region of the People's Republic of China, the National Grand Fundamental Research 973 Program of China under Grant No. 2013CB329404 and the China NSFC projects under Contracts 61373114, 61661166011, 11690011, 61721002.

Author Contributions: Liang-Yong Xia, Hua Chai and Yong Liang designed the simulations. Liang-Yong Xia and De-Yu Meng provided the mathematical proof. Liang-Yong Xia, Xiao-Jun Yao and Yu-Wei Wang contributed to collecting the datasets and analyze the data. Liang-Yong Xia and Yong Liang designed and implemented the algorithm. Liang-Yong Xia, Yu-Wei Wang, De-Yu Meng, Xiao-Jun Yao and Yong Liang contributed to the interpretation of the results. Liang-Yong Xia took the lead in writing the manuscript. Yu-Wei Wang, De-Yu Meng, Xiao-Jun Yao, Hua Chai and Yong Liang revised the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

QSAR	Quantitative structure-activity relationship
QSRR	Quantitative structure-(chromatographic) retention relationships
QSPR	Quantitative structure-property relationship
QSTR	Quantitative structure-toxicity relationship
MLR	Multiple linear regression
MCP	Maximum concave penalty
SCAD	Smoothly clipped absolute deviation
L_1	LASSO
BTAZD	(Benzo-)Triazoles toxicity in Daphnia magna
EDCER	EDC estrogen receptor binding
GHLI	Global half-life index
BCL2	Apoptosis regulator Bcl-2

Appendix A. Proof

We first consider the situation $\beta_j > 0$:

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}(-x_{ij}(y_i - \sum_{k\neq j} x_{ij}\beta_k - x_{ij}\beta_j)) + \lambda\frac{1}{\beta_j + \varepsilon} = 0 \tag{A1}$$

Based on Equation (A1), the gradient of the log-sum regularization at β_j can be expressed as:

$$\frac{\partial R}{\partial \beta_j} = \beta_j - \omega_j + \lambda\frac{1}{\beta_j + \varepsilon} = 0 \tag{A2}$$

Denote $\widetilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k, \widetilde{r}_i^{(j)} = y_i - \widetilde{y}_i^{(j)}, w_j = \sum_{i=1}^{n} x_{ij}\widetilde{r}_i^{(j)}$, which is equivalent to:

$$\beta_j^2 - (\omega_j - \varepsilon)\beta_j + (\lambda - \omega_j\varepsilon) = 0 \tag{A3}$$

$$\beta_j = \frac{\omega_j - \varepsilon \pm \sqrt{(\omega_j - \varepsilon)^2 - 4(\lambda - \omega_j\varepsilon)}}{2} \tag{A4}$$

let: $c_1 = \omega_j - \varepsilon$,$c_2 = c_1^2 - 4(\lambda - \omega_j\varepsilon)$ Thus, we have:

(1) if $c_2 < 0$, Equation (A3) has no real solution.
(2) if $c_2 = 0$, Equation (A3) has the solution $\beta_j = \frac{c_1}{2}$.
(3) if $c_2 > 0$, Equation (A3) has the two solutions $\beta_{j1} = \frac{c_1 - \sqrt{c_2}}{2}$ and $\beta_{j2} = \frac{c_1 + \sqrt{c_2}}{2}$:

$$\begin{aligned} c_2 &= (\omega_j - \varepsilon)^2 - 4(\lambda - \omega_j\varepsilon) \\ &= \omega_j^2 - 2\omega_j\varepsilon + \varepsilon^2 - 4\lambda + 4\omega_j\varepsilon \\ &= (\omega_j + \varepsilon)^2 - 4\lambda > 0 \\ &\omega_j + \varepsilon > 2\sqrt{\lambda} \\ &\omega_j - \varepsilon > 2\sqrt{\lambda} - 2\varepsilon \\ &c_1 > 0 \end{aligned}$$

Thus, $\beta_{j2} > \beta_{j1} > 0$, and it is then easy to obtain that $f'(\beta_j) > 0$ when $0 < \beta_j < \beta_{j1}$ or $\beta_{j2} > \beta_j$ and $f'(\beta_j) < 0$ when $\beta_{j1} < \beta_j < \beta_{j2}$. Therefore, Equation (16) has a local minimum. For $\beta_j < 0$, we can prove it in a similar way.

References

1. Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, M.; Kahn, I.; Dobchev, D.A. Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction. *Chem. Rev.* **2010**, *110*, 5714–5789.
2. Shahlaei, M. Descriptor selection methods in quantitative structure-activity relation-ship studies: A review study. *Chem. Rev.* **2013**, *113*, 8093–8103.
3. Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. Vsmp: A novel variable selection and modeling method based on the prediction. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964–969.
4. Xu, L.; Zhang, W.-J. Comparison of different methods for variable selection. *Anal. Chim. Acta* **2001**, *446*, 475–481.
5. Wegner, J.K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.
6. Khajeh, A.; Modarress, H.; Zeinoddini-Meymand, H. Modified particle swarm optimization method for variable selection in qsar/qspr studies. *Struct. Chem.* **2013**, *24*, 1401–1409.
7. Meissner, M.; Schmuker, M.; Schneider, G. Optimized particle swarm optimization (OPSO) and its application to artificial neural network training. *BMC Bioinform.* **2006**, *7*, 125.
8. Ghosh, P.; Bagchi, M. QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. *Curr. Med. Chem.* **2009**, *16*, 4032–4048.
9. Burden, F.; Winkler, D. Bayesian regularization of neural networks. *Artif. Neural Netw. Methods Appl.* **2009**, *458*, 23–42.
10. Dorigo, M.; Birattari, M.; Stutzle, T. Ant colony optimization. *IEEE Comput. Intell. Mag.* **2006**, *1*, 28–39.
11. Zheng, W.; Tropsha, A. Novel variable selection quantitative structure- property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
12. Mercader, A.G.; Duchowicz, P.R.; Fern'andez, F.M.; Castro, E.A. Modified and enhanced replacement method for the selection of molecular descriptors in qsar and qspr theories. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 138–144.
13. Ara'ujo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65–73.
14. Put, R.; Daszykowski, M.; Baczek, T.; Heyden, Y.V. Retention prediction of peptides based on uninformative variable elimination by partial least squares. *J. Proteome Res.* **2006**, *5*, 1618–1625.
15. Daghir-Wojtkowiak, E.; Wiczling, P.; Bocian, S.; Kubik, L.; Koslinski, P.; Buszewski, B.; Kaliszan, R.; Markuszewski, M.J. Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography. *J. Chromatogr. A* **2015**, *1403*, 54–62.
16. Goodarzi, M.; Chen, T.; Freitas, M.P. QSPR predictions of heat of fusion of organic compounds using Bayesian regularized artificial neural networks. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 260–264.
17. Aalizadeh, R.; Peter, C.; Thomaidis, N.S. Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization-Support Vector Machine QSTR models. *Environ. Sci. Process. Impacts* **2017**, *19*, 438–448.
18. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *73*, 267–288.
19. Algamal, Z.; Lee, M. A new adaptive l1-norm for optimal descriptor selection of high-dimensional qsar classification model for anti-hepatitis c virus activity of thiourea derivatives. *SAR QSAR Environ. Res.* **2017**, *28*, 75–90.
20. Xu, Z.; Chang, X.; Xu, F.; Zhang, H. l1/2 regularization: A thresholding repre-sentation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1013–1027.
21. Algamal, Z.; Lee, M.; Al-Fakih, A.; Aziz, M. High-dimensional qsar modeling using penalized linear regression model with l1/2-norm. *SAR QSAR Environ. Res.* **2016**, *27*, 703–719.
22. Liang, Y.; Liu, C.; Luan, X.-Z.; Leung, K.-S.; Chan, T.-M.; Xu, Z.B.; Zhang, H. Sparse logistic regression with a l1/2 penalty for gene selection in cancer classification. *BMC Bioinform.* **2013**, *14*, 198.

23. Candes, E.J.; Wakin, M.B.; Boyd, S.P. Enhancing sparsity by reweighted l1 minimization. *J. Fourier Anal. Appl.* **2008**, *14*, 877–905.
24. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320.
25. Donoho, D.L.; Johnstone, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455.
26. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
27. Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942.
28. Gramatica, P.; Papa, E. Screening and ranking of pops for global half-life: Qsar approaches for prioritization based on molecular structure. *Environ. Sci. Technol.* **2007**, *41*, 2833–2839.
29. Li, J.; Gramatica, P. The importance of molecular structures, endpoints values, and predictivity parameters in qsar research: Qsar analysis of a series of estrogen receptor binders. *Mol. Divers.* **2010**, *14*, 687–696.
30. Cassani, S.; Kovarich, S.; Papa, E.; Roy, P.P.; van der Wal, L.; Gramatica, P. Daphnia and fish toxicity of (benzo) triazoles: Validated qsar models, and interspecies quantitative activity-activity modeling. *J. Hazard. Mater.* **2013**, *258*, 50–60.
31. Zakharov, A.V.; Peach, M.L.; Sitzmann, M.; Nicklaus, M.C. Qsar modeling of imbalanced high-throughput screening data in pubchem. *J. Chem. Inf. Model.* **2014**, *54*, 705–712.
32. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. *J. Comput. Chem. Softw. News Updates* **2014**, *35*, 1036–1044.
33. Golbraikh, A.; Tropsha, A. Beware of q2. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.