mathematics

MDPI

Article

A More Accurate Estimation of Semiparametric Logistic Regression

Xia Zheng †, Yaohua Rong *,†, Ling Liu and Weihu Cheng

Faculty of Science, College of Statistics and Data Science, Beijing University of Technology, Beijing 100124, China; zhengx@emails.bjut.edu.cn (X.Z.); liuling@bjut.edu.cn (L.L.); chengweihu@bjut.edu.cn (W.C.)

* Correspondence: rongyaohua@bjut.edu.cn

† These authors contributed equally to this work.

Abstract: Growing interest in genomics research has called for new semiparametric models based on kernel machine regression for modeling health outcomes. Models containing redundant predictors often show unsatisfactory prediction performance. Thus, our task is to construct a method which can guarantee the estimation accuracy by removing redundant variables. Specifically, in this paper, based on the regularization method and an innovative class of garrotized kernel functions, we propose a novel penalized kernel machine method for a semiparametric logistic model. Our method can promise us high prediction accuracies, due to its capability of flexibly describing the complicated relationship between responses and predictors and its compatibility of the interactions among the predictors. In addition, our method can also remove the redundant variables. Our numerical experiments demonstrate that our method yields higher prediction accuracies compared to competing approaches.

Keywords: logistic model; kernel machine; variable selection; semiparametric model

Citation: Zheng, X.; Rong, Y.; Liu, L.; Cheng, W. A More Accurate Estimation of Semiparametric Logistic Regression. *Mathematics* **2021**, *9*, 2376. https://doi.org/10.3390/math9192376

Academic Editor: Jin-Ting Zhang

Received: 25 August 2021
Accepted: 22 September 2021
Published: 24 September 2021

1. Introduction

Logistic regression, a well-known classification prediction model, is an extension of linear models in dealing with binary responses. However, its prediction accuracies can be decreased when redundant predictors are included. Therefore, a high demand has emerged for a flexible classification prediction model which has high prediction accuracies and can select important predictors. Recently, regularization methods have had great success in improving stability of estimation and increasing accuracies of prediction. Popular regularization methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) [1], the Smoothly Clipped Absolute Deviation (SCAD) [2], adaptive LASSO [3], and Elastic Net [4], have been widely used in logistic regression models with linear structure [5–7]. In addition, different types of regularization function have also been explored to be rolled in logistic regression models. For example, a $\ell_{\frac{1}{2}}$ penalty and a hybrid $\ell_{\frac{1}{2}+2}$ regularization function have been used in sparse logistic regression models [8,9]. Likewise, in [10], the ℓ_0-norm has been used in a sparse generalized linear model. By shrinking some regression coefficients to exactly zero, most of the above methods can realize model estimation and variable selection simultaneously.

However, linear structure may be insufficient to capture the complicated relationship between responses and predictors. To release the linear assumption, some nonparametric and semiparametric methods have been presented. Among these methods, a Generalized Additive Model (GAM) [11] has drawn increasing attention. Meier et al. [12] estimated a high-dimensional GAM via a sparsity-smoothness penalty. Ravikumar et al. [13] proposed a variable selection procedure for GAM and extended it to an additive logistic model. Li et al. [14] developed a SCAD-based method for simultaneous variable selection and estimation in GAM with non-polynomial dimensionality. These additive models are generally sensible. However, the additive structure may be not flexible enough for some

real data. In addition, the spline based methods used in these models, which requires the specification of the smoothness condition of the unknown function, may be involved and awkward for multidimensional data.

Known as another function approximation method, a kernel machine method starts with a kernel function which implicitly determines the smoothness of the unknown function rather than decides a predetermined form or basis, and hence greatly simplifies specification of a nonparametric function especially for multidimensional data. More recently, many promising kernel machine methods were proposed. For example, the Kernel Logistic Regression (KLR) [15], which replaces the loss function of a support-vector machine with a negative log-likelihood of binomial distribution, can obtain a natural estimate of the posterior probability. Compared to nonparametric models, semiparametric models are more widely used to deal with real data. This is because the semiparametric models not only retain the flexibility of the nonparametric models but also maintain the interpretability of the parametric models. The Least-Squares Kernel Machine method (LSKM) [16] is a popular semiparametric model and has been extended to a generalized semiparametric model [17]. To overcome the limitation of LSKM for small sample cases, a Bayesian hierarchical modeling approach has been proposed by Kim et al. [18]. In addition, Freytag et al. [19,20] proposed two modified kernel functions based on LSKM to improve identification of meaningful associations. However, the prediction accuracies of these methods would be decreased when redundant predictors were contained in the models.

Variable selection based on kernel machine has attracted a lot of attention. The COmponent Selection and Smoothing Operator (COSSO) method [21] is proposed for variable selection via using an ℓ_1 penalty in nonparametric kernel models, which is designed in the framework of smoothing splines ANOVA model, and then is extended to a logistic model by Zhang and Lin [22]. Allen [23] showed that a fully nonparametric KerNel Iterative Feature Extraction (KNIFE) method could also achieve feature selection via using a linearized weighted kernel. In addition, Xu et al. [24] modified the penalty form of the KLR with $\ell_{\frac{1}{2}}$ regularization to gain more sparse solutions. The above variable selection methods are based on fully nonparametric models. Moreover, the general kernel function assumes that the predictors have the same marginal effect on the response variable. This assumption is not flexible enough to describe the relationship between predictors and responses. Alternatively, considering the presence of possible gene–gene interactions, a garrote kernel machine method which uses a score test for variable selection was proposed in Maity and Lin [25]. However, it is a backwards selection method and thus not efficient for modeling. By using a regularization method, a Penalized Garrotized Kernel Machine method (PGKM) which can select important predictors and process modeling simultaneously is proposed in Rong et al. [26]. This is an efficient method for continuous outcomes.

A semiparametric logistic model is a widely used model to predict health outcomes in terms of clinical information and gene expression measurements. The accuracy of modeling estimation is the core problem. Thus, our task is to construct a method which can guarantee the estimation accuracy by removing redundant variables. Specifically, in this paper, we propose a novel Penalized Logistic Garrotized Kernel Machine (PLGKM) method for binary outcomes. PLGKM can eliminate irrelevant predictors in both the parametric and nonparametric part, while allowing for a complex nonlinear relationship between the responses and predictors. The remainder of this paper is organized as follows. In Section 2, we present a novel garrotized kernel machine method for a semiparametric logistic model and investigate the model estimation and variable selection problem. In addition, we propose a "one-group-at-a-time" cyclical coordinate descent algorithm for the solution path of the tuning parameters. The performance of our proposed method is evaluated by several simulation examples in Section 3. We apply our method to a study of the breast cancer in Section 4. Section 5 considers the extension to a generalized kernel machine model. Section 6 contains concluding remarks.

2. Methods

2.1. Kernel Machine for Semiparametric Logistic Regression

We are interested in the relationship between predictors (clinical and demographic information, gene expression measurements) and disease outcomes. A popular technique to explore the relationship between these predictors and responses is semiparametric models which treat clinical and demographic variables in a linear part and set the gene variables in a nonparametric part. By this way, not only can we interpret the margin effect of the clinical and demographic variables on disease outcomes, but we also allow a nonlinear relationship between gene variables and outcomes.

Suppose we have n observations constructed with healthy individuals and patients. We take y_i to denote the health situation of the i-th individual. Specifically, $y_i = 1$ refers to a patient and $y_i = 0$ refers to a healthy individual. For each individual, we collect P clinical and demographic predictors and Q gene predictors. Thus, we have clinical and demographic predictors $\boldsymbol{x}_i \in \mathcal{R}^P$ and gene predictors $\boldsymbol{z}_i \in \mathcal{R}^Q$. Then, we can construct a semiparametric model as follows:

$$\log(\frac{p_i}{1-p_i}) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i), \tag{1}$$

where $p_i = prob(y_i = 1|\boldsymbol{x}_i, \boldsymbol{z}_i)$, $\boldsymbol{\beta}$ is an unknown $P \times 1$ vector of regression coefficients, and $h(\cdot)$ is an unknown and possibly complicated function.

The first part in the model (1) is just a linear model which can be estimated easily. Thus, in this paper, we focus on exploring the nonparametric part of (1). We assume that $h(\cdot)$ lies in a Reproducing Kernel Hilbert Space $\mathcal{H}_K$ and use a linear combination of positive definite kernel functions to construct $h(\cdot)$. Exactly, as Mercer's theorem [27] claims, a positive definite kernel function $K(\cdot, \cdot)$ produces a Reproducing Kernel Hilbert Space $\mathcal{H}_K$. In addition, these positive definite kernels are also called Mercer kernels. Two popular Mercer kernels are:

- Polynomial Kernel: $K(\boldsymbol{z}_i, \boldsymbol{z}_j) = (\boldsymbol{z}_i^T \boldsymbol{z}_j + c)^d$, where c is tuning parameter. The corresponding feature vector contains all terms up to degree d.
- Gaussian Kernel: $K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp\{-\sum_{q=1}^{Q}(z_{iq} - z_{jq})^2/\gamma\}$, where γ is known as the bandwidth. The Gaussian kernel function can be seen as a function of Euclidean distance. In addition, its feature map lies in an infinite dimensional space.

There are also some other commonly used kernel functions like the neural network and smoothing spline kernels [28]. One can choose an appropriate kernel function according to a real situation.

2.2. Penalized Garrotized Kernel Machine for Semiparametric Logistic Regression

In some real data analysis, the redundant predictors contained in $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ may decrease the estimation and prediction accuracies of the semiparametric logistic regression model. Therefore, variable selection is necessary and especially challenging in nonparametric components. In order to facilitate the variable selection in nonparametric components, Rong et al. [26] propose a "garrotized" kernel $K^{(g)}$ which can describe the influence of different marginal effects on the response y_i. In addition, $K^{(g)}$ is defined by

$$\begin{aligned} &K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\delta}) = K(\boldsymbol{z}_i^*, \boldsymbol{z}_j^*), \\ &\boldsymbol{z}_u^* = (\delta_1^{1/2} z_{u1}, \cdots, \delta_Q^{1/2} z_{uQ})^T, u = i, j, \\ &\delta_q \geq 0,\ q = 1, \cdots, Q. \end{aligned} \tag{2}$$

For example, the dth garrotized Polynomial Kernel is $K(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\delta}) = (\boldsymbol{z}_1^T \Delta \boldsymbol{z}_2 + c)^d$, where the $\Delta = \text{diag}(\delta_1, ..., \delta_Q)$ and the garrotized Gaussian kernel is $K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\delta}) = \exp\{-\sum_{q=1}^{Q} \delta_q (z_{iq} - z_{jq})^2\}$.

The unknown δ can be estimated from data. Each δ_q modulates the effect of nonparametric predictors z_q on response. $\delta_q = 0$ implies that z_q is not predictive of the response. Thus, the garrotized kernel formulation provides a flexible way to select variables in a semiparametric setting. In addition, with some proper $K(\cdot,\cdot)$, the nonparametric predictors' interactions will be allowed in the complicated function $h(\cdot)$.

To eliminate the interference of variable dimension, we first standardize the predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. Then, to estimate the parameters of model (1) with the garrotized kernel (2), we maximize the following popular penalized log-likelihood function:

$$\begin{aligned} & f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) \\ =& \frac{1}{n}\sum_{i=1}^{n}\Big\{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)\Big\} - \lambda_1 \sum_{p=1}^{P}|\beta_p| - \lambda_2 \sum_{q=1}^{Q}\delta_q - \frac{1}{2}\lambda_3 \|h\|^2_{\mathcal{H}_{K^{(g)}}} \\ =& \frac{1}{n}\sum_{i=1}^{n}\Big\{y_i[\boldsymbol{x}_i^T\boldsymbol{\beta} + h(\boldsymbol{z}_i)] - \log[1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + h(\boldsymbol{z}_i))]\Big\} - \lambda_1 \sum_{p=1}^{P}|\beta_p| \\ & - \lambda_2 \sum_{q=1}^{Q}\delta_q - \frac{1}{2}\lambda_3 \|h\|^2_{\mathcal{H}_{K^{(g)}}}, \end{aligned} \tag{3}$$

where λ_1 and λ_2 are non-negative regularization parameters, λ_3 is a trade-off parameter between goodness of fit and complexity of the model, and $\|h\|_{\mathcal{H}_{K^{(g)}}}$ denotes the functional norm in the Reproducing Kernel Hilbert Space $\mathcal{H}_{K^{(g)}}$ generated by the garrotized kernel. The linear mapping of the predictors in the kernel function implies that the garrotized Gaussian kernel is essentially a Gaussian kernel. In addition, the garrotized Gaussian kernel is therefore positive definite. Then, with the representer theorem [29], the nonparametric function $h(\boldsymbol{z}_i)$ can be expressed as

$$h(\boldsymbol{z}_i) = \sum_{j=1}^{n}\alpha_j K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\delta}) = k_i(\boldsymbol{\delta})\boldsymbol{\alpha}, \tag{4}$$

where $k_i(\boldsymbol{\delta}) = \{K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_1; \boldsymbol{\delta}), \cdots, K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_n; \boldsymbol{\delta})\}$ is a $n \times 1$ vector, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T$ is an unknown parameter vector. By substituting (4) into (3), we have

$$\begin{aligned} f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) =& \frac{1}{n}\sum_{i=1}^{n}\Big\{y_i[\boldsymbol{x}_i^T\boldsymbol{\beta} + k_i(\boldsymbol{\delta})\boldsymbol{\alpha}] - \log\Big[1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + k_i(\boldsymbol{\delta})\boldsymbol{\alpha})\Big]\Big\} \\ & - \lambda_1 \sum_{p=1}^{P}|\beta_p| - \lambda_2 \sum_{q=1}^{Q}\delta_q - \frac{1}{2}\lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}, \end{aligned} \tag{5}$$

where $\boldsymbol{K}(\boldsymbol{\delta})$ is a $n \times n$ Gram matrix whose (i,j)-th element is $K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\delta})$.

Equivalently, we can rewrite the maximization of (5) in matrix form as

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}} & \frac{1}{n}\Big\{\boldsymbol{Y}^T[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}] - \boldsymbol{1}_n^T \boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})\Big\} \\ & - \lambda_1\|\boldsymbol{\beta}\|_1 - \lambda_2\|\boldsymbol{\delta}\|_1 - \frac{1}{2}\lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}, \end{aligned} \tag{6}$$

where $\boldsymbol{1}_n = (1, 1, \cdots, 1)^T$, $\boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) = (\log[1+\exp(\boldsymbol{x}_1^T\boldsymbol{\beta} + \boldsymbol{k}_1(\boldsymbol{\delta})\boldsymbol{\alpha})], \cdots, \log[1+\exp(\boldsymbol{x}_n^T\boldsymbol{\beta} + \boldsymbol{k}_n(\boldsymbol{\delta})\boldsymbol{\alpha})])^T$ is an n-dimensional vector and $\|\cdot\|_1$ is the ℓ_1 norm. The solution to (6) is called a PLGKM estimate for a semiparametric logistic regression model.

2.3. Algorithm

To solve (6), we propose a "one-group-at-a-time" cyclical coordinate descent algorithm. With a pre-specified regularization parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, this algorithm is processed along a regularization path to update $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}$ cyclically.

Specifically, giving $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t-1)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(t-1)}$, problem (6) can be written as

$$\begin{aligned} \arg\max_{\boldsymbol{\beta}} \frac{1}{n}\Big\{\boldsymbol{Y}^T\Big[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}\Big] - \mathbf{1}_n^T \boldsymbol{A}(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}, \boldsymbol{\delta}^{(t-1)})\Big\} \\ - \lambda_1\|\boldsymbol{\beta}\|_1 - \lambda_2\|\boldsymbol{\delta}^{(t-1)}\|_1 - \frac{1}{2}\lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}, \end{aligned} \tag{7}$$

to get the solution to (7), we adapt the penalized reweighed least squares approach proposed in [1]. We denote the log-likelihood as $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \boldsymbol{Y}^T[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}] - \mathbf{1}_n^T \boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$. Taking $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}$, the log-likelihood can be seen as a function of $\boldsymbol{\eta}$. Thus, we can get the gradient and Hessian of $\ell_n(\cdot)$ with respect to $\boldsymbol{\eta}$ as $\ell_n'(\boldsymbol{\eta})$ and $\ell_n''(\boldsymbol{\eta})$, respectively. Given the current estimate $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\boldsymbol{\beta}^{(t-1)} + \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}$, maximizing log-likelihood $\ell_n(\boldsymbol{\eta})$ is equivalent to maximizing the following two-term Taylor expansion of the log-likelihood

$$\frac{1}{2n}[\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}]^T \ell_n''(\tilde{\boldsymbol{\eta}})[\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}],$$

where $\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) = \tilde{\boldsymbol{\eta}} - \ell_n''(\tilde{\boldsymbol{\eta}})^{-1}\ell_n'(\tilde{\boldsymbol{\eta}})$. Similar to [1], to reduce the difficulty of computation, we approximate the Hessian matrix with a diagonal one. Thus, we can get an approximate solution to (7) by minimizing the penalized reweighed least squares with respect to $\boldsymbol{\beta}$

$$\frac{1}{2n}[\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}]^T[\boldsymbol{W}(\tilde{\boldsymbol{\eta}})][\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}] + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{8}$$

where $\boldsymbol{W}(\tilde{\boldsymbol{\eta}})$ is a positive definite diagonal matrix which has the same diagonal elements as $-\ell_n''(\tilde{\boldsymbol{\eta}})$. Hence, we can transfer (7) to a convex optimization problem. Treating $\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}$ as the working response, $\boldsymbol{W}(\tilde{\boldsymbol{\eta}})$ as weight, and the standard computational procedures for estimation of LASSO regression [30,31] can be used to estimate $\boldsymbol{\beta}$.

Similarly, to update the estimate of $\boldsymbol{\alpha}$ with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(t-1)}$, we can rewrite problem (6) as

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha}} \frac{1}{n}\Big\{\boldsymbol{Y}^T\Big[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}\Big] - \mathbf{1}_n^T \boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\delta}^{(t-1)})\Big\} \\ - \lambda_1\|\boldsymbol{\beta}^{(t)}\|_1 - \lambda_2\|\boldsymbol{\delta}^{(t-1)}\|_1 - \frac{1}{2}\lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}, \end{aligned} \tag{9}$$

in addition, problem (9) is equivalent to minimizing the following function with respect to $\boldsymbol{\alpha}$,

$$\begin{aligned} &\frac{1}{2n}[\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}]^T \boldsymbol{W}(\tilde{\boldsymbol{\eta}})[\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}] \\ &+ \frac{1}{2}\lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}, \end{aligned} \tag{10}$$

which is a quadratic function of $\boldsymbol{\alpha}$. The stationary point of this function is the solution to the following linear equation:

$$\Big[\frac{1}{n}\boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{W}(\tilde{\boldsymbol{\eta}})\boldsymbol{K}(\boldsymbol{\delta}^{(t-1)}) + \lambda_3 \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\Big]\boldsymbol{\alpha} = \frac{1}{n}\boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{W}(\tilde{\boldsymbol{\eta}})(\boldsymbol{V}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{X}\boldsymbol{\beta}^{(t)}), \tag{11}$$

which is straightforward to solve. When the left-hand side of the above equation is singular, a diagonal matrix with small entries such as 1×10^{-5}, can be added to stabilize the estimate.

Finally, with $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$, updating $\boldsymbol{\delta}$ is equivalent to solving a nonlinear optimization problem under the constraints of $\delta_q > 0$, $q = 1, \cdots, Q$. To this end, we

propose to use a one-at-a-time coordinate descent algorithm. Specifically, for δ_{q_0}, $q_0 = 1, \cdots, Q$, with $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and $\delta_q = \delta_q^{(t-1)}, q \neq q_0, q = 1, \cdots, Q$, (6) can be written as

$$\begin{aligned} \arg\max_{\delta_{q_0}} \frac{1}{n}\Big\{ Y^T \Big[X\boldsymbol{\beta}^{(t)} + \boldsymbol{K}(\delta)\boldsymbol{\alpha}^{(t)} \Big] - \mathbf{1}_n^T \boldsymbol{A}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \delta) \Big\} \\ - \lambda_1 \|\boldsymbol{\beta}^{(t)}\|_1 - \lambda_2 \sum_{q \neq q_0}^{Q} \delta_q^{(t-1)} - \lambda_2 \delta_{q_0} - \frac{1}{2}\lambda_3 {\boldsymbol{\alpha}^{(t)}}^T \boldsymbol{K}(\delta)\boldsymbol{\alpha}^{(t)}, \end{aligned} \tag{12}$$

the estimate of δ_{q_0} can be obtained by using the L-BFGS-B algorithm [32] on a standard univariate nonlinear constrained optimization software program. The coordinate descent algorithm for PLGKM is outlined in Algorithm 1:

Algorithm 1 Coordinate descent algorithm [31] for PLGKM

1: Initialization: $t = 0$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(0)}$.
2: **repeat**
3: Set $t = t + 1$.
4: Set $\boldsymbol{\beta}^{(t)}$ to the solution to maximizing function (8).
5: Set $\boldsymbol{\alpha}^{(t)}$ to the solution to (11).
6: Set each element $\delta_{q_0}^{(t)}$ of $\boldsymbol{\delta}^{(t)}$ to the solution to maximizing function (12).
7: **until** convergence

2.4. Selection of Tuning Parameters

We use a decreasing sequence of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ to fit models on the training set. This scheme not only provides us a path of solutions but also exploits warm starts and thus makes the algorithm more stable and faster. To select the optimal tuning parameters, the cross-entropy loss (CEL) is calculated on the validation dataset. The CEL is defined as

$$\text{CEL} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]. \tag{13}$$

Finally, we choose the estimated model that gives the lowest CEL on the validation set. In practice, we start with a rough search in a large range to find a relatively reasonable value for $\boldsymbol{\lambda}$. In addition, then, according to this value, we launch a finer local search to get an appropriate value for $\boldsymbol{\lambda}$.

3. Simulation

3.1. Comparison with LSKM

We conduct a simulation study to compare our PLGKM method with the LSKM method of Liu et al. [17]. This simulation study is started with the generation of 500 observations from the following semiparametric logistic model:

$$\text{logit}(p_i) = \boldsymbol{x}_i\boldsymbol{\beta} + h(\boldsymbol{z}_i), \tag{14}$$

where $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$. x_{ip} and z_{iq} are independently generated from $U(-1, 1)$ and $U(0, 1)$, respectively. To mimic the complicated relationship between gene expression predictors and outcomes, we employ a nonparametric function $h(\cdot)$ which is able to describe the nonlinearity of predictors and is compatible with the interactions among predictors.

To fairly compare the PLGKM method and LSKM method, we randomly choose four different models (exactly the following setting 1 to setting 4) which vary in the number of predictors and the form of the nonparametric function. In addition, in practice, without prior information, redundant predictors are often contained in the models. To show that our PLGKM method can remove the redundant predictors, we add some irrelevant predictors to two models (exactly the following setting 2 and setting 4):

Setting 1: $P = 1$, $Q = 5$, $\boldsymbol{\beta} = 1$, $h(\boldsymbol{z}) = \cos(z_1) - 2z_1^2 + 3.5\exp(z_1)z_2 - 3.5\sin(z_2)\cos(z_3) + z_2z_3 + z_2\sin(z_4) + 1.5z_3\sin(z_4)\sin(z_5) - 1.5z_4^3 - 1.5\exp(z_4)\cos(z_5)$.

Setting 2: $P = 2$, $Q = 15$, $\boldsymbol{\beta} = (1,0)^T$, $h(\cdot)$ is the same as the nonparametric function in setting 1. Thus, in this setting, there is 1 irrelevant $\boldsymbol{X}$ predictor and 10 irrelevant $\boldsymbol{Z}$ predictors.

Setting 3: $P = 1$, $Q = 10$, $\boldsymbol{\beta} = 1$, $h(\boldsymbol{z}) = \cos(z_1) - 2z_2^2 + 1.5\exp(z_3)z_4 - 2\cos(z_3)\sin(z_5) + 2z_1z_5 + z_6\sin(z_7) - 1.5\cos(z_6)z_7 - 1.5z_8^3 - z_8z_9 + \exp(z_9)z_{10} - 1.5\exp(z_9)\cos(z_{10}) + 1.5z_8\sin(z_9)\sin(z_{10})$.

Setting 4: $P = 2$, $Q = 30$, $\boldsymbol{\beta} = (1,0)^T$, $h(\cdot)$ is the same as the nonparametric function in setting 3. Thus, in this setting, there is one irrelevant $\boldsymbol{X}$ predictor and 20 irrelevant $\boldsymbol{Z}$ predictors.

For each setting, we split the 500 observations into a training set with 100 observations, a validation set with 300 observations, and a test set with the remaining 100 observations. To fit the generated data with a semiparametric model, we use our PLGKM method with a garrotized Gaussian kernel and the LSKM method with a Gaussian kernel separately.

The prediction performances of the fitted models are measured using the CEL and misclassification rates (MR) based on the test sets. The MR stands for the proportion of misclassified observations

$$\mathrm{MR} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i), \quad \hat{y}_i = \begin{cases} 1 & \hat{p}_i \geq 0.5, \\ 0 & \hat{p}_i < 0.5. \end{cases} \tag{15}$$

In the following Table 1, for each setting, we compare the mean prediction performances of PLGKM and LSKM based on 500 simulation experiments.

Table 1. MR and CEL of PLGKM and LSKM with standard deviations in parentheses.

	PLGKM	**LSKM**
Setting 1		
MR	**0.2777** (0.0478)	0.2803 (0.0473)
CEL	**0.5648** (0.0364)	0.5693 (0.0334)
Setting 2		
MR	**0.2964** (0.0514)	0.3589 (0.0627)
CEL	**0.5836** (0.0371)	0.6405 (0.0391)
Setting 3		
MR	**0.2864** (0.0469)	0.3136 (0.0514)
CEL	**0.5718** (0.0389)	0.6002 (0.0322)
Setting 4		
MR	**0.2949** (0.0553)	0.3914 (0.0616)
CEL	**0.5839** (0.0333)	0.6585 (0.0306)

From Table 1, we can see that our PLGKM method outperforms the LSKM method for these four settings. Specifically, for settings 1 and 3, our PLGKM method produces slightly smaller average MR and average CEL than the LSKM method. These small advantages are guaranteed by the flexibility of the garrotized kernel in the value of δ_q. For settings 2 and 4, our PLGKM method shows significant advantages in both average MR and average CEL. This is because the PLGKM method can eliminate the irrelevant variables by shrinking the estimation of the corresponding δ_q and β_p to be very small. Therefore, in practice, as it is unknown whether there are irrelevant variables, we recommend the PLGKM method for modeling.

Variable selection is a byproduct of our PLGKM method. The PLGKM method, allowing complicated nonlinear $h(\cdot)$, can simultaneously estimate the parameters of model (14)

and select the important variables. When the estimation of δ_q is zero, we conclude that the corresponding predictor z_q is not related to the outcome. In fact, similar to the SCAD procedure in [2], we consider δ_q to be zero when its estimation is less than 10^{-5}. In addition, for the clinical and demographic predictors in $\boldsymbol{X}$, we eliminate the irrelevant ones of which the estimation is exactly zero.

In the following Table 2, we show the variable selection ability of our PLGKM method based on settings 2 and 4.

Table 2. Variable selection results for the PLGKM methods, with different numbers of irrelevant $\boldsymbol{z}$.

		X			**Z**		
P	**Q**	**C** [1]	**O** [2]	**U** [3]	**C**	**O**	**U**
Setting 2							
2	15	0.5604	0.3626	0.0769	0.0110	0.2308	0.7582
Setting 4							
2	30	0.3188	0.5606	0.1515	0	0	1

[1] *C* (correct selection) denotes the percentage of 500 simulations in which the true model was exactly selected. [2] *O* (over-selection) denotes the percentage of 500 simulations in which the true model was nested in the selected model. [3] *U* (under-selection) denotes the percentage of 500 simulations in which the true model was not a subset of the selected model.

From Table 2, we can see that nearly all relevant $\boldsymbol{X}$ can be selected by PLGKM with a fairly high probability. This is reflected in the low under-selection rates of $\boldsymbol{X}$ given in Table 2. In addition, when the dimension of $\boldsymbol{Z}$ is not very high, our PLGKM method still has the opportunity to select all relevant $\boldsymbol{Z}$. In addition, this opportunity will become less when the dimension of $\boldsymbol{Z}$ increases.

3.2. Comparison with COSSO

The COSSO method [22], a popular method for the nonparametric model, shows an impressive ability in model estimation and variable selection. Thus, in this section, we would like to compare the performance of our PLGKM method based on Gaussian kernel for a semiparametric logistic model with the performance of the COSSO method for a nonparametric model (exactly an additive ANOVA model only containing the main effects). We generate data as in Section 3.1 and process 500 replications of each setting.

We evaluate the performances of these two methods in terms of MR and CEL which are reported in the following Table 3.

Table 3. MR and CEL of PLGKM and COSSO with standard deviations in parentheses.

	PLGKM	**COSSO**
Setting 1		
MR	**0.2790** (0.0389)	0.2924 (0.0943)
CEL	**0.5646** (0.0389)	0.6121 (0.0943)
Setting 2		
MR	**0.2963** (0.0504)	0.3162 (0.0599)
CEL	**0.5840** (0.0347)	0.6267 (0.0922)

From Table 3, we can see that the PLGKM method always produces considerably smaller average MR and CEL than the COSSO method. Exactly, compared with the COSSO method which is limited by its additive structure, the PLGKM method is more flexible as it can describe the more complicated relationships between the predictors and the outcomes by allowing the interactions among predictors. In addition, the outperformances are therefore guaranteed.

4. Analysis of the Breast Cancer Dataset

To explore the performance of our PLGKM method on microarray data, we consider a breast cancer data (GSE70947) obtained from the CuMiDa database (The microarray datasets in CuMiDa database were examined from more than 30,000 microarray experiments from the Gene Expression Omnibus using a rigorous filtering criteria [33]). We have expression data on 35,981 genes from 289 observations which are composed of 143 patients with breast cancer and 146 healthy individuals.

The high dimension of this breast cancer data makes modeling difficult. Thus, we first conduct a dimension reduction using the nonparametric independence screening method proposed in [34]. After this dimension reduction, we get adjusted breast cancer data with only 12 genes. Considering that there is no clinical nor demographic predictors, we fit this adjusted breast cancer data with a nonparametric model using our PLGKM method based on a garrotized Gaussian kernel. In addition, for comparison, we also apply the LSKM method based on Gaussian kernel and the LASSO method for logistic regression to this adjusted breast cancer data.

Specifically, we randomly split the adjusted breast cancer data into a training set with 100 observations for estimation, a validation set with 100 observations for tuning parameters selection, and a test set with the remaining 89 observations for prediction. We repeat this random split for 1000 times. For each repetition, we calculate the CEL and MR for these three methods, respectively. In addition, then we compare these three methods in terms of the average values of the CEL and MR which are collected in the following Table 4.

Table 4. Average prediction error of each method for 1000 replications, with their standard deviations in parentheses.

Methods	CEL (SD)	MR (SD)
PLGKM	**0.3349** (0.0282)	**0.1168** (0.0336)
LSKM	0.3464 (0.0274)	0.1224 (0.0344)
LASSO	0.4012 (0.0764)	0.1678 (0.0556)

From Table 4, we can see that our PLGKM outperforms the other two methods. Compared to the LSKM method, the PLGKM method benefits from the garrotized kernel and thus gives smaller average CEL and average MR. Compared to the LASSO-COX, our PLGKM method makes a significant improvement in prediction. In addition, this implies that the linear structure is not sufficient to describe the complicated relationships between the genes and the outcome and gene–gene interactions.

5. PLGKM Method for Outcomes Having an Exponential Family Distribution

In this paper, we introduce our PLGKM method based on binary outcomes. Exactly, when outcomes have an exponential family distribution, we can also use our PLGKM method to estimate the corresponding semiparametric model. The distribution of y_i follows the exponential family [35]:

$$\mathcal{P}(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i,\phi)\right\}, \tag{16}$$

where $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + h(\boldsymbol{z}_i)$ is a canonical parameter, ϕ is a dispersion parameter, w_i is a known weight, $b(\cdot)$, and $c(\cdot)$ are known functions. From (16), we can obtain $\mu_i = E(y_i) = b^{'}(\theta_i)$, $Var(y_i) = b^{''}(\theta_i)\phi/w_i$. Inspired by the generalized linear model, we can use link functions and garrotized kernel to construct a generalized semiparametric model as

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + h(\boldsymbol{z}_i), \tag{17}$$

where $g(\cdot)$ is a link function. Different response distributions correspond to different link functions $g(\cdot)$. For example, $g(\mu)$ for binomial distribution is equal to $\log \frac{\mu}{1-\mu}$, while $g(\mu)$ for Poisson distribution is equal to $\log(\mu)$.

Analogous to (5), the unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can be obtained by maximizing the penalized log-likelihood function

$$\begin{aligned} f^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) = & \frac{1}{n} \sum_{i=1}^{n} l\Big\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}\Big\} \\ & - \lambda_1 \sum_{p=1}^{P} |\beta_p| - \lambda_2 \sum_{q=1}^{Q} \delta_q - \frac{1}{2} \lambda_3 \boldsymbol{\alpha}^T \boldsymbol{K}(\boldsymbol{\delta}) \boldsymbol{\alpha}, \end{aligned} \tag{18}$$

where $l(\cdot) = \log(\mathcal{P})$ is the log-likelihood function, $\mathcal{P}$ is the density function given in (16), and $\boldsymbol{K}(\boldsymbol{\delta})$ is a $n \times n$ Gram matrix whose (i, j)-th element is $K^{(g)}(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\delta})$.

The "one-group-at-a-time" cyclical coordinate descent algorithm can also be used to solve $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}$. This process is essentially the same as that described in Section 2.3. Specifically, we denote the log-likelihood as $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} l\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}\}$ and adapt the penalized reweighted least squares approach by performing a two-term Taylor expansion on the log-likelihood. Thus, standard computational procedures for solving LASSO regression estimates, like the R package **glmnet**, could be used to estimate $\boldsymbol{\beta}$. Then, by directly solving a linear system of equations, we can estimate $\boldsymbol{\alpha}$ with the stationary point of a quadratic function similar to (10). Finally, with given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, updating $\boldsymbol{\delta}$ is equivalent to solving a nonlinear optimization problem under the constraints that $\delta_q > 0$, $q = 1, \cdots, Q$. The estimate of $\boldsymbol{\delta}$ could be obtained by using the standard univariate nonlinear constrained optimization software.

6. Conclusions

In this paper, we have proposed a PLGKM method for semiparametric logistic model based on the LASSO method and an innovative class of garrotized kernel functions, suitable for the data measured on a strong, metric scale. In addition, we have also devised an efficient coordinate descent computational algorithm for implementation. To explore the performance of our PLGKM method, we have conducted some numerical experiments with simulation data sets and the breast cancer data (GSE70947).

In the numerical experiments, our PLGKM method consistently outperforms the other methods (LSKM, COSSO, and LASSO) as our PLGKM method always provides small average CEL and average MR. In addition, the advantage is especially significant when there are redundant predictors in the models. For breast cancer data (GSE70947), our PLGKM method obviously outperforms the LASSO method.

Our proposed method can not only capture the complicated relationships between predictors and outcomes, but also allow predictors' interactions. In addition, the variable selection ability of our PLGKM method helps to achieve high prediction accuracies. These advantages of our PLGKM method guaranteed its outperformances in the numerical experiments.

In the future, we will explore the extension of the PLGKM method to other models (such as quantile regression model), and improve the handlability of the PLGKM method for complex structured data (such as categorical data).

Author Contributions: Conceptualization, X.Z. and Y.R.; methodology, X.Z. and Y.R.; software, X.Z. and Y.R.; validation, X.Z. and Y.R.; formal analysis, X.Z. and Y.R.; investigation, X.Z. and Y.R.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, L.L., Y.R., X.Z. and W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Science and Technology Project of Beijing Education Commission (Grant No. KM202110005013) and the National Natural Science Foundation of China (No.11701021), (No.11971001), (No.11801019).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://sbcb.inf.ufrgs.br/cumida (accessed on 3 June 2020).

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

References

1. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
2. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
4. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc.* **2005**, *67*, 301–320. [CrossRef]
5. Algamal, Z.Y.; Lee, M.H. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. *Adv. Data Anal. Classif.* **2019**, *13*, 753–771. [CrossRef]
6. Lee, S.; Kwon, S.; Kim, Y. A modified local quadratic approximation algorithm for penalized optimization problems. *Comput. Stat. Data Anal.* **2016**, *94*, 275–286. [CrossRef]
7. Qian, J.; Payabvash, S.; Kemmling, A.; Lev, M.H.; Schwamm, L.H.; Betensky, R.A. Variable selection and prediction using a nested, matched case-control study: Application to hospital acquired pneumonia in stroke patients. *Biometrics* **2014**, *70*, 153–163. [CrossRef] [PubMed]
8. Liang, Y.; Liu, C.; Luan, X.; Leung, K.S.; Chan, T.; Xu, Z.; Zhang, H. Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinform.* **2013**, *14*, 198. [CrossRef]
9. Huang, H.; Liu, X.; Liang, Y. Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization. *PLoS ONE* **2016**, *11*, e0149675. [CrossRef] [PubMed]
10. Liu, Z.; Sun, F.; McGovern, D.P. Sparse generalized linear model with L_0 approximation for feature selection and prediction with big omics data. *BioData Min.* **2017**, *10*, 1–12. [CrossRef]
11. Hastie, T.; Tibshirani, R. *Generalized Additive Models*; Chapmand and Hall: London, UK, 1990.
12. Meier, L.; Van de Geer, S.; Bühlmann, P. High-dimensional additive modeling. *Ann. Stat.* **2009**, *37*, 3779–3821. [CrossRef]
13. Ravikumar, P.; Lafferty, J.; Liu, H.; Wasserman, L. Sparse additive models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 1009–1030. [CrossRef]
14. Li, G.; Xue, L.; Lian, H. SCAD-penalised generalised additive models with non-polynomial dimensionality. *J. Nonparametr. Stat.* **2012**, *24*, 681–697. [CrossRef]
15. Zhu, J.; Hastie, T. Kernel logistic regression and import vector machine. *J. Comput. Graph. Stat.* **2005**, *14*, 185–205. [CrossRef]
16. Liu, D.; Lin, X.; Ghosh, D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **2007**, *63*, 1079–1088. [CrossRef] [PubMed]
17. Liu, D.; Ghosh, D.; Lin, X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinform.* **2008**, *9*, 292. [CrossRef] [PubMed]
18. Kim, I.; Pang, H.; Zhao, H. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. *Stat. Med.* **2012**, *31*, 1633–1651. [CrossRef] [PubMed]
19. Freytag, S.; Bickeboeller, H.; Amos, C.I.; Kneib, T.; Schlather, M. A novel kernel for correcting size bias in the logistic kernel machine test with an application to rheumatoid arthritis. *Hum. Hered.* **2012**, *74*, 97–108. [CrossRef] [PubMed]
20. Freytag, S.; Manitz, J.; Schlather, M.; Kneib, T.; Amos, C.I.; Risch, A.; Jenny, C.C.; Heinrich, J.; Bickeboeller, H. A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Hum. Hered.* **2013**, *76*, 64–75. [CrossRef]
21. Lin, Y.; Zhang, H.H. Component selection and smoothing in multivariate nonparametric regression. *Ann. Stat.* **2006**, *34*, 2272–2297. [CrossRef]
22. Zhang, H.H.; Lin, Y. Component selection and smoothing for nonparametric regression in exponential families. *Stat. Sin.* **2006**, *16*, 1021–1041.
23. Allen, G.I. Automatic feature selection via weighted kernels and regularization. *J. Comput. Graph. Stat.* **2013**, *22*, 284–299. [CrossRef]
24. Xu, C.; Peng, Z.; Jing, W. Sparse kernel logistic regression based on $L_{1/2}$ regularization. *Sci. China-Inf. Sci.* **2013**, *56*, 71–86. [CrossRef]
25. Maity, A.; Lin, X. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics* **2011**, *67*, 1271–1284. [CrossRef]
26. Rong, Y.; Zhao, S.D.; Zhu, J.; Yuan, W.; Cheng, W.; Li, Y. More accurate semiparametric regression in pharmacogenomics. *Stat. Interface* **2018**, *11*, 573–580. [CrossRef]
27. Cristianini N, S.T.J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, MA, USA, 2000.
28. Schölkopf, B.; Smola, A.J. *Learning with Kernels*; MIT Press: Cambridge, MA, USA, 2002.
29. Kimeldorf, G.; Wahba, G. Some results on tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–95. [CrossRef]

30. Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.* **2007**, *1*, 302–332. [CrossRef]
31. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
32. Zhu, C.; Byrd, R.H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560. [CrossRef]
33. Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *J. Comput. Biol.* **2019**, *26*, 376–386. [CrossRef]
34. Fan, J.; Feng, Y.; Song, R. Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Stat. Assoc.* **2011**, *106*, 544–557. [CrossRef] [PubMed]
35. McCullagh, P.; Nelder, J. *Generalized Linear Models*; Chapmand and Hall: New York, NY, USA, 1989.