STATISTICS AND ITS INTERFACE Volume 11 (2018) 573–580

More accurate semiparametric regression in pharmacogenomics*

YAOHUA RONG, SIHAI DAVE ZHAO, JI ZHU, WEI YUAN, WEIHU CHENG, AND YI LI†

A key step in pharmacogenomic studies is the development of accurate prediction models for drug response based on individuals' genomic information. Recent interest has centered on semiparametric models based on kernel machine regression, which can flexibly model the complex relationships between gene expression and drug response. However, performance suffers if irrelevant covariates are unknowingly included when training the model. We propose a new semiparametric regression procedure, based on a novel penalized garrotized kernel machine (PGKM), which can better adapt to the presence of irrelevant covariates while still allowing for a complex nonlinear model and gene-gene interactions. We study the performance of our approach in simulations and in a pharmacogenomic study of the renal carcinoma drug temsirolimus. Our method predicts plasma concentration of temsirolimus as well as standard kernel machine regression when no irrelevant covariates are included in training, but has much higher prediction accuracy when the truly important covariates are not known in advance. Supplemental materials, including R code used in this manuscript, are available online (http://intlpress.com/site/pub/files/_supp/SII-2018-11-4-s2.zip).

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 62G05, 62H20, 62J07, 62P10; secondary 62G08, 62H12, 62J02, 92B15.

KEYWORDS AND PHRASES: Kernel machine, Semiparametric regression, Model selection.

1. INTRODUCTION

Pharmacogenomics studies the role of genomics in drug response by correlating gene expression with drug absorption, distribution, metabolism and elimination. An important problem is to develop accurate drug response prediction models using individuals' genomic, clinical and demographic information, as well as statistical learning methods for investigating the biological mechanisms underlying the outcome. The investigation that motivated our present work was a study of the anticancer agent temsirolimus (CCI-779), which targets renal cell carcinoma. Our goal is to predict, using an individual's gene expression levels, the expected concentration of temsirolimus in the patient's blood plasma. Plasma concentrations reflect the amount of the drug absorbed by the body, so accurate predictions can allow us to identify the patients for whom temsirolimus would be most efficacious.

Standard methods for predictive modeling usually posit a model for the outcome that is linear in the predictors. However, because the relationship between genes and drug plasma concentration may be very complex, e.g., due to gene-gene interactions, linear models may not suffice. Xue et al. [16] proposed a penalized regression method allowing for some nonlinearity, but required that the nonlinearity take the form of a generalized additive models, which is still restrictive. Allen [1] proposed the fully nonparametric KNIFE method to achieve feature selection using linearized weighted kernel. He et al. [9] extended the KNIFE procedure to a semiparametric setting. Alternatively, Liu et al. [11] proposed a least-squares kernel machine (LSKM) method based on semiparametric support-vector machine regression, which can allow for flexible modeling of the role of gene expression values, in addition to controlling for clinical and demographic covariates. However, these methods become inaccurate if the models contain many irrelevant predictors, so methods are needed to select predictors while still allowing for complicated nonlinear effects in semiparametric model.

There are few solutions that can maintain the flexibility of the LSKM while ameliorating the impact of the irrelevant predictors. Popular variable selection methods like the LASSO [14] and SCAD [6] cannot be applied here because they are designed for parametric, and usually linear, models. One possible approach was recently proposed by Maity and Lin [12], which can test whether a single predictor in a LSKM is associated with the outcome in the presence of the other predictors, while allowing for a nonlinear model. Predictors that are not significantly related can be removed from the model, reducing the dimension. Though sensible,

*We would like to thank the Editor, the Associate Editor and the two referees for their constructive comments and suggestions. Rong's work was partially supported by National Natural Science Foundation of China (No. 11701021), National Statistical Science Research Project (No. 2017LZ35), Fundamental Research Foundation of Beijing University of Technology, Introduction of Talent Research Start-up Foundation and Beijing Outstanding Talent Foundation (No. 2014000020124G047); Zhao's work was partially supported by NSF grant DMS-1613005; Li's work was partially supported by NIH grant U01CA209414.

†Corresponding author.

this approach is akin to backwards selection and is not an efficient way of model-building.

We propose a new kernel machine regression approach using a "garrotized" kernel, which generalizes the idea of Maity and Lin [12]. Our method allows for gene-gene interactions and other complex relationships between the gene expression values and the plasma concentration of temsirolimus.

2. METHODS

2.1 Least-squares kernel machine (LSKM) regression

For subjects $i = 1, \ldots, n$, let Y_i be the plasma concentration of temsirolimus, $X_i = (X_{i1}, \ldots, X_{iP})^T$ be a set of clinical and demographic covariates such as age, and $Z_i = (Z_{i1}, \ldots, Z_{iQ})^T$ be expression levels associated with Q genes. These genes may constitute a gene set, e.g., a genetic pathway/network. We assume the following partial linear semiparametric model to relate the response $\mathbf{Y}$ to the covariates:

$$\mathbf{Y} = \mathbf{X}\beta + h(\mathbf{Z}) + \epsilon, \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)^T$ is the $n \times 1$ vector of response variables, $X = (X_1, \ldots, X_n)^T$ is the $n \times P$ non-genomic covariate matrix, $Z = (Z_1, \ldots, Z_n)^T$ is the $n \times Q$ gene expression matrix, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is an $n \times 1$ random error vector with independent components where $\epsilon_i \sim N(0, \sigma^2)$. The regression parameter vector β quantifies the effect of the non-genomic covariates on the outcome, and $h(\cdot)$ is an unknown and possibly complicated function that describes the relationship between genes and the plasma drug concentration. This flexibility is desirable because the true relationship between genes and the outcome is likely very complex.

It is common to assume $h(\cdot)$ lies in a reproducing kernel Hilbert space $\mathcal{H}_K$ generated by some positive definite kernel function $K(\cdot, \cdot)$. According to the Mercer's theorem [4], under some regularity conditions the kernel function $K(\cdot, \cdot)$ implicitly specifies a unique function space $\mathcal{H}_K$, which is spanned by a particular set of orthogonal basis functions $\phi_j(Z)$, $j = 1, \ldots, J$ with J possibly being infinity. The mathematical properties of $\mathcal{H}_K$ imply that any function $h(\cdot) \in \mathcal{H}_K$ can be represented using a set of basis function as $h(z) = \sum_{j=1}^{J} \phi_j(z)\eta_j$ for some coefficients η_j, which is called the primal or basis representation of the function; or as $h(z) = \sum_{m=1}^{M} K(z_m^*, z; \rho)\alpha_m$, a linear combination of the given kernel function $K(\cdot, \cdot)$ evaluated at points $\{z_1^*, \ldots, z_M^*\} \in R^Q$, for some integer M and some constants α_m, which is called the dual representation.

The space $\mathcal{H}_K$ can be implicitly defined by choosing a kernel function. A commonly used one is the Gaussian kernel: $K(z_1, z_2) = \exp\{-\sum_{q=1}^{Q}(z_{1q} - z_{2q})^2/\rho\}$, where ρ is a tuning parameter. The Gaussian kernel generates the function space spanned by the radial basis functions [3], which contains many nonlinear functions and allows for gene-gene interactions. There are also other choices of kernel functions including the dth polynomial, neural network, sigmoid and smoothing spline kernels [13]. The choice of the kernel function thus determines the particular functional space in which the unknown function $h(\cdot)$ is assumed to lie.

Model (1) is the least-squares kernel machine (LSKM) regression model of Liu et al. [11]. It parametrically specifies the effects of the X_i and nonparametrically specifies the effects of Z_i using a unified kernel machine framework [13]. It is simple to fit and closely related to classical linear mixed models [11].

2.2 Garrotized kernel machines

LSKMs become less accurate when Z_i contain more irrelevant genes. This is true of any nonparametric method as the dimension increases, and in the case of LSKM is illustrated in our simulations in Table 1. Here we propose a new "garrotized" kernel that can automatically eliminate irrelevant genes from the model. Given a base kernel $K(\cdot, \cdot)$, our garrotized version $K^{(g)}$ is defined by

$$\begin{aligned} &K^{(g)}(Z_i, Z_j; \delta) = K(Z_i^*, Z_j^*), \\ &Z_u^* = (\delta_1^{1/2} Z_{u1}, \ldots, \delta_Q^{1/2} Z_{uQ})^T, u = i, j, \\ &\delta_q \geq 0, q = 1, \ldots, Q. \end{aligned} \tag{2}$$

For example, the garrotized version of the Gaussian kernel is $K^{(g)}(Z_i, Z_j; \delta) = \exp\{-\sum_{q=1}^{Q} \delta_q (Z_{iq} - Z_{jq})^2\}$. Our family of garrotized kernels includes the kernel of Maity and Lin [12] as a special case.

The δ are unknown and will be estimated from the data. Each δ_q modulates the effect of gene Z_q on drug response. For example, $\delta_q = 0$ implies that Z_q is not predictive of the response. Thus, our garrotized kernel formulation provides a flexible way to select variables in a semi-parametric setting, and compared to the LSKM may be better adapt to the presence of irrelevant genes in Z. The function $h(\cdot)$ can still be very complicated, for example allowing for gene-gene interactions, depending on the chosen base kernel $K(\cdot, \cdot)$. The δ_q are similar to the regression coefficients in a linear model, except that our model does not need to be linear in Z_q.

To estimate the parameters of model (1) with our garrotized kernel (2), we first standardize the non-genomic covariates and each of the gene expression levels to have zero mean and unit variance and then solve the following minimization problem:

$$\begin{aligned} \arg\min_{\alpha, \beta, \delta} &\frac{1}{2n} \sum_{i=1}^{n} (Y_i - X_i^T \beta - h(Z_i))^2 + \\ &\lambda_1 \sum_{p=1}^{P} |\beta_p| + \lambda_2 \sum_{q=1}^{Q} \delta_q + \frac{1}{2} \lambda_3 \|h\|_{\mathcal{H}_K}^2, \end{aligned} \tag{3}$$

where λ_1 and λ_2 are nonnegative regularization parameters, λ_3 is a tuning parameter which controls the trade-off between goodness of fit and complexity of the model, and

$\|h\|_{\mathcal{H}_K}$ denotes the functional norm in the space $\mathcal{H}_K$ generated by the garrotized kernel. The penalty functions involving β and δ are inspired by the LASSO penalty function, and are appropriate under the assumption that only a small number of the P non-genomic covariates and the Q genes are actually associated with the response. The penalty function involving $h(\cdot)$ is standard in the estimation of kernel machine regression models [11, 13].

The representer theorem of Kimeldorf and Wahba [10] allows us to convert (3) into a more manageable optimization problem. The theorem states that the solution can be written as

$$h(Z) = \sum_{j=1}^{n} \alpha_j K^{(g)}(Z, Z_j; \delta),$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ is an unknown vector and the $K^{(g)}(\cdot,\cdot)$ is our garrotized kernel. Minimization of (3) is thus equivalent to minimizing

$$f(\alpha,\beta,\delta) = \frac{1}{2n}\|Y - X\beta - K(\delta)\alpha\|_2^2 + \lambda_1 \sum_{p=1}^{P} |\beta_p| + \lambda_2 \sum_{q=1}^{Q} \delta_q + \frac{1}{2}\lambda_3 \alpha^T K(\delta)\alpha, \quad (4)$$

where $K(\delta)$ is an $n \times n$ matrix, called the Gram matrix, with ij-th element given by $K_{ij}(\delta) = K(Z_i, Z_j; \delta)$. We refer to the solution of (4) as our penalized garrotized kernel machine (PGKM) estimate.

Indeed, our method stands out from the competing methods, in particular, KNIFE [1] and He et al.'s methods [9] in the following aspects. First, our method deals with partial linear models and our framework is general, encompassing the models considered by KNIFE (fully nonparametric models) as special cases. Second, our proposed PGKM approach is flexible and enables identification of important covariates regardless of whether they are parametrically or nonparametrically modeled. In contrast, neither KNIFE nor He et al.'s method can select both types of variables. Our numerical studies suggest the utility of our proposal in selecting important variables even with moderate sample sizes. Finally, of a technical note, KNIFE and the method of He et al. are based on linear approximations to the kernel, while our PGKM method directly uses the original nonlinear garrotized kernel which is more powerful and robust indicated by simulations in Section 3.

2.3 Algorithm

We propose solving (4) for the unknown parameters α, β, δ by using a "one-group-at-a-time" cyclical coordinate descent algorithm, which is computed along a regularization path.

(1) Set initial estimates $\alpha^{ini}, \beta^{ini}, \delta^{ini}$. For example, take the ordinary least square estimates for β and let $\alpha^{ini} = 0.1, \delta^{ini} = 0.1$.

(2) Update α, β, δ cyclically. Specifically,

- Fix α, δ at values $\tilde{\alpha}, \tilde{\delta}$ and write (4) as

$$f(\tilde{\alpha},\beta,\tilde{\delta}) = \frac{1}{2n}\|Y - X\beta - K(\tilde{\delta})\tilde{\alpha}\|_2^2 + \lambda_1 \sum_{p=1}^{P} |\beta_p| + \lambda_2 \sum_{q=1}^{Q} \tilde{\delta}_q + \frac{1}{2}\lambda_3 \tilde{\alpha}^T K(\tilde{\delta})\tilde{\alpha}.$$

  The β can be estimated using standard procedures for computing LASSO regression estimates [8, 7], giving an update $\tilde{\beta}$.

- Holding the values of β, δ fixed at $\tilde{\beta}, \tilde{\delta}$, our optimization problem (4) can be written as

$$f(\alpha,\tilde{\beta},\tilde{\delta}) = \frac{1}{2n}\|Y - X\tilde{\beta} - K(\tilde{\delta})\alpha\|_2^2 + \lambda_1 \sum_{p=1}^{P} |\tilde{\beta}_p| + \lambda_2 \sum_{q=1}^{Q} \tilde{\delta}_q + \frac{1}{2}\lambda_3 \alpha^T K(\tilde{\delta})\alpha,$$

  which is a quadratic form in α. Differentiating the right side of the above equation with respect to α and letting it equal 0, we find that the update for α is the solution to

$$\left[\frac{1}{n}K^T(\tilde{\delta})K(\tilde{\delta}) + \lambda_3 K(\tilde{\delta})\right]\alpha = \frac{1}{n}K^T(\tilde{\delta})(Y - X\tilde{\beta}),$$

  which is straightforward to obtain. If the left-hand side of the previous equation is a singular matrix, a diagonal matrix with small entries can be added to stabilize the estimate.

- Given the estimates of α, β, updating δ is equivalent to solve a nonlinear optimization problem under the constraints $\delta_q \geq 0, q = 1, \ldots, Q$. The δ_q can be updated one at a time. For δ_t, $t = 1, \ldots, Q$, given the estimates of α, β, (4) can be expressed as

$$f(\tilde{\alpha},\tilde{\beta},\delta) = \frac{1}{2n}\|Y - X\tilde{\beta} - K(\delta)\tilde{\alpha}\|_2^2 + \lambda_1 \sum_{p=1}^{P} |\tilde{\beta}_p| + \lambda_2 \sum_{q\neq t}^{Q} \delta_q + \lambda_2 \delta_t + \frac{1}{2}\lambda_3 \tilde{\alpha}^T K(\delta)\tilde{\alpha},$$

  where the δ_q for $q \neq t$ are held fixed at values $\tilde{\delta}_q(\lambda)$. The update for δ_t can be derived using standard univariate nonlinear constrained optimization software.

(3) Repeat Step (2) until the change in the objective function after any coefficient update is less than a threshold, say 1E-5, or the number of iteration reaches a prespecified number.

Cross-validation is often used for tuning parameter selection but can be computationally inconvenient. Instead, we

divide a given dataset into a training set and a validation set. We use the training set to fit models using various prespecified values for $\lambda = (\lambda_1, \lambda_2, \lambda_3)$. The solutions are computed for a decreasing sequence of values for λ. This scheme not only gives us a path of solutions, but also exploits warm starts and leads to a more stable and faster algorithm. We next calculate the prediction error of each fitted model using the validation set. We used the mean squared prediction error (MSPE), defined as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta} - \hat{h}(Z_i))^2. \tag{5}$$

Finally, we choose the estimated model that gives the lowest MSPE on the validation set. In practice we first perform a coarse search through a large range of λ in order to find a reasonable values before conducting a finer localized search.

3. SIMULATION

3.1 Comparison with LSKM

We first compare our proposed PGKM method to that of the LSKM method of Liu et al. [11]. We generate continuous responses Y_i from

$$Y_i = X_i\beta + h(Z_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{6}$$

We independently generate the P covariates X_{ip} from $U(-1,1)$ and the Q covariates Z_{iq} from $U(0,1)$. The random errors ϵ_i follow $N(0,\sigma^2)$, with σ equal to either 0.1, 0.5, or 1. We allow the nonparametric function $h(\cdot)$ to have a complex form with nonlinear functions of the Z's and interactions among the Z's in order to mimic the complex relationships between gene expression values and the plasma concentration of temsirolimus.

We consider four configurations by varying the sample size n, the number of predictors and the number of irrelevant predictors included when training the model.

Setting 1: $n = 60$, $P = 1$, $Q = 5$, $\beta = 1$, $h(Z) = \cos(Z_1) - 1.5Z_2^2 + \exp(-Z_3)Z_4 - 0.8\sin(Z_5)\cos(Z_3) + 2Z_1Z_5$. Model (6) is fit without any additional irrelevant predictors.

Setting 2: $n = 100$, $P = 2$, $Q = 15$, $\beta = (1,0)^T$, $h(\cdot)$ is the same as in setting 1. Model (6) is fit with 1 additional irrelevant X predictor and 10 additional irrelevant Z predictors.

Setting 3: $n = 200$, $P = 1$, $\beta = 1$, $Q = 10$, $h(Z) = \cos(Z_1) - 1.5Z_2^2 + \exp(-Z_3)Z_4 - 0.8\sin(Z_5)\cos(Z_3) + 2Z_1Z_5 + 0.9Z_6\sin(Z_7) - 0.8\cos(Z_6)Z_7 + 2Z_8\sin(Z_9)\sin(Z_{10}) - 1.5Z_8^3 - Z_8Z_9 - 0.1\exp(Z_{10})\cos(Z_{10})$. Model (6) is fit without any additional irrelevant predictors.

Setting 4: $n = 200$, $P = 2$, $\beta = (1,0)^T$, $Q = 30$, $h(\cdot)$ is the same as setting 3. Model (6) is fit with 1 additional irrelevant X predictor and 20 additional irrelevant Z predictors.

For each simulation setting we generate training, validation, and testing datasets of n observations, each according to model (6). We then use the training and validation sets to fit the model using either LSKM with the Gaussian kernel, or our proposed PGKM with our garrotized Gaussian kernel. We perform 500 replications for each setting.

Table 1. Prediction errors of PGKM and LSKM. The last two columns provide the average MSPEs over 500 replications, with standard deviations in parentheses

	PGKM	LSKM
Setting 1		
$\sigma = 0.1$	0.0345 (0.0160)	0.0379 (0.0121)
$\sigma = 0.5$	0.0778 (0.0280)	0.0797 (0.0399)
$\sigma = 1.0$	0.1617 (0.0903)	0.1639 (0.0739)
Setting 2		
$\sigma = 0.1$	0.0430 (0.0133)	0.0928 (0.0487)
$\sigma = 0.5$	0.0693 (0.0196)	0.1398 (0.0609)
$\sigma = 1.0$	0.1746 (0.0525)	0.2369 (0.0513)
Setting 3		
$\sigma = 0.1$	0.0689 (0.0165)	0.0790 (0.0166)
$\sigma = 0.5$	0.1015 (0.0189)	0.1045 (0.0182)
$\sigma = 1.0$	0.1608 (0.0319)	0.1708 (0.0320)
Setting 4		
$\sigma = 0.1$	0.0748 (0.0146)	0.1622 (0.0264)
$\sigma = 0.5$	0.1204 (0.0251)	0.2174 (0.0348)
$\sigma = 1.0$	0.2403 (0.0448)	0.3319 (0.0587)

Table 1 reports the average mean squared prediction errors of PGKM and LSKM in each setting. In settings 1 and 3, the average MSPE obtained by the PGKM method is very close to that obtained by LSKM method for every configuration of $h(\cdot)$ and σ. In other words, our proposed PGKM method has very similar prediction performance compared to the LSKM method when used without any irrelevant variables. The proposed PGKM method in fact incurs a slightly smaller average MSPEs than LSKM for different underlying nonparametric functions and levels of variation. This may be because our modified garrote kernel is more flexible than the base Gaussian kernel as a result of allowing all δ_q to be unequal.

In contrast, in settings 2 and 4, our proposed PGKM always yields much smaller average MSPEs than LSKM. That is mainly because the proposed PGKM method can recognize the irrelevant variables by estimating the corresponding δ_q and β_p to be small. This dramatically improves prediction accuracy. Again, the MSPEs calculated by PGKM are less variable than those calculated using LSKM. Furthermore, the PGKM prediction errors with irrelevant variables are very close to the LSKM results using only the relevant variables. Thus the proposed PGKM method can perform nearly as well as if we knew the true set of relevant variables.

One byproduct of our proposed PGKM method is that while estimating the parameters of model (6), it can simultaneously select variables while still allowing for a complicated

Table 2. Variable selection results for the PGKM methods, with different numbers of irrelevant Z. The percentage of 500 simulations in which the true model was exactly selected is denoted by C (correct selection), the percentage in which the correct model was nested in the selected model is denoted by O (over-selection), and the percentage in which the true model was not a subset of the selected model is denoted by U (under-selection)

			X			Z		
σ	P	Q	C	O	U	C	O	U
Setting 2								
0.1	2	15	0.9474	0.0526	0.0000	0.1316	0.7522	0.1162
0.5	2	15	0.6842	0.1842	0.1316	0.0000	0.7959	0.2041
1.0	2	15	0.2889	0.3111	0.4000	0.0000	0.7818	0.2182
Setting 4								
0.1	2	30	0.9268	0.0000	0.0732	0.0366	0.3585	0.6049
0.5	2	30	0.6047	0.1047	0.2906	0.0000	0.3140	0.6860
1.0	2	30	0.3608	0.3608	0.2784	0.0000	0.3196	0.6804

nonlinear regression model. When the garrote parameters δ_q is estimated as zero, we can conclude that the corresponding covariate Z_q is not related to the response. In practice, we use 10^{-5} as the threshold to decide whether δ_q is estimated as zero. A similar thresholding principle is used in the SCAD procedure of Fan and Li [6]. At the same time, our algorithm can estimate components of β to be exactly zero, which effects variable selection among the covariates X.

We report the variable selection performance of PGKM on simulation settings 2 and 4 in Table 2. The results show that nearly all relevant X and Z covariates can be selected by PGKM with fairly high probability. This is reflected in the low under-selection rates given in Table 2. PGKM can achieve reasonable variable selection without requiring a linear model, and to our knowledge there are few other methods that can accomplish this.

3.2 Comparison with He's method

We next compare the performance of our PGKM method to that of the method of He et al. [9] based on Gaussian kernel. We generate data as in Section 3.1 and consider the following simulation settings. As before, we preform 500 replications of each setting.

Setting 1: $n = 100$, $P = 1$, $Q = 15$, $\beta = 1$, $h(\cdot)$ as in setting 1 of Section 3.1. Model (6) is fit with 10 additional irrelevant Z predictors.

Setting 2: $n = 100$, $P = 2$, $Q = 15$, $\beta = (1,0)^T$, $h(\cdot)$ as in setting 1 of Section 3.1. Model (6) is fit with 1 additional irrelevant X predictor and 10 additional irrelevant Z predictors.

Table 3 reports the results. The method He et al. [9] assumes that the X covariates contain no irrelevant variables,

Table 3. Prediction errors and running times of PGKM and He's method. The second and fourth columns provide the average MSPEs over 500 replications, with standard deviations in parentheses. The third and fifth columns provide the average running times in seconds

	PGKM		He's Method	
Setting 1				
$\sigma = 0.1$	0.0308 (0.0099)	50	0.3742 (0.2642)	560
$\sigma = 0.5$	0.0849 (0.0333)	79	0.7873 (0.7808)	566
$\sigma = 1.0$	0.1732 (0.0585)	88	1.4956 (1.4561)	398
Setting 2				
$\sigma = 0.1$	0.0430 (0.0133)	25	0.4540 (0.3820)	2764
$\sigma = 0.5$	0.0693 (0.0196)	41	0.7814 (0.6853)	1269
$\sigma = 1.0$	0.1746 (0.0525)	30	1.9578 (1.9320)	525

but setting 1 shows that even when this holds, the average MSPEs of the PGKM method are much smaller and less variable compared to those achieved by He's method. This may be because He et al. [9] use linear approximations to the complex nonlinear kernel, whereas PGKM does not use approximations. Furthermore, when X contains some irrelevant covariates, setting 2 shows that our PGKM again yields much smaller average MSPEs than He's method.

Table 3 also provides the average running times of PGKM and He's methods. PGKM is tuned over $(\lambda_1, \lambda_2, \lambda_3)$ on a grid of triplets while He's method is tuned over λ_2 on a grid of scalars. PGKM is considerably faster.

The variable selection results of PGKM and He's method in setting 2 are reported in Table 4. He's method always underselects much more frequently than PGKM, which may be one reason for its poorer predictive performance in Table 3. Furthermore, unlike He's method, PGKM is capable of variable selection among the X covariates, which it does quite successfully.

3.3 Comparison with KNIFE

In this section, we compare the proposed PGKM method with the KNIFE of Allen [1] based on Gaussian kernel. Because PGKM is designed for semiparametric models while the KNIFE is designed for fully nonparametric models, for a fair comparison we conduct the following simulation, using data generated from setting 2 in Section 3.1.

1. Fit a semiparametric model using the PGKM method.
2. Fit a nonparametric model using the KNIFE method directly.
3. Use the KNIFE method fitted to the residuals of a penalized linear regression of Y on X. We refer to this two-step method as "Linear-KNIFE". We use a LASSO penalty to realize variable selection in the X covariates.

Table 5 shows that our proposed PGKM method always yields much smaller average prediction errors than KNIFE.

Table 4. Variable selection results for PGKM and He's method in setting 2. C, O, U are defined the same as those in Table 2

			PGKM						He's Method		
				X			Z			Z	
σ	P	Q	C	O	U	C	O	U	C	O	U
0.1	2	15	0.9474	0.0526	0.0000	0.1316	0.7522	0.1162	0.0000	0.2609	0.7391
0.5	2	15	0.6842	0.1842	0.1316	0.0000	0.7959	0.2041	0.0000	0.1429	0.8571
1.0	2	15	0.2889	0.3111	0.4000	0.0000	0.7818	0.2182	0.0000	0.0909	0.9091

Table 5. Prediction errors and running times of PGKM, KNIFE, and Linear-KNIFE using data from setting 2 of Section 3.1. The second, fourth, sixth columns provide the average MSPEs over 500 replications, with standard deviations in parentheses. The third, fifth and seventh columns provide the average running times in seconds

	PGKM		KNIFE		Linear-KNIFE	
$\sigma = 0.1$	0.0430 (0.0133)	25	7.8800 (5.1100)	15	2.5696 (1.7865)	16
$\sigma = 0.5$	0.0693 (0.0196)	41	6.9600 (5.1600)	17	2.4310 (1.3180)	18
$\sigma = 1.0$	0.1746 (0.0525)	30	9.0400 (4.9900)	16	2.8036 (1.1905)	17

Table 6. Variable selection results for PGKM, KNIFE and Linear-KNIFE. C, O, U are defined the same as those in table 2

	PGKM			KNIFE			Linear-KNIFE		
	C	O	U	C	O	U	C	O	U
X									
$\sigma = 0.1$	0.9474	0.0526	0.0000	0.0875	0.0125	0.9000	0.3125	0.2875	0.4000
$\sigma = 0.5$	0.6842	0.1842	0.1316	0.0750	0.0250	0.9000	0.2125	0.4000	0.3875
$\sigma = 1.0$	0.2889	0.3111	0.4000	0.0750	0.0250	0.9000	0.3000	0.3125	0.3875
Z									
$\sigma = 0.1$	0.1316	0.7522	0.1162	0.0000	0.0500	0.9500	0.1750	0.3150	0.5000
$\sigma = 0.5$	0.0000	0.7959	0.2041	0.0250	0.0750	0.9250	0.0500	0.5500	0.4000
$\sigma = 1.0$	0.0000	0.7818	0.2182	0.0000	0.0250	0.9750	0.0500	0.4750	0.4750

This may be because the proposed PGKM method correctly specifies a partially linear model, while the fully nonparametric model of KNIFE is much more difficult to estimate. In order to eliminate the influence of model misspecification, we compare the prediction accuracy of PGKM to that of Linear-KNIFE. The average MSPE of PGKM is still much smaller than that of Linear-KNIFE. This may be due to the fact that PGKM directly uses the garrotized Gaussian kernel whereas both KNIFE procedures use linear approximations.

Table 5 also provides the average running times of PGKM and KNIFE methods. The results reveal that the running times of PGKM are comparable with those of KNIFE.

The variable selection results for PGKM, KNIFE and Linear-KNIFE methods are reported in Table 6. The percentage of under-selection of irrelevant X and Z covariates based on KNIFE is much larger than that based on our PGKM, which may be due to model misspecification. The performance of selecting nonparametric Z covariates of PGKM are very similar to those based on Linear-KNIFE. Furthermore, our PGKM does a much better job of selecting the parametric X covariates than Linear-KNIFE.

4. ANALYSIS OF THE PHARMACOKINETICS OF TEMSIROLIMUS

We apply the proposed PGKM method to clinical pharmacokinetics data on temsirolimus (CCI-779) from renal cell carcinoma subjects collected by Boni et al. [2]. The data are publicly online. Temsirolimus is an intravenous anticancer agent and has demonstrated inhibitory effects on tumor growth. Renal cell carcinoma subjects received weekly treatments of temsirolimus until they demonstrated evidence of disease progression. We have expression data on 12,626 genes from 39 subjects measured at baseline, as well as plasma drug concentration across time and each subject was measured 1 to 4 times. A total of 58 observations were made. The concentration measurements were summarized using the area under the curve (AUC), a standard pharmacokinetic measure of the body's exposure to a drug. Our goal is to construct a predictive model for the expected CCI-779 cumulative AUC in terms of an individual's gene expression measurements. An accurate model can allow us to identify

Table 7. Average prediction error of each method for 1000 replications, with standard deviations in parentheses

Methods	MSPE (SD)
PGKM	0.4120 (0.2077)
LSKM	0.5842 (0.3537)
LASSO	2.0343 (1.1360)

patients for whom a given dosage level of temsirolimus would be most effective.

To improve the accuracy of our predictions we first perform dimension reduction using the nonparametric independence screening method proposed by Fan and Song [5], which leaves 14 genes remaining. We then apply our proposed PGKM method. For comparison we also apply the LSKM method and the LASSO for linear regression [14] using the same 14 genes.

In order to compare the prediction errors of these three methods, we randomly selected 40 observations for estimation, 9 observations for searching for the best estimated model of each method and the remaining 9 observations for prediction. We calculated the average MSPE of each method over 1000 replications. Table 7 reports the averaged MSPEs and shows that our PGKM is by far the most predictively accurate. Its superior performance compared to LASSO suggests that the standard linear model is not sufficient to explain the highly nonlinear and complex relationship between the genes and the drug plasma concentration. Its superior performance compared to LSKM demonstrates the benefits of our new garrotized kernel.

5. CONCLUSION

We have proposed a flexible variable selection procedure for semiparametric regression based on a new class of garrotized kernels. It can capture complicated relationships between predictors and outcome and possesses more predictive power than the existing methods in the presence of irrelevant predictors. A key advantage of the proposed PGKM method is that it can achieve variable selection while allowing for a complex nonlinear model. Simulations and our analysis of the plasma concentration of the anticancer drug temsirolimus demonstrate the advantages of our method compared to competing approaches.

In this article we considered only continuous outcomes using a Gaussian base kernel. However, our garrotized kernel machine framework can be extended to estimation and variable selection for a much larger class of models and a much wider range of base kernels. We are pursuing extensions into generalized semiparametric models, for example logistic regression and exponential class models, and other kernels, such as the identity-by-state kernel popular in genome-wide association studies [15]. We are also planning to extend the results to accommodate correlated data. The results will be reported elsewhere.

Finally, we have so far only studied situations where the number of covariates is smaller than the sample size. In principle, our framework can also be used in the high-dimensional setting where there are more covariates than observations. In practice this requires overcoming significant computational hurdles, and we are currently investigating more efficient algorithms for fitting our PGKM estimate.

Received 4 September 2017

REFERENCES

[1] ALLEN, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics* **22** 284–299. MR3173715

[2] BONI, J. P., LEISTER, C., BENDER, G., FITZPATRICK, V., TWINE, N., STOVER, J., DORNER, A., IMMERMANN, F. and BURCZYNSKI, M. E. (2005). Population pharmacokinetics of CCI-779: correlations to safety and pharmacogenomic responses in patients with advanced renal cancer. *Clinical Pharmacology & Therapeutics* **77** 76–89.

[3] BUHMANN, M. D. (2003). *Radial basis functions: theory and implementations* **12**. Cambridge University Press. MR1997878

[4] CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

[5] FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106** 544–557.

[6] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

[7] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1–22.

[8] FRIEDMAN, J., HASTIE, T., HÖFLING, H., TIBSHIRANI, R. et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332. MR2415737

[9] HE, Q., CAI, T., LIU, Y., ZHAO, N., HARMON, Q. E., ALMLI, L. M., BINDER, E. B., ENGEL, S. M., RESSLER, K. J., CONNEELY, K. N., LIN, X. and WU, M. C. (2016). Prioritizing individual genetic variants after kernel machine testing using variable selection. *Genetic Epidemiology* **40** 722–731.

[10] KIMELDORF, G. S. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33** 82–95.

[11] LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088. MR2414585

[12] MAITY, A. and LIN, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using garrote kernel machines. *Biometrics* **67** 1271–1284.

[13] SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with kernels*. MIT Press.

[14] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.

[15] WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.

[16] XUE, L., QU, A. and ZHOU, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association* **105** 1518–1530. MR2796568

Yaohua Rong
College of Applied Sciences
Beijing University of Technology
#100 Pingleyuan
Beijing
China
E-mail address: rongyaohua@bjut.edu.cn

Sihai Dave Zhao
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign
U.S.A.
E-mail address: dave.zhao@gmail.com

Ji Zhu
Department of Statistics
University of Michigan
Ann Arbor
U.S.A.
E-mail address: jizhu@umich.edu

Wei Yuan
School of Statistics
Renmin University of China
Beijing
China
E-mail address: wyuan@ruc.edu.cn

Weihu Cheng
College of Applied Sciences
Beijing University of Technology
#100 Pingleyuan
Beijing
China
E-mail address: chengweihu@bjut.edu.cn

Yi Li
West China Hospital at Chengdu
China
Department of Biostatistics
University of Michigan
Ann Arbor
U.S.A.
E-mail address: yili@med.umich.edu
url: http://www-personal.umich.edu/~yili/resindex.html