Published in final edited form as:

Stat Methodol. 2015 January 1; 22: 47–57. doi:10.1016/j.stamet.2014.08.001.

Incorporating auxiliary information for improved prediction using combination of kernel machines

Xiang Zhan^{a,*} and Debashis Ghosh^{a,b}

^aDepartment of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

^bDepartment of Public Health Sciences, Pennsylvania State University, University Park, PA 16802, U.S.A.

Abstract

With evolving genomic technologies, it is possible to get different measures of the same underlying biological phenomenon using different technologies. The goal of this paper is to build a prediction model for an outcome variable Y from covariates X . Besides X , we have surrogate covariates W which are related to X . We want to utilize the information in W to boost the prediction for Y using X . In this paper, we propose a kernel machine-based method to improve prediction of Y by X by incorporating auxiliary information W . By combining single kernel machines, we also propose a hybrid kernel machine predictor, which can yield a smaller prediction error than its constituents. The prediction error of our kernel machine predictors is evaluated using simulations. We also apply our method to a lung cancer dataset and an Alzheimer's disease dataset.

Keywords

Auxiliary information; Combination of kernels; Hybrid predictor; Kernel ridge regression; Mean squared prediction error

1. Introduction

Biomarkers in cancer research are considered to be central to the prevention, detection and monitoring of disease. With continual development of genomic technologies, one consequence is that different data with biomarkers measured by different technologies are available. As a motivating example, we consider data from a lung cancer study in Chen et al. [8]. One of the main scientific goals in Chen et al. [8] focuses on predicting survival time in patients with lung cancer. Affymetrix gene expression data were obtained on 439 tumor samples. As a follow-up, a subset of 47 samples was measured again using a quantitative

© 2014 Elsevier B.V. All rights reserved.

*Corresponding author. Tel.: +1 8143213493. xyz5074@psu.edu (Xiang Zhan).
ghoshd@psu.edu (Debashis Ghosh)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

real-time polymerase chain reaction (qRT-PCR) platform. While both technologies measure gene expression, the A ymetrix data is regarded as a noisy version the qRT-PCR data since qRT-PCR technology is more generalizable and clinically applicable. The goal is to develop prognostic models for the survival outcome from qRT-PCR data. The question we consider in this paper is how the auxiliary information in the A ymetrix data can be used to improve the prediction of survival time given qRT-PCR data. Let Y denote the survival time, X denote qRT-PCR measurement of gene expression and W be A ymetrix measurement. Depending on whether a tumor sample is measured by qRT-PCR, samples are divided into two parts A and B. Subsample A consists of the 47 tumor samples which are measured by qRT-PCR in a follow-up study. And the remaining tumor samples form subsample B. They are denoted by (Y^A, X^A, W^A) and (Y^B, W^B) respectively. The goal in this paper is to use auxiliary information in W to boost the prediction of $Y|X$.

Boonstra et al. [2] first considered this non-standard prediction problem assuming the following models:

$$Y = \beta_0 + X^T \beta + \epsilon; W = \phi I_p + \nu X + \varepsilon, \quad (1)$$

where Y is a continuous response, X and W are p -dimension biomarker measurements, β are $p \times 1$ vector, I_p is $p \times p$ identity matrix, $\epsilon \sim N(0, \sigma^2)$, $\varepsilon \sim N_p(0, \tau I_p)$, $\beta_0, \phi, \nu, \sigma$ and τ are scalars. They proposed a general class of targeted ridge (TR) estimators which include ridge regression (Hoerl and Kennard [11]) as a special case. Ridge regression estimator shrinks the least squares estimator toward zero. And TR estimators shrink the least squares estimator to some certain targets derived from W^B and Y^B , which is how the auxiliary information in subsample B is used. More details can be found in Boonstra et al. [2]. Generally speaking, TR estimators are biased. It is possible that TR can have a better prediction performance by largely reducing variance to offset the introduced bias (Boonstra et al. [2]).

However, we observe two major disadvantages of TR. First, TR fails when the dimension of X is not equal to that of W . The formulas proposed in section 2 of [2] do not work when X and W are of different dimensions. Second, the prediction performance of TR may not be good when the true underlying functional relationship in Eq. (1) is not linear. To address those two issues, we propose a kernel method based on kernel ridge regression (Cristianini and Shawe-Taylor [9]) to solve the aforementioned prediction problem. One general model consistent with the context is:

$$Y = f(X) + \epsilon; X = h(W) + \varepsilon, \quad (2)$$

where $\epsilon \sim N(0, \sigma^2)$ and $\varepsilon \sim N(0, \tau^2)$. Functions $f(\cdot)$ and $g(\cdot)$ are in some Hilbert functional spaces spanned by kernel functions. More details can be found in Section 2. If one takes W as an error-prone version of X , $X = h(W) + \varepsilon$ can be viewed as a weakly parametric measurement error model (Carroll et al. [6]). Kernel machine regression has been widely used in recent works (See [5,14-16] for more details). It is flexible and allows for complicated relationships between response and predictor, which is desirable in practice. The fact that kernel machine makes few assumptions can give it an advantage in certain scenarios. For example, the TR class of estimators are inefficient when the linearity

assumption is violated. The goal of this paper is to establish a prediction model for new observations X_{new} . The performance of the predictive model is typically measured by the mean squared prediction error (MSPE):

$$MSPE(\hat{f}) = E[Y_{new} - \hat{f}(X_{new})]^2 = \sigma^2 + E[f(X_{new}) - \hat{f}(X_{new})]^2. \quad (3)$$

The rest of the paper is organized as follows. The main result of this paper is presented in Section 2. We first review some useful facts about reproducing kernel Hilbert space (RKHS) and kernel ridge regression. Based on that, we propose a kernel machine predictor for the prediction problem considered in this paper. In the end, a hybrid kernel machine predictor is also proposed based on combination of single kernel machine predictors. In Section 3, we present a simulation study to compare our kernel method with the TR method proposed in [2]. A lung cancer dataset and a GRIN2B gene dataset are analyzed using our combination of kernels method in Section 4. This paper concludes with some discussion in Section 5.

2. Prediction Based on Combination of Kernels

2.1. RKHS and Kernel Ridge Regression

For the purpose of this paper, a kernel machine or a kernel $k(\cdot, \cdot)$ is a symmetric real-valued function of two variables, $k: \Omega \times \Omega \rightarrow \mathbb{R}$, with

$$\int_{\Omega} k(x, y) g(x) g(y) dx dy \geq 0,$$

for all squared integrable function $g(\cdot)$ on Ω , where Ω is the input space (usually a compact subset of $\mathbb{R}^p$). One of the most commonly used kernels is the Gaussian kernel $k(z_1, z_2) = \exp\{-\rho^{-1}\|z_1 - z_2\|^2\}$, where $\rho > 0$ and $\|\cdot\|$ is the L^2 norm. The Gaussian kernel generates a function space spanned by the radial basis functions, and ρ is called the shape parameter (see Buhmann [4] for more detail). Another widely used kernel is the d th polynomial kernel

given by $k(z_1, z_2) = (z_1^T z_2 + c)^d$, where c is nonnegative and d is a positive integer. A special polynomial kernel $k(z_1, z_2) = z_1^T z_2$ is called the linear kernel. More examples of kernel machine include spline kernel, tree kernel and graph kernel (Hofmann et al. [12]). A famous theorem involving kernel machine is the Moore-Aronszajn theorem, which states that, for each symmetric and positive definite kernel $k(\cdot, \cdot)$, there is a unique Hilbert space K of functions for which $k(\cdot, \cdot)$ is a reproducing kernel (Aronszajn [1]). The following two facts characterize RKHS and reproducing kernel.

The first fact deals with the representation of functions in K . From Mercer's Theorem (Cristianini and Shawe-Taylor [9]), under some regularity conditions, a symmetric positive definite kernel function $k(\cdot, \cdot)$ implicitly specifies a unique Hilbert functional space K

spanned by a particular set of orthogonal basis functions $\{\phi_j(z)\}_{j=1}^J$. Then any function $f(z) \in \mathcal{H}$ can be represented as a linear combination of those functions

$f(z) = \sum_{j=1}^J a_j \phi_j(z)$ (primal representation), where a_j' are coefficients. Moreover,

$k(x, y) = \sum_{j=1}^J \lambda_j \phi_j(x) \phi_j(y)$, where λ_j are eigenvalues of kernel k . For multi-dimensional z , explicit basis functions may be complicated to specify and J can be large or even infinite. Hence, sometimes it is more convenient to represent $f(z)$ through the dual representation:

$$f(z) = \sum_{i=1}^L b_i k(z_i, z) \text{ for some coefficients } b_i \text{'s and } \{z_1, \dots, z_L\} \in \Omega.$$

The second fact is the reproducing property. For any $f \in \mathcal{H}$,

$$\langle k(x, \cdot), f(\cdot) \rangle_{\mathcal{H}} = f(x), \quad (4)$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in K . As a special case, $k(\cdot, y) \in \mathcal{H}$; so

$$\langle k(x, \cdot), k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y),$$

i.e., k can reproduce itself under the inner-product. In view of this property, k is called reproducing kernel, and K is called the RKHS.

Given observations $(x_1, y_1), \dots, (x_n, y_n)$, we assume an additive error model $y = f(x) + \epsilon$, where $f(\cdot)$ is in RKHS K spanned by kernel $k(\cdot, \cdot)$. One way to estimate $f(\cdot)$ is to minimize penalized least squares functions as follows:

$$\mathcal{L}(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (5)$$

where λ is a regularization parameter which controls the trade-off between goodness of fit and model complexity, $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product in K . Note that for a goal of prediction, the second penalized term is necessary in order to avoid over-fitting.

Using the dual representation, we have that $f(\cdot) = \sum_{i=1}^n b_i k(\cdot, x_i)$. Define the kernel map $\Phi(x_i) \equiv k(\cdot, x_i) \in \mathcal{H}$. Using the two basic facts, Eq. (5) can be written as

$$\mathcal{L}(b) = \sum_{i=1}^n \left(y_i - \sum_{j=1}^n b_j \Phi(x_j) \right)^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n b_i b_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}},$$

which can be viewed as a ridge regression on the kernel map vector $\Phi(x_i)$'s if they are normalized in the sense that $\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} = \delta_{ij}$, where δ_{ij} is the Kronecker delta.

Taking derivatives with respect to b_j 's, some calculations give the solution of kernel ridge regression:

$$\hat{\mathbf{b}} = (K + \lambda I)^{-1} \mathbf{y}, \hat{f}(x) = \mathbf{k}(x)^T (K + \lambda I)^{-1} \mathbf{y}, \quad (6)$$

where $\mathbf{b} = (b_1, \dots, b_n)^T$, K is the kernel matrix with $K_{ij} = k(x_i, x_j)$, I is the n th order identity matrix, $\mathbf{y} \equiv (y_1, \dots, y_n)^T$ and $\mathbf{k}(x) = (k(x, x_1), \dots, k(x, x_n))^T$.

2.2 Combination of Kernels

In our kernel-based framework, the auxiliary information contained in subsample B is used in the following way. First, a kernel ridge regression is fit to the data (X^A, W^A) to estimate h in (2). Second, we want impute X_B using the estimated h , that is $\hat{X}^B = \hat{h}(W^B)$. Last, we estimate f in (2) using kernel ridge regression with augmented data, $Y = \{Y^A, Y^B\}$ and $X = \{X^A, \hat{X}^B\}$. This whole procedure is accomplished using a combination of kernel machines. The “combination” here stems from the special hierarchical structure $W \rightarrow X \rightarrow Y$ (here $A \rightarrow B$ denotes using A to predict B, the same notation applies in what follows). In the above procedure, we need to use different kernel machines to model both $W \rightarrow X$ and $X \times Y$ to get a final predictor $\hat{f}$. Each predictor $\hat{f}$ corresponds to a combination of kernel machines. Moreover, as we will see later in this paper, the word “combination” has a second meaning. different predictors $\hat{f}$'s are combined to form a hybrid predictor.

Returning the prediction problem described in the Introduction, consider a more general gene-specific kernel machine measurement error model $X_{ij} = h_j(W_i) + \varepsilon$, where $i = 1, \dots, n_A$ and $j = 1, \dots, p$. First, based on formula (6), h_j can be estimated by using information in sample A as $\hat{h}_j(w) = \mathbf{k}(w)^T (K + \lambda_j^A I)^{-1} x_j^A$, where $j = 1, \dots, p$, $\mathbf{k}(w) = (k(w, w_1^A), \dots, k(w, w_{n_A}^A))^T$, I is the $n_A \times n_A$ identity matrix, K is $n_A \times n_A$ matrix with $K_{ij} = k(w_i^A, w_j^A)$ and x_j^A is a $n_A \times 1$ vector. Second, based on this estimate, X^B can be imputed as $\hat{x}_{ij}^B = \hat{h}_j(w_i^B)$, $i = 1, \dots, n_B$ and $j = 1, \dots, p$. Last, using the augmented data, the estimate of $\hat{f}$ is of the form:

$$\hat{f}(x) = \sum_{i=1}^{n_A} \hat{\alpha}_i K(x, x_i^A) + \sum_{j=1}^{n_B} \hat{\beta}_j K(x, \hat{x}_j^B), \quad (7)$$

which gives a means for predicting an individual with a new value of x . The influence of subsample B in the prediction is twofold based on Eq. (7). First, subsample B is included in the second term in Eq. (7). Second, by incorporating subsample B, it changes the estimate of $\hat{\alpha}_i$'s in (7).

Finally, note that prediction based on Eq. (7) or Eq. (6) relies on tuning parameter λ and possibly kernel parameter ρ . Moreover residual variance σ^2 is also needed to be estimated for further inference. Liu et al. [15] showed the equivalence between least squares kernel machine models and linear mixed effects models. By this equivalence, (τ, ρ, σ^2) are treated as variance component parameters in the linear mixed effects model that are estimated by solving the score equations (see Liu et al. [15] for details).

The problem considered here is different from the one considered in multi-task learning. A typical data structure considered in a multi-task learning problem is like $W \rightarrow X, W \rightarrow Y$ (Caruana [7] page 44, Evgeniou et al. [10] page 619). Because two outputs/tasks (X and Y) share a common input layer W , it is possible for internal representations that arise in the

input layer for one task to be used by the other task. Multi-task learning benefits from sharing what is learned by different tasks while tasks are trained in parallel (Caruana [7], Evgeniou et al [10]). However the data structure considered in our combination of kernel machines is $W \rightarrow X \rightarrow Y$. We do not use W to predict Y because W is from old technology and is regarded as a noisy version of X .

2.3. Hybrid Predictors

A particular kernel machine predictor may perform well in a certain scenario but is not likely to give a small prediction error under all settings. One way to fix such a problem is by using a hybrid predictor, which is an adaptive combination of multiple kernel machine predictors. Given m kernel machine predictors $\hat{f}_1, \dots, \hat{f}_m$, we consider the combination of the form

$$\hat{f}(x) = \sum_{k=1}^m \alpha_k \hat{f}_k(x),$$

subject to $\alpha_k \geq 0$ and $\sum_{k=1}^m \alpha_k = 1$. The mean squared prediction error for the hybrid predictor is:

$$MSPE(\hat{f}) = \sigma^2 + E[f(x_{new}) - \hat{f}(x_{new})]^2 = \sigma^2 + E\left[f(x_{new}) - \sum_{k=1}^m \alpha_k \hat{f}_k(x_{new})\right]^2.$$

A hybrid predictor $\hat{f}$ has better performance than its constituents $\hat{f}_1, \dots, \hat{f}_m$ in terms of MSPE. This is because the single kernel machine predictor $\hat{f}_k$ is obtained from $\hat{f}$ by setting $\alpha_k = 1$ and $\alpha_i = 0, i \neq k$. And the biggest gain in prediction performance from a single predictor to a hybrid predictor is obtained when dissimilar sets of predictors were combined (See Breiman [3] for more detail).

The weight coefficients α_k 's are estimated by minimization of the mean squared prediction error $MSPE = E\left[y_{new} - \sum_{k=1}^m \alpha_k \hat{f}_k(x_{new})\right]^2$. And we use a cross-validation approach to estimate the weight coefficients α_k 's. In practice, since y_{new} is unknown, we use

$$\frac{1}{n_A} \sum_{i=1}^{n_A} \left[y_i^A - \sum_{k=1}^m \alpha_k \hat{f}_k^{(-i)}(x_i^A) \right]^2$$

as an estimate of MSPE, where $\hat{f}_k^{(-i)}$ is the k th single estimator trained without the i th observation in dataset A. Estimations of α 's are obtained by minimizing this formula. Note that the numerical performance of this hybrid predictor depends on how accurately α 's are estimated. If the stochastic error in estimating α 's is large enough, then it is possible that a single predictor has a smaller prediction error than a hybrid predictor.

3. Simulation

In this section, we conducted simulations to evaluate the prediction performance of our proposed methods and to compare them to those in Boonstra et al. [2], in which four targeted ridge shrinkage methods (RIDG, SRC, FRC, HYB) were introduced. A targeted ridge estimator shrinks the least squares estimator to some certain targets derived from W^B and Y^B , which is how the auxiliary information in subsample B is used. RIDG refers to applying ridge regression only for dataset A, in which the auxiliary information contained in dataset B is completely ignored. SRC (structural regression calibration) and FRC (functional regression calibration) are different in how the target is constructed. HYB, a hybrid predictor, is an adaptive linear combination of RIDG, SRC and FRC. A common assumption in those methods is that both the relationship between Y and X and the relationship between W and Z are linear (see Eq. (1) and Eq. (2)).

In our combination of kernels approach, more general and complicated relationships are allowed, which can give us an edge over the target ridge shrinkage methods while maintaining computational feasibility. In this simulation, both nonlinear and linear relationships of f in Eq. (2) were considered. For the linear biomarker effect, the true model was $f(z) = 5z_1 - 4z_2 - 3z_3 + 2z_4 + z_5$. For the nonlinear biomarker effect, the true model was $f(z) = 2(z_1 - z_2)^2 + z_2z_3 + 3\sin(2z_3)z_4 + z_5^2 + 2\cos(z_4)z_5$. As for h in Eq. (2), like Boonstra et al. [2], W was taken as a noisy version of X and was of the same dimension. The true relationship between W and X was:

$$W = X + \tau \xi, \xi \sim N_p(0, I_p). \quad (8)$$

For the purposes of comparison, we had a similar simulation setup to that in [2]. We fixed the sample size $n_A = 50$ and $n_B = 300$, the dimension of the covariates $p = 5$. First, X_A , X_B and 100 X_{new} 's were generated from a p -dimensional normal distribution $N_p(0, X)$ with $(X)_{ij} = 0.75^{|i-j|}$, where $i, j = 1, \dots, p$. The variance of error term σ^2 , was either 1 or 4. Second, Y_A/X_A , Y_B/X_B and Y_{new}/X_{new} were drawn under each combination of function f and σ^2 . There were 4 simulation settings in total: two choices each for σ^2 and function f . We repeated each setting at a bunch of τ 's in $(0, 2]$. For each τ , W_A/X_A and W_B/X_B were drawn from Eq. (8).

Throughout this simulation study, two kernel machines were used. One was the Gaussian kernel $G(k(z_1, z_2) = \exp(-\rho^{-1}\|z_1 - z_2\|^2))$ and the other was linear kernel

$L(k(z_1, z_2) = z_1^T z_2)$. The quadratic kernel machine $k(z_1, z_2) = (z_1^T z_2 + 1)^2$ was also used. However it did not improve the prediction, so the results of quadratic kernel were not reported. Seven kernel machine models (G, L, GG, GL, LG, LL, HYBKM) were considered in this paper. Methods G and L were similar to the RIDG method in [2], in the sense that a single-kernel regression was fitted only using dataset A to model (Y, X) . Method GG was the model applying two Gaussian kernels to model (Y, X) and (X, W) respectively. The other three methods, GL, LG and LL, were explained in the same way. The last method HYBKM was a hybrid predictor consisting of the first six kernel predictors. The weight coefficients were determined in the way described in section 2.3. For all methods, the MSPE was estimated by averaging the squared prediction error over 100 new individuals. The

numerical values were reported in the Table 1 and Table 2. Based on these two tables, the following conclusions can be drawn.

Effect of kernel choice

Based on the results in Table 1, the MSPE of GG is close to that of GL while LG and LL have much larger prediction errors. This indicates that a Gaussian kernel is almost capable of capturing the linear effect while a linear kernel fails to capture the non-linear effect. This conclusion is supported by the numerical results in Table 2 as well. In both Table 1 and Table 2, our hybrid predictor HYBKM is competitive with the best-performing of its component methods. In most simulation settings, HYBKM provides the best prediction.

Effect of the variance component σ^2

The effect of σ^2 on MSPE is twofold. First, σ^2 is a component of $MSPE$ since $MSPE = \sigma^2 + MSE$. Second, a large σ^2 indicates that larger uncertainty of the relationship between Y and X , which is associated with larger mean squared error of estimation. Therefore it is easy to see that, the larger the σ^2 , the larger the prediction error.

Effect of the variance component τ

Methods G, L and RIDG are not affected by τ because they are modeling the relationship between X and Y directly without using any information in W . For the other methods, the smaller the τ , the smaller the prediction error. Note that τ is the error variance in Eq. (8), smaller τ values mean stronger association between X and W . Thus it will greatly boost the prediction when imputing X using W . Moreover, when τ is big enough, the prediction performance of the two-kernel methods, like GG and GL, is worse than that of single-kernel predictors (G and L). This is because in the case of large τ , the stochastic error accumulated from modeling X from W will deteriorate the prediction of Y using X . Finally, as also seen in Boonstra et al. [2], in Table 2, the MSPE of SRC and FRC rises sharply with τ . However such a trend is not observed in the kernel machine-based methods, and numerical results of kernel methods are quite stable with respect to changes in τ . It can be observed that larger values of τ favor kernel machine predictors relative to TR predictors.

In conclusion, Table 1 shows that, under the assumption of nonlinear biomarker Effect, our kernel-based methods outperform those targeted ridge methods in different simulation scenarios. This is not surprising since those targeted ridge methods are based on the linearity assumption. When the true underlying model is far from linear, those methods quickly lose power in terms of prediction. Table 2 shows that when the true relationship between biomarker and response is linear, our kernel-based methods are competitive with those targeted ridge methods. Both the Gaussian kernel and the linear kernel are able to capture such a linear relationship. Moreover, the results of kernel machine based methods are more stable in Table 2. To summarize, the kernel machine methods could be used for modeling the relationship between outcome and biomarkers. This feature is especially desirable for biomarker data because the relationship between clinical outcome and biomarkers is usually complicated and unknown.

4. Examples

4.1. Lung cancer data

Our dataset comes from Chen et al. [8] and consists of 439 lung cancer samples. It contains measurements on $p=91$ highly correlated genes representing a broad spectrum of biological functions. Their expressions were measured on the log-scale using Affymetrix microarrays (W). A subset of these samples (47) were measured again using qRT-PCR technology (X). Clinical covariates, age, gender and stage of cancer (I-III) were also available. An independent set of 101 tumor samples with qRT-PCR measurements and clinical covariates was used for validation. All 439 samples were divided into two parts, termed dataset A, dataset B. Dataset A consisted of 47 tumor samples with both Affymetrix and qRT-PCR measurements. Dataset B contained 392 samples with only Affymetrix measurements. Moreover, a validation set with 101 samples with qRT-PCR measurements was also available.

In dataset A, 11 measurements in 9 samples were missing. In order to use all the 47 samples, these missing values were imputed using chained equations (Van Buuren [18]) and thereafter assumed known. There were no missing values in dataset B and the validation set. In dataset B, three tumors had event time less than 1 month after surgery. These three tumor samples were removed before analysis; so $n_B = 389$. There was 1 tumor in the validation set which had event time less than 1 month after surgery and was removed as well. The final validation set contained 100 observations.

Because our methodology was developed for continuous outcomes, some pre-processing of the censored data was needed. Here we adopted the same adjustment as in Boonstra et al. [2] (see section 6 in Boonstra et al. [2] for more details). Boonstra et al. [2] also discussed two potential model misspecification issues of model (1). The first one was that gene-specific measurement-error models were desired, which was shown by the fact that 91 LOESS curves for each genes having different patterns. The other issue was the violation of the assumption of constant τ . Some related adjustments were made in [2]. However, these two issues are not problems in our kernel machine setting. For each gene X_j , $j = 1, \dots, 91$, we only used the same kind of kernel (either G or L) to model the relationship between X_j and W as $X_j = h_j(W) + \varepsilon_j$, $\varepsilon_j \sim N(0, \tau_j)$. Note that both h_j 's and τ_j 's are estimated adaptively and can be different from each other. Hence kernel machines estimators are robust to those two potential violations.

In Table 3, we present results of predicting survival time in the validation data using the Gaussian kernel (G) and the linear kernel (L). The same seven kernel machine methods were considered as in the preceding section. The names of these methods are explained in the same way as we did before. We used the 2nd order polynomial kernel as well; it did not improve the results and hence are not reported. The first row in Table 3 presents MSPE results for each method. Following Boonstra et al. [2], we applied their bootstrap algorithm using 100 bootstrap samples to assess the uncertainty in our predictions. The second row in Table 3, is the proportion of the 100 validation samples which are contained in the 95% prediction interval. The third row is the average length of the 95% prediction interval for each validation sample. For the hybrid kernel machine predictor, the weight coefficients are

(0.02, 0, 0.02, 0.02, 0.70, 0.24) corresponding to G, L, GG, GL, LG and LL respectively. In terms of MSPE, LG kernel performs the best followed by the hybrid kernel predictor. The 95% prediction interval of each kernel predictor has an actual coverage rate of 0.95 or 0.94, which is almost equal to the nominal level. Moreover, the length of 95% prediction interval of LG, LL and HYBKM are shorter than that of all TR methods in [2]. The length of the 95% prediction interval of G, L, GG and GL are short than most TR predictors except RIDG. Therefore, a kernel machine method fits a better prediction model than a TR method for this particular dataset.

4.2. Alzheimer's disease data

In an earlier study, Stein et al. [17] found that one SNP located in the GRIN2B gene was related to Alzheimer's disease. Motivated by their study, we extracted SNPs based on the physical position of gene GRIN2B from National Center for Biotechnology Information (NCBI). We extracted the SNPs from 13714kb to 14133kb and found 119 SNPs. A more detailed description of this dataset can be found in Zhan et al. [20]. Among those 119 SNPs, Zhan et al. [20] found that a 11 SNPs were associated with the disease outcome using a liner kernel (see Zhan et al. [20] for more details). To simplify our analysis, we used these 11 SNPs as our predictors.

The GRIN2B gene dataset contains 741 samples, of which 33 samples having missing values in 9 of those 11 SNPs variables. Because these 11 SNPs are located in the genome near to each other, these 9 SNPs with missing values may still contain auxiliary information which are helpful in predicting the disease outcome using the other 2 SNPs. Hence we want to tailor this GRIN2B gene dataset to the combination of kernels framework. Recalling the data structure (Y^A, X^A, W^A) and (Y^B, W^B) considered in this paper, subsample B contained missing values in X variables. Hence, we took those 33 samples as subsample B. For the remaining 708 samples, we divided it almost evenly into a subsample A ($n_A = 350$) and a validation sample ($n = 358$). Moreover, we took these 9 SNP variables as X , and the rest 2 SNPs as W . By doing so, the combination of kernels method provided an alternative way to deal with missing values in SNP variables rather than the mean imputation used in Zhan et al. [20]. Unlike mean imputation, this combination of kernels way of dealing with missing values took advantage of the correlation between SNPs. Note that the dimensions of X and W are not equal now. Hence the TR method proposed in Boonstra et al. [2] no longer applies. We still use the ridge regression method as a benchmark to compare with our combination of kernel methods. The results are reported in Table 4:

The first row of Table 4 presents the MSPE of the various methods. For the hybrid kernel machine predictor, the weight coefficients are (0, 0, 0, 1, 0, 0) corresponding to G, L, GG, GL, LG and LL respectively. Our combination of kernels method has a smaller prediction error than the RIDG method, which implies that the underlying relationship between Y and X is probably is nonlinear. The comparison between a Gaussian kernel and a linear kernel also implies this conclusion. Another finding is that using an addition kernel to model (X, W) does not help much in the prediction of $Y|X$ in this particular dataset. This may be because the size of subsample B (33) is relatively small compared with subsample A (350). It does reduce the length of the 95% confidence interval. Note that the third row of Table 4 is now

the median of the length of confidence intervals, rather than the average as reported in Table 3. This is because the confidence interval length of the RIDG method is extremely large on some new samples, which makes the average as high as 1.78×10^{12} . This might be due to issues like collinearity. As for other kernel-based methods, the medians are close to the means; hence they are more robust. The actual coverage rates of the 95% confidence interval are all good and close to the nominal level. To sum it up, our combination of kernel machine methods provides a stable and generalizable approach dealing with the kind of prediction problems described in this paper.

5. Discussion

In this paper, we have described how to use data from other sources to help build and stabilize prediction models using kernel machine-based methods. These methods do not require an explicit analytical specification on the nonparametric function and make minimal assumptions. It provides a general and flexible approach of modeling the genomic Effect, which allows for the possibility that each individual gene Effect may be nonlinear or interacts with each other in a complicated way. Moreover, by applying combination of kernels to model relationships between different variables, we have seen this to provide improvement in prediction in our simulation studies.

While many kernel machine methods have been used in a testing context, this paper primarily dealt with a prediction problem. The prediction problem is special in the sense that besides the predictor variables X , another set of surrogate variables W are available. If variables (X, W) can be divided into many groups and the relationship between the outcome and each group of variables share a common sparsity pattern, then group-lasso may be a good candidate to incorporate auxiliary information (Kim and Xing [13], Vogt and Roth [17]). However, we do not know whether those assumptions hold in the special prediction problem considered in this paper. Hence we do not pursue the group-lasso method in this paper.

Two popular kernel machines, the Gaussian kernel and the linear kernel were used in this paper. The Gaussian kernel is able to capture a nonlinear Effect while the linear kernel performs slightly better when a linear relationship is presented. We have also proposed a hybrid kernel machine predictor, which is a linear combination of multiple kernel machine predictors. In the simulation studies, we showed that a hybrid kernel machine predictor was flexible and always had relatively small prediction errors. We compared our method with the TR method of Boonstra et al. [2]. The gain in prediction performance across all designs under a nonlinear configuration is encouraging. Even under the setting assumed by the TR class of predictors, the performance of our kernel machine predictors are comparable with those of TR predictors. The R code for implementing the proposed kernel machine method will be available upon request.

In the lung cancer data analysis, for a purpose of comparison, we followed Boonstra et al. [2] and transformed the outcome variable from censored survival time to continuous regression errors. Thus it is valid to apply this kernel machine framework which was developed assuming that the response variable was continuous. Some authors extended the

kernel machine approach to the setting of a binary response (Liu et al. [16]) and a censored response (Cai et al. [5]). How to utilize their framework to help build prediction models is currently under investigation.

Acknowledgments

The research of the authors is supported by NIH R01 CA 129102. They would like to thank both Phil Boonstra for providing the lung cancer data and Wen-Yu Hua for providing the GRIN2B gene data.

References

- [1]. Aronszajn N. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* 1950; 68:337–404.
- [2]. Boonstra PS, Taylor JMG, Mukherjee B. Incorporating auxiliary information for improved prediction in high dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics*. 2013; 14:259–272. [PubMed: 23087411]
- [3]. Breiman L. Stacked regression. *Machine Learning*. 1996; 24:49–64.
- [4]. Buhmann, MD. *Radial Basis Functions*. Cambridge, U.K.: 2003.
- [5]. Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011; 67:975–986. [PubMed: 21281275]
- [6]. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC Press; Boca Raton, FL: 2006. *Measurement Error in Nonlinear Models*.
- [7]. Caruana R. Multitask learning: a knowledge-based source of inductive bias. *Machine Learning*. 1997; 28:41–75.
- [8]. Chen G, Kim S, Taylor JMG, Wang Z, Lee O, Ramnath N, Reddy RM, Lin J, Chang AC, Orringer MB, et al. Development and validation of a qRT-PCR classifier for lung cancer prognosis. *Journal of Thoracic Oncology*. 2011; 6:1481–1487. [PubMed: 21792073]
- [9]. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*. Cambridge, U.K.: 2000.
- [10]. Evgeniou T, Theodoros CA, Pontil M, Shawe-Taylor J. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*. 2005; 6:615–637.
- [11]. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970; 12:55–67.
- [12]. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *The Annals of Statistics*. 2008; 36:1171–1220.
- [13]. Kim, S.; Xing, EP. Tree-guided group lasso for multi-task regression with structured sparsity; *Proceedings of the 27th International Conference on Machine Learning*; 2010; p. 543–550.
- [14]. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.* 2008; 82:386–397. [PubMed: 18252219]
- [15]. Liu D, Lin X, Ghosh D. Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machine and linear mixed models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
- [16]. 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 Bioinformatics*. 2008; 9:292. [PubMed: 18577223]
- [17]. Stein JL, Hua X, Morra JH, et al. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. *NeuroImage*. 2010; 51:542–554.
- [18]. Van Buuren, S. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Press; Boca Raton, FL: 2012.
- [19]. Vogt, J.; Roth, V. A Complete Analysis of the $l_{1,p}$ Group-Lasso. 2012. arXiv preprint arXiv:1206.4632

- [20]. Zhan X, Epstein M, Ghosh D. An adaptive genetic association test using double kernel machines. *Statistics in Biosciences*. 2014 DOI: 10.1007/s12561-014-9116-2.

Table 1

Empirical mean squared prediction error (MSPE) of the simulation study for nonlinear biomarker effect.

σ^2	Method	$\pi=0.01$	0.25	0.5	0.75	1	1.25	1.5	1.75	2
1	G	7.18	7.18	7.18	7.18	7.18	7.18	7.18	7.18	7.18
	L	23.48	23.48	23.48	23.48	23.48	23.48	23.48	23.48	23.48
	GG	5.54	6.20	6.31	6.55	6.93	7.64	7.92	7.67	7.64
	GL	5.55	5.99	6.38	6.64	7.20	7.34	7.57	7.56	7.30
	LG	20.74	21.83	21.54	22.12	31.92	29.18	23.85	33.15	26.42
	LL	20.80	20.75	20.01	20.84	22.78	21.42	21.19	22.95	21.13
	HYBKM	5.55	5.99	6.38	6.63	6.93	7.64	7.21	7.18	7.18
	RIDG	26.17	26.17	26.17	26.17	26.17	26.17	26.17	26.17	26.17
	SRC	20.79	20.53	20.02	20.04	20.12	20.13	20.13	20.17	20.06
	FRC	20.80	20.70	20.08	20.12	20.23	20.20	20.24	20.42	19.78
4	HYB	22.94	22.81	22.73	22.78	22.81	22.78	22.83	22.77	22.75
	G	27.31	27.31	27.31	27.31	27.31	27.31	27.31	27.31	27.31
	L	44.13	44.13	44.13	44.13	44.13	44.13	44.13	44.13	44.13
	GG	24.85	25.55	25.50	25.89	26.34	27.21	27.06	27.17	26.61
	GL	24.92	25.32	25.63	25.96	26.73	26.92	27.12	26.84	26.24
	LG	41.14	42.41	42.05	42.70	55.18	50.35	44.94	56.34	47.00
	LL	41.22	41.11	40.20	41.14	44.46	41.80	41.83	44.19	40.95
	HYBKM	24.91	25.32	25.63	26.05	26.34	27.21	27.09	27.24	27.31
	RIDG	47.69	47.69	47.69	47.69	47.69	47.69	47.69	47.69	47.69
	SRC	41.21	40.82	39.87	39.79	39.91	39.85	39.86	39.90	39.73
	FRC	41.21	41.12	40.12	40.07	40.38	40.01	40.12	40.26	39.31
	HYB	43.84	43.69	43.53	43.62	43.73	43.61	43.65	43.57	43.54

Table 2

Empirical mean squared prediction error (MSPE) of the simulation study for linear biomarker effect.

σ^2	Method	$\tau=0.01$	0.25	0.5	0.75	1	1.25	1.5	1.75	2
1	G	2.52	2.52	2.52	2.52	2.52	2.52	2.52	2.52	2.52
	L	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10
	GG	2.09	3.01	2.37	3.33	3.78	3.63	4.08	5.13	4.83
	GL	2.08	2.68	2.38	2.99	3.04	2.95	4.19	5.43	4.69
	LG	2.05	3.02	3.21	2.51	2.19	2.82	2.20	2.55	2.37
	LL	2.04	2.78	2.79	2.87	2.30	2.21	2.34	3.39	2.69
	HYBKM	2.04	2.06	2.04	2.08	2.06	2.09	2.13	2.10	2.04
	RIDG	2.29	2.29	2.29	2.29	2.29	2.29	2.29	2.29	2.29
	SRC	2.04	2.80	8.15	12.62	14.48	15.32	15.86	16.07	16.27
	FRC	2.04	2.18	4.03	7.25	9.14	10.44	12.11	13.20	14.06
	HYB	2.04	2.11	2.10	2.12	2.12	2.11	2.11	2.10	2.10
4	G	22.13	22.13	22.13	22.13	22.13	22.13	22.13	22.13	22.13
	L	20.87	20.87	20.87	20.87	20.87	20.87	20.87	20.87	20.87
	GG	20.44	20.62	20.15	21.38	22.38	21.85	22.38	24.04	23.37
	GL	20.44	20.43	20.29	20.93	21.60	21.11	22.58	24.34	22.98
	LG	20.43	21.00	21.02	20.17	20.39	20.73	20.19	21.03	20.34
	LL	20.43	20.77	20.77	20.52	21.20	20.25	20.48	22.50	20.65
	HYBKM	20.42	20.31	20.31	20.63	20.62	20.42	20.69	20.86	20.46
	RIDG	21.26	21.26	21.26	21.26	21.26	21.26	21.26	21.26	21.26
	SRC	20.43	20.74	25.20	29.47	31.18	31.98	32.50	32.70	32.88
	FRC	20.43	20.44	21.50	24.54	26.24	27.43	28.98	30.07	30.89
	HYB	20.43	20.44	20.67	20.80	20.81	20.79	20.79	20.78	20.78

Table 3

Results for predicting survival time from gene expression measurements.

	G	L	GG	GL	LG	LL	HYBKM
MSPE	0.732	0.715	0.732	0.732	0.661	0.665	0.662
Avg.coverage	0.95	0.95	0.95	0.95	0.94	0.94	0.94
Avg.length	3.4105	3.4057	3.4098	3.4098	3.1791	3.2731	3.2011

Table 4

Results for GRIN2B gene data.

	RIDG	G	L	GG	GL	LG	LL	HYBKM
MSPE	1.019	0.991	0.998	0.991	0.991	1.001	1.001	0.991
Avg.coverage	0.92	0.94	0.94	0.93	0.93	0.92	0.92	0.93
Med.length	3.46	3.66	3.63	3.47	3.47	3.45	3.45	3.49