Published in final edited form as:

Stat Med. 2011 July 20; 30(16): 2015–2027. doi:10.1002/sim.4233.

Parametric Variable Selection in Generalized Partially Linear Models with an Application to Assess Condom Use by HIV-infected Patients

Chenlei Leng,

Department of Statistics and Applied Probability, National University of Singapore, Singapore

Hua Liang, and

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

Neil Martinson

Perinatal HIV Research Unit, University of the Witwatersrand, Johannesburg, South Africa, and Johns Hopkins University, School of Medicine, Baltimore, MD 21231

Abstract

To study significant predictors of condom use in HIV-infected adults, we propose the use of generalized partially linear models and develop a variable selection procedure incorporating a least squares approximation. Local polynomial regression and spline smoothing techniques are used to estimate the baseline nonparametric function. The asymptotic normality of the resulting estimate is established. We further demonstrate that, with the proper choice of the penalty functions and the regularization parameter, the resulting estimate performs as well as an oracle procedure. Finite sample performance of the proposed inference procedure is assessed by Monte Carlo simulation studies. An application to assess condom use by HIV-infected patients gains some interesting results, which can not be obtained when an ordinary logistic model is used.

Keywords

AIDS; Condom use; LASSO; Least squares approximation; Local linear regression; Profile likelihood; Quasilikelihood; SCAD; Sexual behavior; Spline smoothing

1 Introduction

An estimated 5.7 million South Africans are living with HIV, the largest number of HIV-infected individuals in a single country. The predominant mode of HIV transmission in South African is unprotected heterosexual intercourse [1]. HIV prevention programs have focused on a variety of behavioral, medical and surgical interventions. Although there are no clinical trials to support its use, male condoms markedly reduce the risk of sexual transmission of HIV especially in people who report consistent condom use [2–5]. More recent data from Africa [6] supports this point. Identification of the most important predictors of consistent condom use is therefore critical to assist policy makers in estimating the future course of the HIV epidemic. However, the study of predictors of condom use is complex and unsurprisingly, the literature reports a broad variety of predictors of condom use. For example, [7] suggests that the most significant predictors of condom use among young South African adults were: discussions about condom use and using condoms with the first sexual partner. Whereas in Ghana [8] perceptions of risk for HIV and self-efficacy of condoms were the most important predictors of any condom use. [9] using first-visit data from the same cohort we are studying concluded that a higher household income and a

history of alcohol consumption were significant predictors of condom use in men with a regular sexual partner. In these studies, a commonly used model for study of the relationship between the probability of historical condom use (with $Y = 1$ defined as yes) and its predictors is the ordinary logistic model; i.e., the logit of ($Y = 1$)

$$\text{logit}\{E(Y|X)\} = X^T \beta, \quad (1)$$

where, $X = (x_1, \dots, x_d)'$ are the predictors, because this model is easily computed and interpreted for the model parameter. We may say that a unit change in exposure variable x_1 increases the odds of a positive response multiplicatively with the magnitude $\exp(\alpha_1)$. A concern, however, is that this model may fail to account for the nonlinear effects of some continuous predictors, for example that of the CD4 cell count in the example below.

To carefully explore significance of predictors of condom use, we investigate the data set from an operational cohort of 864 HIV-infected adults conducted at Perinatal HIV Research Unit (PHRU), in Soweto South Africa. Participants are provided with a package of pre-antiretroviral care by primary health care nurses and referred for antiretroviral therapy once they become eligible. We are primarily concerned with factors that could significantly predict consistent condom use. The binary response variable, whether men consistently use condoms with their regular partners, takes the value 1 or 0, meaning a “yes” or “no” for the answer. Independent variables from this study include: x_1 , marital status (yes: married or living together; no: never married, divorced, separated or widowed); x_2 , education status (yes: if the total number of schooling years is higher than 8); x_3 , ever consumed alcohol; x_4 , ever disclosed HIV status; x_5 , STI symptom; x_6 , smoked status; x_7 , age; x_8 , household income; and z , CD4 cell count. Three variables z , x_7 and x_8 are continuous. We explore the impact of these factors on the probability of history of consistent condom use for a patient with HIV at their first visit.

In order to avoid model mis-specification and to assess the possible nonlinearity of the continuous covariates, we fit a generalized additive model (GAM) by treating all discrete predictors as linear components, and all continuous predictors as nonlinear components.

That is, we fit $\text{logit}\{E(Y|X, z)\} = \sum_{k=1}^6 \beta_k x_k + g_1(z) + g_2(x_7) + g_3(x_8)$ using function *gam* in R library *mgcv* with the default to choose the smoothing parameters. This preliminary analysis indicates that only CD4 cell has nonlinear effect. See Figure 1. Motivated by this finding, we propose to use the following generalized partially linear models (GPLMs)

$$E(Y|X, Z) = \mu\{X^T \beta + \theta(Z)\}, \quad (2)$$

where $\mu(\cdot)$ is a link function, β is a $d_x \times 1$ vector, and $\theta(\cdot)$ is an unknown univariate smooth function for covariate Z . Throughout the paper, we use Z to denote the univariate nonparametric covariate and X to denote the parametric covariates.

The GPLMs are an extension of the classical generalized linear models by incorporating nonparametric models for some covariate Z . In this article, we confine our attention to a univariate Z , although the model can obviously be extended to allow multiple covariates in the nonparametric part. The models have been studied by [10–13]. A special class of the GPLMs, the partially linear models, has been intensively studied by [14]. In dealing with a more general case where $\theta(\cdot)$ is a function of a single-index, [15] proposed two different estimation algorithms based on quasi-likelihood and local kernel methods, while [16]

proposed a penalized spline estimation of $\theta(z)$ when the link function $\mu(\cdot)$ is an identity function. As they pointed out, their method is appealing because of its rapid and stable computation. [17] further studied this topic for the case of the generalized link function.

When X is multi-dimensional ($d_x > 1$), it is usually of great interest to infer which covariates are significant. After all, statistical inference on the importance of the variables plays an instrumental role in statistical practice. A judicious variable selection procedure not only helps to interpret the results, but also brings more accurate estimation. However, the traditional variable selection procedures, such as subset selection, have heavy computational requirements and suffer from relative instability of the estimation, see, for example, [18]. In a seminal paper, [19] proposed the least absolute shrinkage and selection operator (Lasso), which conducts variable selection and coefficient estimation simultaneously. Fan and Li [20] recognized the bias inherent in the Lasso formulation and advocated the smoothly clipped absolute deviation (SCAD) approach, which allows one to fit the model as if the correct submodel were known in advance. This is referred to as the oracle property in variable selection by [20]. Thus, the asymptotic distribution of the estimators based on the full model and the true model coincide for the parametric part. [21,22] proposed an adaptive Lasso, which also has this property. For semiparametric models, [23] studied the profiled likelihood-based SCAD method. [24] further studied semiparametric model selection for quasi-likelihood. For linear models with diverging numbers of parameters, [25] showed that the SCAD-based penalized likelihood method has the oracle property. [26] proposed a unified Lasso approach based on least squares approximation (LSA), which takes advantage of the adaptive Lasso and is applicable to many general linear models. [27] proposed an *efficient* and adaptive shrinkage method for variable selection in the Cox model.

In this article, we propose a novel approach to variable selection for GPLMs. This approach capitalizes on preliminary estimates of β and θ and then feeds these estimates of β and its estimated covariance matrix into the LSA. The preliminary estimates can be obtained either via a kernel-based or spline-based approach. We show that the resulting estimator possesses the oracle property and that the resulting estimates achieve the semiparametric efficiency bound.

The rest of the article is organized as follows. Section 2 presents the preliminary estimates using the kernel-based or spline-based estimation, and justifies the asymptotics of the resulting estimators. The LSA-based variable selection procedure is then studied. We also investigate the oracle property of the proposed procedure. We present a real-world example from a HIV epidemic study in Section 3, and illustrate the proposed method via simulation studies in Section 4. Section 5 gives some concluding remarks. The technical proofs are sketched in the Appendix.

2 LSA-based Variable Selection

Before proceeding further, we assume $\text{var}(Y|X, Z) = \sigma^2 V(\mu)$, where $\mu = \mu\{\theta(z) + X^T\beta\}$. With n observations, we introduce the following notation: $\rho_k(t) = \{d\mu(t)/dt\}^k V^{-1}\{\mu(t)\}$ for $k = 1, 2$; $f(z)$ as the density of Z ; $q_1(t, y) = \{y - \mu(t)\} \rho_1(t)$; and $\varepsilon_i = \rho_1\{X_i^T\beta + \theta(Z_i)\} [Y_i - \mu\{X_i^T\beta + \theta(Z_i)\}] = q_1\{X_i^T\beta + \theta(Z_i), Y_i\}$.

In the following two subsections, we briefly introduce the kernel smoothing and regression spline methods. These two methods may be used to obtain a preliminary estimate of β and its associated asymptotic covariance matrix for LSA-based variable selection.

2.1 Kernel-based estimation

We first propose the profile likelihood method to obtain preliminary estimates of β and its asymptotic covariance matrix, which are needed for the LSA procedure. The profile likelihood procedure iterates between solving for β for fixed θ and solving for θ for fixed β .

Estimation of β —We use the profile likelihood approach and the quasi-likelihood method to estimate the parametric components. We first locally estimate $\theta(z_0)$ as a function of β and obtain $\hat{\theta}(\beta, z_0)$. Then we globally estimate β similarly to the likelihood method. The procedure can be described by the following algorithm ([13]).

Step 1—For each fixed z_0 and β , $\hat{\theta}(\beta, z_0)$ denotes the solution in α_0 of the equation

$$0 = \frac{1}{n} \sum_{i=1}^n K_h(z_0 - Z_i) \left(\frac{1}{\frac{z_i - z_0}{h}} \right) \rho_1 \{ \alpha_0 + \alpha_1 (Z_i - z_0) / h + X_i^T \beta \} [Y_i - \mu \{ \alpha_0 + \alpha_1 (Z_i - z_0) / h + X_i^T \beta \}], \quad (3)$$

where $K_h(\cdot) = K(\cdot/h)/h$ for $K(\cdot)$ that is a kernel function and h that is a bandwidth.

Step 2—Given the estimator $\hat{\theta}(\beta, z)$, an estimator of β , $\hat{\beta}_n$ is then obtained by solving the following equation:

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{\partial \mu \{ \hat{\theta}(Z_i, \beta) + X_i^T \beta \}}{\partial \beta} V^{-1} \{ \hat{\theta}(Z_i, \beta) + X_i^T \beta \} [Y_i - \mu \{ \hat{\theta}(Z_i, \beta) + X_i^T \beta \}]. \quad (4)$$

We claim the following conditions, which are standard in the literature for the GPLMs ([14]). These conditions hold throughout the remainder of the article.

Conditions for the kernel-based approach

- (a1) The density function $f(z)$ of Z is positive and continuous at the point z_0 .
- (b1) The derivative function $\theta^{(2)}(\cdot)$ is continuous at the point z_0 .
- (c1) With $R = \theta(Z) + X^T \beta$, $E\{q_1^2(R, Y)|z_0\}$, $E\{q_1^2(R, Y)X|z_0\}$, and $E\{q_1^2(R, Y)XX^T|z_0\}$ are twice differentiable in z_0 .
- (d1) $E\{q_1^{2+\delta}(R, Y)\} < \infty$, for some $\delta > 2$.

Proposition 2.1: When the conditions described above hold and when $nh_4 \rightarrow 0$, the estimator $\hat{\beta}_n$ is asymptotically normal with covariance matrix Σ_β , which is equal to $\sigma^2[E\{p_2(X^T \beta + \theta(Z))\tilde{X}\tilde{X}^T\}]^{-1}$. Here $\tilde{X}$ is defined in (A.3) in the Appendix.

A consistent estimator of Σ_β is available in Appendix A1. In theory, only the rate of convergence for the bandwidth h is necessary because a bandwidth with the rate of convergence in Proposition 2.1 leads to the same limit distribution for the estimators of β .

2.2 Spline-based estimation

An alternative method for approximating the nonparametric term is the spline approach ([28]) as a global smoothing method, which approximates the unknown functions via splines characterized by only a finite number of coefficients. After the spline basis is chosen, the coefficients can be estimated by an efficient one-step method of maximizing the likelihood

function. It is in contrast to the kernel-based method, where the maximizing has to be conducted repeatedly at every local data point. Thus the advantage of the application of the regression spline is its computational efficiency. Use of regression spline smoothing in nonparametric models has been investigated in various contexts. [29] first obtained the rate of convergence of the polynomial spline estimates for the additive model. See [30] for a systematical survey for semiparametric modeling using regression spline, and [31,32] for its application in partly linear models with longitudinal data. [16] studied penalized spline estimation for partially linear single-index models, while [17] further extended their idea for generalized partially linear single-index models. Under mild assumptions, [17] established asymptotic normality for the penalized estimators of the parametric component she derived. After scaling Z_i onto the unit interval $[0, 1]$, we approximate the nonparametric function $\theta(Z)$ by a set of B-spline basis functions $B_1(z), \dots, B_{k+l}(z)$, depending on the knots sequence $0 = z_0 < z_1 < \dots < z_{k+l} = 1$, which are used as internal knots to form a basis for the linear spline space. For example, if $k = 4$, then z_1, z_2, z_3 , and z_4 are the four quartiles of $\{Z_i\}_{i=1}^n$. The constant l is the order of the B-spline basis, and the choices of $l = 2$ (piecewise linear), $l = 3$ (quadratic spline), and $l = 4$ (cubic spline) are common. Then we can approximate $\theta(z)$ as a

linear combination of the B-spline basis functions $\theta(z, \alpha) = \sum_{j=1}^{k+l} \alpha_j B_j(z)$. An attractive property of the spline-based method is that once the basis functions are specified, the estimation problem becomes a linear one and we can use existing generalized linear model-fitting software to estimate β and $\alpha = (\alpha_1, \dots, \alpha_k)^T$.

As in [32], we assume that the dependence of X and Z satisfies the following relationship $X_{ij} = g_j(Z_i) + \delta_{ij}$, where $\{\delta_{ij}\}$ are mean zero independent noises. We assume that $g_j(\cdot)$ are functions whose r th derivatives are bounded for some $r \geq 2$. To study the asymptotic

properties, we denote $V = \text{diag}[\rho_2\{X_i^T \beta + \theta(Z_i)\}]$, $B_i = \{B_1(Z_i), \dots, B_{k+l}(Z_i)\}^T$, $M = (B_1, \dots, B_n)$, $X^* = (I - P)X$, $P = M(M^T V M)^{-1} M^T V$, $K_n = X^{*T} V X^*$, and denote Δ_n as an $n \times d_x$ matrix whose s th column is $\delta_s = (\delta_{1s}, \dots, \delta_{ns})^T$.

Conditions for the spline-based approach

- (a2) The distinct value of Z_i forms a quasi-uniform sequence that grows dense on $[0, 1]$.
- (b2) The derivative of $\theta(\cdot)$ is bounded for some $r \geq 2$.
- (c2) For sufficiently large n , $k(M^T V M)$ is nonsingular and the eigenvalues of $M^T V M(k/n)$ are bounded from 0 and infinity.
- (d2) For random noises, we have: 1) $E(\Delta_n) = 0$ and $\sup_n n^{-1} E\|\Delta_n\|^2 \leq \infty$; 2) $n^{-1} K_n \rightarrow K > 0$ in probability from some positive definite matrix K .

Because $\theta(\cdot)$ is an unknown function, we can always make a transformation to make Z being uniformly distributed. The assumption a2 is generally made in the literature of generalized partially linear/additive models.

Proposition 2.2: Under conditions (a2)–(d2), (d1) and if $k = O(n^{1/(2r+1)})$, the estimator $\hat{\beta}_s$ is asymptotically normal such that $\sqrt{n}(\hat{\beta}_s - \beta) \rightarrow N(0, \Sigma)$, where $\Sigma = \Sigma_\beta$ and Σ_β is defined in Proposition 2.1.

Note that the asymptotic covariance of $\hat{\beta}_s$ is the same as that of $\hat{\beta}_n$. In principle, the number of knots can be chosen by using cross validation or information theoretic criteria such as AIC and BIC ([32]). A quick and heuristic choice is to fix $k = n^{1/(2r+1)}$. For $r = 2$ and cubic splines, we can take the integer part of $n^{1/5}$ as suggested by [32]. For a consistent estimator

of Σ_{β} when the spline-based approach is used, we use $\widehat{\Sigma}_{\beta} = \widehat{\sigma}^2 \widehat{K}_n^{-1} / n$, where $\widehat{\sigma}^2$ is a consistent estimator of σ^2 , as given in Appendix A1, and $\widehat{K}_n = X^{*T} V X^*$ with V evaluated at $(\beta^T, \alpha^T)^T$.

Remark 1: The conditions imposed in Proposition 2.1 indicate that the optimal bandwidth is inappropriate for estimating parameter β and undersmoothing is required. As a consequence, we adopted a strategy suggested by [15] for our bandwidth selection. That is, we first select an optimal bandwidth and then multiply $n^{-5/3}$. This manipulation ensures the correct order of the bandwidth.

2.3 Least squares approximation

Using either the kernel-based or the spline-based approach, we have preliminary estimates of β and its asymptotic covariance matrix. For variable selection, we propose to minimize the following least squares approximation criterion [26]

$$(\beta - \widehat{\beta})^T \widehat{\Sigma}_{\beta}^{-1} (\beta - \widehat{\beta}) + \sum_{j=1}^{d_x} \lambda_{\text{LSA},j} |\beta_j|, \quad (5)$$

where $\lambda_{\text{LSA},j} = \lambda / \widehat{\beta}_j^{-\gamma}$ and λ is the tuning parameter. As in [26], we use $\gamma = 1$ for simplicity. Here λ is selected according to the BIC-type criterion ([26])

$$\text{BIC}(\lambda) = (\widehat{\beta}_{\lambda} - \widehat{\beta})^T \widehat{\Sigma}_{\beta}^{-1} (\widehat{\beta}_{\lambda} - \widehat{\beta}) + \log n \times df_{\lambda} / n,$$

where df_{λ} is the number of nonzero coefficients of the $\widehat{\beta}_{\lambda}$ with the tuning parameter λ .

In order to show the oracle properties, we verify the conditions in [26]. We note that the estimate $\widehat{\Sigma}_{\beta}$ is consistent, though its convergence rate can be arbitrary. First, the covariance assumption in [26] is satisfied. To see that, recall $\Sigma_{\beta} = \sigma^2 [E\{\rho_2(X^T \beta + \theta(Z)) \tilde{X} \tilde{X}^T\}]^{-1}$. For any model S containing the true model, it is easy to check that the asymptotic covariance

matrix for the estimator of β_S denoted as $\widehat{\beta}_{n,S}$ is the inverse of a submatrix of $\widehat{\Sigma}_{\beta}$ for the full model. Thus, as long as $n^{1/2}\lambda \rightarrow 0$ and $n^{(1+\gamma)/2}\lambda \rightarrow \infty$ are satisfied, $\widehat{\beta}_n$ has the oracle properties ([26] Theorems 1–3). Namely, with probability tending to 1, $P(\widehat{\beta}_j = 0) \rightarrow 1$ for $\beta_j = 0$. And if we denote the nonzero coefficients set as $\mathbf{A} = \{j: \beta_j \neq 0\}$, we have the asymptotic normality for $\widehat{\beta}_{n,\mathbf{A}}$ as

$$\sqrt{n}(\widehat{\beta}_{n,\mathbf{A}} - \beta_{\mathbf{A}}) \rightarrow_d N(0, \Sigma_{\mathbf{A}})$$

where $\widehat{\beta}_{n,\mathbf{A}}$ and $\beta_{\mathbf{A}}$ are subvectors of $\widehat{\beta}_n$ and β respectively by taking the elements in $\mathbf{A}$, and $\Sigma_{\mathbf{A}}$ is a submatrix of $\widehat{\Sigma}_{\beta}$ with entries corresponding to elements in $\mathbf{A}$. We see that this asymptotic covariance is the same as would be if the true model was known. Thus, the oracle properties are proved ([20]). Furthermore, it is straightforward to show that the BIC

method yields oracle properties (Theorem 4, [26]). In addition, the resulting estimate of β is efficient [14].

3 Revisit PHRU Study

The positive impact of antiretroviral therapy (ART) at the individual level is well known: treatment leads to suppressing viral load, decreased morbidity and mortality, fewer opportunistic infections, increased survival, and allows many patients to return to productive, healthy lives [33]. But the impact of ART on population HIV dynamics is largely unknown. This is an important question for policy makers because planning ART service delivery, prevention programs and expenditure requires a solid estimate of the future course of the HIV epidemic. The introduction of ART has made it increasingly difficult to predict the course of the epidemic because ART reduces plasma viral load [34] and the amount of virus in the genital tract [35] and therefore infectivity [36]. On the other hand, treatment might reduce HIV transmission. Moreover, improved access to ART may result in more people accessing voluntary counseling and testing, increasing levels of HIV disclosure and higher rates of safe sex. Exploration of the impact of ART on the attention to condom use, along with other risk factors among the HIV-infected patients, is therefore very practically important. CD4 cell, called helper cell, initiates the body's response to invading micro-organisms such as viruses. HIV attacks these types of cells and uses them to make more copies of HIV. As a consequence, HIV weakens the immune system, making it unable to protect the body from infection. CD4 cell is one of the most important biomarkers in assessment of ART treatment. HIV infected people are considered to have "normal" (non-pathologic) CD4 counts if the number is above 500. This biomarker should be paid a special attention on the basis of its importance and our preliminary results obtained by fitting the generalized additive model.

We now use the following GPLM for analysis of the PHRU data

$\text{logit}\{E(y|X, z)\} = \theta(z) + \sum_{k=1}^8 \beta_k x_k$, and apply the proposed variable selection procedure to the model. We use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ here and in the simulation study later. The estimated coefficients and their standard errors are listed in the right panel of Table 1. The selected model based on the kernel-LSA procedure is: $\text{logit}\{E(y|X, z)\} = \hat{\theta}(z) + 0.585x_1 - 0.623x_2 + 0.192x_8$, where $\hat{\theta}(z)$ is depicted (the dashed line) in Figure 2. For comparison, we have also fitted the data set using a generalized linear model and the estimated results are listed in the left panel of Table 1. These results indicate that only x_1 , x_2 , x_8 , and z are significant, while x_3 - x_7 are not significant with 10% nominal level.

The results, not reported here, show that the magnitude of the estimated values for the linear parameters from GPLM and generalized additive models is similar. However, the GPLM fitting gives smaller standard errors. Again, the fitted smooth functions in Figure 2 show that the relationship between the log odd of $y = 1$ and the CD4 cell count may be nonlinear. A GPLM fitted with the proposed variable selection procedure further confirms what we obtained from the GLM fitting; that is, HIV-infected adults with higher levels of education, higher income and married status have higher levels of condom use with their regular partners. But the estimated values for CD4 count from two models convey different messages. The estimation result based on logistic regression indicates that CD4 positively affects the consistent condom use. However, the estimated curve for CD4 based on the GPLM showed an initial increasing trend, followed by a decreasing one after around 500. That is, for people with CD4 count smaller than 500, the higher CD4 count corresponds to higher levels of condom use with their regular partners. This pattern becomes converse in the group with CD4 count larger than 500. The results based on the ordinary logistic fitting indicates that with the higher CD4 cell counts, the patients prefer condom use. This

contradicts to the GPLM fits for which this trend changes when the CD4 cell counts are over 500. This may be because they feel good and are therefore reluctant to use condoms. This finding supports the point that “treatment optimism” might also lead to an increase in sexual risk behavior conjectured by [37]. It is therefore necessary to advise the patients with “normal” CD4 counts to use condoms in order to prevent the spread of HIV infection.

We also compare the performance of the proposed method to the standard GPLM using partial splines but without variable selection. Using V-fold cross validation to assess the misclassification errors, the average misclassification error of GPLM is 0.375 with standard error 0.015 and that of our method is 0.376 with standard error 0.016, when $V = 50$. This demonstrate that the proposed method is competitive to the standard model but gives more parsimonious models. The high misclassification error is not unusual in such studies.

4 Simulation study

We conduct simulation studies to investigate the finite-sample performance of the proposed method, and repeat 200 times for each simulation set-up. We measure the accuracy of the parametric parts in terms of the mean square error (MSE), defined by $MSE(\hat{\beta}) = (\hat{\beta} - \beta)^T E(X^T X)(\hat{\beta} - \beta)$ for an estimator $\hat{\beta}$.

Example 1

We first consider the partly linear logistic regression model by generating a Bernoulli random response with success probability from the following model

$$P(X, Z) = \exp\{X^T \beta + f(Z)\} / [1 + \exp\{X^T \beta + f(Z)\}],$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0)^T$ and $f(z) = 4.26\{\exp(-3.25z) - 4 \exp(-6.5z) + 3 \exp(-9.75z)\}$. The covariate Z is generated from the uniform distribution on $[0, 1]$. The j th covariate X_j is generated to be $U_j + aZ$ such that $(U_1, \dots, U_p)$ follows a multivariate normal distribution with a zero mean and covariance matrix $(\sigma_{ij})_{10 \times 10}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The constant a is used such that X and Z are correlated. In the simulation, we take $a = 0, 1, 2$ so the correlations between X_j and Z are 0, 0.28, and 0.50, respectively. In this setting, X_j is linearly correlated to Z . Nonlinear correlations between the covariates can be considered and the conclusion is qualitatively similar. Two sample sizes, $n = 300$ and $n = 600$, are considered. We record the average number of correctly identified zero coefficients (C) and that of the incorrectly identified zero (IC). Additionally, we record the percentage of the correctly identified models (CME). Also reported is the median relative mean square errors (MRMSE), defined as the median of the ratio of the MSE for an estimator to that of the full model. The median of absolute deviation (MAD) of the relative MSE divided by 0.6745 is also included. This measure can be regarded as the true standard error [20]. The results are reported for LSA and the oracle estimator, assuming that we know the true model structure. We report the kernel estimates and the spline estimates in Table 2.

Comparing the model complexity of models selected by the proposed procedure, we note that the results based on kernel and spline approaches are similar in terms of model complexity. From Table 2, it can be seen that with an increase of sample sizes the proportion of correctly fitted model increases and the number of true zero coefficients identified is closer to 7, whereas a general converse trend is observable as a gets large. When $n = 300$ and a increases from 1 to 2, the loss in the proportion of correctly fitted models can be a round 35%, while this inferiority can be smaller when the sample size increases to 600. With a small sample size and an increase of a from 1 to 2, the values of the MRMSE of the

proposed procedure also dramatically increase twice, and are substantially bigger than those of the oracle. Finally, for this logistic regression example, we observe that the spline-based approach gives estimates with smaller MRMSEs.

Example 2

We now conduct another simulation following more closely with the PHRU study. Suggested by the final model in the previous section, we fit a partial linear model with three parametric components x_1, x_2, x_8 and a nonparametric component in z , using function *gam* in R library *mgcv* with the default choice. Taking the fitted probabilities $\hat{p}_i, i = 1, \dots, n$ from this model, we generate the response variables $\tilde{y}_i, i = 1 \dots, n$ from Bernoulli distributions with these probabilities. We then sample without replacement N times from $\{\tilde{y}_i, Z_i, X_i\}_{i=1}^n$, with $N = n, N = n/2$ or $N = n/3$. To study a large dimensional model, we add to the covariate X_i an additional D iid standard normal random variables. Thus, there are $8 + D$ covariates for the parametric part and we take $D = 6$ or 12 . We repeat this data generating procedure 200 times and fit each data using GPLM with spline smoothing and variable selection. To save space, we omit the result with kernel smoothing. An interesting question raised by one referee in using GPLM is how to compare the performance of the proposed variable selection with that of just doing regression and testing whether the variables are zero. Theoretically investigating such a comparison is by no means straightforward. Thus we compare our method with this hypothesis test based method using simulation. Particularly, we use *gam* to fit the GPLM and then delete variables in the parametric part which have p-values larger than 0.05. We then fit another GPLM using *gam*. This approach is referred to as “Test”. To study the performance of our method under different signal noise ratios, we also try to multiple the fitted logit values by a factor “S” with $S = 1.5$ and $S = 2$. The results are summarized in Table 3. We see clearly that the model selection result improves when either the sample size increases or the signal strength, indicated by S , increases. The same phenomenon is observed for the estimation accuracy in terms of MRMSEs. With a strong signal ($S = 2$) and a large sample size ($N = 846$), the proposed method performs quite close to the oracle procedure with the model structure known in advance. In general, the proposed method outperforms the hypothesis test based procedure in terms of both model selection and estimation accuracy, especially when signal is strong and sample size is large.

5 Conclusion and Discussion

Variable selection for GPLMs is relatively overlooked in statistical literature, although estimation for the models has been studied well [12,13,38] and widely used in various disciplines (for example, [39] in a study of air pollution and [40] in a study of chronic kidney disease). In this paper, we have proposed efficient and simple kernel-based and spline-based variable selection procedures to deal with this problem. We have compared these two methods and found out that the performance of the proposed methods is similar. The spline-based one is easier to implement and achieves significant benefits in computational cost, based on our simulation studies and real data analysis, as expected. For both methods, we have established asymptotic normality of the proposed estimators.

The proposed variable selection method is motivated by the PHRU study in which the effect of CD4 cell was found to be nonlinear. It is worth pointing out that this pattern may not apply to other studies since population-level impact of ART on epidemic dynamics is multi-dimensional, not fully understood, and needs more research. In practice, however, there may exist multiple covariates which are nonparametrically related to the response. In this case, we may use the spline-based estimation method by parameterizing each covariate effect using splines. Then it is possible to use a penalized likelihood approach where the whole set

of the spline coefficients of each covariate is penalized. This is a much more challenging problem and deserves more attention for future research.

An interesting question is how to select nonparametrically modeled variables. [41] proposed a nonparametric component selection procedure in a penalized likelihood framework. [42] applied COSSO to the Cox proportional hazards model with nonparametric relative risk. Although a similar extension to generalized additive models is of interest, it is not clear and extremely challenging whether the theoretical properties of the COSSO method such as the existence and the convergence rate of the COSSO estimator can be transferred to the estimation of the additive nonparametric functions.

Our method extends the existing literature on semiparametric model selection developed for continuous outcomes ([23,43]). In particular, the doubly penalized estimator in [43] may be developed for variable selection for the GPLM. It is of great interest to study the proposed method when d_x , the dimension of the parametric covariate X , grows with the sample size. Since our method relies on a preliminary estimate of β , our method, however, can not allow $d_x > n$. This is a limitation of the current method. However, with suitable conditions on d_x , we expect that the results in [25] and [43] may be used for generalizing the proposed method to problems with diverging dimensionality.

Acknowledgments

Leng's research is supported partially by NUS grants. Liang's research was partially supported by NIH/NIAID grant AI59773 and NSF grant DMS 0806097. Patient care is supported by a PEPFAR grant (674-A-00-05-00003-00) through USAID. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and USAID. We thank the Editor and two anonymous referees for constructive comments.

References

1. WHO. Report on the global aids epidemic. Geneva, Switzerland: 2008.
2. A meta-analysis of condom effectiveness in reducing sexually transmitted hiv. *Social Science & Medicine*. 1993; 36:1635–1644. [PubMed: 8327927]
3. Pinkerton S, Abramson PR, Holtgrave D. What is a condom worth? *AIDS and Behavior*. 1999; 3:301–312.
4. Weller S, Davis K. Condom effectiveness in reducing heterosexual hiv transmission 2002. *Cochrane Database of Systematic Reviews: Reviews*. 2002; (1)
5. Cherutich P, Brentlinger P, Nduati R, Kiari J, Farquhar C. Condom use among sexually active kenyan female adolescents at risk for hiv-1 infection. *AIDS and Behavior*. 2008; 12:923–929. [PubMed: 18688705]
6. Morris C, Ferguson A. Estimation of the sexual transmission of HIV in Kenya and Uganda on the trans-Africa highway: the continuing role for prevention in high risk groups. *Sexually Transmitted Infections*. 2006; 82:368–371. [PubMed: 16854995]
7. Hendriksen ES, Pettifor A, Lee SJ, Coates TJ, Rees HV. Predictors of Condom Use Among Young Adults in South Africa: The Reproductive Health and HIV Research Unit National Youth Survey. *American Journal of Public Health*. 2007; 97:1241–1248. [PubMed: 17538062]
8. Adih WK, Alexander CS. Determinants of condom use to prevent HIV infection among youth in Ghana. *Journal of Adolescent Health*. 1999; 24:63–72. [PubMed: 9890367]
9. Lurie M, Pronyk P, de Moor E, Heyer A, de Bruyn G, Struthers H, McIntyre J, Gray G, Marinda E, Klipstein-Grobusch K, et al. Sexual behavior and reproductive health among HIV-infected patients in urban and rural South Africa. *Journal of Acquired Immune Deficiency Syndromes*. 2008; 47:484–493. [PubMed: 18209685]
10. Liang H. Generalized partially linear models with missing covariates. *Journal of Multivariate Analysis*. 2008; 99:880–895. [PubMed: 19412320]

11. Liang H, Ren H. Generalized partially linear measurement error models. *Journal of Computational and Graphical Statistics*. 2004; 14:237–250.
12. Hunsberger S. Semiparametric regression in likelihood-based models. *Journal of the American Statistical Association*. 1994; 89:1354–1365.
13. Severini TA, Staniswalis JG. Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*. 1994; 89:501–511.
14. Härdle, W.; Liang, H.; Gao, JT. *Partially Linear Models*. Springer Physica; Heidelberg: 2000.
15. Carroll R, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association*. 1997; 92:477–489.
16. Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*. 2002; 97:1042–1054.
17. Yu, Y. Unpublished manuscript. Penalized spline estimation for generalized partially linear single index models 2007.
18. Fan, J.; Li, R. *International Congress of Mathematicians. Vol. III. Eur. Math. Soc; Zürich: 2006. Statistical challenges with high dimensionality: feature selection in knowledge discovery; p. 595–622.*
19. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
20. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
21. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
22. Zhang HH, Lu W. Adaptive-LASSO for Cox's proportional hazards model. *Biometrika*. 2007; 94:691–703.
23. Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*. 2004; 99:710–723.
24. Li R, Liang H. Variable selection in semiparametric regression modeling. *The Annals of Statistics*. 2008; 36:261–286.
25. Fan J, Peng H. On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics*. 2004; 32:928–961.
26. Wang H, Leng C. Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*. 2007; 101:1418–1429.
27. Zou H. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*. 2008; 95:241–247.
28. Stone CJ. Additive regression and other nonparametric models. *The Annals of Statistics*. 1985; 13:689–705.
29. Stone CJ. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*. 1986; 14:590–606.
30. Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*. Cambridge University Press; New York: 2003.
31. He X, Zhu Z, Fung WK. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*. 2002; 89:579–590.
32. He X, Fung WK, Zhu ZY. Robust estimation in generalized partially linear models for clustered data. *Journal of the American Statistical Association*. 2005; 472:1176–1184.
33. Palella F, Delaney K, Moorman A, Loveless M, Fuhrer J, Satten G, Aschman D, Holm-berg S, the HIV Outpatient Study Investigators. Declining morbidity and mortality among patient with advanced human immunodeficiency virus infection. *New England Journal of Medicine*. 1998; 338:853–860. [PubMed: 9516219]
34. Gray R, Wawer M, Brookmeyer R, Sewankambo N, Serwadda D, Wabwire-Mangen F, Lutalo T, Li X, van Cott T, Quinn T, et al. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1 discordant couples in Rakai, Uganda. *Lancet*. 2001; 357:1149–1153. [PubMed: 11323041]

35. Cu-Uvin S, Snyder B, Harwell J, Hogan J, Chibwesha C, Hanley D, Ingersoll J, Kurpewski J, Mayer K, Caliendo A. The association between paired plasma and cervicovaginal lavage fluid HIV-1 RNA levels during 36 months. *Journal of Acquired Immune Deficiency Syndrome*. 2006; 42:584–587.
36. Zhang L, Ramratnam B, Tenner-Racz K, He Y, Vesanen M, Lewin S, Talal A, Racz P, Perelson A, Korber B, et al. Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. *New England Journal of Medicine*. 1999; 340:1605–1613. [PubMed: 10341272]
37. Blower S, Bodine E, Kahn J, McFarland W. The antiretroviral roll-out and drug resistant HIV in africa: insights from empirical data and theoretical models. *AIDS*. 2005; 19:1–14. [PubMed: 15627028]
38. Härdle, W.; Müller, M.; Sperlich, S.; Werwatz, A. *Nonparametric and Semiparametric Models*. Springer-Verlag; New York: 2004.
39. Dominici F, McDermott A, Zeger S, Samet J. On the use of generalized additive models in time-series of air pollution and health. *American Journal of Epidemiology*. 2002; 156:193–203. [PubMed: 12142253]
40. Hsu C, Kao LWH, Coresh J, Pankow J, Marsh-Manzi J, Boerwinkle E, Bray MS. Apolipoprotein e and progression of chronic kidney disease. *Journal of the American Medical Association*. 2005; 293:2892–2899. [PubMed: 15956634]
41. Lin Y, Zhang H. Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics*. 2006; 34:2272–2297.
42. Leng C, Zhang HH. Model selection in nonparametric hazard regression. *Journal of Nonparametric Statistics*. 2006; 18:417–429.
43. Ni X, Zhang HH, Zhang D. Automatic model selection for partially linear models. *Journal of Multivariate Analysis*. 2009; 100:2100–2111. [PubMed: 20160947]

Appendix A1: A sketch of proof of Proportion 2.1

From (3) and use of the Taylor expansion, we get the following expression:

$$\widehat{\theta}(z_0, \widehat{\beta}_n) - \theta(z_0) = \frac{1}{nf(z_0)\mathcal{A}(z_0)} \sum_{i=1}^n K_h(Z_i - z_0) \varepsilon_i - \frac{E\{X\rho_2(R)|Z=z_0\}}{\mathcal{A}(z_0)} (\widehat{\beta}_n - \beta) + o_p(n^{-1/2}) + \frac{1}{2} h^2 \theta^{(2)}(z_0), \quad (\text{A.1})$$

where $\mathcal{A}(z_0) = E\{\rho_2(R)|Z = z_0\}$. A direct calculation yields $\partial \widehat{\theta}(z_0, \beta) / \partial \beta = -\mathcal{B}(z_0) + o_p(1)$ with $\mathcal{B}(z_0) = E\{\rho_2(R)X|Z = z_0\} / \mathcal{A}(z_0)$. On the other hand, it follows from (4) that $\widehat{\beta}_n$ is the solution of

$$0 = \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{\beta}_n + \frac{\partial \widehat{\theta}(Z_i, \beta)}{\partial \beta} \right\} [Y_i - \mu\{\widehat{\theta}(Z_i, \beta) + X_i^T \beta\}] \rho_1\{\widehat{\theta}(Z_i, \beta) + X_i^T \beta\}. \quad (\text{A.2})$$

Write $\widetilde{X}_i = X_i - \mathcal{B}(Z_i)$ and

$$\widetilde{X} = X - \mathcal{B}(Z). \quad (\text{A.3})$$

We can rewrite (A.2) as

$$\frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i = \frac{1}{n} \sum_{i=1}^n \rho_2 \{X_i^T \beta + \theta(Z_i)\} \tilde{X}_i \{ \tilde{X}_i^T \hat{\beta}_n + \hat{\theta}(Z_i, \hat{\beta}_n) - X_i^T \beta - \theta(Z_i) \}.$$

On the basis of the fact given in (A.1), the term in the bracket of the right-hand side may be simplified as follows

$$\tilde{X}_i^T (\hat{\beta}_n - \beta) + h^2 / 2 \theta^{(2)}(Z_i) + \frac{1}{n f(Z_i) \mathcal{A}(Z_i)} \sum_{j=1}^n K_h(Z_i - Z_j) \varepsilon_j. \quad (\text{A.4})$$

Combining the above statements and a direct but cumbersome simplification yields

$$\frac{1}{n} \sum_{i=1}^n \rho_2 \{X_i^T \beta + \theta(Z_i)\} \tilde{X}_i \tilde{X}_i^T (\hat{\beta}_n - \beta) = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \varepsilon_i + \frac{1}{2} n h^2 E \{ \rho_2(R) \tilde{X} \theta^{(2)}(Z) \}.$$

Using a central limiting theorem and recalling the definition of ε_i , we know that $\sqrt{n}(\hat{\beta}_n - \beta)$ is asymptotically normal and the asymptotic covariance matrix is $\sigma^2 [E \{ \rho_2(R) \tilde{X} \tilde{X}^T \}]^{-1}$. We therefore complete the proof of Proposition 2.1.

Estimation of Σ_β

Set $\hat{\rho}_{2i} = \rho_2 \{X_i^T \hat{\beta}_n + \hat{\theta}(Z_i)\}$. Let $\hat{\mathcal{A}}(z)$ and $\hat{E}(\rho_{2i}|z)$ be the local linear estimators of $\mathcal{A}(z)$ and $E(\rho_2|z)$ based on the “data” $\hat{\rho}_{2i}$. Write $\hat{\mathcal{B}}(z) = \hat{E}(\rho_2|z)/\hat{\mathcal{A}}(z)$ and $\hat{X}_i = X_i - \hat{\mathcal{B}}(Z_i)$. The asymptotic covariance matrix Σ_β can be estimated by

$$\widehat{\Sigma}_\beta = \widehat{\sigma}^2 \left\{ \frac{1}{n} \sum_{i=1}^n \hat{\rho}_{2i} \hat{X}_i \hat{X}_i^T \right\}^{-1},$$

where

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left[Y_i - \mu \{X_i^T \hat{\beta}_n + \hat{\theta}(Z_i)\} \right]^2 V^{-1} [\mu \{X_i^T \hat{\beta}_n + \hat{\theta}(Z_i)\}],$$

It is easy to show that $\widehat{\Sigma}_\beta$ is a consistent estimator of Σ_β .

Appendix A2: A sketch of proof of Proposition 2.2

First note that this proposition deals with a simplified model in [32]. Following the arguments in Theorem 1 of [32], it is not difficult to show that $\hat{\theta}(\cdot)$ converges at the optimal rate of convergence $n^{-2r/(2r+1)}$ and that $\hat{\beta}$ follows a normal distribution asymptotically. The asymptotic covariance matrix of $\sqrt{n}(\hat{\beta}_s - \beta)$ can be verified by noticing that $n^{-1}S_n$ in [32] converges to $\sigma^2 K$, where K is the limit of K_n in the condition (d2).

Figure 1.
The fitted nonparametric terms using the generalized additive model with shaded pointwise 95% confidence bands.

Figure 2.

The estimated function $\theta(CD4)$ against CD4 cell count for the PHRU data. The solid, dashed, and dotted lines indicate the estimated results based on the spline, linear, and kernel methods, respectively.

Results for the PHRU data. Left panel: Estimated values, associated standard error, and P-value by using the ordinary logistic model. Right panel: Estimates, associated standard errors of the coefficients using the GPLM with the proposed variable selection procedures.

Table 1

	logistic model			GPLM		
	Est.	s.e	Pr(> z)	kernel+LSA (s.e.)	Spline+LSA (s.e.)	
Mar-Status (x_1)	0.881	0.198	$<10^{-3}$	0.585(0.199)	0.637(0.197)	
Edu (x_2)	-0.736	0.170	$<10^{-3}$	-0.623(0.171)	-0.648(0.169)	
Alcohol (x_3)	0.147	0.171	0.393	0(0)	0(0)	
Disclosed (x_4)	0.401	0.262	0.126	0(0)	0(0)	
STISymptoms (x_5)	0.006	0.175	0.973	0(0)	0(0)	
Smoked (x_6)	-0.088	0.219	0.689	0(0)	0(0)	
Age (x_7)	0.047	0.074	0.526	0(0)	0(0)	
Income (x_8)	0.237	0.075	0.002	0.192(0.075)	0.183(0.075)	
CD4 (z)	0.137	0.075	0.070			

Example 1: Logistic regression using the kernel method ($h = 0.2$) and the spline method ($df = 5$). “C” is the average of corrected estimated zeros (the true C is 7); “IC” is the average number of incorrectly estimated zeros; “CME” is the percentage of correct models; “MRMSE” is median relative mean square errors; “LSA” is the proposed method; “Oracle” is the method assuming the true model structure is known.

Table 2

method	n	a	C	IC	CME (%)	LSA-MRMSE (MAD)	Oracle-MRMSE (MAD)
Kernel	300	0	6.845	0.000	84.5	0.465 (0.546)	0.419 (0.255)
		1	6.795	0.055	83.0	0.572 (0.726)	0.353 (0.270)
	600	2	6.660	1.025	54.5	1.426 (1.503)	0.332 (0.273)
		0	6.860	0.000	87.0	0.562 (0.669)	0.460 (0.264)
	600	1	6.890	0.000	89.0	0.482 (0.570)	0.444 (0.260)
		2	6.835	0.000	84.5	0.396 (0.750)	0.389 (0.252)
Spline	300	0	6.815	0.000	83.0	0.381 (0.645)	0.379 (0.292)
		1	6.840	0.110	82.0	0.526 (0.663)	0.297 (0.227)
	600	2	6.955	1.525	45.0	1.393 (1.650)	0.231 (0.199)
		0	6.875	0.000	88.5	0.474 (0.570)	0.442 (0.302)
	600	1	6.900	0.000	90.0	0.382 (0.469)	0.391 (0.272)
		2	6.805	0.030	81.5	0.397 (0.501)	0.331 (0.269)

Table 3

Example 2: Logistic regression using the spline method ($df = 5$). The true C is 11 for $D = 6$ and is 17 for $D = 12$. The column “S” indicates the signal strength.

D	S	N	Method	C	IC	CME (%)	LSA-MRMSE (MAD)	Oracle-MRMSE (MAD)
6	1	282	Spline	10.875	1.965	8.0	0.469 (0.391)	0.138 (0.131)
			Test	10.245	1.065	12.0	0.494 (0.363)	
		423	Spline	10.825	1.290	21.5	0.442 (0.373)	0.117 (0.117)
			Test	10.370	0.630	29.0	0.423 (0.356)	
	1.5	846	Spline	10.725	0.295	53.5	0.309 (0.235)	0.117 (0.112)
			Test	10.340	0.170	47.0	0.321 (0.258)	
		282	Spline	10.615	0.710	30.0	0.367 (0.338)	0.139 (0.152)
			Test	10.280	0.385	29.5	0.385 (0.337)	
	2	423	Spline	10.625	0.260	51.5	0.319 (0.255)	0.143 (0.121)
			Test	10.285	0.150	43.0	0.370 (0.323)	
		846	Spline	10.800	0.020	80.5	0.237 (0.232)	0.130 (0.125)
			Test	10.430	0.000	54.5	0.296 (0.259)	
12	1	282	Spline	10.650	0.230	52.5	0.309 (0.264)	0.144 (0.117)
			Test	10.400	0.115	47.5	0.331 (0.319)	
		423	Spline	10.735	0.080	71.5	0.279 (0.255)	0.139 (0.119)
			Test	10.385	0.025	52.5	0.286 (0.241)	
	1.5	846	Spline	10.890	0.015	87.5	0.226 (0.194)	0.177 (0.169)
			Test	10.540	0.000	60.0	0.304 (0.238)	
		282	Spline	16.940	2.245	3.5	0.415 (0.266)	0.081 (0.071)
			Test	15.965	1.235	3.0	0.484 (0.301)	
	2	423	Spline	16.845	1.505	12.5	0.405 (0.331)	0.089 (0.071)
			Test	16.070	0.620	19.0	0.417 (0.256)	
		846	Spline	16.745	0.360	49.5	0.250 (0.210)	0.087 (0.076)
			Test	16.025	0.175	30.0	0.312 (0.260)	
1.5	282	423	Spline	16.840	1.035	27.0	0.336 (0.296)	0.103 (0.088)
			Test	16.215	0.430	30.0	0.316 (0.255)	
	423	282	Spline	16.690	0.295	51.0	0.258 (0.212)	0.106 (0.108)
			Test	16.030	0.110	33.0	0.325 (0.234)	

D	S	N	Method	C	IC	CME (%)	LSA-MRMSE (MAD)	Oracle-MRMSE (MAD)
		846	Spline	16.800	0.030	81.0	0.204 (0.172)	0.115 (0.096)
			Test	16.130	0.005	45.0	0.281 (0.220)	
2		282	Spline	16.725	0.380	52.0	0.213 (0.197)	0.097 (0.083)
			Test	15.995	0.140	34.0	0.290 (0.244)	
		423	Spline	16.615	0.085	64.0	0.241 (0.186)	0.106 (0.104)
			Test	16.000	0.015	38.5	0.311 (0.257)	
		846	Spline	16.805	0.000	83.0	0.177 (0.154)	0.099 (0.090)
			Test	16.175	0.000	45.0	0.270 (0.252)	