Kernel Dimension Reduction with Missing Data

Ziyu Zhou

Supervisor: Prof. Kerby Shedden

An honors thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science (Honors Statistics) at the University of Michigan, 2024

April 2024

Abstract

Kernel dimension reduction (KDR), a form of sufficient dimension reduction (SDR), is a framework for identifying potentially nonlinear multivariate relationships between high-dimensional predictors X and outcomes Y, both of which may be multivariate. Here we propose a way to accommodate missing data in either the predictors or the outcomes, enabling KDR to be applied in a much broader range of settings. We cast the problem as that of predicting the missing elements of the kernel matrices using their conditional expected values given all observed data, based on an auxiliary model. We present simulation studies showing that our method is computationally tractable for moderate-sized data sets and has good statistical performance. To aid in interpretation of the nonlinear sufficient predictors, we use Multivariate Adaptive Regression Splines (EARTH/MARS) to estimate the unknown link functions. We illustrate the approach by presenting an analysis of longitudinal data of height for age z-score (HAZ) and systolic blood pressure (SBP) in a sample of people from the Dogon population of Mali.

1 Introduction

Multivariate regression of potentially high dimensional data remains a methodological challenge for data scientists, especially when aiming to accommodate substantial nonlinearity in the conditional mean relationship. As a motivating example, consider the association between height and systolic blood pressure (SBP) in an under-nourished population, where height is a proxy for nutrition, and SBP in early adulthood is an indicator of cardiovascular health. When both HAZ and SBP are measured over time, we can view the data as following continuous latent stochastic processes (random continuous functions) which are infinite dimensional. Recently-developed methods for kernel dimension reduction [12], exploiting methods and theories from Reproducing Kernel Hilbert Spaces (RKHS), naturally accommodate multivariate and process-valued data. This method provides a promising approach for data analysis in our setting, and is the basis of this thesis.

Kernel Dimension Reduction utilizes kernels [8] to capture pairwise relationships among observations, separately for the predictors and the responses. This in turn yields a "kernelized" estimator of the Sliced Inverse Regression (SIR) operator $\mathrm{Cov}E[X|Y]$, whose dominant eigenfunctions can be viewed as "sufficient nonlinear predictors" for X, that capture all information in X relevant for $P(Y|X)$. An important feature of this approach is that both X and Y may be vectors, possibly of infinite dimension, and in particular there is no need for Y to be scalar as is the case in many other approaches to regression analysis. This foundational kernel-based SDR method is described in detail in Section 3 below.

Data from human longitudinal studies are usually sporadically observed, so that the data are missing for all but a finite number of measurement occasions. Thus, the methods of KDR, which require computing the kernel at each pair of X values, cannot be directly applied. Devising and assessing a method for overcoming this methodological challenge is the main goal of this thesis. At a high level, our approach involves specifying a working model for the data (X and Y separately), and imputing the missing values of the kernel matrix as their means conditioned on all observed data.

An outline of this thesis is as follows. In Section 2 we briefly review the human biology of height and health and some classical methods for dimension reduction regression with scalar responses and finite-dimensional predictors. Then in Section 3 we review recent work on kernel dimension reduction (KDR) approaches to nonparametric regression. Section 4 develops our proposed approach for employing kernel dimension reduction regression in settings where there may be missing values in either the predictors or the outcomes. In Section 5 we rigorously assess the enhanced KDR methodology using simulated datasets. In Section 6 we apply the enhanced KDR method to data from a longitudinal study of blood pressure in relation to height in the Dogon of Mali. Our goal is to delineate the relationship between height in youth and adult systolic blood pressure, both of which are assessed repeatedly over time. We will consider the plausibility of our findings in this dataset, and assess the sensitivity of the results to the modeling choices and tuning parameters.

2 Background

2.1 Human biology of height and health

Height being related to the risk of disease and mortality was first noted in the late 19th century [5]. Data from the early 20th century, particularly from the insurance industry, suggested that taller individuals generally had longer lifespans compared to shorter individuals [2]. The correlation between height and various health conditions, such as Alzheimer's disease, cardiovascular diseases, and different forms of cancer, has been extensively studied. A comprehensive investigation utilizing both epidemiological and genetic methodologies assessed adult height in connection with 50 diseases, concluding that height was correlated with 32 diseases, while genetically influenced height had associations with 12 diseases [6].

A pioneering study by Gertler et al. in 1951 revealed that young men who were at risk for coronary artery disease were, on average, about 5 centimeters shorter than those not at risk [4]. This observation was corroborated by subsequent research from Paffenbarger and Wing [9], who found that university students who suffered strokes

were typically 2-3 centimeters shorter than those who did not. While there has been considerable research exploring the connection between blood pressure and height in more developed regions of the world [1, 10], there is a lack of data from underdeveloped regions. The data analyzed below in Section 6 are from an under-resourced population in Mali. In this context, the relationship between height and blood pressure may predominantly reflect the consequences of undernutrition, rather than overnutrition and obesity. Active debate remains around the consequences of childhood undernutrition. One point of view is that smaller children grow up to become smaller adults, with consequently lower blood pressure. Another perspective is that undernutrition in childhood is a risk factor for higher adult blood pressure, as undernourished children may suffer from developmental abnormalities such as lower kidney nephron density and impaired vasculature. The Dogon longitudinal study data provide a unique resource for studying the relationship between childhood growth and adult blood pressure in a generally undernourished population.

2.2 Dimension Reduction Regression

The increasing prevalence of high-dimensional data in numerous fields calls for effective dimension reduction strategies for data analysis. Sufficient Dimension Reduction (SDR) is a leading set of techniques in this realm, providing an array of methods that simplify the data while preserving essential information. Conventionally this "simplification" involves a linear dimension reduction, but recently methods using nonlinear dimension reduction have been developed, and that is what we employ here.

A key method among the classical approaches to dimension reduction regression is Sliced Inverse Regression (SIR) [7], which identifies the effective dimension reduction subspace through examination of the nested moment matrix $M \equiv \mathrm{Cov}E[X|Y]$. Another interpretation of this approach is that it conducts an inverse regression of each predictor on the response variable, thereby circumventing the curse of dimensionality. In doing so it identifies a low dimensional affine space containing the inverse regression function $E[X|Y]$.

For a scenario involving a p-dimensional predictor variable X and a univariate response Y, SDR distills a more manageable representation of X. This is generally achieved through linear combinations expressed as $\beta^T X$, where β is a $p \times d$ matrix whose columns are vectors $\beta_1 \dots \beta_d$, and where $d < p$. The identifying condition for SDR is as below:

$$Y \perp\!\!\!\perp X \mid \beta_1^T X, \dots, \beta_d^T X \tag{1}$$

This condition ensures that $\beta^T X$ encompasses all the regression information necessary

to explain Y given X, effectively reducing the dimensionality as long as d is small compared to p. In the classical SIR algorithm, the vectors β_j are estimated as the dominant eigenvectors of an orthogonalized version of M.

3 Kernel Dimension Reduction

The Kernel Dimension Reduction method that forms the basis of this work was presented in 2022 [12], and derives from Sliced Inverse Regression [7]. It employs the "kernel trick" along with ideas from reproducing kernel Hilbert spaces. This results in a nonlinear reduction of X to functions $f_j(X), j = 1, \dots d$, such that

$$Y \perp\!\!\!\perp X \mid f_1(X), \dots, f_d(X). \tag{2}$$

Although the reduced variates f_j are nonlinear functions of X, they are estimable using linear methods applied to kernel matrices, analogous to the familiar kernel ridge regression (KRR) method. The use of kernels has several advantages, one being that the domains of X and Y need not be Euclidean spaces, as long as a suitable kernel function can be constructed for their domains. This allows, for example, X and Y to lie on Riemannian manifolds, or, as in our case, to be projections of stochastic processes to finite sets of observation occasions.

To begin, kernel functions κ_X and κ_Y are selected, with the squared exponential radial basis kernel being a popular choice. Note that this kernel, like nearly all kernels, involves selection of a bandwidth parameter, which will be explored in our simulation studies and data analyses below. Subsequently, matrices $K_X = (\kappa_X(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and $K_Y = (\kappa_Y(Y_i, Y_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ are calculated. The centering matrix, denoted as $Q = I - n^{-1}1_n 1_n^T$, where 1_n is vector of ones, is used to compute centered versions of these matrices, resulting in $G_X = QK_XQ$ and $G_Y = QK_YQ$.

Next, the coordinate representation of the sample metric Sliced Inverse Regression (SIR) operator, Λ_{SIR}, is calculated along with a ridge-regularized version to enhance numerical stability. This is represented as

$$\Lambda_{\text{SIR}} = (G_X + \tau_1 I_n)^{-1} G_Y (G_Y + \tau_2 I_n)^{-1} G_X \tag{3}$$

where τ_1 is defined as $c \cdot \phi_1(G_X)$, with $\phi_1(G_X)$ being the largest eigenvalue of the designated matrix, and $c = 0.2$. Similarly, τ_2 is defined as $c \cdot \phi_1(G_Y)$.

To estimate the range of Λ_{SIR}, eigen-decomposition is performed on its coordinates as given in equation (3). By identifying the d leading eigenvectors, $v_1, \dots, v_d$ of $(\Lambda_{SIR})(\Lambda_{SIR})^T$, the sufficient predictors for an observation X within are estimated

as $v_1^T Q k_X(X), \ldots, v_d^T Q k_X(X)$, where $k_X(X) = (\kappa_X(X, X_1), \ldots, \kappa_X(X, X_n))^T$.

4 A method for accommodating missing data in Kernel Dimension Reduction

Missing values in datasets pose a considerable challenge and are especially prevalent in studies involving field data for human populations. If we view our data in idealized form as a continuous stochastic processes, nearly all data will be missing for each individual. In Section 6 below, we will consider data from the Dogon population of Mali, treating height and blood pressure as finite sequences of longitudinal measurements of infinite dimensional processes. These data will be used to illustrate our approach to handling missing data in KDR analysis.

A key step of the KDR framework is calculation of the distances between pairs of X variables and between pairs of Y variables. Missing values complicate the computation of these distances. To address this, we propose using an imputation method for the squared distance matrix $E[||X_i - X_j||^2 \mid X_i^{\text{obs}}, X_j^{\text{obs}}]$, based on a working model for $P(X)$ that is used to calculate conditional means. The same approach is used to impute pairwise distances for Y. This approach can make use of all observed data, without the need to discard any partial observations.

Let Z denote a random vector, and let I and J denote the index sets of missing and observed values within Z, respectively. Thus $Z[I]$ and $Z[J]$ are the subvectors of missing and non-missing values, respectively. We are interested in understanding the distribution of Z given the non-missing values $Z[J]$. We achieve this using a working model in which the conditional distribution $P(Z \mid Z[J])$ follows a normal distribution, $N(\theta, \Phi)$, where θ and Φ represent the mean and the covariance of this distribution, respectively. Under Gaussianity of Z, the parameters θ, Φ can be expressed in terms of the marginal moments of Z:

$$\begin{aligned} \theta &\equiv E[Z \mid Z[J]] \\ &= E[Z] + \text{Cov}(Z, Z[J]) \cdot \text{Cov}(Z[J])^{-1} \cdot (Z[J] - E[Z[J]]) \end{aligned} \tag{4}$$

This formula tells us the expected value of Z, given the observed data, can be adjusting the overall expected value of Z based on how the missing and observed values relate to each other (their covariance) and how the observed values themselves vary from their expected value.

The covariance Φ of the conditional distribution reflects how the values of Z vary

with respect to each other:

$$\begin{aligned}\Phi &\equiv \mathrm{Cov}(Z \mid Z[J]) \\ &= \mathrm{Cov}(Z) - \mathrm{Cov}(Z, Z[J]) \cdot \mathrm{Cov}(Z[J])^{-1} \cdot \mathrm{Cov}(Z[J], Z)\end{aligned} \tag{5}$$

This equation adjusts the overall covariance of Z by removing the part that can be predicted from the observed values, again taking into account the relationship between missing and observed data.

For imputation purposes, take the X_i to be exchangeable, and let $Z = (X_i', X_j')' \in \mathbb{R}^{2q}$. Take Z to follow a normal distribution with mean $(E[X]', E[X]')'$ and covariance matrix

$$\mathrm{Cov}(Z) = \begin{pmatrix} \mathrm{Cov}(X) & 0 \\ 0 & \mathrm{Cov}(X) \end{pmatrix}.$$

Let $B = [I_{p\times p} \ -I_{p\times p}]$, so that $BZ = X_i - X_j$. The distribution of BZ given the observed data also follows a normal distribution, but with modified parameters,

$$P(BZ \mid Z[J]) = N(B\theta, B\Phi B'). \tag{6}$$

The expectation of the squared norm of BZ, which is the expected square distance between X_i and X_j, given the observed data, can be computed as:

$$\begin{aligned}E[||X_i - X_j||^2 \mid X_i^{\mathrm{obs}}, X_j^{\mathrm{obs}}] &= E[||BZ||^2 \mid Z[J]] \\ &= \|B\theta\|^2 + \mathrm{tr}(B\Phi B').\end{aligned} \tag{7}$$

This formula calculates the expected squared distance by summing up the squared expected values of the transformed variables $(B\theta)_i^2$ and their variances $(B\Phi B')_{ii}$. This sum provides a measure of how spread out the transformed points are, taking into account both the location and spread of the underlying normal distribution shaped by the observed data.

The calculations above require estimates of the marginal moments $E[Z]$ and $\mathrm{Cov}(Z)$, for which we use the methods implemented in the R package *fastimputation*. We note that other methods for estimating marginal means and covariance matrices in the presence of missing data exist, and moreover a non-Gaussian working model could be employed, although this would considerably complicate the estimation of marginal moments.

Based on (7), we impute all values within the $n \times n$ kernel matrices K_X and K_Y,

then proceed with the KDR analyis as if these kernel matrices were calculated directly from the data.

5 Simulation Studies

In this section we use simulation to aid in understanding the performance of the KDR approach from three perspectives. We first consider the setting in which no data are missing, and we evaluate the recovery of the mean structure in terms of the nonlinear sufficient predictors. We focus especially on the effect of bandwidth specification, and on the roles of sample size, predictor and outcome dimension, and signal-to-noise ratio. We then consider a simulation study in which the data are missing completely at random, so that the missingness rate becomes another important factor to consider. Finally, we consider the cost of using kernel methods in a setting where the actual mean structure is linear. This allows us to compare the loss of efficiency due to kernelization, in comparison to an "oracle" linear model which is unbiased when the population mean structure is in fact linear.

5.1 Simulation study of KDR under nonlinear situation without missing values

The purpose of this simulation study is to evaluate the effectiveness of Kernel Dimension Reduction (KDR) in enhancing regression analysis by distilling predictor variables into two sufficient predictors, denoted as z_1 and z_2. These predictors are then employed to predict the dependent variable, Y, with an emphasis on preserving essential regression features that provide insight into the relationship between X and Y. For the simulation, a Gaussian radial basis function kernel is selected, defined by $k(X_i, X_j) = \exp(-\sigma||X_i - X_j||^2)$, where various scale values σ are evaluated through the following simulation process.

The study begins with the generation of a predictor matrix X of size $n \times p$, where n represents the number of observations and p indicates the dimension of predictors. The elements of X are drawn independently from a standard normal distribution. A response matrix Y with dimension $n \times 2$ is subsequently constructed. The expected responses $E[Y_1]$ and $E[Y_2]$ are modeled as $E[Y_1] = (1+X_1)^2$ and $E[Y_2] = (1+X_1+X_2)^2$, respectively. To maintain a controlled signal-to-noise ratio, a predetermined R^2 value is utilized to establish the residual variance for the response variables Y. Specifically, for homoscedastic errors,

$$
\begin{aligned}
R^2 &= \mathrm{Var}(E[Y|X])/(E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])) \\
&= \mathrm{Var}(E[Y|X])/(\tau^2 + \mathrm{Var}(E[Y|X])) \qquad (8)
\end{aligned}
$$

and $\mathrm{Var}(E[Y|X])$ can be estimated directly from the expected values in our simulation studies. We can rearrange the expression (8) to obtain the additive error variance

$$\tau^2 = \mathrm{Var}(E[Y|X]) \cdot (1 - R^2)/R^2. \qquad (9)$$

Upon the completion of this setup, we employ the KDR method to extract the sufficient predictors, denoted as z_1, z_2, from the training data. Using these predictors, we then apply Multivariate Adaptive Regression Splines (MARS) [3] to model and predict each component of Y in a sample of validation data having the same size as the training data. MARS is capable of estimating nonlinear and non-additive conditional mean structures, and the residual mean squared error from MARS should approximate the residual error variance, which is a known quantity in the setting of a simulation study.

To evaluate the efficacy of our methodology, we compute the correlation coefficient R^2 for the predicted versus the actual values of Y. This assessment aims to measure the agreement with the predetermined R^2 value set for Y, thereby offering a quantitative measure of the accuracy achieved by the KDR and MARS models in comparison to established benchmarks.

The simulation was executed a total of 100 times ($k = 100$), with each iteration involving 500 observations ($n = 500$) and adhering to a preset R^2 value of 0.75 across different values for the kernel scale parameter σ (the larger the value of σ, the smaller the bandwidth). This simulation protocol is designed to approximate the characteristics of the human biology dataset analyzed below in Section 6.

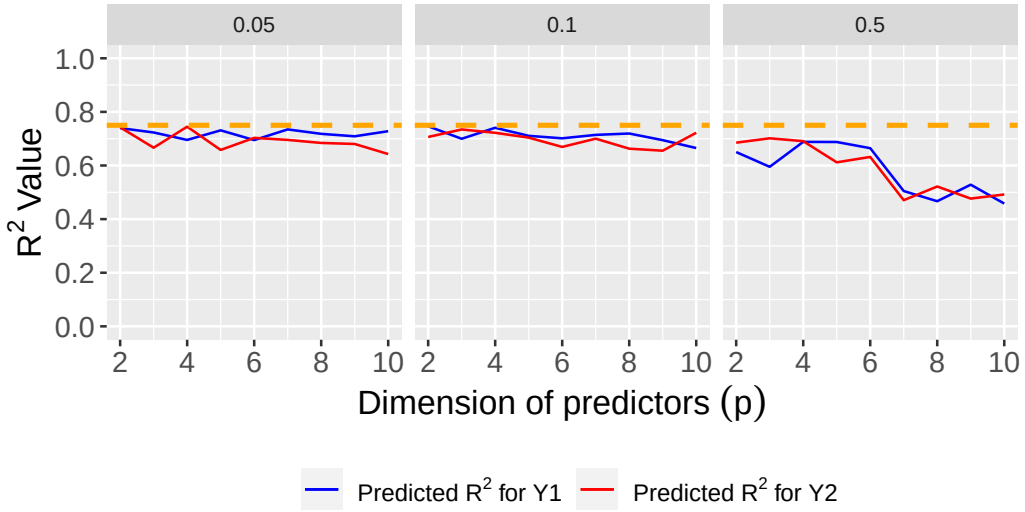


Figure 1: KDR performance with Gaussian radial basis function (RBF) kernel in a nonlinear situation without missing data across different kernel scale values σ and different dimension of predictors, measured by the predicted R^2, compared to the true $R^2 = 0.75$ in orange dashed line.

Figure 1 presents the predictive accuracy across various σ values and various predictor dimensions p. The depiction includes an orange dashed line that serves as a benchmark, symbolizing the ideal predetermined R^2 value of 0.75. The results overall show that the residual variance closely matches the population error variance, as long as the bandwidth σ is well-chosen. Specifically, at a near-oracle bandwidth of $\sigma = 0.05$, the R^2 values for both Y_1 and Y_2 remain near the target value of 0.75. With a smaller than oracle bandwidth $\sigma = 0.1$, the results are effectively equivalent to when $\sigma = 0.05$, showing a degree of robustness in the bandwidth selection. However when the bandwidth is too small, such as setting $\sigma = 0.5$, the R^2 values for both response variables Y_1 and Y_2 diminish as the dimension increases, although the method continues to work well in lower dimensions. Taken together, these findings highlight the method's robustness in maintaining a stable level of predictive accuracy, even as the challenge of higher dimensionality introduces a greater volume of non-essential information. All these results depend on the bandwidth being well-chosen.

5.2 Simulation study of KDR under nonlinear situation with missing values

Using a Gaussian radial basis kernel, denoted $k(X_i, X_j) = \exp(-\sigma||X_i - X_j||^2)$, the computation of the distance between the predictor vector X for each pair of observations is a key step. However, the presence of missing values in the X vectors complicates the direct application of KDR. To circumvent this issue, the approach described in Section

4 is incorporated to impute the distance matrix, thereby rendering the data amenable for subsequent use in KDR. In this simulation study, the dataset is generated using the method described in Section 5.1. Missing values are randomly distributed across the entire dataset, with each matrix, X and Y, having 80 percent of its values present and 20 percent missing. After assigning these missing values, any row in both the X and Y matrices that contains only missing values is removed, while rows containing at least one non-missing value are retained. Subsequently, our imputation approach is applied to compute the kernel matrices for X and Y separately. Once these matrices are established, they are introduced into the KDR framework, allowing for the construction of sufficient operators for the predictor variables X. Prediction of each component of Y is then performed using MARS. The effectiveness of this approach is evaluated by comparing the recovered R^2 of the data set after missing value imputation with the true predetermined R^2 of the original complete data set. This comparison aims to demonstrate the efficacy of the missing value algorithm within the KDR framework in preserving the integrity of the data and the accuracy of subsequent predictions.

The simulation was executed a total of 100 times ($k = 100$), with each iteration involving 500 observations ($n = 500$) and adhering to a preset R^2 value of 0.75 across different σ values.

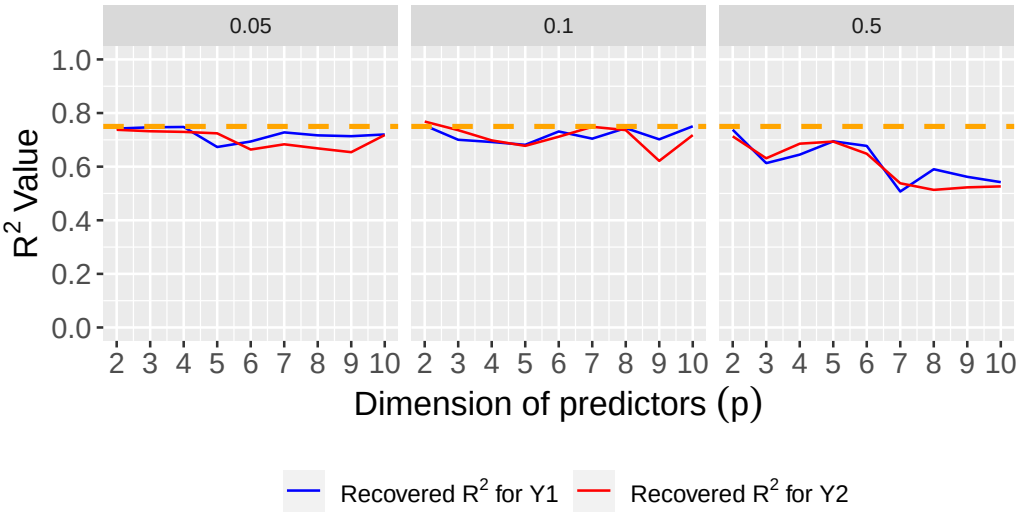


Figure 2: KDR performance with Gaussian RBF kernel in nonlinear situation without missing data across different σ values under different dimension of predictors, measured by the predicted R^2, compared to the true $R^2 = 0.75$ in orange dashed line.

Figure 2 displays the predictive accuracy across various kernel scale parameters (σ) and predictor dimensions (p). An orange dashed line serves as a benchmark, representing an ideal R^2 value of 0.75. The results strongly resemble those shown in Figure 1, suggesting that we were able to overcome the presence of missing data without sacrificing predictive accuracy.

5.3 Simulation study of KDR under linear situation without missing values

In the preceding simulation studies, we posited a nonlinear association between the predictor variables X and the response variable Y, for which we employed KDR and MARS to forecast Y. However, when the relationship between X and Y is linear, we have the opportunity to evaluate the efficiency of KDR in relation to conventional multivariate ordinary linear regression models (multivariate OLS).

We begin by generating the predictor variables X $(n \times p)$ and the error terms e $(n \times 2)$. These are created using the multivariate normal distribution. Specifically, X is generated with mean of 0 and exchangeable covariance of 0.6, while e is produced with a mean of 0.5 and exchangeable covariance of 0.2. We then define a matrix B $(p \times 2)$ which follows the uniform distribution on $(-0.3, 0.3)$ to synthesize the response variable Y $(n \times 2)$ via the relationship $Y = BX + e$. Subsequently, we estimate the matrix $\hat{B}$ using the standard least squares prediction formula $\hat{B} = (X^T X)^{-1} X^T Y$. The predicted values of Y, denoted as $\hat{Y}$, are obtained by calculating $\hat{Y} = \hat{B}X$.

Next, we generated a testing dataset using a shifted version of the covariate model used to produce the training data. The use of such a *covariate shift* introduces a greater challenge in that a certain degree of extrapolation is taking place. Specifically, we shifted the mean of all columns of X by 0.5 and scaled the variances of the predictors by a factor 1.5. The true responses for the test set were then derived using the formula $Y_{\text{test}} = X_{\text{test}} B + e$. Predictions for Y_{test} were generated by applying the coefficient matrix $\hat{B}$, obtained from the linear regression model trained on the original dataset, to X_{test} through $\hat{Y_{\text{test}}} = \hat{B} X_{\text{test}}$. Subsequently, we calculated the Mean Squared Error (MSE) for both response variables Y_1 and Y_2 within the multivariate linear regression framework.

For a comparative analysis, we employed KDR and MARS techniques on the same training data and made predictions on X_{test}. We then examined the ratio of MSEs between these methods and the multivariate linear regression, specifically calculating

$$\frac{\text{MSE(KDR + MARS)}}{\text{MSE(Multivariate Linear Regression)}}$$

By comparing this ratio against the benchmark of 1, we sought insights into the efficiency of KDR under conditions akin to linear scenarios. This comparative metric sheds light on the relative benefits or costs of employing KDR in situations where a linear model might be adequate, thus facilitating a more informed selection of predictive methodologies based on the intrinsic characteristics of the data.

We conducted the simulation, running it a total of $k = 300$ times, each time with

$n = 500$ observations. The KDR reduce the dimension of the X to two, with sufficient predictors z_1 and z_2.

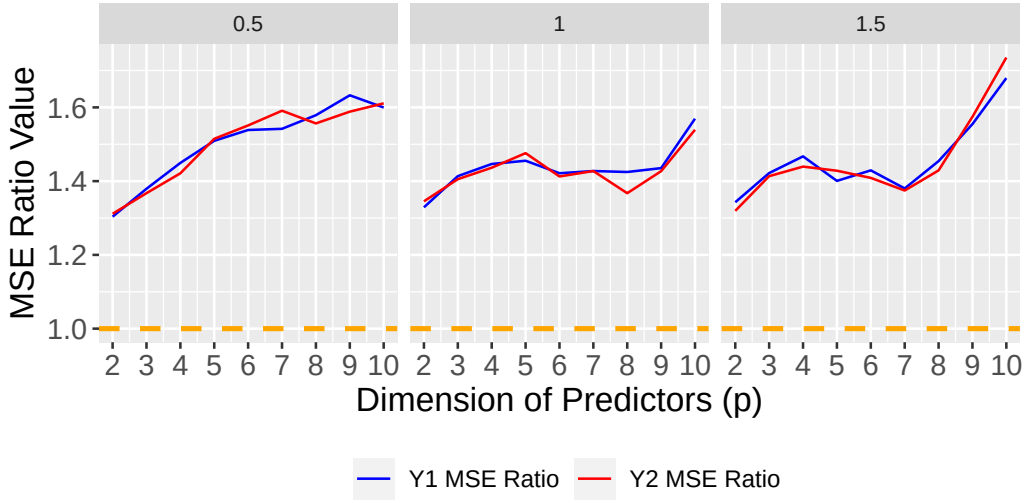


Figure 3: KDR performance with Gaussian RBF kernel in linear situation without missing data cross different σ values under different dimension of predictors, compared to the benchmark MSE ratio = 1 in orange dashed line.

According to the results shown in Figure 3, the impact of varying the bandwidth σ on the mean squared error (MSE) ratio for responses Y_1 and Y_2 is clearly depicted. When $\sigma = 0.5$, the MSE ratio for both responses starts at 1.3 and gradually increases with the number of predictors, reaching 1.6 as the dimensionality expands. For $\sigma = 1$, the MSE ratio remains relatively stable at approximately 1.45 as the predictor dimension increases from 2 to 9, but rises to 1.55 when the number of predictors reaches 10. At a bandwidth of $\sigma = 1.5$, the MSE ratio maintains stability at 1.4 for predictor dimensions between 2 and 7; however, it escalates to 1.7 as the dimension extends from 7 to 10. These observations indicate that with an appropriately selected bandwidth, the MSE of KDR is no more than 2 times greater than that of the "oracle" multivariate OLS, with a narrower gap in lower dimensions.

6 Blood pressure and anthropometry in the Dogon Population of Mali

We illustrate the trajectory regression approaches discussed above using data from a study of human growth and health outcomes. The central question is whether childhood undernutrition, reflected in growth curves based on childhood measures of height, associate with trajectories of blood pressure in adulthood. Since these function-on-function regression relationships are likely to be non-linear, and are based on data that

were measured repeatedly for each individual, but at irregularly spaced time points, this dataset provides a suitable illustration of the methods developed above.

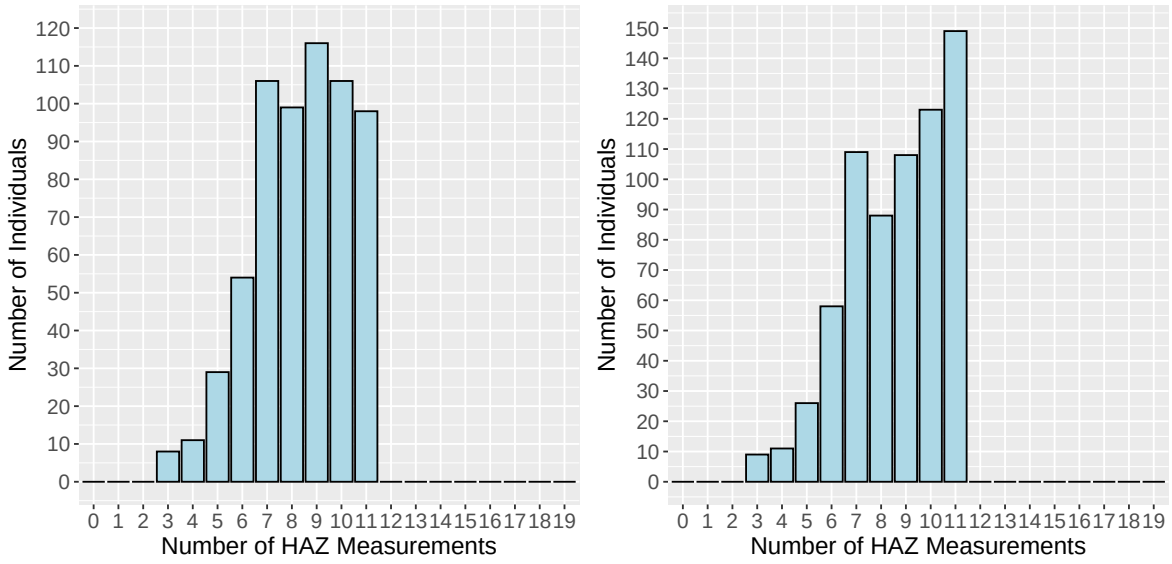
6.1 Data Description

The Dogon Longitudinal Study (DLS) began in the late 1980's in the Bandiagara escarpment region of Mali, west Africa, led by Professor Beverly Strassmann of the University of Michigan [11]. Between 1998 and 2002, approximately 1700 "F1" individuals were recruited, consisting of children born in nine villages between 1993 and 2000. These individuals have been followed longitudinally and the vast majority of surviving individuals continue to be followed as of 2024. Anthropometry, including the height-for-age Z-score (HAZ) was collected on all measurement occasions, and systolic blood pressure (SBP) was collected for most individuals older than age 12.

These data are clustered by individual, with repeated measures for both HAZ and SBP taken over multiple years of follow-up. These repeated measures are used to define the kernel-based similarity matrices K_X and K_Y that in turn are used to identify the sufficient dimension reduction subspace.

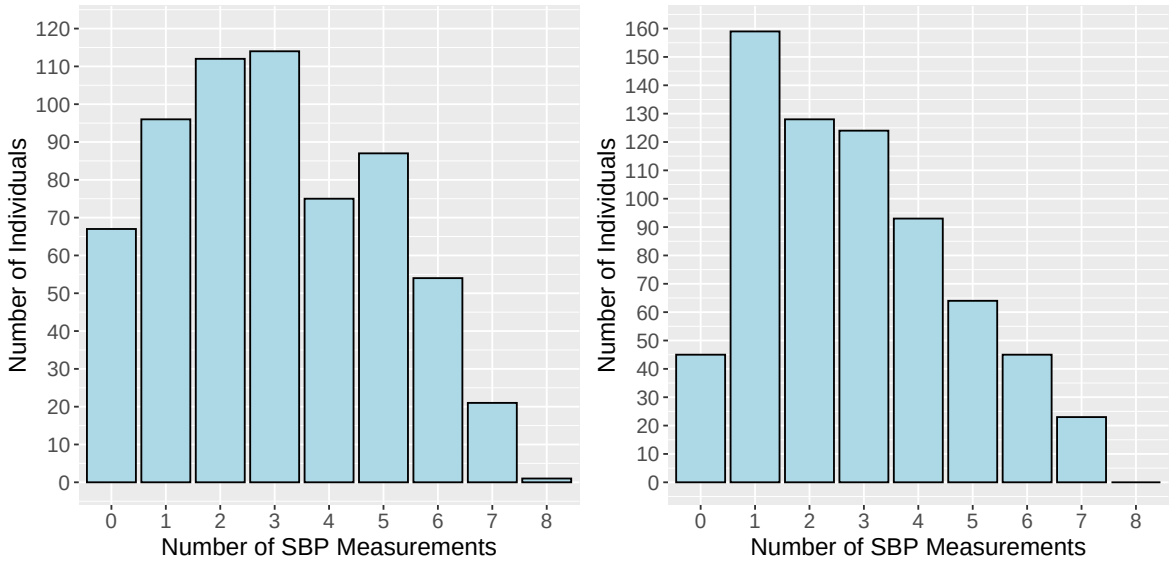
In this study, the dependent variable is systolic blood pressure (SBP) and the independent variable is the height-for-age Z-score HAZ. Both quantities are measured repeatedly for most subjects, with typically a year or more between consecutive measurement occasions. For this analysis, we use HAZ measures taken from birth up to 18 years of age, and blood pressure measures taken between the ages of 19 and 26. Note that this implies that the HAZ measures always precede the SBP measures in our analysis.

Figure 4 shows the distribution of the number of HAZ measurement occasions per individual, and Figure 5 shows the distribution of the number of SBP measures per individual.



(a) Females (b) Males

Figure 4: Distribution of HAZ measurement occasions per individual from age 0 to 18.



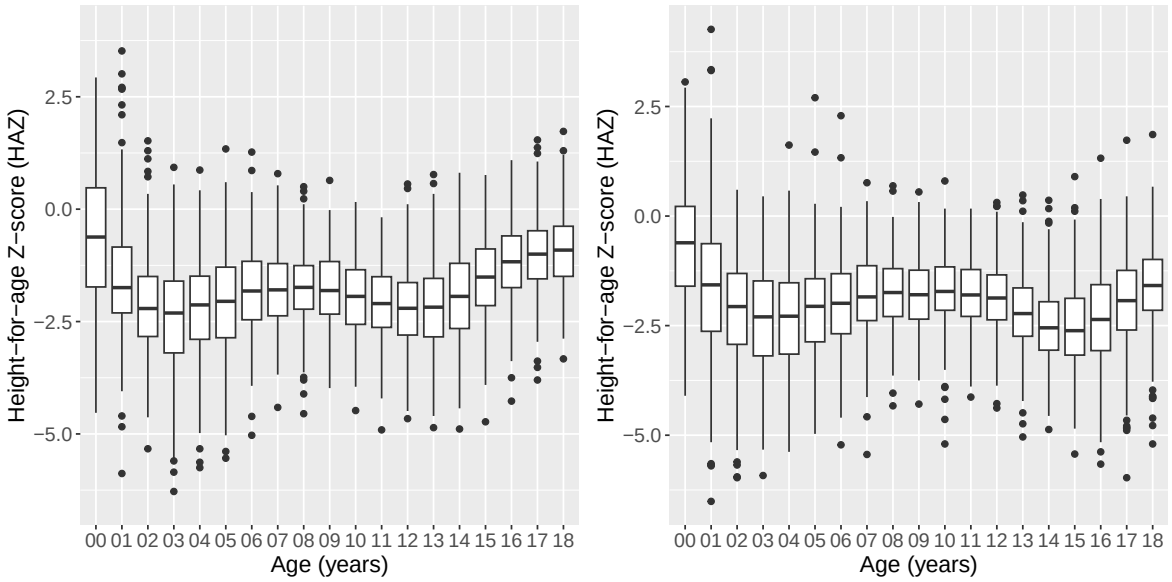
(a) Females (b) Males

Figure 5: Distribution of SBP measurement occasions per individual from age 19 to 26.

The initial participant pool comprised 627 females and 681 males. For the purposes of analysis, an evaluation of the dataset was conducted to locate any records where an individual's entire HAZ or SBP data were missing, as such instances lack the necessary information for any form of association-based analysis. Among the females, all individuals had at least one HAZ observation between birth to 18 years. However, 67 individuals had no available SBP data from ages 19 to 26 and were subsequently excluded from our study, leaving a final sample of 560 females for analysis. The same

process was carried out for the male participants, revealing that 45 individuals had no SBP data for the 19 to 26 age range. Following their exclusion, the final male sample size was 636 for subsequent analyses.

The height-for-age Z-scores are defined with respect to a reference population by the World Health Organization (WHO). They are not Z-scores for this population specifically, and in fact the Z-scores in our data tend to be negative since this population is relatively undernourished. The age-specific HAZ distributions are depicted separately for females and males in Figure 6. Dogon children are born at approximately median length, but rapidly fall behind the WHO reference distribution, reaching a nadir at around age 3. Notably, females exhibit a higher incidence of potential outliers in both the positive and negative directions between the ages of 1 to 4. Beyond the age of 13, there is an upward trend in the median HAZ values among females, surpassing the median values observed in their male counterparts within the same age range.

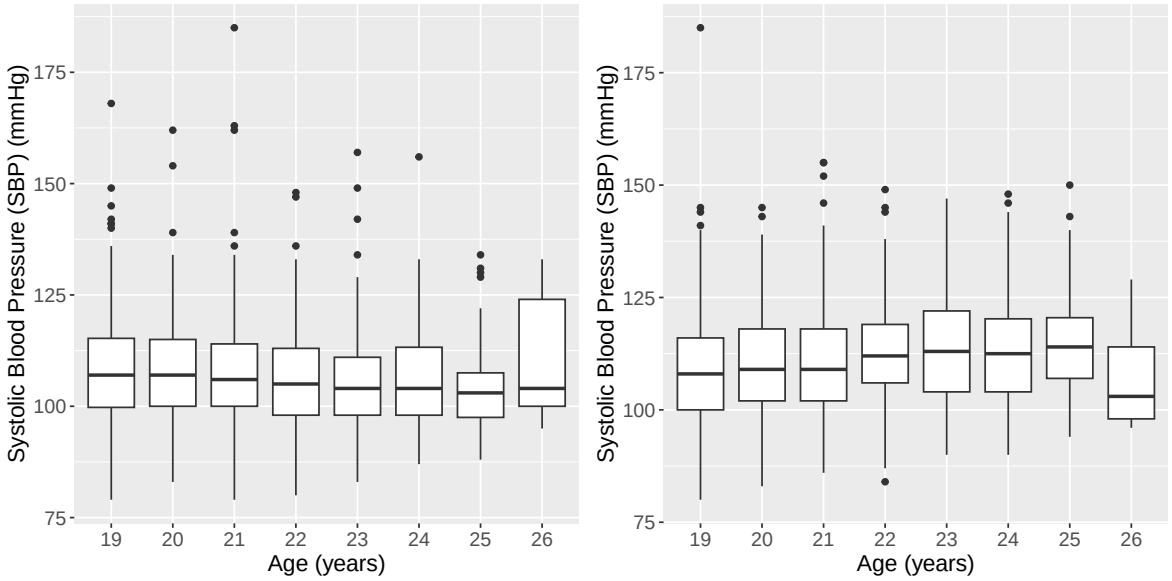


(a) Distribution of HAZ in females

(b) Distribution of HAZ in males

Figure 6: Distribution of HAZ from birth to 18 years old in females and males.

The range of systolic blood pressure (SBP) is delineated for females and males in Figure 7. For females, the median SBP remains relatively constant from ages 19 to 26. In contrast, males exhibit an ascending trend in median SBP from ages 19 to 25, with values exceeding those of females in the corresponding age bracket. Across most ages, the 25th percentiles of SBP values in females tend to be lower compared to those in males.



(a) Distribution of SBP in females

(b) Distribution of SBP in males

Figure 7: Distribution of SBP from ages 19 to 26 in females and males.

6.2 Kernel Dimension Reduction for partially observed functions

Since the Dogon longitudinal study is longitudinal, both the exposure (HAZ) and the outcome (SBP) are measured repeatedly over time. Our goal here is to conduct a *function on function* trajectory analysis, using the kernel dimension reduction methods discussed above. Doing so focuses the analysis on the shapes of the trajectories rather than on the individual observations, and accommodates the dependencies among the observations. Each subject has trajectories for HAZ and for SBP. We wish to proceed as above by constructing kernel matrices K_X and K_Y that capture the similarities between all pairs of subjects, for their HAZ and SBP data respectively. Since these measurements are made at distinct sets of ages for each subject, we cannot apply the kernel dimension reduction technique of Virta, Lee and Li directly [12]. Instead, we consider integer grids of ages 0 through 18 for HAZ and 19 through 26 for SBP, with unavailable observations treated as missing. We then proceed to apply kernel dimension reduction using the missing data imputation approaches discussed above.

In the subsequent analysis, we chose to utilize a Gaussian (squared exponential) radial basis function (RBF) kernel, which is expressed as

$$k(X_i, X_j) = \exp(-\sigma||X_i - X_j||^2). \tag{10}$$

Since the squared exponential RBF (10) is a function of the Euclidean distances between

pairs of observations, the imputation approach discussed in Section 4 can be used to provide complete kernel matrices K_X and K_Y.

The squared exponential RBF depends on a scale parameter σ. Note that in the parameterization (10), σ is inversely related to the bandwidth – that is, greater values of σ correspond to weights that decay more rapidly with distance. We begin with a single scale parameter since the HAZ variables are approximately standardized at each age by construction, and we standardized the SBP variables to have equal variance at each age. We also consider and ultimately adopt estimates based on differing scale parameters for HAZ and SBP.

6.3 Dimension Diagnostics

We use the approach developed in Section 4 to estimate kernel matrices K_X and K_Y for the partially observed HAZ and SBP data, respectively. These estimates are constructed separately for females and for males. Here we consider the extent to which the SIR operator Λ_{SIR} (3) is approximately a low rank matrix, for various settings of the kernel scale parameter σ. When Λ_{SIR} is approximately low rank, the non-linear dimension reduction approach is likely to yield informative results.

The singular values of Λ_{SIR} capture the extent to which regression information is concentrated in the dominant predictors. Let λ_j denote the j^{th} eigenvalue of Λ_{SIR}. We consider two canonical patterns of decay for the eigenvalues: power-law and exponential. We do not expect either of these to hold exactly throughout the range of eigenvalues, but these models can be informative benchmarks for interpreting the eigenvalues.

If the eigenvalues follow a power law distribution, they follow the pattern

$$\lambda_j = cj^{-\alpha}, \tag{11}$$

so that

$$\log(\lambda_j) = \log(c) - \alpha \log(j). \tag{12}$$

Thus, under a power law relationship, λ_j and j will exhibit a linear pattern when plotted in log space.

An alternative possibility is that the eigenvalues exhibit exponential decay

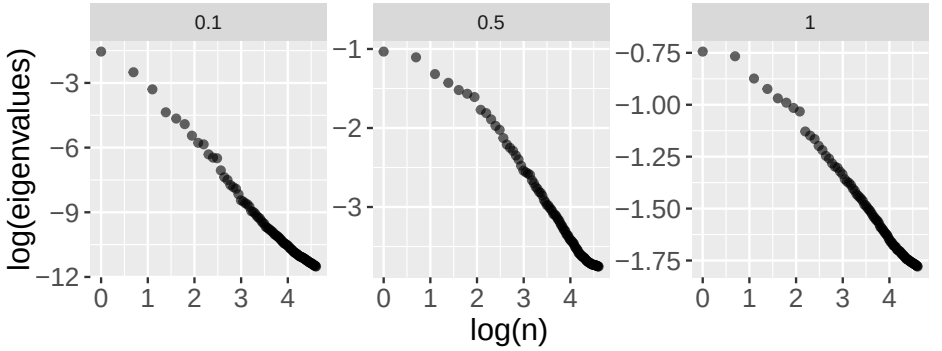
$$\lambda_j = ce^{-\alpha j} \tag{13}$$

so that

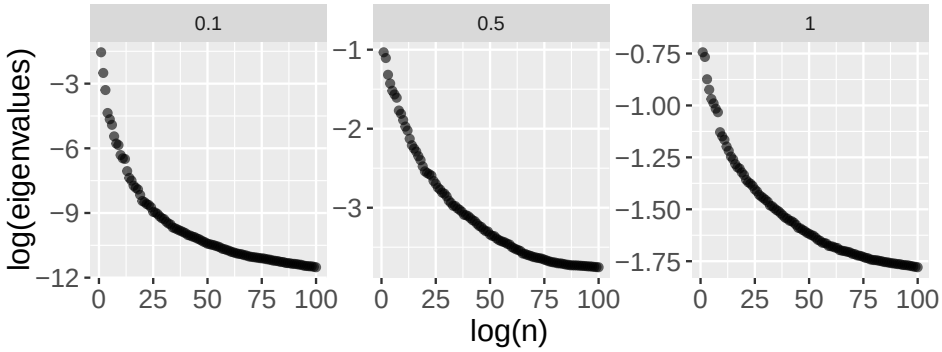
$$\log(\lambda_j) = \log(c) - \alpha j \tag{14}$$

Under an exponential relationship, λ_j and j will exhibit a linear pattern when plotted in semi-log space.

Figures 8 and 9 show plots of the eigenvalues of Λ_{SIR} for females and for males under different σ values, respectively. The exponential pattern, reflected in the two panel b plots, evidently does not fit well for either sex, for any bandwidths. The power-law behavior (panel a plots) is arguably a better fit. For large bandwidths (small σ) the relationship is quite linear, but becomes *biphasic* for the smaller bandwidths (larger σ). A biphasic pattern exhibits a shallower slope for the first few terms followed by a steeper slope thereafter. We notice that the eigenvalue patterns for females and males are quite similar. In the subsequent analyses involving the boy subpopulation, we use $\sigma = 0.1$ for constructing the Gaussian RBF kernel. For girls, we use $\sigma = 0.1$ for the HAZ kernel, but we modified the SBP bandwidth to $\sigma = 1/2$ to obtain smoother results.

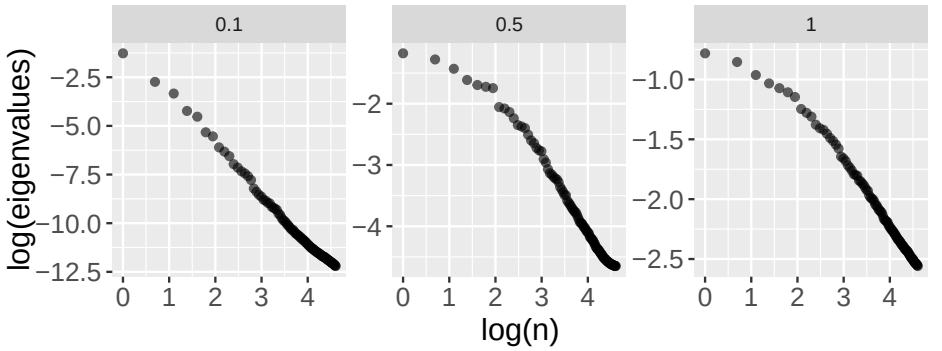


(a)

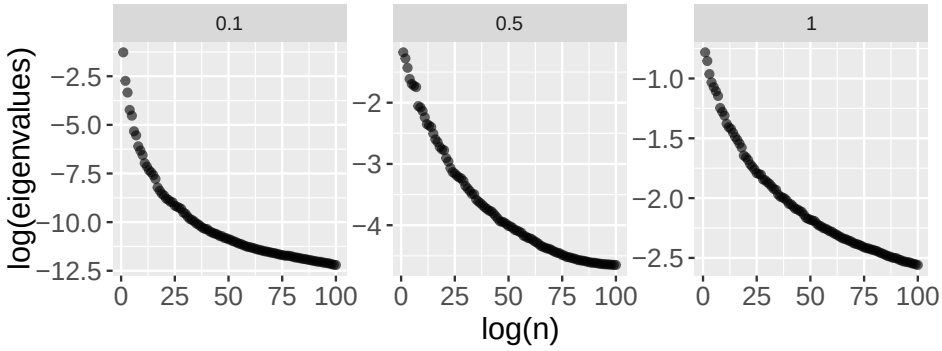


(b)

Figure 8: Diagnostic plots for eigenvalues of females, in log space (a) and semi-log space (b). The bandwidth parameter σ is shown at the top of each plot.



(a)

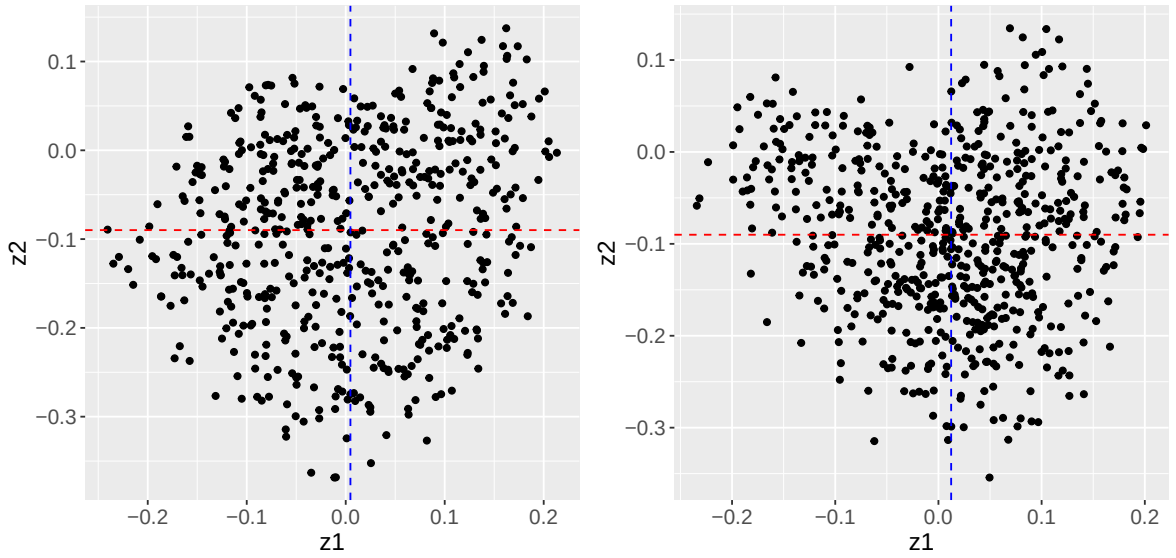


(b)

Figure 9: Diagnostic plots for eigenvalues of males, in log space (a) and semi-log space (b). The bandwidth parameter σ is shown at the top of each plot.

6.4 Investigation of KDR scores

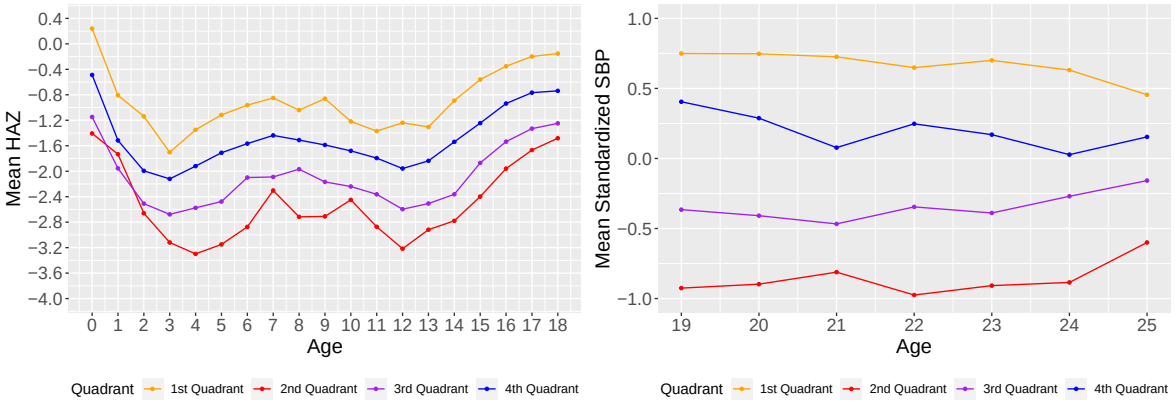
We focus on the leading two estimated nonlinear sufficient predictors, z_1 and z_2. These predictors are designed to encapsulate complementary information about the underlying trajectories, and to capture the associations between the HAZ and SBP trajectories. Figure 10 shows scatterplots of z_2 versus z_1 for females and for males. The lack of evident structure in these scatterplots suggests that the two extracted components of HAZ and of SBP are complementary as desired.



(a) Females (b) Males

Figure 10: Relationship between two sufficient predictors.

To understand how the nonlinear sufficient predictors z_1 and z_2 relate to the observable data, we stratified the observations into four quadrants based on the scores shown in Figure 10. For each quadrant and for both genders, we calculated the mean HAZ and SBP for all points whose sufficient predictors fall into a given quadrant, using available data at each age to obtain the means. In Figures 11 and 12 we plot these conditional mean HAZ and conditional mean standardized SBP trajectories against age. This approach allows us to examine more closely any trends or patterns that may be captured by the sufficient predictors.



(a) Mean HAZ by quadrant (b) Mean standardized SBP by quadrant

Figure 11: Comparative trends in mean HAZ and mean standardized SBP by quadrants among females.

Female subcohort: In the female subcohort (Figure 11), all four quadrant-mean HAZ

trajectories show a rapidly declining pattern from birth to age four. This largely reflects that fact that Dogon children are only slightly smaller than WHO norms at birth, but tend to rapidly fall behind WHO norms in the first few years of life. This is largely due to undernutrition and the prevalence of diarrheal diseases. Those children who survive to age 12 frequently return to near normal height by age 18, a phenomenon known as *catch-up growth*. Preceding this period of catch-up growth is a period from ages 9-12 when each subgroup remains flat or falls further behind the WHO standards.

Unlike the quadrant-mean HAZ trajectories, the quadrant-mean SBP values for females are nearly constant with respect to age, reflecting the fact that in females, blood pressure does not present systematic trends during early adulthood. The spread of these four quadrant-mean SBP trajectories is almost 2 SBP standard deviations, indicating that these four trajectories span most of the range of the data.

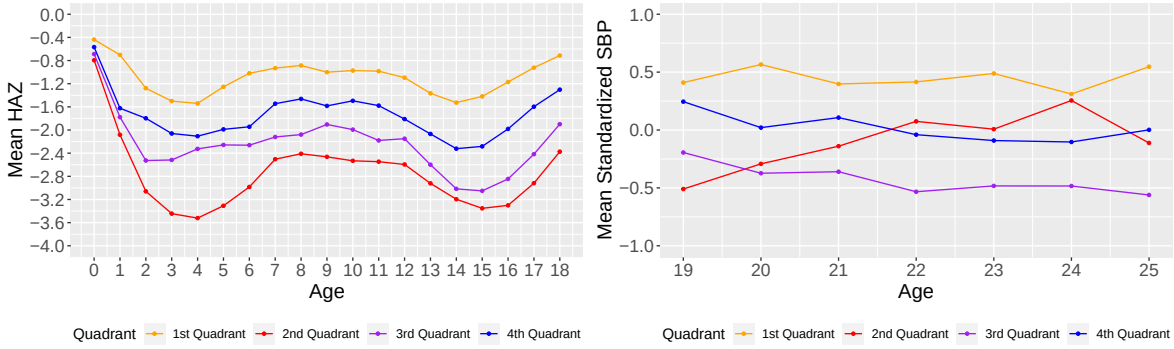
Our focus is on the relationship between HAZ and SBP trajectories. Since the left and right panels of Figure 11 are linked by color, we see that the quadrant-means exhibiting greatest (orange), second greatest (blue), third greatest (purple) and least (red) childhood height Z-scores also had the greatest, second greatest, third greatest and least adult blood pressures, across all ages. This suggests that greater childhood height in girls tracks into greater adult SBP. Of note is that the two quadrants with the lowest HAZ at ages 0-2 (quadrants 2 and 3) are essentially indistinguishable during this age period, but separate into distinct groups after age 2, with one of them (quadrant 2) exhibiting especially low HAZ values from ages 3-6.

Male subcohort: For the male subcohort (Figure 12), the quadrant-mean HAZ trajectories are quite similar to those of the females, with two notable differences. First, the shortest quadrant (2) has an even deeper nadir at around age 4 in boys compared to girls. Second, the boys have a prominent deficit with respect to WHO norms during the teenage years (which is much less discernable in the females), and begin catch-up growth later. Moreover, the boys are somewhat further behind WHO norms than the girls at age 18.

The quadrant-mean SBP trajectories for males are more diverse than those for females. One of the quadrants is stable with age, two are slightly declining, and one is substantially increasing. This reflects less stationarity in male blood pressure during this decade of life.

In terms of the relationship between HAZ and SBP, the quadrants with largest (orange), second largest (blue) and third largest (purple) childhood height Z-scores (Figure 12) also had the greatest, second greatest, and third greatest adult blood pressures. This suggests that for these subjects, greater childhood height tracks into greater adult SBP. However, the children with least height (red) have a markedly ascending pattern

of adult blood pressure. It appears that males experiencing severe childhood undernutrition have rapidly increasing SBP in adulthood, although the levels at any given age are not necessarily higher than those of males who did not experience childhood undernutrition. Nevertheless, the rapidly increasing trajectory could be a concern if it continues into subsequent decades of life. A possible mechanism for this association is that individuals experiencing severe childhood undernutrition may suffer developmental consequences to the kidneys and vascular tissues from their adverse childhood experiences.



(a) Trends in mean HAZ

(b) Trends in mean standardized SBP

Figure 12: Comparative trends in mean HAZ and mean standardized SBP by quadrants among males.

7 Discussion

In this study, we introduced and assessed an approach for nonparametric multivariate regression with data that may be missing at random in predictors and/or outcomes. The approach builds on recent innovations in kernel dimension reduction (KDR), and utilizes an auxiliary prediction model to impute the unobserved kernel elements. Following estimation of the sufficient predictors, a low-dimensional nonparametric regression can be carried out to obtain an estimate of the regression function.

We carried out simulation studies to assess the robustness of our approach with both complete and with partially missing data. As with most kernel methods, our approach requires selection of bandwidth parameters (σ), so we explored the role of bandwidths as well as different dimensions for the predictors and outcomes. The findings underscore the resilience of the KDR method, which maintains a stable level of predictive accuracy, even as dimensionality increases.

Furthermore, we conducted a comparison of kernel methods against an "oracle" linear model in scenarios where the underlying mean structure is inherently linear. This comparison helps quantify the costs associated with using kernel methods in a

setting where they are not needed. The results demonstrate that the cost of employing KDR is modest, with at most 1.6 times greater MSE when using the kernel method.

We illustrated the approach using data from a longitudinal study of anthropometry in children and subsequent blood pressure in adulthood, in the Dogon population of Mali. Stratifying on sex, we found that smaller children grow up to have lower blood pressure, with the notable exception being that the smallest boys have an increasing blood pressure trajectory as adults. This shows that smaller children tend to track into being smaller adults, and smaller adults have lower SBP. However the smallest boys exhibit a concerning pattern where potentially severe childhood undernutrition confers risk for cardiovascular health in adulthood, possible due to deficiencies in development. This pattern raises alarms about potential long-term health implications of inadequate childhood nutrition, suggesting that early nutritional interventions could be crucial in preventing future cardiovascular issues, especially in men.

Our analysis also revealed the nonlinear "factors" of childhood growth that are most associated with adult SBP trajectories. While height shortly after birth was approximately normal by WHO standards, both females and males exhibited very small stature in early childhood, followed by a phase of catch-up growth after the age of 12. This catch-up growth underscores the resilience of human growth trajectories and the potential for recovery from early growth deficits, however it could be that rapid catch-up growth is adverse for adult health outcomes, as suggested by our findings involving SBP.

By linking the trajectories of HAZ and SBP, our study not only provides insights into the specific health dynamics of the Dogon population but also contributes to the broader understanding of how early life growth patterns can influence long-term health outcomes. These findings are crucial for public health strategies aimed at mitigating the long-term consequences of childhood undernutrition and for designing interventions that target the critical windows of growth and development. Furthermore, taken as a case study, our analysis of the Dogon population indicates that dimension reduction, both kernel-based and more conventional approaches, can be an effective tool for analysis of complex datasets that arise in longitudinal observational studies.

References

[1] Brianna Bourgeois et al. "Associations between height and blood pressure in the United States population". In: *Medicine* 96.50 (2017), e9233.

[2] Louis I Dublin, Alfred James Lotka, and Mortimer Spiegelman. "Length of life: A study of the life table". In: *(No Title)* (1949).

[3] Jerome H Friedman. "Multivariate adaptive regression splines". In: *The annals of statistics* 19.1 (1991), pp. 1–67.

[4] Menard M Gertler, Stanley M Garn, and Paul D White. "Young candidates for coronary heart disease". In: *Journal of the American Medical Association* 147.7 (1951), pp. 621–625.

[5] Elise Juzda Smith. "Class, health and the proposed British anthropometric survey of 1904". In: *Social History of Medicine* 28.2 (2015), pp. 308–329.

[6] Florence Y Lai et al. "Adult height and risk of 50 diseases: a combined epidemiological and genetic analysis". In: *BMC medicine* 16 (2018), pp. 1–18.

[7] Ker-Chau Li. "Sliced Inverse Regression for Dimension Reduction". In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327. ISSN: 01621459. URL: http://www.jstor.org/stable/2290563.

[8] Jens Nilsson, Fei Sha, and Michael I. Jordan. "Regression on manifolds using kernel dimension reduction". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: Association for Computing Machinery, 2007, pp. 697–704. ISBN: 9781595937933. DOI: 10.1145/1273496.1273584. URL: https://doi.org/10.1145/1273496.1273584.

[9] RS Paffenbarger Jr and AL Wing. "Characteristics in youth predisposing to fatal stroke in later years". In: *The Lancet* 289.7493 (1967), pp. 753–754.

[10] Lulu Song et al. "Height and prevalence of hypertension in a middle-aged and older Chinese population". In: *Scientific reports* 6.1 (2016), p. 39480.

[11] B. I. Strassmann. "Cooperation and competition in a cliff-dwelling people". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.Suppl 2 (2011), pp. 10894–10901. DOI: 10.1073/pnas.1100306108.

[12] Joni Virta, Kuang-Yao Lee, and Lexin Li. "Sliced inverse regression in metric spaces". In: *Statistica Sinica* 32 (2022), pp. 2315–2337.