Published in final edited form as:

Biometrics. 2018 September ; 74(3): 834–844. doi:10.1111/biom.12866.

SILFM: Single Index Latent Factor Model Based on High-dimensional Features

Hojin Yang, Hongtu Zhu, and Joseph G. Ibrahim

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420, USA

Summary

The aim of this paper is to develop a single-index latent factor modeling (SILFM) framework to build an accurate prediction model for clinical outcomes based on a massive number of features. We develop a three-stage estimation procedure to build the prediction model. SILFM uses an independent screening method to select a set of informative features, which may have a complex nonlinear relationship with the outcome variables. Moreover, we develop a latent factor model to project all informative predictors onto a small number of local subspaces, which lead to a few key features that capture reliable and informative covariate information. Finally, we fit the regularized empirical estimate to those key features in order to accurately predict clinical outcomes. We systematically investigate the theoretical properties of SILFM, such as risk bounds and selection consistency. Our simulation results and real data analysis show that SILFM outperforms many state-of-the-art methods in terms of prediction accuracy.

Keywords

Dimension reduction; independent screening; latent factor model; prediction; regularized empirical risk

1. Introduction

We consider a high-dimensional prediction problem based on a set of n independent observations $\{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$, where $\mathbf{x}_i$ is a $p_x \times 1$ vector of all candidate features and y_i is an outcome variable, such as diagnostic status. Without loss of generality, we consider a nonparametric prediction model given by

$$y_i = f(\mathbf{x}_i) + \sigma_y \varepsilon_{iy} = f(x_{i1}, \dots, x_{ip_x}) + \sigma_y \varepsilon_{iy}, \quad (1)$$

where $f(\dots)$ is a generic link function and $\varepsilon_{iy} \sim \mathcal{N}(0, 1)$. In the classical setting with $n \gg p_x$, various parametric and nonparametric regression models have been developed to find a linear/nonlinear combination of predictors which can efficiently characterize y_i (Hastie et al., 2009; Clarke et al., 2009; Zhang and Singer, 2010). Although there is a large literature on

the development of supervised learning methods for prediction problems (Hastie et al., 2009; Clarke et al., 2009; Zhang and Singer, 2010), most of them suffer from the curse of dimensionality due to diverging spectra and noise accumulation in the high dimensional feature space with $p_x \gg n$ (Bickel and Levina, 2004; Fan and Lv, 2008). For instance, in medical imaging studies, it is interesting to study the predictive value of image signals at millions of locations (or voxels) ($p_x \sim 10^6$) for clinical outcomes. High variance and overfitting have been major concerns in this setting. Therefore, it is imperative to use dimension reduction and/or regularization methods, such as projection, screening methods, or the Lasso, to extract and select 'low-dimensional' and 'informative' features, while avoiding overfitting (Zou and Hastie, 2005; Bair et al., 2006; Fan and Fan, 2008; Liu et al., 2011). Although many marginal variable screening techniques, such as the Sure Independence Screening (SIS) procedure, are shown to be able to filter out many uninformative variables in many scenarios (Fan and Lv, 2008; Fan and Song, 2010; Li et al., 2012; Mai and Zou, 2013), these 'informative' features selected from such screening methods can be highly correlated and non-sparse.

Throughout the paper, we use $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{i\tilde{p}_x})$ to denote a $\tilde{p}_x \times 1$ vector of relatively low-dimensional informative features for predicting y_i . In this case, model (1) reduces to

$$y_i = f(\tilde{\mathbf{x}}_i) + \sigma_y \varepsilon_{iy} = f(\tilde{x}_{i1}, \dots, \tilde{x}_{i\tilde{p}_x}) + \sigma_y \varepsilon_{iy}. \quad (2)$$

In many applications, such as genetics or neuroimaging, $\mathbf{x}_i$ and/or $\tilde{\mathbf{x}}_i$ can be highly correlated, and moreover, the number of important features can be *non-sparse*, that is, $p_x \gg \tilde{p}_x \gg n$. Such a high correlation structure and non-sparsity are notoriously difficult for existing dimension reduction and regularization methods (Tibshirani, 1996; Zou, 2006; Fan and Fan, 2008; Fan and Lv, 2010; Buhlmann et al., 2012). For instance, almost all regularization methods for high-dimensional regression strongly depend on some assumptions on the correlation structure of $\mathbf{x}_i$ and the sparsity (Zhao and Yu, 2006; Candès and Tao, 2007; Buhlmann et al., 2012). Moreover, individual features can be weakly correlated with the response, whereas their joint effect can be strong. Therefore, it is imperative to aggregate these correlated and informative features into p_z key features with $p_z \ll n$.

Let $\mathbf{z}_i = (z_{i1}, \dots, z_{ip_z})^T$ be a $p_z \times 1$ vector of such key features. Finally, model (1) may be approximated by

$$y_i = f(\mathbf{z}_i) + \sigma_y \varepsilon_{iy} = f(z_{i1}, \dots, z_{ip_z}) + \sigma_y \varepsilon_{iy}. \quad (3)$$

When $\mathbf{z}_i = \Gamma_z \mathbf{x}_i$, in which Γ_z is a $p_z \times p_x$ matrix, model (3) reduces to the well-known semi parametric index model in the dimensional reduction literature (Li, 1991; Cook and Ni, 2005; Zhang and Yin, 2014; Yang, 2016). Most existing dimension reduction methods focus on the scenario when p_x is smaller than n . See Ma and Zhu (2013) for a comprehensive review on dimension reduction. Little has been done on the scenario when p_x is much larger

than n and/or $\mathbf{x}_i$ is highly correlated due to many statistical and computational challenges (Li, 2007; Ma and Zhu, 2013; Yu et al., 2013; Yin and Hilafu, 2015). For instance, many sufficient variable selection methods require the calculation of a large sample covariance matrix of $\mathbf{x}_i$ and its inverse, which can be non-trivial (Chen et al., 2010; Yu et al., 2013; Yin and Hilafu, 2015).

The aim of this paper is to develop a single index latent factor model (SILFM) framework using (3) to predict y_i using $\mathbf{x}_i$. SILFM can be regarded as an extension/integration of the well-known single index model, the high-dimensional linear model (HLM), and the latent factor model in the literature. Compared with the existing literature, we make at least four major contributions in this paper:

- Model (3) differs from most single index models considered in the literature, in which $\mathbf{z}_i = \Gamma \mathbf{x}_i$. Specifically, we introduce a latent factor model to characterize the potential relationship between $\mathbf{z}_i$ and $\mathbf{x}_i$. Such a latent factor model can be useful and powerful for handling *weak and correlated* individual signals, but *strong joint effects*.
- Moreover, model (3) differs from those models considered in many contemporary works on variable selection, where the signals are mostly *rare but strong*. For instance, to deal with the “curse-of-dimensionality”, it is common to assume an additive structure with $f(\mathbf{x}_i) = \sum_{j=1}^{p_x} f_j(x_{ij})$ and a sparse signal $\#\{j: f_j(\cdot) \neq \text{constant}\} \ll n$.
- A comprehensive three-stage estimation procedure is developed to adaptively and sequentially improve prediction accuracy. Our estimation procedure includes *screening*, *aggregating*, and *nonlinear fitting*. Each step is computationally efficient even for the high-dimensional scenario with $p_x \gg n$.
- We investigate several theoretical properties of SILFM, such as the sure independence screening property and risk bounds.

This paper is organized as follows. In Section 2, we introduce the general SILFM framework. In Sections 3, simulation studies are conducted to evaluate the small-sample performance of SILFM. In Section 4, we apply SILFM to the analysis of hippocampus data obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset. In Section 5, we systematically investigate the theoretical properties of SILFM. Concluding remarks are given in Section 6.

2. SILFM: Single Index Latent Factor Model

2.1 Model Setup

The measurement models of SILFM are specified by

$$y_i = f(\mathbf{z}_i) + \sigma_y \varepsilon_{iy}, \quad (4)$$

$$\tilde{\mathbf{x}}_i = F_R(\mathbf{x}_i) = G(\mathbf{z}_i) + \mathbf{e}_{ix}, \quad (5)$$

where $\mathbf{e}_{ix}$ is a $\tilde{p}_x \times 1$ vector of measurement errors with zero mean, $F_R(\cdot) : R^{p_x} \rightarrow R^{\tilde{p}_x}$ is a dimension reduction function of $\mathbf{x}_i$, and $G(\cdot) : R^{p_z} \rightarrow R^{\tilde{p}_x}$ is a smooth function of $\mathbf{z}_i$. SILFM includes many well-known models as special cases. For instance, if $\mathbf{z}_i = \mathbf{x}_i$ and $f(\mathbf{x}_i) = \sum_{j=1}^{p_x} f_j(x_{ij})$, then SILFM reduces to additive models. Furthermore, if $f(\mathbf{x}_i) = \sum_{j=1}^{p_x} x_{ij}\beta_j$, then SILFM reduces to a high-dimensional linear model. Moreover, when $\tilde{\mathbf{x}}_i = \mathbf{x}_i = \mathbf{z}_i + \mathbf{e}_{ix}$, SILFM reduces to a measurement error model.

Model (5) is a generalized version of standard latent factor models when $G(\mathbf{z}_i) = \Lambda_G \mathbf{z}_i$ in which Λ_G is a $\tilde{p}_x \times p_z$ matrix. Model (5) includes many well-known models as special cases. Specifically, if $G(\mathbf{z}_i) = \Lambda_G \mathbf{z}_i$ and Λ_G is full column rank, then $\mathbf{z}_i$ can be rewritten as

$$\mathbf{z}_i = \Gamma_G \tilde{\mathbf{x}}_i - \Gamma_G \mathbf{e}_{ix} = (\Lambda_G^T \Lambda_G)^{-1} \Lambda_G^T F_R(\mathbf{x}_i) - (\Lambda_G^T \Lambda_G)^{-1} \Lambda_G^T \mathbf{e}_{ix}, \quad (6)$$

where $\Gamma_G = (\Lambda_G^T \Lambda_G)^{-1} \Lambda_G^T$. Furthermore, if $\mathbf{e}_{ix} = \mathbf{0}$ and $F_R(\mathbf{x}_i) = \Lambda_R \mathbf{x}_i$ in which Λ_R is a $\tilde{p}_x \times p_x$ matrix, then $\mathbf{z}_i$ can be written as $(\Lambda_G^T \Lambda_G)^{-1} \Lambda_G^T \Lambda_R \mathbf{x}_i$ and model (4) reduces to the well-known single-index model. When $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ and $G(\cdot)$ is a nonlinear function, model (5) reduces to a standard model for the nonlinear dimension reduction.

A unique feature of SILFM is that (5) integrates both a selection process and dimension reduction into a single formulation. Specifically, $F_R(\cdot)$ and $G(\cdot)$ can be regarded as a feature selection map and a dimension reduction map, respectively. It may allow us to efficiently deal with *weak and correlated* individual signals, that may have *strong joint effects* on y_i . By using $F_R(\cdot)$, we may be able to eliminate many individual signals unrelated to prediction. The use of $G(\cdot)$ allows us to aggregate many weak and correlated individual signals into a few strong and independent signals.

2.2 Estimation Procedure

We develop a three-stage estimation procedure in order to sequentially estimate $F_R(\cdot)$, $G(\cdot)$, and $f(\cdot)$, while achieving better prediction accuracy. Our estimation procedure SILFM is a three-stage process consisting of *screening*, *aggregating*, and *nonlinear fitting* as follows:

$$(y_i, \mathbf{x}_i) \xrightarrow{\text{screening}} \tilde{\mathbf{x}}_i \xrightarrow{\text{aggregating}} \mathbf{z}_i \xrightarrow{\text{nonlinear fitting}} y_i = \hat{f}(\mathbf{z}_i). \quad (7)$$

An overview of our procedure for three stages is given as follows.

- Stage (I). Use a Sure Independence Screening (SIS) procedure based on a Hilbert-Schmidt Independence Criterion (HSIC) to select a set of important features $\mathbf{x}_j$.
- Stage (II). Extract the key features $\mathbf{z}_i$ from the selected important features.
- Stage (III). Use an empirical minimization method such as the kernel ridge or/and support vector regression to build a prediction method based on the extracted key features.

Stage (I) is a fully nonparametric robust screening method based on HSIC. The key steps of Stage (I) include three steps as follows.

- Step (I.1). Use HSIC and its associated p -value to measure the relationship of each feature individually to the response.
- Step (I.2). Rank marginal HSIC values or their p -values according to their size (or their degree of dependence to the response).
- Step (I.3). Filter out all noisy features whose size is smaller than a given threshold.

The HSIC statistic is a two-variable independence test in Reproducing Kernel Hilbert Spaces (RKHS) (Gretton et al., 2005). As shown in (Sejdinovic et al., 2013), the HSIC statistic is consistent when a characteristic kernel is used and is equivalent to the distance covariance (DC) test of multivariate independence when the distance-induced kernel in HSIC is chosen (Székely et al., 2007). Moreover, the HSIC test can be more sensitive than DC when other kernels are used, and the HSIC test can be readily extended to many metric spaces. It should be noted that the use of HSIC is not critical in Stage (I) and any other independence test, such as the fused Kolmogorov filter developed in (Mai and Zou, 2013), can be used here.

We review the key ideas of HSIC for testing the independence between two random variables. Let $Z \sim \mathbb{P}_Z$ and $Y \sim \mathbb{P}_Y$ be, respectively, random variables on $\mathcal{Z}$ and $\mathcal{Y}$, which are two nonempty topological spaces. Let $\mathbb{P}_{Z,Y}$ be the joint probability measure of (Z, Y) . Let $K_{\mathcal{Z}}$ and $K_{\mathcal{Y}}$ be kernels on $\mathcal{Z}$ and $\mathcal{Y}$ with respective RKHSs $\mathcal{H}_{K_{\mathcal{Z}}}$ and $\mathcal{H}_{K_{\mathcal{Y}}}$. Then, it is well known that $K_{\mathcal{Z} \times \mathcal{Y}}((z, y), (z', y')) = K_{\mathcal{Z}}(z, z')K_{\mathcal{Y}}(y, y')$ is a kernel on the product space $\mathcal{Z} \times \mathcal{Y}$ with RKHS $\mathcal{H}_{K_{\mathcal{Z}} \times \mathcal{Y}}$ that is isomorphic to the tensor product $\mathcal{H}_{K_{\mathcal{Z}}} \otimes \mathcal{H}_{K_{\mathcal{Y}}}$. The HSIC of Z and Y is defined as

$$\text{HSIC}(Z, Y)^2 = \int \int K_{\mathcal{Z} \times \mathcal{Y}} d([\mathbb{P}_{Z,Y} - \mathbb{P}_Z \mathbb{P}_Y] \times [\mathbb{P}_{Z,Y} - \mathbb{P}_Z \mathbb{P}_Y]). \quad (8)$$

A fundamental result is that if $K_{\mathcal{Z}}$ and $K_{\mathcal{Y}}$ are universal kernels, then $\text{HSIC}(Z, Y) = 0$ if and only if $\mathbb{P}_{Z,Y} = \mathbb{P}_Z \mathbb{P}_Y$.

We construct an empirical estimate of HSIC. Let H_n be a centering matrix $\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{1}_n = (1, \dots, 1)^T$ is an $n \times 1$ vector with all elements 1. Let $K_{\mathcal{Z},n}$ be an $n \times n$ matrix with the (i, i') th element $K_{\mathcal{Z}}(z_i, z_{i'})$, and let $K_{\mathcal{Y},n}$ be an $n \times n$ matrix with the (i, i') th element $K_{\mathcal{Y}}(y_i, y_{i'})$. Given an independently and identically

distributed sample $\{(z_i, y_i)\}_{i=1}^n$, we can construct an empirical estimate of HSIC as the sum of U-statistics given by

$$\widehat{\text{HSIC}}(Z, Y)^2 = n^{-2} \text{tr}(K_{\mathcal{Z}, n} H_n K_{\mathcal{Y}, n} H_n).$$

The estimated $n\widehat{\text{HSIC}}(Z, Y)^2$ has some nice statistical properties, which form the theoretical foundation of the HSIC screening procedure. Statistically, as $n \rightarrow \infty$, $n\widehat{\text{HSIC}}(Z, Y)^2$ converges to the weighted sum of $\chi^2(1)$ random variables in distribution (Gretton et al., 2005; Sejdinovic et al., 2013; Székely et al., 2007). Since different features may have different patterns, such as scale, we use a computationally fast approach based on a spectral method to approximate the p -value of $\widehat{\text{HSIC}}$ for each feature. Specifically, for the j -th component of $\mathbf{x}_j$, we calculate its HSIC and p -value. However, for computational simplicity, it is more convenient to directly use the value of the estimated HSIC to filter out 'noisy' features. In this case, for a given threshold γ_n , we can form the set of important features according to

$$\hat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p_X : |n\widehat{\text{HSIC}}(X_j, Y)^2| \geq \gamma_n\},$$

where X_j and Y are, respectively, the random variables for the j -th component of $\mathbf{x}$ and y . Theoretically, we will show that our variable screening procedure enjoys the sure independence screening property under some mild conditions. Compared to test marginal screening methods, Stage (I) aims to use a relatively small γ_n in order to increase the chance of keeping *all important and/or weak signals*. Specifically, we set the number of the selected features to be $[nN]$ for some constant N , where $[a]$ denotes the greatest integer less than a . For example, we may set the number of the selected features to be $5n$, $7n$, and $10n$ rather than using $n-1$ or $[n/\log(n)]$ in the sparse signal case (Fan and Lv, 2010).

Stage (II) is not only a dimension-reduction method, but it is also an information aggregation method. Consider the true active set $\mathcal{M} = \{1, \dots, p_X\}$ for the variables in $\mathbf{x}_j$. Stage (II) includes three steps as follows:

- Step (II.1). Calculate the (kernel) correlation matrix of the selected features, denoted by $R_{\mathbf{x}} = (r_{jk})_{1 \leq j, k \leq p_X}$.
- Step (II.2). Use the covariance thresholding method introduced by Bickel and Levina (2008) and the hierarchical clustering method to partition $\mathcal{M}$ into $p_{Z,s}$ multiple disjoint clusters $\mathcal{M} = \cup_{k=1}^{p_{Z,s}} \mathcal{M}_{k,s}$ with $\mathcal{M}_{k,s} \cap \mathcal{M}_{k',s} = \emptyset$ for $k \neq k'$ and $s = 1, \dots, S$, where $\mathcal{M}_{k,s}$ is a subset of $\mathcal{M}$ and $p_{Z,s}$ is an integer, which may vary across s . For each s , let $\tilde{r}_s$ be a given thresholding value and $T_{\tilde{r}_s}$ be thresholding operator such that $T_{\tilde{r}_s}(R_{\mathbf{x}}) = (r_{jk} \mathbb{I}(|r_{jk}| - \tilde{r}_s))$, where $\mathbb{I}(\cdot)$ is an indicator function of an event. Let Π be a hierarchical clustering function that maps each $j \in \hat{\mathcal{M}}_{\gamma_n}$ into a unique cluster $\mathcal{M}_{k,s}$ based on $T_{\tilde{r}_s}(R_{\mathbf{x}})$. That is, $\Pi(\cdot, \cdot)$ is defined as $\Pi(j, T_{\tilde{r}_s}(R_{\mathbf{x}})) \in \mathcal{M}_{k,s}$ for each $j \in \hat{\mathcal{M}}_{\gamma_n}$.

- Step (II.3). For each $\mathcal{M}_{k,s}$, we calculate the sample (kernel) covariance matrix of these features with their indices in $\mathcal{M}_{k,s}$, denoted as $S_{\mathbf{x},k,s}$, and the eigenvalue-eigenvector pairs of $S_{\mathbf{x},k,s}$. Finally, we extract the key features $\mathbf{z}_i$ based on the scores from the eigenvectors corresponding to the $r_{k,s}$ algebraically largest eigenvalues of $S_{\mathbf{x},k,s}$.

Stage (II) can be regarded as a novel generalization of the supervised principal component (PC) method Bair et al. (2006), since it conducts PC method on marginally selected features with their indices in each cluster. A key difference is that in Step (II.2), we choose a series of $0 < \tilde{r}_1 < \dots < \tilde{r}_S < 1$ so that we can threshold the correlation matrix at different levels. It is expected that the larger $\tilde{r}_s$ is, the larger $p_{z,s}$ is. Equivalently, for large $\tilde{r}_s$, we only use group features that are highly correlated with each other. As varying a series of thresholds, we extract information from the selected features at different degrees of correlation, which allow us to select the most informative projected features that have the largest prediction power in Stage (III). To reduce the complexity of selecting optimal thresholds, we consider a set of fixed thresholds, such as $\{0.0, 0.25, 0.5, 0.75\}$, which may be sufficient for producing distinct clusters, and use the first PC for each cluster since each cluster reveals the highly correlated and low rank structure. We denote $\hat{\theta}^{(k,s)}$ to be the first PC from the k th cluster at different degrees of correlation ($S_{\mathbf{x},k,s}$) and construct the key feature, $\mathbf{z}_i = (\hat{\theta}^{(s_1)})^T \mathbf{x}_i, \dots, \hat{\theta}^{(s_S)}^T \mathbf{x}_i)^T$, where $\hat{\theta}^{(s)} = P(s) \{ \bigoplus_{k=1}^{p_{z,s}} \hat{\theta}^{(k,s)} \}$, $P(s)$ is a $\tilde{p}_x \times \tilde{p}_x$ matrix permuting the rows of the $\tilde{p}_x \times p_{z,s}$ matrix, $\bigoplus_{k=1}^{p_{z,s}} \hat{\theta}^{(k,s)}$ back to the initial order of the selected features and the direct sum is denoted by $\bigoplus$.

Stage (III) is to estimate the unknown link function $f(\cdot)$ which predicts the best value of the target function on any test point $\mathbf{x}$. Let $K_{\mathcal{Z}}(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a positive definite kernel. Therefore, the gram matrix $K_{\mathcal{Z},n} = K(\mathbf{z}_i, \mathbf{z}_j)$ is a positive definite $n \times n$ matrix. Given the kernel $K_{\mathcal{Z}}$, we can construct a unique RKHS $\mathcal{H}_{\mathcal{Z}}$ on $\mathcal{Z}$ such that $K_{\mathcal{Z}}(\cdot, \cdot)$ is an inner product and $f(\mathbf{z}) = \langle f, K_{\mathcal{Z}}(\cdot, \mathbf{z}) \rangle$ for all $f \in \mathcal{H}_{\mathcal{Z}}$ and $\mathbf{z} \in \mathcal{Z}$. We consider the regularized empirical estimate of f with respect to the loss function L defined as

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}_{\mathcal{Z}}} \{ \lambda \|f\|_{\mathcal{H}_{\mathcal{Z}}}^2 + n^{-1} \sum_{i=1}^n L(y_i, f(\mathbf{z}_i)) \}, \quad (9)$$

where $\lambda > 0$ is a regularization parameter and $\|\cdot\|_{\mathcal{H}_{\mathcal{Z}}}^2$ denotes the norm in $\mathcal{H}_{\mathcal{Z}}$. The use of the penalty term encourages smoothness and avoids overfitting. By using the representer theorem, any f minimizing (9) can be written as

$$\hat{f}_{\lambda}(\mathbf{z}) = \sum_{i=1}^n \hat{\alpha}_i K_{\mathcal{Z}}(\mathbf{z}_i, \mathbf{z}) \text{ for } \mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}^n. \quad (10)$$

When L is the squared loss function, $\hat{f}_\lambda(\mathbf{z})$ is the kernel ridge regression (KRR) (Schölkopf and Smola, 2001) estimate, where for a fixed λ , $\hat{\mathbf{a}} = (\mathbf{K}_{\mathcal{Z},n} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}$ and $\mathbf{y} = (y_1, \dots, y_n)^T$. We can utilize the support vector regression (SVR) (Drucker et al., 1997) by replacing the squared loss by the hinge loss to estimate f as alternative method. We focus on the KRR and SVR in this paper. Following the suggestion in Meyer and Wien (2014), we set the kernel width to be the dimension of the observed predictor, since the dimension of the key features may not be small. Moreover, we specify $\lambda = 0.001 \times n^{-2/3}$ based on the theoretical and numerical results in Fukumizu et al. (2007).

3. Simulation Studies

In this section, we conducted two simulation studies in order to examine the finite sample performance of SILFM. In order to compare other competing methods, we examined two types of performance measures for dimension reduction and prediction accuracy. For each scenario, 100 simulated data sets were generated, while each simulated data set consists of a training set with $n = 100$ and a test set with $n = 100$. First, for dimension reduction, we consider the true positive rate defined as $P_{tp} = |\hat{\mathcal{M}}_{\gamma_n} \cap \mathcal{M}|/|\mathcal{M}|$, the screening accuracy defined as $P_A = |\hat{\mathcal{M}}_{\gamma_n} \cap \mathcal{M}|/|\hat{\mathcal{M}}_{\gamma_n}|$, and the true negative rate defined as $P_{tn} = |\hat{\mathcal{M}}_{\gamma_n}^c \cap \mathcal{M}^c|/|\mathcal{M}^c|$, where $\mathcal{M}^c$ and $\hat{\mathcal{M}}_{\gamma_n}^c$ are, respectively, the complement of $\mathcal{M}$ and $\hat{\mathcal{M}}_{\gamma_n}$. Second, for prediction accuracy, we computed the empirical squared prediction error of the test data set as $n^{-1} \sum_{i=1}^n (y_i^{test} - \hat{f}(x_i^{test}))^2$, where $\hat{f}(\cdot)$ is the prediction model built from the training set and $(x_i^{test}, y_i^{test})_s$ are observations in the test set.

3.1 Simulation 1: Continuous Response (I)

We generated $\mathbf{x}_i$ from a multivariate normal distribution $N_{p_X}(0, \Sigma)$ with $n = 100$ and $p_X = 3000$. Moreover, we set $\Sigma = \Sigma_{\tilde{p}_X} \oplus \mathbf{I}_{2400}$ and $\Sigma_{\tilde{p}_X} = \Sigma_{300} \oplus \Sigma_{300} + \Omega_2 \otimes \rho_b J_{300}$, where $\tilde{p}_X = 600$, $\otimes$ is the Kronecker product, $J_{300} = \mathbf{1}_{300} \mathbf{1}_{300}^T$, $\Omega_2 = ((0, 1)^T, (1, 0)^T)$, and Σ_{300} is a 300×300 correlation matrix with the element $\Sigma_{300}(j, j') = \rho_w$ if $j = j'$. In this case, we set the number of informative features to be the first 600 features, that is $\mathcal{M} = \{1 \dots j = 600\}$. We set $\rho_w = 0.9$, $\rho_b = 0.7$, and $F_R(\mathbf{x}_i) = \mathbf{x}_i$, and then consider a nonlinear model as follows:

$$y_i = \sqrt{z_{i1}^2 + z_{i2}^2 + z_{i3}^2} + \log(\sqrt{z_{i1}^2 + z_{i2}^2 + z_{i3}^2}) + \sigma_y \varepsilon_{iy},$$

where $\sigma_y = 0.2$ and $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^T$ is a 3×1 vector of key features specified by $z_{i1} = \mathbf{1}_{600}^T \tilde{\mathbf{x}}_i / 600$, $z_{i2} = (\mathbf{1}_{300}^T, \mathbf{0}_{300}^T) \tilde{\mathbf{x}}_i / 300$, and $z_{i3} = (\mathbf{0}_{300}^T, \mathbf{1}_{300}^T) \tilde{\mathbf{x}}_i / 300$ in which $\mathbf{0}_k$ is a $k \times 1$ vector of zeros. We see that the number of such features is distinguished as *non-sparse* ($p_X = 3000 \gg \tilde{p}_X = 600 \gg n = 100$).

In Stage (I) of SILFM, we compare the HSIC-SIS procedure with other SIS procedures based on distance correlation (DC-SIS), Pearson correlation (SIS), Spearman correlation (SP-SIS), and Kendall correlation (KD-SIS). We set the number of selected features $\hat{\mathcal{M}} = \hat{\tilde{p}}_X$

to be 450, 650, and 850, respectively. Table 1 presents five selected quantiles of indices for the true important features, empirical true positive, accuracy, true negative and false positive measures for each estimated size as top features. It is observed that HSIC outperforms most of existing correlation measures.

In Stage (II) of SILFM, we set the number of thresholdings $S = 4$, $(\tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_3, \tilde{\tau}_4) = (0.0, 0.25, 0.50, 0.75)$, and the desired number of groups (1, 2, 4, 8, 10), estimated the cluster group for each $\tilde{\tau}_k$ by using a hierarchical clustering method and used the first PC from each cluster for all k to estimate the latent factors, $\hat{\mathbf{z}}_k$, where we kept these configurations in Stage (II) as fixed in the analysis. Furthermore, to enhance the outcome information in Stage (II), we applied either the additional HSIC screening or sliced inverse regression (Li, 1991).

In Stage (III), we used KRR and SVR as our learning methods based on the latent factors, denoted by SILFM1 and SILFM2. Similarly, if we apply the HSIC screening (or sliced inverse regression) in Stage (II), we denote SILFM estimates corresponding to KRR and SVR, denoted as SILFM1_y and SILFM2_y (or SILFM1_s and SILFM2_s), respectively.

As a comparison, we also consider six competing learning methods, including Lasso (Tibshirani, 1996), PC regression (PCR) (Jolliffe, 2005), partial least squares (PLS) (Helland, 1988), generalized additive model (GAM) (Hastie and Tibshirani, 1990), single index model (SIDX) (Ichimura, 1993), and sparse single index model (SSIDX) (Alquier and Biau, 2013). Due to the issue of the degree of freedom, we used PCR and PLS as predictors of SIDX and GAM models when the number of the selected features is greater than n . We used the Gaussian RBF kernel $\exp(-\|x_1 - x_2\|^2/\sigma)$ with $\sigma = p_z$, and selected $\lfloor p_z/2 \rfloor$ factors for SILFM1_y and SILFM2_y. For other prediction methods, we used an optimized tuning parameter that minimizes the corresponding cross-validation error. Table 2 reports the average of prediction error for each method calculated from the 100 test data sets. Our SILFMs outperform all other competing methods for the nonsparse case.

As the sensitivity analysis, we also conduct the sparse scenario by replacing three latent factors, z_{j1} , z_{j2} , and z_{j3} by the first three features, x_{j1} , x_{j2} , and x_{j3} for the same nonlinear model. We used the identical SILFM method mentioned above and did not report the screening result based on HSIC-SIS procedure, since all true features were selected for any screening methods. In spite of utilizing the nonsparse approach with SILFM in this sparse scenario, Table 2 shows that our SILFMs have good performance. Specifically, SILFM1 works as well as GAM, SIDX, and SSIDX, while SILFM2_y outperforms all other competing methods for all feature selection cases.

3.2 Simulation 2: Continuous Response (II)

Similar to Simulation 1, we also simulated $\mathbf{x}_i$ from $N(0, \Sigma = \Sigma_{\tilde{p}_x} \oplus \mathbf{I}_{2400})$ and set $\tilde{p}_x = 600$ and $F_R(\mathbf{x}_i) = \mathbf{z}_i$. Moreover, we divide the 600 informative features into three blocks with 200 features in each block and define $\Sigma_{\tilde{p}_x} = \Sigma_{200} \oplus \Sigma_{200} \oplus \Sigma_{200} + \Omega_3 \otimes \rho_b J_{200}$, where $\rho_b = 0.6$, $J_{200} = \mathbf{1}_{200} \mathbf{1}_{200}^T$, $\Omega_3 = ((0, 1, 1)^T, (1, 0, 1)^T, (0, 1, 1)^T)$, and Σ_{200} is a 200×200 correlation matrix with the element $\Sigma_{200}(j, j') = \rho_w = 0.9$ if $j = j'$. We consider a nonlinear model as follows:

$$y_i = 5 \sin(9z_{i3}\pi/10) + z_{i2}^2 + 3 \exp(z_{i1}) + \sigma_y \varepsilon_{iy},$$

where $\varepsilon_{iy} \sim N(0, 1)$. Moreover, $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^T$ is a 3×1 vector of key features specified as $\mathbf{z}_i = \Gamma_G(\mathbf{x}_i - \mathbf{e}_{ix})$, where $\mathbf{e}_{ix} \sim N(\mathbf{0}_{600}, \sigma_x^2 \mathbf{I}_{600})$ and

$$\Gamma_G = \begin{pmatrix} B_{i_1} & \cdots & B_{i_{200}} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ B_{i_1} & \cdots & B_{i_{200}} & B_{i_{201}}, \cdots, B_{i_{400}} & 0 & \cdots & 0 \\ B_{i_1} & \cdots & B_{i_{200}} & B_{i_{201}}, \cdots, B_{i_{400}} & B_{i_{401}}, \cdots, B_{i_{600}} \end{pmatrix}, \quad (11)$$

in which the $B_{i,j}$ were sampled from $\mathcal{B} = \{2/(600), 4/(600), \dots, 2\} = \{B_1, B_2, B_3, \dots, B_{600}\}$ with the same probability without replacement, and the levels of σ_x^2 and σ_y^2 are set to be 0.2, respectively. This simulation setting is also similar to the first one such that the number of important features is distinguished as *non-sparse* and their effect on y_i is nonlinear. Three panels in Figure 1 present the simulation results based on 100 independent simulation runs. Panels (A) and (B) reveal that HSIC-SIS and other nonparametric SIS methods have similar performance as higher true positive and true negative rates, respectively by varying the number of selected features. As expected, the standard SIS method is much worse than all other SIS methods and has difficulty in finding the true significant features. Panel (C) shows that the SILFM methods have the best overall prediction performance compared to all other competing methods when the number of selected informative features is greater than $\tilde{p}_X$. Moreover, SILFM2_y outperforms SILFM1 and SILFM2, where SILFM2_y derived from SILFM2 is only reported for the brevity in the panel.

4. Real Data Analysis

To illustrate the usefulness of SILFM, we consider a date set collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. To measure cognitive impairment in this study, the mini-mental state examination (MMSE) is commonly used, where any score greater than or equal to 24 points indicates a normal cognition. Below this, scores can indicate severe (< 9 points), moderate (10–18 points) or mild (19–23 points) cognitive impairment. Hence, it is of scientific interest to identify the association between MMSE and other measurements containing imaging, genetic, and clinical variables and to predict behavior scores (MMSE) within an integrated model framework. Such a prediction model not only predicts the cognitive trajectory, but also potentially provides new

approaches for early diagnosis of AD. This earlier identification would allow for more efficient selection of samples for clinical trials and possibilities for earlier disease treatment.

In our data set, 406 subjects were obtained from the ADNI database (adni.loni.usc.edu), along with measurements of covariates, including sex (194 females and 212 males), age (mean 75.0 years), education years (mean 15.7 years) and Apolipoprotein E (APOE) SNPs, rs429358 and rs7412 as the genetic variables, where these two SNPs define a 3 allele haplotype, ϵ_2 (43 yes), ϵ_3 (186 yes), and ϵ_4 (131 yes) variants for APOE status, and the left and right hippocampus surfaces images as 30,000 radial distance. We randomly split independent datasets with 15 folds where each dataset includes 226 individuals for the training data and 180 individuals used for test data. Since the hippocampus, as a part of the limbic system, plays an important role in the consolidation of information from short-term memory to long-term memory and spatial navigation, we related these to the clinical, demographic, and genetic covariates using the following SILFM model:

$$y_{\text{mme},i} = f(x_{\text{sex},i}, x_{\text{age},i}, x_{\text{edu},i}, x_{\epsilon_2,i}, x_{\epsilon_3,i}, x_{\epsilon_4,i}, x_{\text{hippocampus},i}) + \varepsilon_i. \quad (12)$$

We used HSIC-SIS to select the top 100, 200, ..., 1,000 features in Stage (I) and the same approaches introduced in Section 3 in Stages (II) and (III) as SILFM estimation. Table 3 presents the average of prediction errors for all methods. Overall, SILFM methods based on the kernel machine methods dramatically improve the prediction accuracy. SILFM2 and SILFM2_s especially provided the best and the second best models when we selected 700 and 600 features, respectively.

It is known that the hippocampus is a structure that lies deep in the brain and panel (A) in Figure 2 depicts the anatomical image of the hippocampus consisting of four regions: Sub, CA 1, CA2, and CA3. Huang and Kandel (1994) pointed out that CA1 is closer to the output region of the hippocampus and it is important for representing space in the environment, so that individual cells in the CA1 region encode for space and therefore long-term memory for space and attentional modulation of space importantly involves the CA1 region. Panels (B) and (C) in Figure 2 reveal 3D hippocampus images (left and right views, respectively) and the latent factors identified in Stage (II) are marked by different colors. The selected features (700) based on HSIC-SIS are highly concentrated on the CA1 region and these features are especially interpreted that they have the functional relationship with MMSE score. Moreover, the CA1 region that contains the selected features can be divided into multiple layers based on the latent factor scores, indicating that multiple pathways may be involved in memory storage. CA1 neurons contained in each multiple layer identified by the latent factors are jointly involved with the memory ability. To clearly explain this phenomenon, a medical investigation should be pursued and will be pursued as another future research.

Finally, we elaborate more on the connection between the regularity conditions and our real data analysis. As shown in panels (A) and (B) of Figure 3, the selected marginal signals are stronger than the stochastic noise. As shown in panels (C) and (D) of Figure 3, the profile of the outcome is bounded and the nonparametric fitting leads to a better fit. Assumption of the latent factor model is valid since the anatomical hippocampus structure consists of four

biological sub-regions. Panels (E) and (F) of Figure 3 show the correlation matrices of the selected features (1000) after thresholding at 0 and 0.7.

5. Asymptotic Analysis

We investigate several theoretical properties of SILFM including sure independence screening property and risk bound for SILFM. For simplicity, it is assumed that $\{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ are independent and identically distributed. The following conditions are used to facilitate the technical details. Although they may not be the weakest conditions, they do help to simplify the proof. We have stated the following theorems, whose detailed proofs and required conditions can be found in a supplementary document.

First, we will show that the features selected by Stage (I) enjoy the sure screening property under some general conditions.

Theorem 5.1

Let $\mathcal{F}_j = \{f_j \in \mathcal{H}_{X_j} \mid \|f_j\|_{\mathcal{H}_{X_j}} = 1\}$ and $\mathcal{G} = \{g \in \mathcal{H}_Y \mid \|g\|_{\mathcal{H}_Y} = 1\}$ be functional classes with the unit ball in an RKHS on each marginal and outcome domain. Under Conditions 1–2, for any positive constant c_1 , there exists a constant c_2 such that

$$\mathbb{P}(\max_{1 \leq j \leq p_X} \max_{f_j \in \mathcal{F}_j, g \in \mathcal{G}} |\hat{\beta}_j - \beta_j| \geq c_1 n^{-\kappa}) \leq p_X \exp(-2c_2 n^{1-2\kappa}). \quad (13)$$

If Condition 3 also holds, then taking $r_n = c_3 n^{-\kappa}$ with $c_3 = c_0/2$ leads to

$$\mathbb{P}(\mathcal{M} \subset \widehat{\mathcal{M}}_{r_n}) \xrightarrow{P} 1, \quad (14)$$

where $\xrightarrow{P}$ denotes convergence in probability.

Theorem 5.1 establishes the sure screening property of our screening method. Compared with the existing literature (Fan and Lv, 2008; Fan and Song, 2010), our results in Theorem 5.1 focus on a *non-sparse* circumstance ($p_X \gg \tilde{p}_X \gg n$) and thus the maximum dimension holds as $\log p_X/n^{1-2\kappa} \rightarrow 0$ where $0 < \kappa < 1/2$. Therefore, it is unnecessary to assume the standard tail probability condition for each feature.

Next, we investigate the consistency of the estimator of Stage (III) and the prediction performance of SILFM. Details for the notation and conditions are given in Supplementary material.

Theorem 5.2

Let $L(\cdot, \cdot)$ be a squared loss function and P be a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F} \subset L_\infty(X)$ be a non-empty and compact set. Moreover, let H be RKHS of a continuous kernel K on $\mathcal{X}$ such that $L_\infty(X) \subset H$ and K is m times continuously differentiable on $\mathbf{R}^d$. Suppose Conditions 4–10 hold. Then for a fixed $\tau > 0$, the convergence rate of $\hat{f}_{n,\lambda,\theta}^\wedge$ (SILFM) is given by

$$\lambda \|\hat{f}_{n,\lambda,\hat{\theta}}\|^2 + \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\hat{\theta}}^*) \leq O_p(n^{-\frac{2m\beta}{(2m+d)(2\beta+1)}}) + O_p(\sqrt{\log \mathcal{M}_r} n^{-\kappa})$$

as $\|f_{P,\hat{\theta}}^0 - f_{P,\hat{\theta}}^*\| = o_p(1)$ and $\lambda \rightarrow 0$ where $0 < \kappa < 1$.

Theorem 5.2 gives an integrated insight of various SILFM estimators on the prediction performance of SILFM. First, if $d = O(p_X)$, then the SILFM estimator is at the rate of $O_p(n^{-\frac{2m\beta}{(2m+px)(2\beta+1)}})$. However, if d is relatively small, we can obtain a faster rate for the SILFM estimator, but we have to pay the price for using the dimension reduction methods in Stages (I) and (II), which is $O_p(\sqrt{\log \mathcal{M}_r} n^{-\kappa})$. As $\sqrt{\log \mathcal{M}_r} n^{-\kappa} \rightarrow 0$, SILFM can truly achieve a better prediction risk, which has been justified by the simulations and real data analysis. This is one justification for using sequential estimators of SILFM under the challenging situation of ultrahigh dimensionality, highly correlated predictors and a complex functional relationship with the response.

6. Discussion

We have developed a SILFM framework to build an accurate prediction model for clinical outcomes based on a massive number of features. SILFM as a three-stage estimation procedure integrating an independent screening method, a latent factor model, and kernel ridge regression. Theoretically, we have established several theoretical properties of SILFM, such as risk bound and selection consistency. Our simulation results and real data analysis show that SILFM outperforms many other methods in terms of prediction accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Associate Editor and referees, whose questions and insightful comments have led to a much improved paper.

References

- Alquier P, Biau G. Sparse single-index model. *Journal of Machine Learning Research*. 2013; 14:243–280.
- Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *Journal of the American Statistical Association*. 2006; 101:119–137.
- Bickel PJ, Levina E. Some theory for fisher's linear discriminant function, naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*. 2004; 10:989–1010.
- Bickel PJ, Levina E. Covariance regularization by thresholding. *The Annals of Statistics*. 2008; 36:2577–2604.
- Bühlmann P, Rutimann P, Van de Geer S, Zhang CH. Technical report. ETH Zurich: 2012. Correlated variables in regression: clustering and sparse estimation.

- Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*. 2007; 35:2313–2351.
- Chen X, Zou H, Cook RD. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Annals of Statistics*. 2010; 38:1696–1723.
- Clarke BS, Fokoué E, Zhang HH. *Principles and Theory for Data Mining and Machine Learning*. Springer; 2009.
- Cook RD, Ni L. Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*. 2005; 100:410–428.
- Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. *Advances in neural information processing systems*. 1997
- Fan J, Fan Y. High dimensional classification using features annealed independence rules. *Annals of statistics*. 2008; 36:2605–2637. [PubMed: 19169416]
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*. 2008; 70:849–911.
- Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20:101–144. [PubMed: 21572976]
- Fan J, Song R. Sure independence screening in generalized linear models with np -dimensionality. *The Annals of Statistics*. 2010; 38:3567–3604.
- Fukumizu K, Bach FR, Gretton A. Statistical consistency of kernel canonical correlation analysis. *The Journal of Machine Learning Research*. 2007; 8:361–383.
- Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory*. 2005; 3734:63–77.
- Hastie T, Tibshirani R. *Generalized Additive Model*. Chapman and Hall New York: 1990.
- Hastie T, Tibshirani R, Friedman JJH. *The Elements of Statistical Learning*. Springer; New York, NY: 2009.
- Helland IS. On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*. 1988; 17:581–607.
- Huang Y-Y, Kandel ER. Recruitment of long-lasting and protein kinase a-dependent long-term potentiation in the $ca1$ region of hippocampus requires repeated tetanization. *Learning & Memory*. 1994; 1:74–82. [PubMed: 10467587]
- Ichimura H. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*. 1993; 58:71–120.
- Jolliffe I. *Principal component analysis*. Wiley Online Library; 2005.
- Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. *The Annals of Statistics*. 2012; 40:1846–1877.
- Li K-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. 1991; 86:316–327.
- Li L. Sparse sufficient dimension reduction. *Biometrika*. 2007; 94:603–613.
- Liu Y, Zhang HH, Wu Y. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*. 2011; 106:166–177. [PubMed: 22162896]
- Ma Y, Zhu L. A review on dimension reduction. *International Statistical Review*. 2013; 81:134–150. [PubMed: 23794782]
- Mai Q, Zou H. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*. 2013; 100:229–234.
- Meyer D, Wien FT. Support vector machines. R package (e1071). 2014; 1:23–26.
- Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. The MIT Press; 2001.
- Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*. 2013; 41:2263–2291.
- Székel GJ, Rizzo ML, Bakirov NK, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*. 2007; 35:2769–2794.

- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.* 1996; 58:267–288.
- Yang H. PhD thesis. The University of North Carolina at Chapel Hill; 2016. Learning methods in reproducing kernel Hilbert space based on high-dimensional features.
- Yin X, Hilafu H. Sequential sufficient dimension reduction for large p , small n problems. *Journal of the Royal Statistical Society: Series B.* 2015 page in press.
- Yu Z, Zhu L, Peng H, Zhu L. Dimension reduction and predictor selection in semiparametric models. *Biometrika.* 2013; 100:641–654.
- Zhang HP, Singer BH. *Recursive Partitioning and Applications.* 2. Springer; New York: 2010.
- Zhang N, Yin X. Direction estimation in single-index regressions via hilbert-schmidt independence criterion. *Statistica Sinica.* 2014 page in press.
- Zhao P, Yu B. On model selection consistency of lasso. *The Journal of Machine Learning Research.* 2006; 7:2541–2563.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association.* 2006; 101:1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B.* 2005; 67:301–320.

Figure 1.

Results in simulation 2: panels A (true positive rate), B (true negative rate), and C (averages of prediction errors) report the results based on the different SIS methods (HSIC (red), Kendall (green), Spearman (blue), and Pearson (black)), and prediction methods (SILFM1 (red), SILFM1_s (black), SILFM2 (blue), SILFM2_y (purple), Lasso (green), GAM_{pca} (skyblue), GAM_{pls} (darkgreen), SIDX_{pca} (orange), SIDX_{pls} (brown) and SSIDX (pink)).

Figure 2. Hippocampus image in ADNI analysis: panels A (anatomical image), B(right side image in analysis) and C (left side image in analysis) report hippocampus images, where different colors (blue, red, purple, yellow and others) in panels B and C reveal latent factor scores from each cluster and the regions with skyblue are the regions filtered out in Stage (I).

Figure 3.

Assessing validity in ADNI study: panels A (left hippocampus) and B (right hippocampus) report HSIC values, where the red solid lines mark the cutoff values, and panels C, D, E, and F reveal profile of MMSE score, scatter plot between MMSE and one strongest hippocampus feature, the correlation matrix among the selected features, and the thresholding correlation matrix at 0.7.

Table 1

Average of prediction errors in simulation 1: each column indicates the number of $|\hat{\theta}|$ used for the prediction model. Each row presents a learning method, where the number in parenthesis indicates the standard deviation for each corresponding method.

Method	5%	25%	50%	70%	95%	P_{ip}	P_A	P_m	P_{ip}	$ \hat{\theta} $
Oracle	30.95	150.75	300.50	450.25	570.05	1.00	1.00	1.00	0.00	600
HSIC-SIS	24.25 (3.02)	117.75 (16.14)	234.25 (27.03)	384.10 (25.92)	557.89 (8.50)	0.750 (<0.001)	1.000 (<0.001)	1.000 (<0.001)	0.250 (<0.001)	450
DC-SIS	24.45 (3.08)	117.57 (16.21)	234.50 (25.32)	385.97 (24.93)	565.53 (81.86)	0.744 (0.010)	0.993 (0.014)	0.998 (<0.001)	0.255 (0.010)	450
SIS	246.39 (274.90)	679.62 (429.91)	1261.05 (613.36)	1896.17 (693.95)	2734.66 (536.81)	0.231 (0.236)	0.308 (0.315)	0.870 (0.059)	0.768 (0.236)	450
KD-SIS	364.36 (268.81)	765.57 (405.98)	1413.00 (539.14)	2228.52 (553.68)	2839.12 (325.77)	0.148 (0.196)	0.198 (0.262)	0.849 (0.049)	0.851 (0.196)	450
SP-SIS	151.73 (29.12)	747.00 (68.37)	1483.30 (66.05)	2219.42 (53.92)	2837.05 (27.55)	0.154 (0.017)	0.205 (0.022)	0.851 (<0.001)	0.845 (0.017)	450
HSIC-SIS	33.45 (<0.01)	163.25 (<0.01)	325.50 (0.91)	487.75 (4.58)	1389.92 (163.06)	1.000 (0.010)	0.923 (0.010)	0.979 (<0.001)	0.000 (0.010)	650
DC-SIS	33.45 (<0.01)	163.25 (<0.01)	325.80 (1.51)	489.85 (9.08)	1500.51 (246.27)	0.994 (0.020)	0.917 (0.017)	0.977 (<0.001)	0.006 (0.020)	650
SIS	214.48 (247.60)	624.57 (397.54)	1207.15 (539.79)	1891.55 (583.93)	2767.22 (303.97)	0.349 (0.294)	0.322 (0.272)	0.816 (0.073)	0.651 (0.294)	650
KD-SIS	295.58 (250.64)	686.90 (380.21)	1387.80 (476.54)	2203.95 (439.19)	2828.43 (137.34)	0.231 (0.242)	0.213 (0.223)	0.787 (0.061)	0.768 (0.242)	650
SP-SIS	140.58 (24.35)	729.75 (55.48)	1486.25 (53.58)	2229.77 (41.75)	2840.43 (21.36)	0.224 (0.020)	0.207 (0.017)	0.785 (<0.001)	0.775 (0.020)	650
HSIC-SIS	43.45 (<0.01)	213.25 (<0.01)	425.50 (0.70)	938.92 (63.63)	2597.31 (55.67)	1.000 (<0.001)	0.705 (<0.001)	0.895 (<0.001)	0.000 (<0.001)	850
DC-SIS	43.45 (<0.01)	213.25 (<0.01)	425.50 (1.54)	960.32 (53.58)	2569.21 (57.46)	0.999 (<0.001)	0.705 (<0.001)	0.895 (<0.001)	0.001 (<0.001)	850
SIS	174.42 (219.99)	599.02 (360.37)	1210.15 (474.93)	1992.95 (421.86)	2798.43 (91.87)	0.445 (0.305)	0.314 (0.216)	0.757 (0.076)	0.554 (0.305)	850
KD-SIS	265.93 (217.09)	667.95 (351.11)	1406.00 (418.74)	2211.75 (319.27)	2838.35 (74.67)	0.306 (0.263)	0.216 (0.186)	0.722 (0.066)	0.693 (0.263)	850
SP-SIS	149.71 (18.53)	745.27 (48.624)	1496.95 (47.218)	2233.07 (37.79)	2841.87 (17.67)	0.287 (0.022)	0.203 (0.017)	0.717 (<0.001)	0.713 (0.022)	850

Table 2

Average of prediction errors in simulation 1: each column indicates the number of $\|\hat{\eta}\|$ used for the prediction model and each row presents a learning method, where the number in parenthesis indicates the standard deviation for each corresponding method.

Method	Nonsparse case			Sparse case		
	450	650	850	450	650	850
SILEM _{1y}	0.135 (0.125)	0.338 (0.165)	0.278 (0.169)	1.476 (0.672)	1.583 (0.774)	1.669 (0.807)
SILEM ₁	0.275 (0.144)	0.983 (0.173)	1.006 (0.211)	1.179 (0.255)	1.339 (0.309)	1.369 (0.293)
SILEM _{1s}	1.985 (0.429)	2.945 (0.487)	2.827 (0.381)	3.980 (0.563)	3.847 (0.591)	3.623 (0.485)
SILEM _{2y}	0.138 (0.099)	0.377 (0.113)	0.338 (0.112)	0.599 (0.134)	0.665 (0.160)	0.668 (0.185)
SILEM ₂	0.252 (0.138)	1.103 (0.200)	1.112 (0.225)	1.234 (0.240)	1.416 (0.278)	1.476 (0.293)
SILEM _{2s}	1.897 (0.342)	2.757 (0.466)	2.713 (0.379)	3.591 (0.398)	3.562 (0.494)	3.455 (0.464)
LASSO	2.222 (0.323)	2.404 (0.366)	2.637 (0.444)	2.877 (0.448)	2.963 (0.474)	2.987 (0.483)
PCR	2.303 (0.361)	2.475 (0.351)	2.438 (0.361)	3.059 (0.476)	2.995 (0.430)	2.957 (0.425)
PLS	2.386 (0.416)	2.661 (0.410)	2.560 (0.372)	3.263 (0.525)	3.146 (0.457)	3.061 (0.445)
GAM _{pcu}	0.348 (0.129)	0.270 (0.090)	0.392 (0.134)	0.719 (0.192)	1.205 (0.484)	1.494 (0.578)
GAM _{pjs}	0.738 (0.745)	1.248 (0.852)	2.124 (0.558)	1.264 (0.739)	2.636 (3.715)	2.661 (3.006)
SIDX _{pcu}	1.428 (1.155)	1.227 (0.549)	1.136 (0.473)	1.300 (0.792)	1.778 (0.829)	1.894 (0.734)
SIDX _{pjs}	0.900 (0.776)	1.144 (0.920)	1.720 (1.096)	1.088 (1.094)	1.401 (1.325)	1.749 (1.355)
SSIDX	0.459 (0.336)	0.347 (0.100)	0.492 (0.827)	0.833 (1.063)	1.217 (1.475)	1.505 (1.163)

Table 3

Average of prediction errors in hippocampal surfaces data: each column indicates the number of $|\hat{M}|$ used for the prediction model and each row presents a learning method, where the number in parenthesis indicates the standard deviation for each corresponding method and the bold number indicates the best performance.

Method	100	200	300	400	500	600	700	800	900	1000
SILFM _{1y}	79.77 (26.57)	81.25 (26.57)	81.31 (27.08)	81.20 (27.15)	80.32 (24.96)	80.17 (27.13)	79.85 (24.91)	80.05 (25.24)	79.42 (24.24)	79.33 (26.14)
SILFM ₁	77.14 (27.51)	76.54 (26.62)	76.11 (24.87)	75.17 (24.36)	75.69 (22.94)	75.70 (21.77)	75.81 (20.74)	77.07 (22.18)	77.68 (21.28)	77.15 (22.01)
SILFM _{1s}	81.44 (27.92)	81.76 (28.91)	81.95 (27.20)	80.13 (22.63)	80.62 (25.00)	81.02 (21.84)	78.82 (22.12)	81.29 (24.60)	82.78 (24.80)	82.95 (23.72)
SILFM _{2y}	74.62 (23.08)	76.13 (21.40)	75.48 (22.53)	76.85 (22.02)	76.36 (22.96)	76.13 (21.86)	75.14 (23.06)	75.92 (23.67)	75.62 (22.92)	73.23 (23.73)
SILFM ₂	73.76 (26.32)	73.46 (26.02)	72.49 (25.78)	71.93 (25.76)	71.65 (26.18)	70.64 (23.83)	69.97 (23.40)	70.91 (23.28)	71.42 (23.55)	72.24 (25.65)
SILFM _{2s}	72.57 (24.07)	73.21 (23.07)	72.70 (23.92)	74.08 (23.62)	75.12 (27.28)	70.90 (20.17)	71.24 (21.16)	74.29 (23.69)	77.38 (21.65)	78.89 (26.11)
LASSO	86.71 (27.60)	87.36 (29.10)	87.54 (28.84)	87.23 (28.41)	87.26 (27.63)	86.40 (27.86)	85.97 (27.60)	87.50 (28.74)	87.17 (26.55)	88.40 (30.19)
PCR	81.50 (26.93)	81.57 (26.84)	80.85 (24.27)	82.88 (28.02)	81.49 (24.16)	81.08 (23.88)	81.10 (23.06)	80.95 (24.18)	80.35 (23.79)	80.40 (23.36)
PLS	81.94 (25.99)	85.29 (25.94)	82.83 (22.06)	82.42 (23.35)	81.60 (23.29)	80.45 (22.29)	80.09 (21.46)	82.87 (25.64)	79.98 (21.31)	80.30 (21.99)
GAM _{psa}	75.75 (34.06)	84.72 (34.25)	79.17 (32.01)	79.37 (28.33)	78.60 (26.10)	81.23 (26.13)	82.36 (24.56)	78.20 (26.93)	77.51 (26.54)	81.42 (29.43)
GAM _{pls}	77.40 (27.25)	79.17 (28.06)	78.03 (25.46)	78.13 (23.07)	77.85 (24.09)	75.75 (23.40)	75.95 (23.71)	82.96 (27.22)	80.98 (24.02)	80.22 (24.09)
SIDX _{psa}	96.20 (29.52)	114.79 (39.44)	112.92 (36.17)	103.42 (40.06)	103.79 (37.93)	98.62 (25.64)	112.91 (57.88)	106.99 (34.23)	98.03 (31.38)	119.18 (39.78)
SIDX _{pls}	108.20 (41.40)	101.11 (37.58)	107.15 (44.06)	95.84 (29.91)	87.14 (25.86)	82.33 (22.40)	78.51 (21.95)	85.62 (30.64)	81.43 (21.42)	79.40 (22.63)
SSIDX	81.52 (15.84)	91.93 (32.88)	109.08 (43.86)	100.42 (32.97)	95.50 (33.97)	100.06 (34.45)	99.40 (33.78)	102.55 (44.12)	98.68 (28.81)	94.70 (31.11)