Statistica Sinica **32** (2022), 2431-2453
doi:https://doi.org/10.5705/ss.202020.0406

SPARSE SLICED INVERSE REGRESSION VIA CHOLESKY MATRIX PENALIZATION

Linh H. Nghiem[1,3], Francis K. C. Hui[2], Samuel Müller[1,3] and A. H. Welsh[2]

[1] *University of Sydney,* [2] *Australian National University and* [3] *Macquarie University*

Abstract: We introduce a new sparse sliced inverse regression estimator called Cholesky matrix penalization, and its adaptive version, for achieving sparsity when estimating the dimensions of a central subspace. The new estimators use the Cholesky decomposition of the covariance matrix of the covariates and include a regularization term in the objective function to achieve sparsity in a computationally efficient manner. We establish the theoretical values of the tuning parameters that achieve estimation and variable selection consistency for the central subspace. Furthermore, we propose a new projection information criterion to select the tuning parameter for our proposed estimators, and prove that the new criterion facilitates selection consistency. The Cholesky matrix penalization estimator inherits the advantages of the matrix lasso and the lasso sliced inverse regression estimator. Furthermore, it shows superior performance in numerical studies and can be extended to other sufficient dimension reduction methods in the literature.

Key words and phrases: Cholesky decomposition, information criterion, Lasso, sparsity, sufficient dimension reduction.

1. Introduction

In a regression problem with a scalar outcome y and a p-variate predictor $\mathbf{X} = (X_1, \ldots, X_p)^\top$, sufficient dimension reduction refers to a class of methods that try to express the outcome as a function of a few linear combinations of covariates (Li (2018)). In other words, sufficient dimension reduction aims to find a matrix $\mathbf{B}$ of dimension $p \times d$, with $d \ll p$, such that

$$y \perp \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}, \tag{1.1}$$

with $\perp$ denoting statistical independence. Condition (1.1) implies that the d linear combinations $\mathbf{B}^\top \mathbf{X}$ contain all the information about y on $\mathbf{X}$, so we can replace $\mathbf{X}$ by $\mathbf{B}^\top \mathbf{X}$ without loss of information. Dimension reduction is achieved because the number of linear combinations d is usually much smaller than the

Corresponding author: Linh H. Nghiem, Australian National University, Canberra 0200, Australia. E-mail: linh.nghiem@anu.edu.au.

number of covariates p. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ be the columns of $\mathbf{B}$. An alternative formulation of the relationship between y and $\mathbf{X}$ under (1.1) is provided by the multiple index model

$$y = f(\boldsymbol{\beta}_1^\top \mathbf{X}, \ldots, \boldsymbol{\beta}_d^\top \mathbf{X}, \varepsilon), \tag{1.2}$$

where f is an unknown link function and ε is a random noise term independent of $\mathbf{X}$. In (1.1) and (1.2), the matrix $\mathbf{B}$ and the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ are, in general, not unique when $d \geq 1$ (Li (2018)). Therefore, the goal in sufficient dimension reduction is to identify the central subspace, which is defined as the intersection of all subspaces spanned by the column spaces of $\mathbf{B}$ satisfying (1.1). The central subspace, denoted by $\mathcal{S}_{y\perp\mathbf{X}}$, is unique under mild conditions (Li (2018)). The transformations $\boldsymbol{\beta}_j^\top \mathbf{X}$, for $j = 1, \ldots, d$, are called sufficient predictors. The number of indices in the multiple index model, d, is also known as the structural dimension of the central subspace. A variety of sufficient dimension reduction methods have been proposed in the literature, including sliced inverse regression (SIR, Li (1991)), sliced average variance estimation (SAVE, Cook (2000)), principal Hessian direction (pHd, Li (1992); Cook (1998)), minimum average variance estimation (Xia et al. (2002)), and directional regression (Li and Wang (2007)), among others. An overview of these methods can be found in Li (2018).

When the number of covariates is large, we often assume each dimension $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ is sparse; that is, for each $j = 1, \ldots, d$, only a few elements of each dimension $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$ are nonzero. Motivated by this idea, the last few years have seen an emerging body of literature combining sparsity with SIR, as well as with other sufficient dimension reduction methods by adding a regularization term to an appropriate objective function. For example, Yin and Hilafu (2015) proposed a sequential approach for estimating SIR. Lin, Zhao and Liu (2018) first proposed a screening approach to perform variable selection; then, the selected variables were included in the classical SIR. Assuming the covariates follow the standard p-variate Gaussian distribution with at most s nonzero components in each dimension, Lin et al. (2021) established the minimax rate of the risk of the estimated projection matrix when the number of indices d is bounded. Tan et al. (2018) proposed a convex formulation for sparse SIR in a high-dimensional setting by adapting techniques from sparse canonical correlation analysis. While most methods have been able to identify the theoretical values of the regularization parameter necessary to obtain estimation consistency of the central space, few have done so for variable selection consistency in each estimated dimension; one example is Qian, Ding and Cook (2019). Even in that case, the method used to select the regularization parameter in their numerical studies and data application does not guarantee selection consistency.

In this paper, we propose a new approach to constructing a sparse SIR estimator based on the Cholesky decomposition of the sample covariance matrix. This Cholesky matrix penalization (CHOMP) estimator has a close connection to the lasso SIR estimator proposed by Lin, Zhao and Liu (2019), but has several advantages over it. First, while both the CHOMP and the lasso SIR achieve estimation consistency for the central subspace, we generalize the CHOMP estimator to an adaptive version that can achieve both estimation and variable selection consistency. Furthermore, for both the CHOMP and its adaptive version, we propose a new projection information criterion (PIC) to select the regularization parameter in the corresponding objective function. To the best of our knowledge, this is the first data-driven method that is theoretically demonstrated to achieve variable selection consistency for the central subspace. Our simulation studies show that the adaptive CHOMP estimator with the regularization parameter selected by the PIC outperforms the lasso SIR in terms of both estimation error and variable selection. Finally, the CHOMP-type estimator is easily generalized to many other sufficient dimension reduction methods, such as SAVE and pHd, and the corresponding adaptive CHOMP-type estimators are shown empirically to have competitive performance in finite samples as well.

The following notation is used throughout the paper. For any p-dimensional nonzero vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$, let $\mathcal{P}(\boldsymbol{v}) = \boldsymbol{v}(\boldsymbol{v}^\top \boldsymbol{v})^{-1}\boldsymbol{v}^\top$ denote the projection matrix associated with $\boldsymbol{v}$, and let $\|\boldsymbol{v}\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$, $\|\boldsymbol{v}\|_1 = \sum_{j=1}^p |v_j|$, $\|\boldsymbol{v}\|_0 = \sum_{j=1}^p 1(v_j \neq 0)$, and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ denote its ℓ_2, ℓ_1, ℓ_0, and ℓ_∞ norms respectively. For any index set T, the notation $\boldsymbol{v}_T$ and $\boldsymbol{v}_{T^c}$ denote the subvectors consisting of the components of $\boldsymbol{v}$ in T and T^c, respectively. For any $m \times q$ nonsingular matrix $\mathbf{A}$ with entries a_{ij}, let $\mathcal{P}(\mathbf{A}) = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top$ denote the projection matrix associated with $\mathbf{A}$. In addition, we define the Frobenius norm of $\mathbf{A}$ to be $\|\mathbf{A}\|_F = (\sum_{i=1}^m \sum_{j=1}^q a_{ij}^2)^{1/2}$, whereas the ℓ_2 induced norm $\|\mathbf{A}\|_2$ is its largest singular value $\sigma_1(\mathbf{A})$. Finally, for ease of notation, we let μ_j generically denote the tuning parameter used to estimate the jth dimension of the central subspace for all of the penalized methods.

2. A Review of SIR and the Matrix Lasso

We first review SIR, which is the basis for the other methods discussed in this paper. Assuming the predictor vector $\mathbf{X}$ follows an elliptical distribution with location vector zero and scale matrix $\boldsymbol{\Sigma}$, it is demonstrated in Li (1991) that the column space of $\mathbf{B}$ in equation (1.1) satisfies

$$\boldsymbol{\Sigma}\,\mathrm{col}(\mathbf{B}) = \mathrm{col}(\boldsymbol{\Lambda}), \tag{2.1}$$

where $\boldsymbol{\Lambda} = \mathrm{var}\left\{\mathbb{E}(\mathbf{X}|y)\right\}$. If we observe independent and identically distributed (i.i.d.) data pairs $(\boldsymbol{x}_i^\top, y_i)$, for $i = 1, \ldots, n$, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, let $\mathcal{X}$ denote the $n \times p$ design matrix; without loss of generality, assume each predictor is centered at zero, and let $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top = n^{-1}\mathcal{X}^\top\mathcal{X}$ be the sample covariance matrix. Next, if the outcome y is ordered, then the matrix $\boldsymbol{\Lambda}$ is estimated by first dividing the data into H non-overlapping slices of roughly equal sizes, $J_1, \ldots, J_H$, based on the increasing order of y. If y is categorical, each slice may correspond to one category in the outcome. Then, we compute the vector of covariate averages within each slice, $\bar{\boldsymbol{x}}_h^\top = |J_h|^{-1}\sum_{i=1}^n \boldsymbol{x}_i^\top 1(y_i \in J_h)$, with $|J_h|$ being the size of slice J_h. As a result, an estimate for $\boldsymbol{\Lambda}$ is given by $\hat{\boldsymbol{\Lambda}} = H^{-1}\sum_{h=1}^H \bar{\boldsymbol{x}}_h\bar{\boldsymbol{x}}_h^\top$. Let $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d$ be the eigenvectors corresponding to the d largest eigenvalues of $\hat{\boldsymbol{\Lambda}}$. Then, Li (1991) showed that each dimension $\hat{\boldsymbol{\beta}}_j$ of the central subspace can be estimated by

$$\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\eta}}_j,\ j = 1, \ldots, d. \tag{2.2}$$

In a recent paper, Lin, Zhao and Liu (2019) introduced two sparse SIR estimators that are closely connected to the lasso estimator in the regular linear model, namely the matrix lasso and the lasso SIR. Based on the relationship (2.2), the matrix lasso estimator $\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}}$ is defined as

$$\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}} = \underset{\boldsymbol{\beta}_j}{\mathrm{argmin}}\,\frac{1}{2}\left\|\hat{\boldsymbol{\eta}}_j - \hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_j\right\|_2^2 + \mu_j\left\|\boldsymbol{\beta}_j\right\|_1,\ j = 1, \ldots, d. \tag{2.3}$$

Although the matrix lasso estimator was introduced in Lin, Zhao and Liu (2019), it was largely dismissed, and its theoretical properties are yet to be examined. One possible reason for this is because, similarly to any regularization method, the performance of the matrix lasso estimator depends on how the tuning parameters μ_j, for $j = 1, \ldots, d$, are chosen. However, selecting appropriate tuning parameters for the matrix lasso is challenging from both theoretical and practical perspectives, for two reasons. First, the outcome $\hat{\boldsymbol{\eta}}_j$ does not contain independent observations, so regular cross-validation is not guaranteed to work. In addition, unlike the linear model case, the matrix $\hat{\boldsymbol{\Sigma}}$ is a $p \times p$ symmetric matrix, so the first term in (2.3) can be zero if $\hat{\boldsymbol{\Sigma}}$ is invertible, as occurs, for example, in low-dimensional settings where $n > p$. In Section S4 of the Supplementary Material, we demonstrate empirically that common methods for selecting tuning parameters for the matrix lasso do not guarantee good performance.

On the other hand, a main advantage of the matrix lasso that was not emphasized in Lin, Zhao and Liu (2019) is that the formulation (2.3) is convex and directly mimics the population relationship that characterizes the SIR in (2.1). As a result, the formulation of the matrix lasso can be extended to many other sufficient dimension reduction methods that are formed by changing the matrix $\boldsymbol{\Lambda} = \mathrm{var}\left\{\mathbb{E}(\mathbf{X}|y)\right\}$ in equation (2.1) to another quantity. For example, a sliced average variance estimator is obtained with $\boldsymbol{\Lambda} = \mathbb{E}\left\{\boldsymbol{\Sigma} - \mathrm{var}(\mathbf{X}|y)\right\}^2$, and a principal Hessian direction estimator is obtained with $\boldsymbol{\Lambda} = \mathbb{E}\left[\mathbf{X}\,\mathbf{X}^\top\left\{y - \mathbb{E}(y)\right\}\right]$. Hence, understanding the behavior of the matrix lasso estimator, and building upon it to devise improved estimators provides a unified strategy for investigating sparse dimensions of a central subspace, as shown in Section 7.

In fact, the lasso SIR estimator, also proposed by Lin, Zhao and Liu (2019), can be considered a recasting of the matrix lasso. Lin, Zhao and Liu (2019) suggest that we can write the matrix $\hat{\boldsymbol{\Lambda}} = (nc)^{-1}\mathcal{X}^\top\boldsymbol{M}^\top\boldsymbol{M}\mathcal{X}$ for an appropriate $H \times n$ matrix $\boldsymbol{M}$ and $c = \lfloor n/H \rfloor$. As a result, each corresponding eigenvector can be expressed as $\hat{\boldsymbol{\eta}}_j = n^{-1}\mathcal{X}^\top\tilde{\boldsymbol{y}}_j$, where $\tilde{\boldsymbol{y}}_j = (c\hat{\lambda}_j^{-1})\boldsymbol{M}^\top\boldsymbol{M}\mathcal{X}$, for $j = 1, \ldots, d$, and $\hat{\lambda}_j$ is the eigenvalue of $\hat{\boldsymbol{\Lambda}}$ corresponding to $\hat{\boldsymbol{\eta}}_j$; see Section S3 of the Supplementary Material for more detail. If we use the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathcal{X}^\top\mathcal{X}$ to estimate $\boldsymbol{\Sigma}$, then (2.1) can be written as $\mathcal{X}^\top\mathcal{X}\boldsymbol{\beta}_j \propto \mathcal{X}^\top\tilde{\boldsymbol{y}}_j$, and the lasso SIR estimator is defined as

$$\hat{\boldsymbol{\beta}}_j^{\mathrm{L}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\,\|\widetilde{\boldsymbol{y}}_j - \mathcal{X}\boldsymbol{\beta}\|_2^2 + \mu_j\|\boldsymbol{\beta}\|_1,\ \ j = 1, \ldots, d.$$

This formulation depends critically on the special form of $\boldsymbol{\Lambda} = \mathrm{var}\{\mathbb{E}(\mathbf{X}\,|y)\}$ used in SIR, when a good estimator for it can be written in the form $\hat{\boldsymbol{\Lambda}} = \mathcal{X}^\top\mathbf{K}$ for an $n \times p$ matrix $\mathbf{K}$. Therefore, it is not straightforward to extend it to other sufficient dimension reduction methods that are obtained by changing $\boldsymbol{\Lambda}$. For example, if we want to perform sufficient dimension reduction using the SAVE method, it is not obvious how to find a good estimator in the form $\hat{\boldsymbol{\Lambda}} = \mathcal{X}^\top\mathbf{K}$ for $\boldsymbol{\Lambda} = \mathbb{E}\left\{\boldsymbol{\Sigma} - \mathrm{var}(\mathbf{X}|y)\right\}^2$ so that the idea of lasso SIR can be applied. In the next section, we provide another reformulation of the matrix lasso that both inherits desirable properties of the lasso SIR and can be applied to other methods in a more straightforward way.

3. Cholesky Matrix Penalization for SIR

3.1. Estimators

Recall that at the population level, the SIR estimator satisfies the relationship (2.1). Let $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d$ be the eigenvectors associated with the d largest eigenvalues of $\boldsymbol{\Lambda}$. Then, the vector $\boldsymbol{\beta}_j$ satisfies

$$\boldsymbol{\Sigma}\boldsymbol{\beta}_j = \boldsymbol{\eta}_j,\ j = 1, \ldots, d. \tag{3.1}$$

For each j, equation (3.1) is a system of p linear equations. Because we do not impose any additional structure on the symmetric and positive-definite matrix $\boldsymbol{\Sigma}$, an efficient way to solve the system is to use the Cholesky decomposition. Specifically, letting $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is the Cholesky factor of $\boldsymbol{\Sigma}$, equation (3.1) is equivalent to

$$\boldsymbol{L}^\top\boldsymbol{\beta}_j = \boldsymbol{\kappa}_j,\ \text{where}\ \ \boldsymbol{L}\boldsymbol{\kappa}_j = \boldsymbol{\eta}_j,\ j = 1, \ldots, d.$$

Because $\boldsymbol{L}$ is a lower triangular matrix, the vector $\boldsymbol{\kappa}_j$ is obtained by backward substitution, and the vector $\boldsymbol{\beta}_j$ is obtained by forward substitution. Next, denote $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{\eta}}_j$ as estimators for $\boldsymbol{L}$ and the eigenvector $\boldsymbol{\eta}_j$, respectively. Typically, the vector $\hat{\boldsymbol{\eta}}_j$ is the eigenvector of the matrix $\hat{\boldsymbol{\Lambda}}$.

Let $\hat{\boldsymbol{\kappa}}_j$ be calculated from $\hat{\boldsymbol{L}}\hat{\boldsymbol{\kappa}}_j = \hat{\boldsymbol{\eta}}_j$; for $\hat{\boldsymbol{\kappa}}_j$ to be well defined, the estimator $\hat{\boldsymbol{L}}$ needs to be invertible. For the remainder of the paper, we assume $n > p$, so we can choose $\hat{\boldsymbol{L}}$ as the Cholesky factor of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$. In Section S4 of the Supplementary Material, we investigate a high-dimensional setting ($n < p$), where the Cholesky factor $\boldsymbol{L}$ can be estimated efficiently by imposing an additional structure on $\boldsymbol{\Sigma}$. We define the Cholesky matrix penalization (CHOMP) estimator for the SIR as

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j}{\text{argmin}}\, \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\beta}_j - \hat{\boldsymbol{\kappa}}_j\right\|_2^2 + \mu_j\left\|\boldsymbol{\beta}_j\right\|_1,\ j = 1, \ldots, d, \tag{3.2}$$

where μ_j is a nonnegative tuning parameter. Furthermore, we can penalize each component of $\boldsymbol{\beta}_j$ differently by introducing a vector of adaptive weights $\omega_j = (\omega_{j1}, \ldots, \omega_{jp})^\top$ and defining

$$\hat{\boldsymbol{\beta}}_j^* = \underset{\boldsymbol{\beta}_j}{\text{argmin}}\, \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\beta}_j - \hat{\boldsymbol{\kappa}}_j\right\|_2^2 + \mu_j\sum_{k=1}^{p}\omega_{jk}|\beta_{jk}|,\ j = 1, \ldots, d. \tag{3.3}$$

We refer to this estimator as the adaptive Cholesky matrix penalization (adaptive CHOMP) estimator, in line with the adaptive lasso estimator proposed by Zou (2006). Moreover, similarly to Zou (2006), we set the weights to $\omega_{jk} = |\bar{\beta}_{jk}|^{-\gamma}$, with $\bar{\beta}_{jk}$ being the kth component of an initial consistent estimate $\bar{\boldsymbol{\beta}}_j$ and γ a positive constant. Because $n > p$, we choose $\bar{\boldsymbol{\beta}}_j$ to be the unpenalized estimate $\bar{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\eta}}_j$. In the simulation study presented in Section 5, we find that, as expected, the inclusion of these adaptive weights makes the performance of the adaptive CHOMP superior to that of (the unweighted) CHOMP, the matrix lasso and the lasso SIR estimator in terms of the both estimation error and variable selection for the central subspace.

3.2. Matrix lasso, Cholesky matrix penalization, and Lasso SIR

The matrix lasso, CHOMP and lasso SIR estimators essentially derive from the same relationship (2.1). Moreover, if no regulation is imposed, $\mu_j = 0$, all the estimators are equivalent. However, when regularization is needed to achieve sparse solutions, the behavior of the tuning parameters for the matrix lasso differs fundamentally from that of the other two.

In fact, from the definition of the matrix lasso estimator given in equation (2.3) and the first-order optimality condition, each component of the matrix lasso estimator $\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}} = (\hat{\beta}_{j1}^{\mathrm{ML}}, \ldots, \hat{\beta}_{jp}^{\mathrm{ML}})^\top$ satisfies

$$\hat{\boldsymbol{\Sigma}}_k^\top \left(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}} - \hat{\boldsymbol{\eta}}_j\right) + \mu_j b_{jk}^{\mathrm{ML}} = 0, \quad k = 1, \ldots, p, \tag{3.4}$$

where $\hat{\boldsymbol{\Sigma}}_k^\top$ denotes the kth row of $\hat{\boldsymbol{\Sigma}}$, the scalar $b_{jk}^{\mathrm{ML}} = \mathrm{sign}(\hat{\beta}_{jk}^{\mathrm{ML}})$ if $\hat{\beta}_{jk}^{\mathrm{ML}} \neq 0$, and $b_{jk}^{\mathrm{ML}} \in [-1, 1]$ if $\hat{\beta}_{jk}^{\mathrm{ML}} = 0$. As a result, the entire vector $\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}}$ is set to zero if and only if $\mu_j > \|\hat{\boldsymbol{\Sigma}}^\top \hat{\boldsymbol{\eta}}_j\|_\infty$. Because we do not impose any sparsity structure on $\hat{\boldsymbol{\Sigma}}$, each component $\hat{\boldsymbol{\Sigma}}_k^\top \hat{\boldsymbol{\eta}}_j$ is the sum of p terms, meaning it can diverge to infinity when the dimension p grows. This fact does not change when each covariate is standardized to have variance one and the sample covariance matrix is a correlation matrix. As a result, when p is growing, the range of μ_j that needs to be considered is unbounded.

On the other hand, the range of μ_j for both the CHOMP and the lasso SIR estimator that needs to be considered is the unit interval. Each component of the CHOMP estimate $\hat{\boldsymbol{\beta}}_j$ and of the lasso SIR estimate $\hat{\boldsymbol{\beta}}_j^L$ satisfies

$$\hat{\boldsymbol{L}}_k^\top \left(\hat{\boldsymbol{L}}\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\kappa}}_j\right) + \mu_j b_{jk} = 0, \quad \text{i.e} \quad \hat{\boldsymbol{\Sigma}}_k^\top \hat{\boldsymbol{\beta}}_j - \eta_{jk} + \mu_j b_{jk} = 0 \tag{3.5}$$

$$n^{-1}\boldsymbol{x}_k^\top \left(\mathcal{X}\hat{\boldsymbol{\beta}}_j^L - \tilde{y}_j\right) + \mu_j b_{jk}^L = 0 \quad \text{i.e} \quad \hat{\boldsymbol{\Sigma}}_k^\top \hat{\boldsymbol{\beta}}_j^L - \eta_{jk} + \mu_j b_{jk}^L = 0, \tag{3.6}$$

respectively, where $\hat{\boldsymbol{L}}_k^\top$ denotes the kth row of $\hat{\boldsymbol{L}}$, the scalar $b_{jk} = \text{sign}(\hat{\beta}_{jk})$ if $\hat{\beta}_{jk} \neq 0$, and $b_{jk} \in [-1, 1]$ if $\hat{\beta}_{jk} = 0$, with a similar definition for b_{jk}^L. This implies that the CHOMP and the Lasso SIR estimators have the same estimating equation for every tuning parameter μ_j. As a result, the whole vector $\hat{\boldsymbol{\beta}}_j = 0$ is set to zero if and only if $\mu_j \geq \|\hat{\boldsymbol{L}}\hat{\boldsymbol{\kappa}}_j\|_\infty = \|\hat{\boldsymbol{\eta}}_j\|_\infty$. Because all the components of $\hat{\boldsymbol{\eta}}_j$ are between -1 and 1, to choose an appropriate value for μ_j, we need only consider $\mu_j \in [0, 1]$, regardless of the dimension p. In practice, we usually choose the tuning parameter from a grid of values, so having a fixed upper bound on the grid, regardless of p, is desirable to fine tune the estimator.

One way to restrict the bound of the tuning parameters for the matrix lasso estimator is to work with the standardized covariates $\boldsymbol{z}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}$, for $\boldsymbol{x}_i$, $i = 1, \ldots, n$. In that case, since the sample covariance matrix of the transformed z-data is the identity matrix, the quantity $\|\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\eta}}_j\|_\infty = \|\hat{\boldsymbol{\eta}}_j\|_\infty$ is bounded by one. However, a major disadvantage of this approach is that the matrix lasso estimator on the $\boldsymbol{z}_i$-data, denoted as $\hat{\boldsymbol{\beta}}_j^{\text{ML}(z)}$, can be sparse, but the final estimator $\hat{\boldsymbol{\beta}}_j^{\text{ML}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\beta}}_j^{\text{ML}(z)}$ for $\boldsymbol{\beta}_j$ is not guaranteed to be sparse, because we do not impose any sparsity requirement on the matrix $\hat{\boldsymbol{\Sigma}}^{-1/2}$. As a result, no variable selection is achieved for any dimension of the central subspace.

Finally, unlike the lasso SIR estimator, the CHOMP estimator inherits the flexibility of the matrix lasso in that it is easy to adapt to other sufficient dimension reduction methods. For example, for the sliced average variance estimator, we change $\boldsymbol{\Lambda} = \text{var}\{E(\mathbf{X}\,|y)\}$ in equation (2.1) to $\boldsymbol{\Lambda} = \mathbb{E}\left\{\boldsymbol{\Sigma} - \text{var}(\mathbf{X}\,|y)\right\}^2$, and make the corresponding estimate $\hat{\boldsymbol{\Lambda}}$ its sample version. In this situation, it is not as straightforward to define the vector $\tilde{\boldsymbol{y}}_j$ such that the eigenvector $\hat{\boldsymbol{\eta}}_j$ of $\hat{\boldsymbol{\Lambda}}$ can be written as $\hat{\boldsymbol{\eta}}_j = n^{-1}\mathcal{X}^\top\tilde{\boldsymbol{y}}_j$ to apply the idea of the lasso SIR. Nevertheless, we can still compute the CHOMP estimate and its adaptive version by solving the problem (3.2). We will elaborate on this point in Section 7.

3.3. Projection information criterion

To choose the tuning parameter μ_j for the CHOMP and adaptive CHOMP estimators, we propose minimizing the projection information criterion (PIC), defined as

$$\text{PIC}(\mu_j) = \begin{cases} \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 + \dfrac{\log(p)}{p}\left\|\hat{\boldsymbol{\beta}}_j(\mu_j)\right\|_0, & \text{if } \hat{\boldsymbol{\beta}}(\mu_j) \neq \mathbf{0} \\ \infty, & \text{if } \hat{\boldsymbol{\beta}}_j(\mu_j) = \mathbf{0}, \end{cases} \tag{3.7}$$

where for the jth dimension, the notation $\hat{\boldsymbol{\beta}}_j(\mu_j)$ denotes either the CHOMP or its adaptive version associated with the tuning parameter $\mu_j > 0$. The main difference between the projection and the usual information criteria is in the loss function, and we motivate our choice as follows. In the multiple index model, each vector $\boldsymbol{\beta}_j$ is not unique, but the projection matrix associated with it is unique. Hence, a sensible way of quantifying the goodness of fit is via the estimated projection matrix. The specific form of the loss part, $\|\mathcal{P}\{\hat{\boldsymbol{\beta}}_j(\mu_j)\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\|_F^2$, measures the deviation of $\hat{\boldsymbol{\beta}}(\mu_j)$ from an already established consistent estimator. Because the projection matrix is not well defined for the zero vector, we ignore this case by setting the PIC to infinity when the parameter estimates are zero. In other words, we do not expect the true vector $\boldsymbol{\beta}_j$ to be a zero vector for any dimension. The model complexity penalty term $\tau_j = \log(p)/p$ controls the trade-off between the model loss and the complexity part. Intuitively, this choice of model complexity penalty proceeds as follows. Because the loss part is bounded above by two and the number of nonzero components for each $\hat{\boldsymbol{\beta}}_j(\mu_j)$ can range from zero to p, the denominator of τ_j is set to p to make the two parts have relatively the same magnitude. The numerator of τ_j follows the same spirit as the Bayesian information criterion (BIC) penalty; however, it is set to $\log(p)$ instead of to $\log(n)$ to make τ_j go to zero without imposing any further condition on the growth rates of n and p. For each dimension $j = 1, \ldots, d$, we demonstrate in Section 4 that this model complexity term leads to selection consistency; that is, the PIC asymptotically identifies the nonzero components of each dimension $\boldsymbol{\beta}_j$ correctly with this model complexity term.

Finally, we briefly mention the issue of estimating the number of indices d from the data. In general, if the original data are divided into H slices, the maximum number of dimensions that can be estimated by the SIR methods is $H - 1$ (Li (2018)). When p is fixed, a variety of methods for determining d have been proposed in the literature, including the sequential testing approach of Li (1991) and the bootstrap methods of Ye and Weiss (2003), among others. When p is growing , Lin, Zhao and Liu (2019) proposed a method for choosing the number of indices for the lasso SIR based on a clustering of the eigenvalues of $\hat{\boldsymbol{\Lambda}}$. We anticipate similar methods could be developed for the (adaptive) CHOMP estimators, but the precise choice of d is outside the scope of this study. In the numerical studies below, we assume d to be known.

4. Theoretical Results

We prove several results related to the estimation consistency and variable selection consistency of the estimated projection matrix $\mathcal{P}(\hat{\mathbf{B}})$ and $\mathcal{P}(\hat{\mathbf{B}}^*)$, where $\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}^*$ are $p \times d$ matrices, the columns of which are CHOMP and adaptive CHOMP estimators, respectively. These theoretical results are derived by combining the results for the lasso estimator for the regular linear model with the results of the lasso SIR estimator developed in Lin, Zhao and Liu (2019), showing another advantage of the CHOMP estimator over the matrix lasso estimator. Furthermore, we demonstrate that using the new PIC leads to a selection consistent estimator. In this section, we allow the number of covariates p to grow with the sample size n, but the ratio $p/n \to 0$ when $n \to \infty$. This condition ensures that the Cholesky factor $\hat{\boldsymbol{L}}$ of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ is invertible with probability one, so the vector $\hat{\boldsymbol{\kappa}}_j$ (and functions thereof) is well defined. Proofs of all results can be found in Sections S1 and S2 of the Supplementary Material.

First, we state several technical conditions that are used throughout the development below.

(C1) There exist constants $C_{\min}$ and $C_{\max}$ such that $0 < C_{\min} < \lambda_{\min}(\boldsymbol{\Sigma}) < \lambda_{\max}(\boldsymbol{\Sigma}) < C_{\max} < \infty$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ denote the minimum and maximum eigenvalues, respectively, of $\boldsymbol{\Sigma}$.

(C2) The d largest eigenvalues $\lambda_1, \dots, \lambda_d$ of $\boldsymbol{\Lambda} = \mathrm{var}\{\mathbb{E}(X|y)\}$ satisfy $0 < \lambda_d \leq \cdots \leq \lambda_1 \leq \lambda_{\max}(\boldsymbol{\Sigma}) < \infty$.

(C3) The central curve $m(y) = \mathbb{E}(\mathbf{X} \mid y)$ satisfies the sliced stability condition of Lin, Zhao and Liu (2018).

Condition (C1) is usually imposed in analyses of high-dimensional linear regression models (Wainwright (2019)). This condition implies that the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ satisfies a so-called restricted value condition over a cone set, which is described more clearly below. As discussed in Lin, Zhao and Liu (2019), Condition (C2) is a refined version of a commonly imposed condition in the SIR literature, that is, $\mathrm{rank}(\boldsymbol{\Lambda}) = d$, meaning the dimension of the space spanned by the central curve is equal to the dimension of the central subspace. Finally, Condition (C3) controls the smoothness of the central curve m and the tail distribution of $m(y)$; see Lin, Zhao and Liu (2018) for a detailed discussion of this condition.

Recall for each dimension $j = 1, \dots, d$, the vector $\boldsymbol{\eta}_j = \boldsymbol{\Sigma}\boldsymbol{\beta}_j = \boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{\beta}_j$ is the eigenvector associated with the jth largest eigenvalue of $\boldsymbol{\Lambda} = \mathrm{var}\{\mathbb{E}(\mathbf{X} \mid y)\}$, and $\hat{\boldsymbol{\eta}}_j$ is the same quantity of the estimated matrix $\hat{\boldsymbol{\Lambda}}$. Define $\tilde{\boldsymbol{\eta}}_j = \mathcal{P}(\boldsymbol{\eta}_j)\hat{\boldsymbol{\eta}}_j$ to be the projection of $\hat{\boldsymbol{\eta}}_j$ on $\boldsymbol{\eta}_j$. The projection implies that $\tilde{\boldsymbol{\eta}}_j \propto \boldsymbol{\eta}_j$. As a result, if we define $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\eta}}_j$, then $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Sigma}^{-1}\mathcal{P}(\boldsymbol{\eta}_j)\hat{\boldsymbol{\eta}}_j = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_j(\boldsymbol{\eta}_j^\top\boldsymbol{\eta}_j)^{-1}\boldsymbol{\eta}_j^\top\hat{\boldsymbol{\eta}}_j \propto \boldsymbol{\beta}_j$; in other words, $\tilde{\boldsymbol{\beta}}_j$ has the same projection matrix as the true dimension $\boldsymbol{\beta}_j$. We refer to $\tilde{\boldsymbol{\beta}}_j$ as the "pseudo-true" parameter for the dimension j in the theoretical development, and bound the difference $\delta_j = \hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j$ to establish the consistency of the estimated projection matrix.

Denote $S_j = \{k : \beta_{jk} \neq 0\}$, the set of indices corresponding to nonzero components of the true dimension $\boldsymbol{\beta}_j$, and $s_j = |S_j|$, the cardinality of the set S_j. Furthermore, denote $S = \bigcup_{j=1}^d S_j$, the set of active covariates across all dimensions, and $s = |S|$. Because $\tilde{\boldsymbol{\beta}}_j \propto \boldsymbol{\beta}_j$, then $\tilde{\beta}_{jk} = 0$ for any $j \in S_j^c$. The following theorem establishes the consistency of the estimated projection matrix from the CHOMP estimator.

Theorem 1. *Consider a multiple index model with $n\lambda_d = p^\nu$, for $\nu > 1$. Assume Conditions* (C1)–(C3) *hold and the number of dimensions d is known. Let $\hat{\mathbf{B}}$ be the matrix the columns $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d$ of which are solutions of* (3.2) *with tuning parameter $\mu_j = M\{\log(p)/(n\hat{\lambda}_j)\}^{1/2}$ for a sufficiently large constant M, where $\hat{\lambda}_j$ is the jth largest eigenvalue of the matrix $\hat{\boldsymbol{\Lambda}}$. Then, the estimated projection matrix $\mathcal{P}(\widehat{\mathbf{B}})$ satisfies*

$$\left\|\mathcal{P}(\widehat{\mathbf{B}}) - \mathcal{P}(\mathbf{B})\right\| \leq C\left\{\frac{s\log(p)}{n\lambda_d}\right\}^{1/2}$$

for a sufficiently large constant C with probability tending to one as $n \to \infty$.

For the adaptive CHOMP estimator, let $\rho_n = \min_{j=1,\ldots,d}\min_{k\in S_j}|\beta_{jk}|$, the smallest magnitude of nonzero component across all dimensions. As n grows, we allow ρ_n to converge to a positive finite constant or to zero at a relatively slow rate. In particular, we assume

(C4) For each dimension $j = 1, \ldots, d$, the initial estimator $\bar{\boldsymbol{\beta}}_j$ satisfies $\|\bar{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j\|_\infty = O_p(\delta_n)$, for some sequence $\delta_n \to 0$, such that $\delta_n = o(\rho_n)$.

(C5) (Mutual incoherence) There exists a constant C such that

$$\left\|\mathcal{X}_{S^c}^\top\mathcal{X}_S\left(\mathcal{X}_S^\top\mathcal{X}_S\right)^{-1}\right\|_\infty \leq C < \infty,$$

where the notation $\mathcal{X}_S$ denotes the submatrix of $\mathcal{X}$ with column indices belonging to S.

Condition (C4) focuses on the initial estimator and is critical to ensure the weight vector is appropriately defined such that the weights for nonzero coefficients converge to a finite constant, and the weights for the zero coefficients diverge to infinity as the sample size grows. Similar conditions for the initial estimator have been used extensively in analyses of the adaptive lasso for high-dimensional sparse linear models, such as in Zou (2006) and Huang, Ma and Zhang (2008). The mutual incoherence condition (C5), which is also commonly used in analyses of the adaptive lasso, is a relatively weak condition on the correlatedness between the active and nonactive covariates. With these conditions, we establish the selection consistency of the adaptive CHOMP estimator.

Theorem 2. *Consider a multiple index model with $n\lambda_d = p^\nu$ for $\nu > 1$. Assume Conditions* (C1)–(C5) *hold, and the number of dimensions d is known. For each dimension $j = 1, \dots, d$ assume*

$$\rho_n^{-\gamma} s^{3/2} n^{-1/2} \to 0,\ \frac{\delta_n^\gamma}{\mu_j} \to 0,\ n^{-1}\rho_n^{-2\gamma} s \log(p) \to 0;\ \rho_n^{-2\gamma}\mu_j s^{1/2} \to 0.$$

Then, the adaptive CHOMP estimator $\hat{\boldsymbol{\beta}}_j^*$ defined in (3.3) *is selection consistent: pr$(\hat{\beta}_{jk}^* \neq 0) \to 1$ if $k \in S_j$, and pr$(\hat{\beta}_{jk}^* = 0) \to 1$ if $k \notin S_j$. Furthermore, if $s/n \to 0$, then the projection matrix $\mathcal{P}(\widehat{\mathbf{B}^*})$ associated with the adaptive Cholesky matrix estimator satisfies*

$$\left\|\mathcal{P}(\widehat{\mathbf{B}^*}) - \mathcal{P}(\mathbf{B})\right\|_F \leq C \left\{\frac{s\log(p)}{n\lambda_d}\right\}^{1/2},$$

for a sufficiently large constant C, with probability tending to one as $n \to \infty$.

When the initial estimator $\bar{\boldsymbol{\beta}}_j$ is the unpenalized estimate, the quantity $\delta_n = (p/n)^{1/2} \to 0$. If $\rho_n = O(1)$, Theorem 2 implies selection consistency holds if $s = O(n^\alpha)$ with $\alpha < 1/3$, $s\log(p)/n \to 0$, and the tuning parameter $\mu_j = O(n^\zeta)$ with $\zeta \in [\gamma(\nu^{-1} - 1), -\alpha/2]$.

Next, we study the large-sample properties of using the PIC to select the tuning parameters μ_j for the adaptive CHOMP estimator. To facilitate the theoretical analysis, we study a generalized form of the PIC, defined as

$$\text{PIC}(\mu_j;\tau_j) = \begin{cases} \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 + \tau_j \left\|\hat{\boldsymbol{\beta}}_j(\mu_j)\right\|_0, & \text{if } \hat{\boldsymbol{\beta}}_j(\mu_j) \neq \mathbf{0} \\ \infty, & \text{if } \hat{\boldsymbol{\beta}}_j(\mu_j) = \mathbf{0}, \end{cases} \tag{4.1}$$

where $\tau_j > 0$ is a model complexity term. Now, for a given value of the tuning parameter μ_j, let $\hat{\boldsymbol{\beta}}_j^*(\mu_j)$ be the corresponding solution of the minimization problem (3.3) and $\hat{S}_j\,(\mu_j) = \{k : \hat{\beta}_{jk}^*(\mu_j) \neq 0\}$. Next, we establish the following result for the selection consistency of the PIC.

Theorem 3. *Consider the multiple index model with the same conditions as those in Theorem* 2*. For each dimension* $j = 1, \ldots, d$*, denote*

$$\xi_j = \min \left\{ \frac{\beta_{jk}^2}{\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j}, \beta_{jk} \neq 0 \right\},$$

and assume that ξ_j *goes to zero at a slower rate than* p/n*. For any sequence* τ_j *that goes to zero at a rate slower than* p/n*, but faster than* ξ_j*, that is* $p/n \lesssim \tau_j \lesssim \xi_j$*, the adaptive CHOMP estimator with tuning parameter* μ_j *selected by minimizing* $\mathrm{PIC}(\mu_j; \tau_j)$ *defined in* (4.1)*, with the initial estimate* $\bar{\boldsymbol{\beta}}_j$ *being the unpenalized estimate, satisfies* $pr\{\hat{S}_j\,(\mu_j) = S_j\} \to 1$ *as* $n \to \infty$*.*

In Theorem 3, the quantity ξ_j controls the relative magnitude of the minimum nonzero coefficient compared to the ℓ_2 norm of the jth dimension. The condition that the model complexity term τ_j goes to zero faster than ξ_j ensures that minimizing the PIC does not lead to underfitting; in other words, when ξ_j is small, the model complexity term τ_j has to be small as well. Furthermore, the term τ_j has to go to zero at a rate slower than p/n to avoid overfitting. If all the nonzero components of $\boldsymbol{\beta}_j$ have the same magnitude, that is, $\xi_j = o(s_j^{-1})$, then Theorem 3 implies that the rate of convergence to zero is between p/n and s_j^{-1}, so we can set $\tau_j = \log(p)/p$, as defined in equation (3.7). In the simulation below, we verify that this choice of τ_j leads to strong variable selection in finite sample settings. To the best of our knowledge, our proposed PIC is the first data-driven approach to select a regularization parameter that theoretically guarantees achieving variable selection consistency for a central subspace.

5. Simulation Studies

5.1. Single-index model

We conduct simulation studies to investigate the performance of the proposed estimators in finite-sample settings. In all the settings below, the number of true dimensions d is assumed to be known. For the first simulation, we generate data pairs $(\boldsymbol{x}_i^\top, y_i)$ from one of the following models: (I) $y_i = \sin(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0) + \varepsilon_i$, (II) $y_i = 0.5(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0)^3 + \varepsilon_i$, and (III) $y_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i)$. These models are also considered by Lin, Zhao and Liu (2019) in their simulation study of the single-index model. Each row vector $\boldsymbol{x}_i$ is independently generated from a p-variate Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = \mathbf{D}\tilde{\boldsymbol{\Omega}}\mathbf{D}$, where $\tilde{\boldsymbol{\Omega}} = (\tilde{\sigma})_{ij}$ is a correlation matrix with elements being either (a) $\tilde{\sigma}_{ij} = 0.5^{|i-j|}$

(autoregressive structure) or (b) $\tilde{\sigma}_{ii} = 1$ and $\tilde{\sigma}_{ij} = 0.5$ when $i \neq j$ (homogeneous structure), and $\mathbf{D}$ is a diagonal matrix with elements generated randomly from the uniform distribution Unif$(0.5, 2)$. As a result, each covariate has a different variance. Next, the vector $\boldsymbol{\beta}_0$ is generated with the first $s = 5$ components being nonzero. These nonzero components have random sign and magnitude generated from the uniform distribution Unif$(1, 1.5)$. Finally, each random noise term ε_i is generated independently from the standard normal distribution. The sample size is fixed at $n = 1{,}000$, as in Lin, Zhao and Liu (2019), while the number of covariates varies over $p \in \{100, 200\}$. For each combination of the above parameters, 500 samples are generated. We set the number of slices to $H = 20$ when computing all the estimators as in Lin, Zhao and Liu (2019).

We compare the performance of the matrix lasso, CHOMP, adaptive CHOMP with $\gamma = 1$ and $\gamma = 2$, and the lasso SIR estimators, based on three metrics: the estimation error $\|\mathcal{P}(\boldsymbol{\beta}_0) - \mathcal{P}(\hat{\boldsymbol{\beta}})\|_F$, the false positive rate (FPR), and the false negative rate (FNR). These metrics are averaged across the 500 samples. For the CHOMP-type estimators, the tuning parameters are selected based on the PIC proposed in Section 3.3. For the lasso SIR estimator, we used a tuning parameter chosen using 10-fold cross-validation. In Section S3 of the Supplementary Material, we demonstrate that the lasso SIR with tuning parameter chosen in this way exhibits roughly the same performance as that of the lasso SIR estimator with a tuning parameter chosen to minimize the actual estimation error. The latter is not available in practice, because it requires knowledge of the true projection matrix $\mathcal{P}(\boldsymbol{\beta}_0)$. For the matrix lasso estimator, we examine its performance for different choices of tuning parameters (see Section S4 of the Supplementary Material). We find that a 10-fold cross-validation procedure often leads to the lowest estimation error among the methods that can be used in practice. In general, this tuning parameter selection method does not guarantee optimal performance; however, we use it to compare the performance of matrix lasso with that of other methods. How to select the best tuning parameter for the matrix lasso is outside the scope of this study. The results for the simulation settings when $\tilde{\boldsymbol{\Omega}}$ has the autoregressive structure are presented in Table 1; the results for the settings when $\tilde{\boldsymbol{\Omega}}$ has the homogeneous structure show similar conclusions, and are presented in Section S6 of the Supplementary Material.

Table 1 shows that the adaptive CHOMP estimator consistently has the best performance in terms of all three metrics. In particular, the estimation error of the CHOMP with a tuning parameter selected using the PIC is much lower than that of the matrix lasso estimator; these numerical results confirm the benefit of using the Cholesky decomposition. This benefit is strengthened further by the

Table 1. Performance of the estimators in the single-index model simulation with the correlation matrix $\tilde{\boldsymbol{\Omega}}$ having an autoregressive structure. Standard errors are included in parentheses. The lowest estimation error for each setting is highlighted.

Model	p	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	MLasso
				$\gamma = 1$	$\gamma = 2$		
(I)	100	Error	0.26 (0.12)	0.12 (0.06)	**0.10** (0.05)	0.19 (0.06)	0.39 (0.16)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.19 (0.09)	0.66 (0.23)
		FNR	0.00 (0.03)	0.00 (0.01)	0.00 (0.00)	0.00 (0.01)	0.02 (0.10)
	200	Error	0.29 (0.13)	0.13 (0.07)	**0.12** (0.06)	0.23 (0.08)	0.49 (0.09)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.72 (0.16)
		FNR	0.01 (0.04)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.01)
(II)	100	Error	0.07 (0.04)	**0.03** (0.01)	**0.03** (0.01)	0.06 (0.02)	0.10 (0.08)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.18 (0.09)	0.40 (0.15)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.03)
	200	Error	0.08 (0.04)	**0.03** (0.01)	**0.03** (0.01)	0.06 (0.02)	0.11 (0.02)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.50 (0.12)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
(III)	100	Error	0.11 (0.05)	**0.04** (0.02)	**0.04** (0.02)	0.08 (0.02)	0.13 (0.10)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.19 (0.09)	0.51 (0.16)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.05)
	200	Error	0.11 (0.06)	**0.04** (0.02)	**0.04** (0.02)	0.09 (0.03)	0.17 (0.03)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.62 (0.13)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

adaptive CHOMP; the adaptive estimators with $\gamma = 1$ and $\gamma = 2$ both have the smallest estimation error in all the settings. With regard to variable selection, both the matrix lasso and the lasso SIR estimators tend to overfit. The adaptive CHOMP estimator is able to fully recover the sparsity pattern of $\boldsymbol{\beta}_0$, with the average FPRs and FNRs being zero in all settings.

5.2. Multiple-index model

For the multiple-index model, we generate independent data pairs $(\boldsymbol{x}_i^\top, y_i)$ from the model (IV) $y_i = (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)\left\{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_2) + \varepsilon_i\right\}, i = 1, \ldots, n$. The model is also considered by Lin, Zhao and Liu (2019) in their simulation study of the multiple-index model. The predictors $\boldsymbol{x}_i$ and the random noise ε_i are generated in the same manner as in Section 5.1. We consider two different sparsity patterns in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In the first case, the two vectors have the same sparsity patterns; specifically, both have the first $s_1 = s_2 = 5$ components nonzero. In the second

Table 2. Performance of estimators in the multiple-index model simulation. Standard errors are in parentheses. The lowest estimation error for each setting is highlighted.

p	Sparsity	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	MLasso
				$\gamma = 1$	$\gamma = 2$		
100	Same	Error	0.31 (0.25)	0.22 (0.27)	**0.21** (0.28)	0.28 (0.27)	0.49 (0.32)
		FPR	0.00 (0.01)	0.00 (0.02)	0.01 (0.02)	0.32 (0.11)	0.63 (0.16)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.13)
	Different	Error	0.38 (0.13)	0.24 (0.10)	**0.22** (0.09)	0.26 (0.06)	0.48 (0.26)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.39 (0.11)	0.63 (0.16)
		FNR	0.00 (0.02)	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.03 (0.14)
200	Same	Error	0.32 (0.26)	**0.21** (0.28)	0.22 (0.29)	0.29 (0.26)	0.48 (0.24)
		FPR	0.00 (0.00)	0.00 (0.01)	0.00 (0.02)	0.21 (0.08)	0.66 (0.08)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)
	Different	Error	0.40 (0.13)	0.24 (0.09)	**0.22** (0.09)	0.30 (0.08)	0.46 (0.10)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.26 (0.09)	0.66 (0.08)
		FNR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)

case, the two vectors have different, but overlapping sparsity patterns. Specifically, the first $s_1 = 5$ components of $\boldsymbol{\beta}_1$ are nonzero, while the fourth to the seventh components of $\boldsymbol{\beta}_2$ are nonzero ($s_2 = 4$). The nonzero components are generated in the same way as those of $\boldsymbol{\beta}_0$ in the single-index model simulation.

For each sample, we compute the same estimators for the first and second dimensions separately. All the other parameters, including the tuning parameters for the estimators, are chosen in the same way as in Section 5.1. We assess the estimators based on the estimation error of the projection matrix, FPR, and FNR. Similarly to Tan et al. (2018), for the multiple-index model, the FPR and FNR are assessed based on the diagonal of the projection matrix. For example, a false positive for the jth component is counted when $\mathcal{P}(\hat{\mathbf{B}})_{jj}$ is non-zero but $\mathcal{P}(\mathbf{B})_{jj}$ is zero.

Table 2 shows that the adaptive CHOMP estimator with $\gamma = 2$ performs best overall among the considered estimators. Similarly to the single-index model, the CHOMP has considerably a smaller estimation error than that of the matrix lasso. With regard to variable selection, as in the single-index simulation, the matrix lasso and lasso SIR estimators tend to overfit. The CHOMP and adaptive CHOMP estimators are able to recover all the important covariates in two dimensions by having both average FPRs and FNRs zero or very close to zero in all the considered settings.

In summary, the simulation studies both verify the theoretical results, and demonstrate the superior performance of the adaptive CHOMP estimator, in conjunction with the PIC, in terms of both the estimation error and variable selection for the central subspace. In Section S3.1 of the Supplementary Material, we also examine the numerical performance of an adaptive version of the lasso SIR, and find that its performance is similar or worse than that of the adaptive CHOMP. Furthermore, as demonstrated in Section 7, the adaptive CHOMP estimator has a distinct advantage over the adaptive lasso SIR estimator in that it can be easily extended to other inverse-regression methods for the sparse estimation of a central subspace.

6. Data Application

We apply the methods to a cancer trial data set that contains information about the mean cancer mortality rate and 33 socioeconomic variables over the period 2010–2016 for $n = 3{,}047$ counties in the United States. The data set was created from merging several data, and is publicly available at `https://data.world/nrippner/cancer-trials`. It is of interest to model the mean cancer mortality rate (y) from all of the other variables. For illustration purposes, we remove one interval-censored predictor (*binnedInc*), which represents the median income per capita, binned by decile, and three other predictors having a considerable degree of missingness. This leaves us with $p = 28$ covariates ($\mathbf{X}$), all of which are then standardized before the analysis.

First, we use the `dr` package in R to compute the (unpenalized) SIR estimator and estimate the number of dimensions using the chi-square marginal dimension test of Cook (2004). The number of slices is set to $H = 20$. As a result, the number of dimensions of the central subspace is estimated to be three. Next, we calculate the CHOMP, adaptive CHOMP with $\gamma = 1$, adaptive CHOMP with $\gamma = 2$, and lasso SIR estimators. The tuning parameters for these penalized estimators are selected in the same fashion as in the simulation study. Because the true coefficient is not available for real data, we use the distance correlation between the sufficient predictors $\mathbf{X}\hat{\mathbf{B}}$ and the response y as a performance measure of the methods, where $\hat{\mathbf{B}}$ is a 28×3 estimated matrix of the three dimensions. A higher distance correlation means a stronger association (both linear and nonlinear) between two variables, thus implying a better prediction ability; see Székely, Rizzo and Bakirov (2007) and Wang, Shin and Wu (2018). We also examine the number of covariates selected by each method across all three dimensions.

Figure 1 shows that the methods produce similar sufficient dimensions for the first two dimensions. The response appears to have a strong linear relationship

Table 3. Performance of SDR methods in the cancer trial data set.

Methods	Distance correlation	# of important variables
Unpenalized SIR	0.59	28
CHOMP	0.61	4
Adaptive CHOMP ($\gamma = 1$)	0.59	4
Adaptive CHOMP ($\gamma = 2$)	0.66	16
Lasso SIR	0.42	28

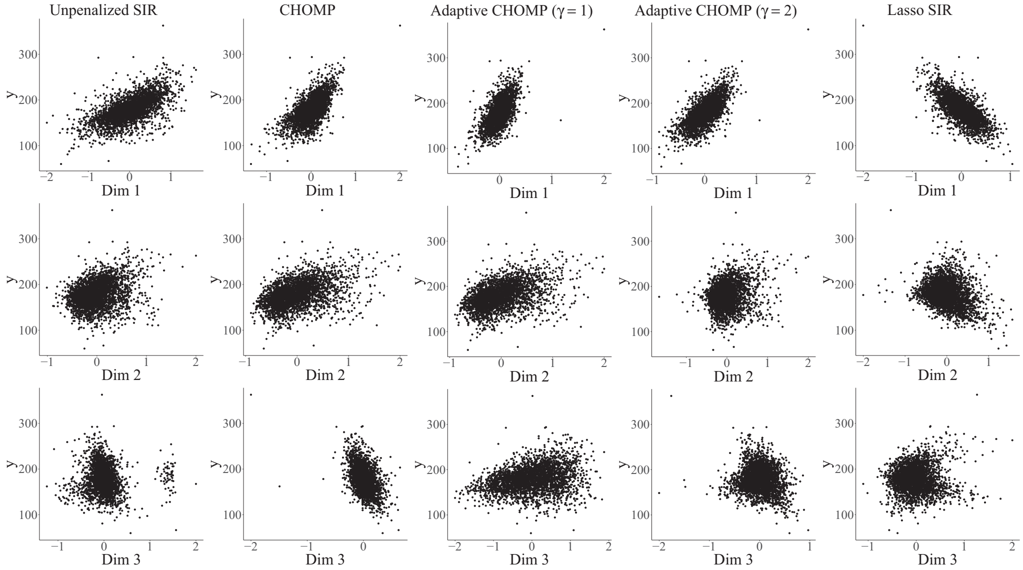

Figure 1. Plots of the response versus each sufficient predictor obtained by each method in the Cancer Trial dataset application.

with the first sufficient predictor, while the relationship between the response and the second sufficient predictor is more varied. For the third dimension, different methods produce sufficient estimators with quite different relationships with the response. Table 3 shows that the adaptive CHOMP estimator with $\gamma = 2$ produces sufficient predictors that have the highest distance correlation with the response, while the lasso SIR leads to sufficient predictors with the lowest distance correlation. For variable selection, on the one hand, the lasso SIR estimator selects all the covariates. As seen in the simulation results, the lasso SIR estimates tend to have a high FPR; this is likely to happen in this data application as well. On the other hand, the CHOMP estimator and adaptive CHOMP estimator with $\gamma = 1$ produce very sparse estimates, with only four covariates across all three dimensions; the adaptive CHOMP estimator with $\gamma = 2$

selects 16 estimates. Three covariates are selected by all three estimators: the mean number of diagnoses per capita, poverty rate, and percentage of residents whose highest education level attained is bachelor degree or higher. Compared with the simulation results, the performance of the adaptive CHOMP estimator in the data application is more sensitive to the choice of γ; this may be because the true nonzero coefficients may have a wider spread than they do in the simulation study. The optimal choice for γ is left as a topic for future research.

7. CHOMP for Other Inverse Regression Methods

As discussed at the end of Section 3, one advantage of the (adaptive) CHOMP estimator is its ability to extend to other sufficient dimension reduction methods. For example, we consider a class of methods that satisfy the population equation $\mathbf{\Sigma}\,\mathrm{col}(\mathbf{B}) = \mathrm{col}(\mathbf{Q})$, where the matrix $\mathbf{Q}$ is a method-specific kernel matrix. For example, the SIR corresponds to $\mathbf{Q} = \mathrm{var}\{\mathbb{E}(\mathbf{X}\,|y)\}$; the SAVE corresponds to $\mathbf{Q} = \mathbb{E}\left\{\mathbf{\Sigma} - \mathrm{var}(\mathbf{X}\,|y)\right\}^2$, and the pHd corresponds to $\mathbf{Q} = \mathbb{E}\left[\mathbf{X}\,\mathbf{X}^\top\{y - \mathbb{E}(y)\}\right]$, among many others. Next, let $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d$ be the eigenvectors associated with the d largest eigenvalues of the kernel matrix $\mathbf{Q}$. Then, the jth method-specific sufficient dimension satisfies $\mathbf{\Sigma}\,\boldsymbol{\beta}_j = \boldsymbol{\eta}_j$, for $j = 1, \ldots, d$. Following the same argument as in Section 3.1, we can then calculate $\hat{\boldsymbol{\kappa}}_j$ such that $\hat{\boldsymbol{L}}\hat{\boldsymbol{\kappa}}_j = \hat{\boldsymbol{\eta}}_j$, where $\hat{\boldsymbol{\eta}}_j$ is the jth eigenvector of the sample counterpart of the $\mathbf{Q}$ matrix. The CHOMP estimator corresponding to each sufficient dimension reduction method can then be constructed as the solution to the minimization problem (3.2). Similarly, the adaptive CHOMP estimator is the solution of the minimization problem (3.3), where the weights are set to $\omega_{jk} = |\bar{\beta}_{jk}|^{-\gamma}$, with $\bar{\beta}_{jk}$ being the kth element of the unpenalized method-specific estimator $\bar{\boldsymbol{\beta}}_j = \hat{\mathbf{\Sigma}}^{-1}\hat{\boldsymbol{\eta}}_j$. We refer to the resulting estimators as, for example, CHOMP-SAVE and adaptive CHOMP-SAVE when the CHOMP and adaptive CHOMP, respectively, are applied to the SAVE; similar definitions hold for the pHd.

We conduct a simulation study to demonstrate the performance of these estimators in scenarios where a SIR is unable to estimate the sufficient dimension. One such common scenario is when the link function f in (1.1) is symmetric around zero (Li (2018, Sec. 3.2)). We generate data from the single-index model (V) $y_i = (\boldsymbol{x}_i^\top\boldsymbol{\beta}_0)^2 + \varepsilon_i$ and the multiple-index model (VI) $y_i = (\boldsymbol{x}_i^\top\boldsymbol{\beta}_1)^2 - (\boldsymbol{x}_i^\top\boldsymbol{\beta}_2)^4 + \varepsilon_i$ for $i = 1, \ldots, n$. Each row vector $\boldsymbol{x}_i$ is generated from a multivariate normal distribution with the autoregressive correlation structure outlined in Section 5.1, and the random noise ε_i is generated from the standard normal distribution. The vector $\boldsymbol{\beta}_0$ is generated as in Section 5.1, while

Table 4. Performance of sufficient dimension estimators in the single-index simulation when the true link function f is symmetric. Standard errors are included in parentheses. The lowest estimation error is highlighted for each setting.

Model	Metric	Lasso SIR	CHOMP			Adaptive CHOMP ($\gamma = 2$)		
			SIR	SAVE	pHd	SIR	SAVE	pHd
(V)	Error	1.41 (0.03)	1.40 (0.03)	0.88 (0.35)	0.54 (0.22)	1.41 (0.02)	0.60 (0.45)	**0.29** (0.20)
	FPR	0.18 (0.11)	0.03 (0.03)	0.02 (0.03)	0.01 (0.02)	0.09 (0.05)	0.02 (0.04)	0.01 (0.02)
	FNR	0.82 (0.20)	0.96 (0.09)	0.36 (0.37)	0.08 (0.15)	0.90 (0.14)	0.20 (0.35)	0.02 (0.10)
(VI)	Error	1.99 (0.05)	2.00 (0.05)	0.84 (0.28)	1.57 (0.12)	1.99 (0.02)	**0.66** (0.32)	1.54 (0.12)
	FPR	0.27 (0.13)	0.06 (0.05)	0.01 (0.03)	0.03 (0.05)	0.16 (0.06)	0.02 (0.04)	0.11 (0.06)
	FNR	0.71 (0.22)	1.00 (0.11)	0.00 (0.13)	0.57 (0.15)	0.86 (0.15)	0.00 (0.09)	0.43 (0.11)

the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are generated to have different sparsity patterns as in the multiple-index model simulation in Section 5.2. We set the sample size to $n = 1000$ and the number of covariates to $p = 100$. We consider the lasso SIR, CHOMP-SIR, CHOMP-SAVE, and CHOMP-pHd estimators, and the adaptive version with $\gamma = 2$ for each CHOMP-based estimator. We use the PIC to select the tuning parameters for all the (adaptive) CHOMP estimators. We run each setting on 500 samples and compare the estimators based on the same performance metrics as those used in Section 5.

Table 4 shows that, in general, the SIR-based estimators do not perform well in terms of either estimation or variable selection when the true link function f is symmetric around zero. Using CHOMP combined with SAVE or pHd leads to a substantially smaller estimation error and improved variable selection. Furthermore, the adaptive CHOMP-SAVE and CHOMP-pHd estimators improve on the performance of the corresponding non-adaptive estimators. For the single-index model (V), these adaptive estimators have low FPRs and FNRs, and the adaptive CHOMP-pHd with $\gamma = 2$ has the lowest estimation error. For the multiple-index model (VI), the adaptive CHOMP-SAVE estimator with $\gamma = 2$ has the lowest estimation error and performs best in terms of variable selection. The CHOMP-pHd estimators have a relatively high FPR. In low-dimensional settings without any sparsity assumption, the empirical behavior of SAVE and pHd are known to be somewhat similar (Li (2018, Chap. 8, p.102)). However, the preliminary simulation results in this section suggest that the corresponding CHOMP-type estimators may have different performance when sparsity is imposed, depending on various factors, such as the true number of dimensions. We leave a full investigation of these CHOMP-type estimators to further research. However, overall, the results demonstrate that the CHOMP estimator can be extended to other sufficient dimension reduction methods, and confirm the advantages of the adaptive

CHOMP-type approach for the sparse estimation of a central subspace.

8. Conclusion

This paper presents three main contributions to the literature on sparse sufficient dimension reduction. First, we introduce the CHOMP approach, which is based on the Cholesky decomposition of the sample covariance matrix, for SIR estimation of a central subspace, along with the first data-driven PIC theoretically guaranteed to achieve variable selection consistency. Second, though the CHOMP estimator alone may not be as good as the lasso SIR in simulation studies, the CHOMP approach can be generalized easily to an adaptive version that not only achieves estimation and variable selection consistency, but also exhibits superior performance to that of the lasso SIR. Finally, the CHOMP approach is easily extended to other inverse regression-based estimators, for which the corresponding adaptive CHOMP estimators show superior empirical performance in terms of both estimation and variable selection.

In this paper, we focus on the CHOMP estimators when $n > p$ and $p/n \to 0$ as $n \to \infty$. In this setting, the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ is positive-definite and invertible, and so is its Cholesky factor $\hat{\boldsymbol{L}}$. In high-dimensional settings when $n < p$, the main challenge when using CHOMP is how to estimate $\boldsymbol{L}$ given that the matrix $\hat{\boldsymbol{\Sigma}}$ is no longer positive-definite and invertible. In Section S5 of the Supplementary Material, we explore using of CHOMP in a high dimensional setting where the Cholesky factor $\boldsymbol{L}$ can be estimated efficiently from regression techniques. Future research may investigate the theoretical properties of the CHOMP estimator in such high-dimensional settings as well as when combining CHOMP with other sufficient dimension reduction methods. Finally, how to estimate the number of dimensions d from the data in a sparse setting remains an open problem.

Supplementary Material

The Supplementary Material contains proofs of all technical results in Section 4, a brief review of the lasso SIR estimator, an adaptive version of it, an extension of the CHOMP technique to a high-dimensional setting, and additional simulation results, including the performance of the matrix lasso estimator under different choices of tuning parameters. Furthermore, the `R` code is available on the Github page of the corresponding author at `github.com/lnghiemum`.

Acknowledgments

We thank the two anonymous Reviewers and the Associate Editor who have provided constructive feedback to help us improve the paper. All authors are supported by the Australian Research Council Discovery Grant DRC180100836.

References

Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics-Theory and Methods* **29**, 2109–2121.

Cook, R. D. (1998). Principal Hessian directions revisited. *Journal of the American Statistical Association* **93**, 84–94.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics* **32**, 1062–1092.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 1603–1618.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R.* Chapman and Hall/CRC, Boca Raton.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025–1039.

Lin, Q., Li, X., Huang, D. and Liu, J. S. (2021). On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics* **49**, 1–20.

Lin, Q., Zhao, Z. and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

Lin, Q., Zhao, Z. and Liu, J. S. (2019). Sparse sliced inverse regression via Lasso. *Journal of the American Statistical Association* **114**, 1726–1739.

Qian, W., Ding, S. and Cook, R. D. (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association* **114**, 1277–1290.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Tan, K. M., Wang, Z., Zhang, T., Liu, H. and Cook, R. D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika* **105**, 769–782.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint.* Cambridge University Press, New York.

Wang, C., Shin, S. J. and Wu, Y. (2018). Principal quantile regression for sufficient dimension reduction with heteroscedasticity. *Electronic Journal of Statistics* **12**, 2114–2140.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 363–410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 879–892.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Linh H. Nghiem

School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia.

E-mail: linh.nghiem@anu.edu.au

Francis K. C. Hui

Research School of Finance, Actuarial Studies and Statistics, Australian National University, Acton, ACT 2600, Australia.

E-mail: francis.hui@anu.edu.au

Samuel Müller

School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW 2109, Australia.

E-mail: samuel.muller@mq.edu.au

A. H. Welsh

Research School of Finance, Actuarial Studies and Statistics, Australian National University, Acton, ACT 2600, Australia.

E-mail: alan.welsh@anu.edu.au

(Received October 2020; accepted June 2021)