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.