Communications for Statistical Applications and Methods
2013, Vol. 20, No. 4, 291–300

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.4.291

Method-Free Permutation Predictor Hypothesis Tests in Sufficient Dimension Reduction

Kyungjin Lee[a], Suji Oh[a], Jae Keun Yoo[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

In this paper, we propose method-free permutation predictor hypothesis tests in the context of sufficient dimension reduction. Different from an existing method-free bootstrap approach, predictor hypotheses are evaluated based on p-values; therefore, usual statistical practitioners should have a potential preference. Numerical studies validate the developed theories, and real data application is provided.

Keywords: Permutation, predictor hypothesis tests, regression, sufficient dimension reduction.

1. Introduction

Sufficient dimension reduction (SDR) in regression of $Y|\mathbf{X} \in \mathbb{R}^p$ pursues the replacement of the original p-dimensional predictors $\mathbf{X}$ by a lower-dimensional linearly transformed predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information about selected aspects of the conditional distribution of $Y|\mathbf{X}$, where $\boldsymbol{\eta}$ is a $p \times d$ matrix with $d < p$. Its equivalent numerical expression is:

$$Y \perp\!\!\!\perp f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence and $f(\mathbf{X})$ varies depending on the selected aspects of $Y|\mathbf{X}$.

A subspace spanned by the columns of $\boldsymbol{\eta}$ satisfying statement (1.1) is called a *dimension reduction subspace* (DRS). Then, naturally, one seeks for the minimal subspace among all possible DRSs. Hereafter, for notational convenience, a subspace spanned by the columns of a $p \times q$ matrix $\mathbf{A}$ will be denoted as $\mathcal{S}(\mathbf{A})$.

We explain changes of statement 1.1 and its meaning depending on the form of $f(\mathbf{X})$. If the main interest in regression is the conditional distribution itself, $f(\mathbf{X})$ is equal to $\mathbf{X}$, and statement 1.1 is:

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}.$$

In such case, the minimal subspace is called the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ (Cook, 1998b). Then $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on $Y|\mathbf{X}$.

If the conditional mean $E(Y|\mathbf{X})$ is of main interest, then $f(\mathbf{X})$ becomes $E(Y|\mathbf{X})$, and statement 1.1 is:

$$Y \perp\!\!\!\perp E(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}.$$

This work was supported by the Basic Science Research Program of the National Research Foundation of Korea (KRF) funded by the Ministry of Education, Science and Technology (2012-040077) for Jae Keun Yoo.

[1] Corresponding author: Associate Professor, Department of Statistics, Ewha Womans University, 11-1 Daehyun-Dong Seodaemun-Gu, Seoul 120-750, Korea. E-mail: peter.yoo@ewha.ac.kr

Then the related minimal subspace is called the *central mean subspace* $\mathcal{S}_{E(Y|\mathbf{X})}$ (Cook and Li, 2002). In this case, $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on $E(Y|\mathbf{X})$.

When the first k conditional moments of $Y|\mathbf{X}$ is of primary focus, $f(\mathbf{X})$ becomes a set of the conditional moments of $Y|\mathbf{X}$ up to k such as $\{E(Y|\mathbf{X}), M^{(2)}(Y|\mathbf{X}), \dots, M^{(k)}(Y|\mathbf{X})\}$, and statement 1.1 is:

$$Y \perp\!\!\!\perp \left\{E(Y|\mathbf{X}), M^{(2)}(Y|\mathbf{X}), \dots, M^{(k)}(Y|\mathbf{X})\right\} \Big| \boldsymbol{\eta}^{\mathrm{T}}\mathbf{X},$$

where $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$ for $k \geq 2$.

Then the minimal space is called the *central k^{th}-moment subspace* $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ (Yin and Cook, 2002). Among the three subspaces, the following relation are easily established: $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}^{(k)}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. For the various target subspaces, their true dimension will be called *structural dimension* and will be denoted as d throughout the rest of the paper.

The estimation of the three target subspaces of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ and $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$, or equivalently and simply $\mathcal{S}(\boldsymbol{\eta})$, should be the primary interest in SDR. Most SDR methods usually connect $\mathcal{S}(\boldsymbol{\eta})$ to certain kernel matrix $\mathbf{M} \in \mathbb{R}^{p\times r}$, fully or partially informative to $\mathcal{S}(\boldsymbol{\eta})$ and estimable under certain conditions, such that $\mathcal{S}(\mathbf{M}) \subseteq \mathcal{S}(\boldsymbol{\eta})$. Usually it is assumed that $\mathcal{S}(\mathbf{M}) = \mathcal{S}(\boldsymbol{\eta})$. Based on this relation, the inference about $\mathcal{S}(\boldsymbol{\eta})$, equivalently, $\boldsymbol{\eta}$ should be done through $\mathbf{M}$.

Selections of predictors significant to regression are often a crucial procedure in model-based regression; however, selections of predictors have been largely out of focus in SDR context, until Cook (2004) recently defined predictor hypothesis such that

$$\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = O_p, \tag{1.2}$$

where $\mathcal{H}$ is an h-dimensional user-selected subspace of predictor space and results in a subset of $\mathbf{X}$, $\mathbf{P}_{\mathcal{H}}$ is an orthogonal projection onto $\mathcal{H}$, and O_p indicates the origin in $\mathbb{R}^p$.

If statement (1.2) is rephrased by a conditional independence statement such as (1.1), it would be more helpful to understand the statement. Let partition the original predictor $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_h = \mathbf{P}_{\mathcal{H}}\mathbf{X}, \mathbf{X}_{-h} = \mathbf{P}_{\mathcal{H}^\perp}\mathbf{X})$, where $\mathcal{H}^\perp$ is the orthogonal complement of $\mathcal{H}$. The statement in (1.2) holds, if and only if $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-h}$. That is, $\mathbf{X}_h$ does not contribute to $Y|\mathbf{X}$. Letting $\mathbf{H} \in \mathbb{R}^{p\times h}$ be an orthonormal basis matrix of $\mathcal{H}$ and setting $\mathbf{H} = \mathrm{e}_i$, $i = 1, \dots, p$, the predictor hypothesis tests can be considered as a variable selection procedure, where e_i represents a canonical basis with the ith element one and elsewhere zeros. In addition, Cook (2004) provided implementation in the context of SIR.

Cook's work greatly contributes to test predictor hypothesis in SDR literature by providing general paradigm for it; however, several limitations encounter its direct application to various SDR methods. First, the statement in (1.2) holds for $\mathcal{S}_{Y|\mathbf{X}}$ alone. Second, it is not clear to derive test statistics for the various target subspaces naturally due to the limitation. Third, related tests statistics are developed; however, it may be problematic in the derivation of their asymptotics.

To overcome these deficits, Yoo (2011) newly defined a unified predictor hypothesis tests applicable to all target subspaces and related SDR methods. In addition, a bootstrap approach to test the predictor hypothesis was proposed. Yoo's work provides an unified paradigm for predictor hypothesis tests through using bootstrapping techniques; however, it has its limitation in practical use because it does not provide p-values for the tests. Instead, it suggests a general guideline on how to determine which predictors are important. Therefore, debates for the decisions may occur and there exists no clear winner in some cases. We will explain this briefly in the later section.

This manuscript develop a way to test the unified predictor hypothesis with supporting p-values. For this we adopt a permutation approach used in sufficient dimension reduction context. We first suggest proper test statistics. In addition, their sampling distributions are empirically derived through samples constructed by permuting null parts of predictors under the null predictor hypothesis. Then finally, p-values are computed from the empirical distribution to evaluate the hypothesis.

The manuscript is organized as follows. Section 2 is devoted to a short review of unified predictor hypothesis and bootstrapping approach. Section 3 develops a permutation unified predictor hypothesis tests. Section 4 presents numerical studies and real data application. Section 5 summarizes our work.

2. Literature Review

2.1. Popular sufficient dimension reduction methodologies

Here four popular SDR methodologies (among others) are briefly introduced, and those methods will be used in numerical studies. Please see the references for more details. Hereafter we denote $\mathbf{M}_\bullet$ to represent kernel matrices used in each method to estimate one of the three subspaces explained in a section of Introduction. We will assume that $\mathbf{M}_\bullet$s are fully informative to their own target subspaces.

Sliced inverse regression (SIR; Li, 1991):

A method of SIR estimates $\mathcal{S}_{Y|\mathbf{X}}$ through the inverse mean of $E(\mathbf{X}|Y)$. Let $\mathbf{M}_{\text{SIR}} = \text{cov}\{E(\mathbf{X}|Y)\}$. Then the relation of $\mathcal{S}(\mathbf{M}_{\text{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$ holds. If Y is categorical, the construction of a sample version of $E(\mathbf{X}|Y)$ is quite straightforward. In case that Y is many-valued or continuous, Y is partitioned by dividing its range into h slices.

Sliced average variance estimation (SAVE; Cook and Weisberg, 1991):

While the SIR uses $E(\mathbf{X}|Y)$ to restore $\mathcal{S}_{Y|\mathbf{X}}$, a method of SAVE considers the second inverse conditional moment of $\text{cov}(\mathbf{X}|Y)$. It can be shown that $\mathbf{M}_{\text{SAVE}} = E\{\mathbf{I}_p - \text{cov}(\mathbf{X}|Y)\}^2$ is informative to $\mathcal{S}_{Y|\mathbf{X}}$, in sense that $\mathcal{S}(\mathbf{M}_{\text{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The quantity of $\text{cov}(\mathbf{X}|Y)$ is estimated within each slice of Y just like the SIR.

Principal Hessian directions (pHd; Li, 1992):

The original proposal of a method of pHd in Li (1992) suggests to construct $\boldsymbol{\Sigma}_{yxx} = E[\{Y - E(Y)\}\mathbf{X}\mathbf{X}^{\text{T}}]$ to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Cook (1998a), however, showed that $\boldsymbol{\Sigma}_{yxx}$ actually estimated $\mathcal{S}_{E(Y|\mathbf{X})}$, not $\mathcal{S}_{Y|\mathbf{X}}$. In practice, instead of Y in $\boldsymbol{\Sigma}_{yxx}$, the OLS residuals of $\epsilon = Y - E(Y) - \boldsymbol{\beta}^{\text{T}}\{\mathbf{X} - E(\mathbf{X})\}$ are usually used, where $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\,\text{cov}(\mathbf{X}, Y)$ and $\boldsymbol{\Sigma} = \text{cov}(\mathbf{X})$. Then we construct $\boldsymbol{\Sigma}_{\epsilon xx} = E(\epsilon\mathbf{X}\mathbf{X}^{\text{T}})$, and this approach is called residual-based pHd. Then we have $\mathbf{M}_{\text{pHd}} = \boldsymbol{\Sigma}_{\epsilon xx}$.

New class dimension reduction (NCM; Ye and Weiss, 2003):

A new class of dimension reduction was proposed by Ye and Weiss (2003). The Is key idea is to construct a weighted mean of two kernel matrices among the three methods introduced in (1)–(3), for example, $\mathbf{M}_{\text{NCM}} = \omega\mathbf{M}_{\text{SIR}} + (1 - \omega)\mathbf{M}_{\text{SAVE}}$, where $0 < \omega < 1$. Then Ye and Weiss (2003) showed that $\mathcal{S}(\mathbf{M}_{\text{NCM}}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. A bootstrapping approach was employed to find the optimal ω.

2.2. Unified Predictor hypothesis and bootstrapping approach

A unified predictor hypothesis is a predictor hypothesis directly applicable to all types of target subspaces and it is: $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p$, where $\mathcal{S}_{f(\mathbf{X})}$ represents an user-selected target subspace. Then the hypothesis is equivalently rephrased as the following conditional independence statement: $Y \perp\!\!\!\perp f(\mathbf{X})|\mathbf{P}_{\mathcal{H}^\perp}\mathbf{X}$. The unified predictor hypothesis implies that the subset $\mathbf{P}_{\mathcal{H}}\mathbf{X}$ does not contribute to the selected aspect of $Y|\mathbf{X}$.

Based on the unified hypothesis, two types of hypothesis forms might be considered depending on application-specific requirements:

Unified marginal predictor hypothesis

$$H_0^M : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p \quad \text{versus} \quad H_1^M : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} \neq \mathcal{O}_p.$$

Unified conditional predictor hypothesis

$$H_0^C : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p \text{ given } d = m \quad \text{versus} \quad H_1^C : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} \neq \mathcal{O}_p \text{ given } d = m.$$

To highlight the clear difference between the marginal and conditional hypotheses, H_0^M and H_0^C can be rewritten as:

$$H_0^M : Y \perp\!\!\!\perp f(\mathbf{X})|\mathbf{P}_{\mathcal{H}^\perp}\mathbf{X} \quad \text{and} \quad H_0^C : Y \perp\!\!\!\perp f(\mathbf{X})|\mathbf{P}_{\mathcal{S}(\mathbf{B})}\mathbf{X} \text{ with } \mathbf{P}_{\mathcal{H}}\mathbf{B} = 0, \tag{2.1}$$

where $\mathbf{B} \in \mathbb{R}^{p\times m}$ stands for an orthonormal basis matrix for any m-dimensional $\mathcal{S}_{f(\mathbf{X})}$. As we can see, the difference is given in the conditioning components. In the marginal hypothesis, no requirement is given, so $\mathbf{P}_{\mathcal{H}^\perp}\mathbf{X}$ must be conditioned, while the specification of d forces that $\mathbf{P}_{\mathcal{S}(\mathbf{B})}\mathbf{X}$ must appear in the conditioning part with restriction of $\mathbf{P}_{\mathcal{H}}\mathbf{B} = 0$ for the conditional hypothesis. Due to the difference, computations of related statistics are different in the bootstrap approach; however, the method to construct bootstrap samples is the same.

For the marginal case, a distance between $\mathcal{S}(\hat{\mathbf{M}})$ and $\mathcal{S}(\mathbf{P}_{\mathcal{H}^\perp}\hat{\mathbf{M}})$ is measured, where a $p\times p$ matrix $\hat{\mathbf{M}}$ stands for a sample version of related kernel matrices constructed through various SDR methods to estimate $\boldsymbol{\eta}$. For the criteria of the distance, Yoo (2011) adopted one minus vector correlation coefficient (Hotelling, 1936). For the conditional case, a distance between $\mathcal{S}(\hat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{P}_{\mathcal{H}^\perp}\hat{\mathbf{B}})$ is used as test statistics, where $\hat{\mathbf{B}}$ is a sample version orthonormal basis matrix of a m-dimensional DRS. In practice $\hat{\mathbf{B}}$ is the eigenvectors of $\hat{\mathbf{M}}$ corresponding to its m largest eigenvalues.

To have empirical distributions of the two statistics, we construct bootstrap samples from pairs of the original data $(Y_i, \mathbf{X}_i)$, $i = 1,\ldots,p$, and compute $\mathbf{P}_{\mathcal{H}^\perp}\hat{\mathbf{M}}$ and $\mathbf{P}_{\mathcal{H}^\perp}\hat{\mathbf{B}}$ from the bootstrap samples. Then the average correlations $\hat{\mathbf{M}}$ and $\hat{\mathbf{B}}$ from the original data and the corresponding quantities from the bootstrap samples are used for the determination. This procedure is done for all predictors and it is determined that predictors with relatively larger distances are significant to the regression. Yoo (2011) suggests 0.8 for the marginal tests and 0.4 or 0.6 for conditional tests; however, the suggested values can be contingently changed depending on the data. For more details on the bootstrap approach, readers are recommended to refer to Yoo (2011).

The bootstrap approach has a major limitation in practical use. Guidelines for the determination are provided; however, the decision should be relative and debatable. In the next section, we develop a permutation approach to provide p-values for the marginal and conditional unified predictor hypothesis. New test statistics are proposed for the new approach.

3. Permutation Unified Predictor Hypothesis Tests

3.1. Marginal unified permutation predictor hypothesis test

In the marginal predictor hypothesis, the dimension of $\mathcal{S}_{f(\mathbf{X})}$, that is, $\mathcal{S}(\boldsymbol{\eta})$, is not defined. Therefore, the null hypothesis of H_0^M : $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p$ implies that $\mathbf{P}_{\mathcal{H}}\mathbf{X}$ is redundant to regression with respect to $\mathcal{S}_{f(\mathbf{X})}$. Based on this, first, we partition $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_{\mathbf{H}_\perp} = \mathbf{H}_\perp^{\mathrm{T}}\mathbf{X}, \mathbf{X}_{\mathbf{H}} = \mathbf{H}^{\mathrm{T}}\mathbf{X})$, where matrices of $\mathbf{H}$ and $\mathbf{H}_\perp$ stand for orthonormal basis matrices of $\mathcal{H}$ and $\mathcal{H}^\perp$ respectively, with $\mathbf{H}_\perp^{\mathrm{T}}\mathbf{H} = 0$. Then the partial predictor $\mathbf{H}^{\mathrm{T}}\mathbf{X}$ alone is randomly permuted and permutation predictors are constructed accordingly, such as $\mathbf{X}^{\mathrm{perm}} = (\mathbf{X}_{\mathbf{H}_\perp}, \mathbf{X}_{\mathbf{H}}^{\mathrm{perm}})$. Next, recalling that $\hat{\mathbf{M}}$ is the related kernel matrix to estimate $\boldsymbol{\eta}$, we compute the ordered eigenvalues of $\hat{\lambda}_i$, $i = 1, \dots, p$, of $\hat{\mathbf{M}}$, with $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. As test statistics to evaluate the marginal unified predictor hypothesis, we consider the following quantity:

$$\mathcal{T}(\mathcal{H}) = \sum_{i=1}^{p} \hat{\lambda}_i.$$

Let $\hat{\lambda}_i^{\text{ref}}$ and $\hat{\lambda}_i^{\text{perm}}$, $i = 1, \dots, p$, be the eigenvalues of $\hat{\mathbf{M}}$ constructed from the original sample and its permuted samples. To derive an empirical distribution of $\mathcal{T}(\mathcal{H})$, sums of eigenvalues of $\mathcal{T}_{\text{perm}}(\mathcal{H}) = \sum_{i=1}^{p} \hat{\lambda}_i^{\text{perm}}$ are computed from the k permutation samples. Then, the proportions of $\mathcal{T}_{\text{perm}}(\mathcal{H}) > \mathcal{T}_{\text{ref}}(\mathcal{H})$ are computed, which are the p-values to evaluate the null hypothesis, where $\mathcal{T}_{\text{ref}}(\mathcal{H}) = \sum_{i=1}^{p} \hat{\lambda}_i^{\text{ref}}$. The reasoning is as follows. If the null hypothesis is true, we normally expect relatively high proportions for the case of $\mathcal{T}_{\text{perm}}(\mathcal{H}) > \mathcal{T}_{\text{ref}}(\mathcal{H})$, because $\mathbf{X}_{\mathbf{H}}^{\text{perm}}$ has the null impacts to the regression and hence gives only random perturbation to have nothing to with the regression. The marginal unified permutation predictor hypothesis tests are summarized as:

- **Marginal unified permutation predictor hypothesis test**

1. Based on $H_0 : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p$, partition $\mathbf{X}$ into $(\mathbf{X}_{\mathbf{H}_\perp} = \mathbf{H}_\perp^{\mathrm{T}}\mathbf{X}, \mathbf{X}_{\mathbf{H}} = \mathbf{H}^{\mathrm{T}}\mathbf{X})$.
2. Construct a sample version of related kernel matrices, $\hat{\mathbf{M}}$, from the original sample, and compute $\mathcal{T}_{\text{ref}}(\mathcal{H}) = \sum_{i=1}^{p} \hat{\lambda}_i^{\text{ref}}$, where λ_i^{ref}s are its eigenvalues.
3. Permute $\mathbf{X}_{\mathbf{H}}$ randomly and construct permuted predictors of $\mathbf{X}^{\text{perm}} = (\mathbf{X}_{\mathbf{H}_\perp}, \mathbf{X}_{\mathbf{H}}^{\text{perm}})$.
4. From the permutation samples, obtain $\mathcal{T}_{\text{perm}}(\mathcal{H}) = \sum_{i=1}^{p} \hat{\lambda}_i^{\text{perm}}$.
5. Repeat steps 3 to 4 k times.
6. Calculate the percentage of $\mathcal{T}_{\text{perm}}(\mathcal{H}) > \mathcal{T}_{\text{ref}}(\mathcal{H})$, and report the percentage as a p-value.

3.2. Conditional unified permutation predictor hypothesis test

Recall the conditional hypothesis of $H_0^C : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = \mathcal{O}_p$ given $d = m$ versus $H_1^C : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} \neq \mathcal{O}_p$ given $d = m$. In H_0^C and H_1^C, the dimension of $\mathcal{S}_{f(\mathbf{X})}$ is given. This implies that we had better test the predictor hypothesis with respect to $\boldsymbol{\eta}$, not the related kernel matrix $\mathbf{M}$. Based on $\mathcal{H}$; therefore, we can partition $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta} = \mathbf{P}_{\mathbf{H}_\perp}\boldsymbol{\eta} + \mathbf{P}_{\mathbf{H}}\boldsymbol{\eta}.$$

Under H_0^C, $\mathbf{P}_{\mathbf{H}}\boldsymbol{\eta} = 0$, and hence we have the following relation:

$$Y \perp\!\!\!\perp f(\mathbf{X}) | (\mathbf{P}_{\mathbf{H}^\perp}\boldsymbol{\eta})^{\mathrm{T}}\mathbf{X}.$$

That is, by given $d = m$, we do not have to consider p-dimensional predictor $\mathbf{X}$, but, instead, we consider $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. Under H_0^C, the regression of $Y|\{(\mathbf{P}_{\mathbf{H}^\perp}\boldsymbol{\eta})^{\mathrm{T}}\mathbf{X}, (\mathbf{P}_{\mathbf{H}}\boldsymbol{\eta})^{\mathrm{T}}\mathbf{X}\}$, with randomly permuting $(\mathbf{P}_{\mathbf{H}}\boldsymbol{\eta})^{\mathrm{T}}\mathbf{X}$, should be equally informative to the regression of $Y|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ in the context of specific SDR methodologies to recover $\boldsymbol{\eta}$. The information of the two regressions can be, naturally, measured by the sum of the m largest eigenvalues computed from the chosen SDR methods. For the conditional tests, we use the following quantity as a test statistic:

$$\mathcal{T}(\mathcal{H}|d) = \sum_{i=1}^{m} \hat{\lambda}_i.$$

Then the reference test statistic of $\mathcal{T}_{\text{ref}}(\mathcal{H}|d)$ is computed from a regression of $Y|\hat{\boldsymbol{\eta}}^{\text{T}}\mathbf{X}$. Let $\mathbf{X}_{\mathbf{1}|d} = (\mathbf{P}_{\mathbf{H}^\perp}\hat{\boldsymbol{\eta}})^{\text{T}}\mathbf{X}$ and $\mathbf{X}_{\mathbf{0}|d} = (\mathbf{P}_{\mathbf{H}}\hat{\boldsymbol{\eta}})^{\text{T}}\mathbf{X}$. Then the permuted test statistics of $\mathcal{T}_{\text{perm}}(\mathcal{H}|d)$ are computed from the regression of $Y|(\mathbf{X}_{\mathbf{1}|d}, \mathbf{X}_{\mathbf{0}|d}^{\text{perm}})$.

- **Conditional unified permutation predictor hypothesis test**

1. Based on $H_0 : \mathbf{P}_{\mathcal{H}}\mathcal{S}_{f(\mathbf{X})} = O_p$ given $d = m$, partition $\hat{\boldsymbol{\eta}}^{\text{T}}\mathbf{X}$ into $(\mathbf{X}_{\mathbf{1}|d} = (\mathbf{P}_{\mathbf{H}^\perp}\hat{\boldsymbol{\eta}})^{\text{T}}\mathbf{X}, \mathbf{X}_{\mathbf{0}|d} = (\mathbf{P}_{\mathbf{H}}\hat{\boldsymbol{\eta}})^{\text{T}}\mathbf{X})$, where $\hat{\boldsymbol{\eta}} \in \mathbb{R}^{p\times m}$ is the estimate of $\mathcal{S}_{f(\mathbf{X})}$ under $d = m$.
2. Compute the reference test statistics from $Y|\hat{\boldsymbol{\eta}}^{\text{T}}\mathbf{X}$ such that
$$\mathcal{T}_{\text{ref}}(\mathcal{H}|d) = \sum_{i=1}^{d} \hat{\lambda}_i^{\text{ref}},$$
where λ_i^{ref}s are its eigenvalues.
3. Permute $\mathbf{X}_{\mathbf{0}|d}$ randomly and construct permuted predictors of $\mathbf{X}_d^{\text{perm}} = (\mathbf{X}_{\mathbf{1}|d}, \mathbf{X}_{\mathbf{0}|d}^{\text{perm}})$.
4. From the permutation samples, obtain $\mathcal{T}_{\text{perm}}(\mathcal{H}|d) = \sum_{i=1}^{d} \hat{\lambda}_i^{\text{perm}}$.
5. Repeat steps 3 to 4 k times.
6. Calculate the percentage of $\mathcal{T}_{\text{perm}}(\mathcal{H}|d) > \mathcal{T}_{\text{ref}}(\mathcal{H}|d)$, and report the percentage as p-values.

4. Numerical Studies and Data Analysis

4.1. Numerical studies

We consider two artificial models for numerical studies. The following predictor configurations were used in the two models. The coordinates of $\mathbf{X} = (X_1, \ldots, X_5)^{\text{T}}$ were independently generated from $N(0, 1)$. A random error ε, independent of $\mathbf{X}$, was also sampled from $N(0, 1)$. Then, the next four models were constructed:

Model 1 $Y|\mathbf{X} = X_1 + \varepsilon$; **Model 2** $Y|\mathbf{X} = X_1^2 + \varepsilon$;

Model 3 $Y|\mathbf{X} = X_1^2 + X_2^2 + \varepsilon$; **Model 4** $Y|\mathbf{X} = X_1 + X_2^2 + \varepsilon$.

We conducted marginal and conditional permutation predictor tests based on SIR (Li, 1991) for Model 1 and pHd (Li, 1992) and SAVE (Cook and Weisberg, 1991) for Model 2. And, for Model 3, the method of pHd alone was considered, and NCM (Ye and Weiss, 2003) to combine the two methods of SIR and SAVE with two weights of $\omega = 0.5$ and $\omega = 0.7$ for SIR was used for Model 4. Five and four slices were considered for SIR and SAVE respectively. For the conditional bootstrap and permutation tests, the true dimensions of $d = 1$ for Models 1–2 and $d = 2$ for Models 3–4 were used. In addition, each model was iterated 100 times with 500 permutations for $n = 50$ and $n = 100$ respectively. In all tests, level 5% was used. As the summary of the studies, we report the rejection percentages of the null hypotheses for testing each coordinate effect. In Models 1–2, X_1 alone contributes to the regression, and X_1 and X_2 do to the regression for Models 3–4. Thus, in Models 1–2, the rejection percentages of X_1 should be close to 100%, which indicate the observed powers. In addition, the rejection percentages for all other predictors should be close to the nominal level 5%, which represent the observed levels. The difference for Models 3–4 is highlighted to X_2 because

Table 1: Percentages of rejection of the null hypothesis for Models 1–4 in section 4.1; M1_{SIR}, Model 1 by SIR; M2_{pHd}, Model 2 by pHd ; M2_{SAVE}, Model 2 by SAVE; M3_{pHd}, Model 3 by pHd; M4_{NCA}, Model 4 by New Class Approach

		Marginal Permutation Tests					Conditional Permutation Tests				
		X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
M1_{SIR}	$n = 50$	100	3.00	2.00	4.00	6.00	100	3.00	6.00	6.00	9.00
	$n = 100$	100	5.00	6.00	5.00	4.00	100	5.00	6.00	5.00	4.00
M2_{pHd}	$n = 50$	100	3.00	7.00	5.00	1.00	100	2.00	7.00	4.00	12.0
	$n = 100$	100	2.00	6.00	6.00	4.00	100	4.00	3.00	6.00	4.00
M2_{SAVE}	$n = 50$	75.0	4.00	5.00	6.00	2.00	100.0	7.00	6.00	8.00	6.00
	$n = 100$	100	5.00	5.00	7.00	3.00	100	7.00	5.00	4.00	6.00
M3_{phd}	$n = 50$	97.0	96.0	3.00	3.00	7.00	98.0	98.0	7.00	3.00	7.00
	$n = 100$	100	100	4.00	4.00	3.00	100	100	6.00	4.00	6.00
M4_{NCA}	$n = 50$	31.0	66.0	3.00	4.00	4.00	75.0	80.0	9.00	6.00	3.00
$\omega = 0.5$	$n = 100$	90.0	100	6.00	4.00	6.00	100	100	5.00	7.00	6.00
M4_{NCA}	$n = 50$	75.0	44.0	7.00	6.00	7.00	85.0	61.0	9.00	8.00	4.00
$\omega = 0.7$	$n = 100$	100	95.0	7.00	8.00	5.00	100	100	5.00	8.00	6.00

the predictor contributes to the regressions; therefore, the rejection percentage for X_2 represents the observed level in the models. Table 1 summarizes the rejection percentages for X_i, $i = 1, \dots 5$, for the four models.

Table 1 shows that the marginal and conditional permutation tests are almost equally good, although the conditional tests turns out more powerful in Model 2 through the SAVE and Model 4. Both tests provide reasonably good observed powers and levels with $n = 100$ in all models. Comparing the test performances between pHd and SAVE, the method of pHd shows more reliable performances than SAVE especially with smaller sample size $n = 50$. This is because of implementation of the methods itself. The method of SAVE methodologically requires covariance matrices for each slice. Thus, with smaller samples, the covariances within each slice may be not well-estimated, which may cause relatively worse test results than pHd. For Model 4, with $n = 50$, the performances for testing X_1 and X_2 are quite different, especially in the marginal test. With $\omega = 0.5$, the effect of X_1 is not tested well, while the tests for X_2 is not good with $\omega = 0.7$. In the model, the effect of X_1 is accounted by SIR and that of X_2 is explained by SAVE. With $\omega = 0.5$, SAVE is more preferable than SIR, so the contribution of X_1 to the regression is relatively weakly measured. However, with $\omega = 0.7$, SIR has more weight than SAVE, so the effect of X_1 is tested better than that of X_2. The same story goes on the conditional tests, but the differences in the rejection percentages are not as dramatic as the marginal tests. With $n = 100$, either of $\omega = 0.5$ or 0.7 provides reliable test results in both tests.

Based on simulation studies with mild sample sizes, it can be concluded that the proposed permutation tests may not be a cause of concern in practice.

4.2. Data analysis - Swiss banknote data

For an illustration purpose, the Swiss banknote data analyzed in Shao *et al.* (SCW; 2007) was considered. In the data, the following seven variables were used. The response is a binary variable to indicate the status of a banknote as genuine or counterfeit. All the other six variables are predictors of length of bill (Length), width of left edge (Left), width of right edges (Right), top margin width (Top), bottom margin width (Bottom), and length of image diagonal (Diagonal), and the six predictors are measured in millimeters. Following SCW, the method of SAVE was adopted and it was considered that $\hat{d} = 2$.

In the analysis, we considered three different cases. In the first case, the method of SAVE was fitted with all 6 predictors, and the dimension tests were conducted. In addition, the five kinds of coordinate tests were considered: the SCW-marginal coordinate tests, the marginal bootstrap coordinate tests, the conditional bootstrap coordinate tests given $d = 2$, the marginal permutation coordinate tests, and the conditional coordinate permutation tests given $d = 2$. According to Yoo (2011), the values of 0.8 and 0.4 for the marginal and conditional bootstrap coordinate tests, respectively, are recommended for making decisions, and we follow a more general rule to select predictors distinguished with the others along with the former guidelines.

The second case was based on the same selection results (Left, Bottom and Diagonal) of predictors by the SCW-marginal tests and the bootstrap tests. So, instead of the original six predictors, the data was refitted through SAVE with the selected three predictors. Then the dimension tests and the five kinds of the coordinate tests under consideration were conducted again.

The third case was based on the selection results (Left, Top, Bottom and Diagonal) by the proposed permutation tests. The method of SAVE was fitted on a regression of the response given the selected four predictors. Moreover, accordingly, the dimension tests and the five coordinate tests were re-performed.

For each case, the p-values by the SCW-coordinate tests and the proposed permutation tests and the distances by the bootstrap tests are summarized in Table 2. In addition, Table 3 reports the p-values for the dimension tests computed in each case. We will use level 5% to make a decision, and a notation of " • " in Table 2 indicates the removals of the corresponding predictors in SAVE fits.

According to Table 3, for case 1, which all six predictors were used in, it can be concluded that $\hat{d} = 2$ with p-value $= 0.217$. Reading Table 2, since all values for the marginal bootstrap coordinate tests are less than 0.8, the criteria value for the tests decreases to 0.7 from 0.8. Then, for the SCW-marginal coordinate tests and the proposed marginal and conditional permutation coordinate tests, p-values for two predictors of Bottom and Diagonal are both 0.000. The distances for Bottom and Diagonal are 0.701 and 0.708 with the marginal bootstrap coordinate tests, and 0.805 and 0.774 for the conditional bootstrap coordinate tests in order. The two predictors of Bottom and Diagonal among the six predictors can be determined to be significant to the regression; in addition, the predictor of Left is determined to be significant by the four coordinate tests except the conditional permutation coordinate test (p-value $= 0.416$). Oddly, the conditional bootstrap coordinate tests decide that all the predictors are significant, because all distances are over the suggested value of 0.4. This implies that the bootstrap conditional coordinate tests clearly overestimate the importance of predictors. Based on the results, it can be reasonably concluded that both the SCW and bootstrap tests select three predictors of Left, Bottom and Diagonal, while the permutation tests determine that the four predictors of Left Top, Bottom and Diagonal are important.

We re-did the tests with the selected three predictors of Left, Bottom, and Diagonal, following the selection results by the SCW and bootstrap tests. According to Table 2, two predictors of Bottom and Diagonal are still determined to be significant in the SCW tests (Bottom, 0.000 and Diagonal, 0.001), the conditional bootstrap tests (0.654 and 0.549), and the two permutation tests (0.000 and 0.000). However, the marginal bootstrap tests determine that no predictors are important to the regression because no values are greater than 0.7. This is clear contradiction to the dimension estimation of $\hat{d} = 2$ in Table 3. The predictor of Left is significant in the SCW test (0.010) and the marginal permutation test (0.001). However, one unexpected issue occurs in the estimation of d, if comparing case 1. Table 3 shows that the hypothesis of $d = 2$ is rejected with p-value $= 0.013$, so $\hat{d}$ is suggested to be greater than 2. This is partially because the SCW and the bootstrap tests eliminate more variables than necessary. Assuming the estimation of $\hat{d} > 2$ to be correct, the selection of two predictors of Bottom and Diagonal by the bootstrap conditional tests result in a contradiction. Based on the discussion, the

Table 2: Coordinate tests for banknote data in section 4.2; Marg_{SCW}, marginal coordinate tests by Shao *et al.* (2007); $\text{Marg}_{\text{boot}}$, marginal bootstrap coordinate tests by Yoo (2011); $\text{Cond}_{\text{boot}}$, conditional bootstrap coordinate tests given $d = 2$ by Yoo (2011) $\text{Marg}_{\text{perm}}$, marginal permutation coordinate tests; $\text{Cond}_{\text{perm}}$, conditional permutation coordinate tests given $d = 2$

		Length	Left	Right	Top	Bottom	Diagonal
Case 1	Marg_{SCW}	0.361	0.002	0.396	0.240	0.000	0.000
	$\text{Marg}_{\text{boot}}$	0.642	0.744	0.679	0.512	0.701	0.708
	$\text{Cond}_{\text{boot}}$	0.643	0.632	0.588	0.587	0.805	0.774
	$\text{Marg}_{\text{perm}}$	0.305	0.016	0.124	0.004	0.000	0.000
	$\text{Cond}_{\text{perm}}$	0.084	0.416	0.334	0.000	0.000	0.000
Case 2	Marg_{SCW}	●	0.010	●	●	0.000	0.001
	$\text{Marg}_{\text{boot}}$	●	0.187	●	●	0.624	0.524
	$\text{Cond}_{\text{boot}}$	●	0.189	●	●	0.654	0.549
	$\text{Marg}_{\text{perm}}$	●	0.001	●	●	0.000	0.000
	$\text{Cond}_{\text{perm}}$	●	0.124	●	●	0.000	0.000
Case 3	Marg_{SCW}	●	0.026	●	0.160	0.000	0.006
	$\text{Marg}_{\text{boot}}$	●	0.728	●	0.243	0.727	0.610
	$\text{Cond}_{\text{boot}}$	●	0.487	●	0.418	0.718	0.547
	$\text{Marg}_{\text{perm}}$	●	0.001	●	0.000	0.000	0.000
	$\text{Cond}_{\text{perm}}$	●	0.036	●	0.000	0.000	0.000

Table 3: Structural dimension tests for banknote data in Section 4.2

	$d = 0$	$d = 1$	$d = 2$
Case 1	0.000	0.001	0.217
Case 2	0.000	0.000	0.013
Case 3	0.000	0.000	0.084

analysis of case 2 seems not reasonable to represent the regression.

Following the guidance by the permutation tests (which is case 3) the data was fitted with four predictors of Left, Top, Bottom, and Diagonal. Table 2 shows that the permutation tests determine that all four predictors are significant, while the SCW tests decide that the predictor of Top alone is not significant with p-value $= 0.160$. The marginal bootstrap tests decide that the two predictors of Left (0.728) and Bottom (0.727) are significant among the four, while the conditional bootstrap tests determine that all four predictors are significant. Therefore, the two permutation tests and the conditional bootstrap tests produce the same results. One important thing is that the structural dimension is still decided to be two with p-value $= 0.084$, which is consistent with case 1. Therefore, we can conclude that, in the banknote data, the structural dimension should be equal to two when involving four predictors of Left, Top, Bottom, and Diagonal to discriminate and classify fake banknotes.

5. Discussion

In this paper, we propose method-free permutation predictor hypothesis tests in the context of sufficient dimension reduction. Marginal and conditional permutation predictor hypothesis tests are suggested; subsequently, one can do the tests adequately in their own context. The proposed permutation tests provides p-values; therefore, usual statistical practitioner can make decisions to evaluate the predictor hypotheses just as they do in other tests.

The tests can also be directly applied to various sufficient dimension reduction methods. Thus it is expected that the tests can enhance real application strengths in practice.

We need to admit that the proposed permutation tests do not overwhelm other predictor hypothesis tests existing in sufficient dimension reduction. However, we believe that the proposed tests can provide reliable additional evidence for variable selection in sufficient dimension context. If one adopts the proposed tests along with the other tests, they will be in the better position to make variable selections correctly. The codes for the proposed permutation tests are available upon request.

Acknowledgments

The authors are grateful to three referees for many helpful comments.

References

Cook, R. D. (1998a). Principal Hessian directions revisited, *Journal of the American Statistical Association*, **93**, 84–100.

Cook, R. D. (1998b). *Regression Graphics*, Wiley, New York.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean, *Annals of Statistics*, **30**, 455–474.

Cook, R. D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction by K.C. Li, *Journal of the American Statistical Association*, **86**, 328–332.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**, 321–377.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 326–342.

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.

Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests of sliced average variance estimation, *Biometrika*, **94**, 285–296.

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 Cook, R. D. (2002). Dimension reduction for the conditional *k*th moment in regression. *Journal of Royal Statistical Society, Series B*, **64**, 159–175.

Yoo, J. K. (2011). Unified predictor hypothesis tests in sufficient dimension reduction: Bootstrap approach, *Journal of the Korean Statistical Society*, **40**, 217–222

Received April 5, 2013; Revised June 11, 2013; Accepted July 8, 2013