In Goal 2, we investigate the statistical properties of SOS. We conduct theoretical studies for SOS in discriminant analysis, dimension reduction, and regression problems. Our research yields theoretical guarantee for SOS and broadens its applicability. Moreover, because of the nonconvex nature of SOS, many existing techniques are inapplicable for its theoretical studies. We develop new generic techniques to show that SOS has a local minimizer with desirable statistical properties. In addition to obtaining statistical properties for SOS, our proof techniques will be applicable to other non-convex methods as well.

In Goals 3 & 4 we consider the analysis of tensor data. Tensor data are often extremely high-dimensional, and we need to take advantage of the tensor structure to perform efficient analysis. In Goal 3 we study the analysis for datasets with both vector and tensor predictors. We propose a novel IVTR model that simultaneously models the relationships among the vector, predictor and the response. Therefore, by fitting this model we achieve two goals. On one hand, we integrate the information from the two types of predictors for the response to obtain accurate prediction. On the other hand, we detect the dependence between the two types of predictors. Moreover, many existing tensor regression methods rely on the low-rank assumption, which may be inefficient when no low-rank approximation is possible. We provide an alternative that does not depend on the low-rank assumption.

In Goal 4 we develop a model-free screening method for tensor data. Our proposal preserves the low computation cost and the high flexibility in screening methods, but incorporates the tensor structure information to achieve more efficient and interpretable selection. Moreover, most, if not all, existing screening methods are marginal, but our proposal considers a voxel along with its neighbors at the same time. We show that this simple adaption can greatly improve marginal methods by exploiting the spatial information. We will also develop theoretical results to show that our proposal enjoys the SURE screening property that all the important voxels are preserved.

2 Research Goals

2.1 Goal 1: High-dimensional sufficient dimension reduction with simultaneous variable and rank selection

2.1.1 Background

Sufficient dimension reduction (SDR) methods find low-rank reduction of predictors that contain all the information for the response without specifying a model between the reduction and the response. An appropriate model can be chosen after the reduction. Hence, SDR methods are flexible alternatives to linear models when the mechanism behind the data is complicated. Consider $(Y, \mathbf{X})$, where $\mathbf{X}$ is the p-dimensional predictor, and $Y \in \mathbb{R}$ is the response. For a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}, 1 \le d \le p$, $\mathbf{X}^{\mathrm{T}}\mathbf{B}$ is a sufficient reduction of $\mathbf{X}$, if

$$Y \perp \mathbf{X} \mid \mathbf{X}^{\mathrm{T}}\mathbf{B}, \tag{1}$$

where (1) means that Y is independent of $\mathbf{X}$ given $\mathbf{X}^{\mathrm{T}}\mathbf{B}$. A sufficient reduction always exists, because $\mathbf{X}$ is a trivial p-dimensional reduction. But of course we want d to be small so that $\mathbf{X}^{\mathrm{T}}\mathbf{B}$ reduces the dimension of $\mathbf{X}$. We denote the smallest possible value of d as d^*, which is unknown in practice. Moreover, even when we know d^*, the matrix $\mathbf{B}$ is not unique. Rather, only the column space of $\mathbf{B}$ is identifiable. Hence, SDR methods look for the d^*-dimensional linear space $\mathcal{S}_{Y|\mathbf{X}}$, such that any of its basis satisfies (1). The space $\mathcal{S}_{Y|\mathbf{X}}$ is referred to as the central subspace [12, 13].

Under many popular models, the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ has natural interpretations. For example, consider the multi-index model $Y = f(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^*) + \epsilon$, where $f : \mathbb{R}^{d^*} \mapsto \mathbb{R}$ is an unknown function,

	$\widehat{\mathbf{V}}$	$\widehat{\mathbf{U}}$
SIR	$\frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{n}\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^{\mathrm{T}}$	$(\sqrt{\frac{n_1}{n}}\hat{\boldsymbol{\mu}}_1,\ldots,\sqrt{\frac{n_K}{n}}\hat{\boldsymbol{\mu}}_K)$
SAVE	$\widehat{\boldsymbol{\Sigma}}^{1/2}[\dfrac{1}{K}\sum_{k=1}^{K}\frac{n_k}{n}\{(\mathbf{I}-\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{\Sigma}}^{-1/2})^2\}]\widehat{\boldsymbol{\Sigma}}^{1/2}$	$(\sqrt{\frac{n_1}{n}}(\widehat{\boldsymbol{\Sigma}}-\widehat{\boldsymbol{\Sigma}}_1),\ldots,\sqrt{\frac{n_K}{n}}(\widehat{\boldsymbol{\Sigma}}-\widehat{\boldsymbol{\Sigma}}_K))$
PHD	$\{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\}\widehat{\boldsymbol{\Sigma}}^{-1}\{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\}$	$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$

Table 2: Choices of $\widehat{\mathbf{V}},\widehat{\mathbf{U}}$ in SIR, SAVE and PHD. We slice Y by choosing constants $a_0 < a_1 < \ldots < a_K$ and letting $H_i = k$ if $a_{k-1} < Y_i < a_k$. Let n_k be the number of observations in Slice k and define $\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{H_i=k}\mathbf{X}_i$, $\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k-1}\sum_{H_i=k}(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k)(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}$.

$\boldsymbol{\beta}^* \in \mathbb{R}^{p\times d^*}$, and ϵ is the statistical error. It is easy to see that the column space of the coefficient matrix $\boldsymbol{\beta}^*$ coincides with $\mathcal{S}_{Y|\mathbf{X}}$. Hence, SDR methods estimate $\boldsymbol{\beta}^*$ without the knowledge of f. However, the concept of central subspace does not involve model assumptions such as the multi-index model. Instead, with SDR, we can explore the possibility of various models in much lower dimensions. Therefore, SDR methods are very flexible and have wide applications. Many SDR methods have been proposed [37, 17, 38, 3, 10, 72, 16, 34, 84, 49]. The two earliest and possibly most popular methods are the sliced inverse regression (SIR) and sliced average variance estimation (SAVE).

SIR and SAVE can be viewed as generalized eigenvalue problems. Let $\widehat{\boldsymbol{\Sigma}}$ be the sample covariance of $\mathbf{X}$. For properly chosen $\widehat{\mathbf{V}} \in \mathbb{R}^{p\times p}$, SIR and SAVE look for $\hat{\boldsymbol{\beta}}_i$ such that

$$\widehat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}_i = \delta_i\widehat{\mathbf{V}}\hat{\boldsymbol{\beta}}_i \tag{2}$$

for some $\delta_i \in \mathbb{R}$. The space spanned by the $\hat{\boldsymbol{\beta}}_i$ corresponding to the top d^* eigenvalues δ_i is an estimate of the space $\mathcal{S}_{Y|\mathbf{X}}$. The choices of $\widehat{\mathbf{V}}$ for SIR and SAVE are listed in Table 2.

Let n be the sample size. In high-dimensional datasets where $p \gg n$, one naturally wishes to perform SDR methods. The model-free nature of SDR methods is especially desirable when parametric models are inadequate. However, many SDR methods are inapplicable when $p > n$ [41, 40, 42, 9, e.g]. Existing high-dimensional SDR methods are typically restricted to SIR [81, 65, 32, 46, 64, 73, 45, e.g]. These methods perform variable selection, assuming that d^* is known. However, the determination of d^* in high dimensions is an open problem. Therefore, in this research goal we propose a unified framework to generalize SDR methods. Our framework extends beyond SIR, and performs variable and rank selection simultaneously.

2.1.2 Our proposal

Our proposal is based on a new constrained quadratic optimization formula for SDR methods. For ease of presentation, we only discuss this formula for SIR and SAVE. For $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, let $\mathrm{tr}(\boldsymbol{\Omega}) = \sum_{j=1}^{p}\omega_{jj}$. In low dimensions, we have the following lemma.

Lemma 1 *Assume that $p < n$. If $\widehat{\mathbf{V}}$ is chosen as those for SIR and SAVE, the top d^* generalized eigenvectors in* (2) *span the same subspace as the columns of $\tilde{\mathbf{B}}$, where*

$$\tilde{\mathbf{B}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times r}}\left\{\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}) - 2\mathrm{tr}(\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{B})\right\} \textit{ subject to } \mathrm{rank}(\mathbf{B}) = d^*, \tag{3}$$

with $\widehat{\mathbf{U}}$ defined as in Table 2 and r being the number of columns of $\hat{\mathbf{U}}$.

To the best of our knowledge, we are the first ones to derive the results in Lemma 1. We construct high-dimensional SDR methods based on (3). The objective function in (3) has the familiar quadratic form, while the constraint is much more difficult to tackle. On one hand, it involves the unknown rank d^*. In low dimensions, there are methods to determine d^* [23, 16], but they do not apply to high-dimensional data. On the other hand, even when we know d^*, the constraint is nonconvex.

Therefore, when $p > n$, we replace the rank constraint with the nuclear norm penalty [22, 57, 26]. For a matrix $\mathbf{\Omega} \in \mathbb{R}^{p\times r}$ with singular values $\eta_1, \ldots, \eta_r$, its nuclear norm is $\|\mathbf{\Omega}\|_* = \sum_{i=1}^r |\eta_i|$. By adding a nuclear norm penalty on $\mathbf{B}$, we obtain a convex relaxation of (3) that produces an estimate for d^*. We further impose the sparsity assumption that only a few variables are involved in the sufficient reduction. To perform variable selection, we note that the sparsity in SDR naturally has a group structure. By the definition of $\mathcal{S}_{Y|\mathbf{X}}$, a variable X_j is not important if and only if for any $\mathbf{B} \in \mathbb{R}^{p\times r}$ that spans $\mathcal{S}_{Y|\mathbf{X}}$, we have $b_{j1} = \ldots = b_{jr} = 0$. Hence, we impose the sparsity through group lasso [74].

Combining the two penalties for variable and rank selection, we propose a general framework for high-dimensional dimension reduction (HDR) as follows:

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times r}} \left\{ \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{B}) - 2\mathrm{tr}(\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{B}) + \lambda_1 \sum_{i=1}^p \sqrt{\sum_{j=1}^r b_{ij}^2} + \lambda_2\|\mathbf{B}\|_* \right\}, \tag{4}$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters chosen by cross validation. The column space of $\widehat{\mathbf{B}}$ is an estimate of $\mathcal{S}_{Y|\mathbf{X}}$, while the rank of $\widehat{\mathbf{B}}$ is an estimate for d^*.

We use the alternating direction method of multipliers (ADMM) [2] to solve (4). Consider the augmented problem:

$$\arg\min_{\mathbf{B},\mathbf{C}\in\mathbb{R}^{p\times r}} \left[\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{B} - 2\mathbf{B}^{\mathrm{T}}\widehat{\mathbf{U}}) + \lambda_1 \sum_{i=1}^p \sqrt{\sum_{j=1}^r b_{ij}^2} + \lambda_2\|\mathbf{C}\|_* + \gamma\|\mathbf{B}-\mathbf{C}\|_F^2 \right] \text{ s.t. } \mathbf{B} = \mathbf{C}, \tag{5}$$

where $\gamma > 0$ is a small constant. The Lagrange for (5) is

$$L_\gamma(\mathbf{B},\mathbf{C},\boldsymbol{\xi}) = \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{B} - 2\mathbf{B}^{\mathrm{T}}\widehat{\mathbf{U}}) + \lambda_1 \sum_{i=1}^p \sqrt{\sum_{j=1}^r b_{ij}^2} + \lambda_2\|\mathbf{C}\|_* + \mathrm{tr}\{\boldsymbol{\xi}^{\mathrm{T}}(\mathbf{B}-\mathbf{C})\} + \frac{\gamma}{2}\|\mathbf{B}-\mathbf{C}\|_F^2, \tag{6}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{p\times r}$ is a matrix to be optimized. The minimizer of $L_\gamma(\mathbf{B},\mathbf{C},\boldsymbol{\xi})$ of $\mathbf{B}$ is the same as the solution to (4). To find $\widehat{\mathbf{B}}$, we iteratively minimize $L_\gamma(\mathbf{B},\mathbf{C},\boldsymbol{\xi})$ over one parameter while fixing the others. With some simplification, this procedure reduces to iteratively solving the following problems over $t = 1, 2, \ldots$ until convergence:

$$\begin{aligned} \mathbf{B}^{t+1} &= \arg\min_{\mathbf{B}} \left[\mathrm{tr}\{\mathbf{B}^{\mathrm{T}}(\widehat{\mathbf{\Sigma}} + \frac{\gamma}{2}\mathbf{I})\mathbf{B} - 2\mathbf{B}^{\mathrm{T}}(\widehat{\mathbf{U}} + \gamma\mathbf{C}^t - \frac{1}{2}\boldsymbol{\xi}^t)\} + \lambda_1 \sum_{i=1}^p \sqrt{\sum_{j=1}^r b_{ij}^2} \right], & (7) \\ \mathbf{C}^{t+1} &= \arg\min_{\mathbf{C}} \{\lambda_2\|\mathbf{C}\|_* + \frac{\gamma}{2}\|\mathbf{C} - (\mathbf{B}^{t+1} + \gamma^{-1}\boldsymbol{\xi}^t)\|_F^2\}, & (8) \\ \boldsymbol{\xi}^{t+1} &= \boldsymbol{\xi}^t + \gamma(\mathbf{B}^{t+1} - \mathbf{C}^{t+1}). & (9) \end{aligned}$$

The problem in (7) can be solved with a groupwise coordinate descent algorithm [50], while $\mathbf{C}^{t+1}$ can be found by soft-thresholding the singular values of $\mathbf{B}^{t+1} + \gamma^{-1}\boldsymbol{\xi}^t$ [26].

As for its statistical properties, we conjecture that HDR is consistent even when p grows at an exponential rate of n in the sense that span($\widehat{\mathbf{B}}$) converges to $\mathcal{S}_{Y|X}$. We further conjecture that HDR achieves variable selection and rank selection consistency under proper conditions. These conjectures will be proved in the funded period.