Statistica Sinica Preprint No: SS-2023-0143

Title	Sufficient Dimension Reduction for Classification
Manuscript ID	SS-2023-0143
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0143
Complete List of Authors	Xin Chen, Jingjing Wu, Zhigang Yao and Jia Zhang
Corresponding Authors	Jia Zhang
E-mails	zhangjia@swufe.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Sufficient Dimension Reduction for Classification

Xin Chen, Jingjing Wu, Zhigang Yao and Jia Zhang

Southern University of Science and Technology, University of Calgary,

National University of Singapore and Southwestern University of Finance and Economics

Abstract: We propose a new sufficient dimension reduction approach designed deliberately for high-dimensional classification problems. This novel method is named as Maximal Mean Variance (MMV), inspired by the mean variance index first proposed by Cui, Li and Zhong (2015). MMV requires reasonably mild restrictions on the predictors, and keeps the model-free advantage without the need to estimate the link function. The consistency of the MMV estimator is established under regularity conditions with possibly diverging number of predictors and categories of the response. We also construct the asymptotic normality for the estimator when the dimension of the predictors keeps fixed. The relationship between MMV and several classical classification algorithms are further elaborated. Moreover, although without any definite theoretical guarantee, our method works pretty well when the sample size is far less than the problem dimension. The surprising classification efficiency gain of MMV is demonstrated by simulation studies and real data analysis.

Key words and phrases: Classification, consistency, mean variance index, sufficient dimension reduction.

1. Introduction

Sufficient dimension reduction fits into what is currently quite a hot area in research of high dimensional data. Large quantities of related articles and studies have appeared in recent decades. However, most of the literature focuses on the regression problem where the response Y is a continuous variable, while little is designed specially for the problem of classification with a categorical response.

The slice-based methods, including but not limited to the seminal sliced inverse regression (SIR, Li (1991)), sliced average variance estimation (Cook and Weisberg, 1991), directional regression (Li and Wang, 2007) and sliced regression (Wang and Xia, 2008), can be naturally applied to the classification problem with the slices determined directly by the categories of the response. It seems to work nicely, but the number of the slices is strictly restricted by the number of the categories, which can be problematic when there are only a few categories. More specifically, faced with a common binary classification problem, the number of the slices is imposed as 2, and then the number of the effective dimension-reduction directions is correspondingly forced to 1, which would directly reduce the accuracy of classification. Moreover, almost all of the above methods require the linearity condition or/and constant covariance condition, which are difficult to verify in practice, and the results may be misleading if these conditions are violated.

Other popular sufficient dimension reduction methods, like minimum average variance estimation (Xia et al., 2002), inverse regression (Cook and Ni, 2005) and distance covariance based sufficient dimension reduction (Sheng and Yin, 2013, 2016), either require the response to be continuous or just treat the response as a numeric variable, which limits the applicability of these methods to classification problems.

To overcome the aforementioned limitations of the classical sufficient dimension reduction methods, we propose a novel sufficient dimension reduction approach – Maximum Mean Variance (MMV), designed deliberately for high-dimensional classification problems. This method utilizes the MV index first proposed by Cui, Li and Zhong (2015) to construct a sequential optimization problem to seek sufficient dimension reduction. Since it is not slice-based, MMV circumvents the restriction on slice number. The method also dispenses with the need for the linearity condition or constant variance condition, and it does not make any assumptions regarding specific distributions for the predictors $\mathbf{X}$, $\mathbf{X}|Y$ or $Y|\mathbf{X}$, which is essential in the methods of Zhu and Zeng (2006), Cook and Forzani (2009), Cook and Li (2009), Bura and Forzani (2015), Bura, Duarte and Forzani (2016) and Zhang, Chen and Zhou (2022). In addition, our method inherits the model-free advantage without estimating the link function. These benefits together broaden the scope of applications for our approach.

The consistency of the MMV estimator is established for both fixed and diverging problem dimensions, and the asymptotic normality is constructed for the case of fixed dimension. The relationship between MMV and classification is more than the usual stepwise heuristics of dimension reduction first and classification next, which is elaborated upon by taking the examples of linear discriminant analysis (LDA) and index models. Moreover, we note that the asymptotic theory of MMV estimator is quite challenging to set up. Since the empirical MV index includes the kernel estimation of conditional and unconditional distribution functions, it cannot be directly expressed by the sum of independent and identically distributed random variables.

The rest of the paper is organized as follows. Section 2 revisits some basic definitions in the literature of sufficient dimension reduction and the definition and properties of the MV index. In Section 3, we propose the MMV approach to high-dimensional classification. In Section 4, we elaborate on the delicate relationship between MMV and several popular classification methods. Consistency and asymptotic normality of the MMV estimator are studied in Section 5. Several simulation studies together with numerical comparisons and a real data example are conducted to illustrate the efficiency and priority of the proposed method in Section 6. Section 7 concludes the article, and the technical proofs are deferred to the supplementary material.

2. Preliminaries

2.1 Sufficient dimension reduction revisited

Let $Y \in \mathbb{R}$ be the response and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be continuous predictors. For a subspace $\mathcal{S} \subset \mathbb{R}^p$, consider a projection $P_{\mathcal{S}} : \mathbb{R}^p \to \mathbb{R}^q$ with $q \leq p$. Denote by $P_{\mathcal{S}}\mathbf{X}$ the projection of $\mathbf{X}$ onto $\mathcal{S}$. Then the subspace $\mathcal{S}$ is called the *dimension reduction subspace* (Li, 1991, 1992) if

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}}\mathbf{X}\,, \tag{2.1}$$

where $\perp\!\!\!\perp$ is the independence notation. If $P_{\mathcal{S}}\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{X}$ for some $p \times q$ matrix $\mathbf{B}$, then (2.1) can be rewritten as

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^{\mathrm{T}}\mathbf{X}\,,$$

which indicates that the regression information of Y on $\mathbf{X}$ are completely contained by $\mathbf{B}^{\mathrm{T}}\mathbf{X}$, a q vector. If $q < p$, then we can regress Y on the q predictor $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ instead of the original p predictor $\mathbf{X}$ without losing any information, and that is why we call the process of seeking $\mathbf{B}$ or $\mathcal{S}$ sufficient dimension reduction.

The intersection of all dimension reduction subspaces is called the *central subspace* (Cook, 1994, 1996), denoted by $\mathcal{S}_{Y|\mathbf{X}}$, which has the minimal dimensions among all dimension reduction subspaces. In this paper, we assume the central subspace exists uniquely, which is quite mild and practical (Cook, 1996; Li, 2018).

2.2 Mean variance index

Let Y be a categorical response with R classes $\{y_1, y_2, \ldots, y_R\}$, and Z be a continuous covariate. The MV index (Cui, Li and Zhong, 2015) is defined as follows

$$\mathrm{MV}(Z|Y) = \mathrm{E}_Z[\mathrm{Var}_Y\{F(Z|Y)\}] = \sum_{r=1}^{R} p_r \int [F_r(z) - F(z)]^2 \mathrm{d}F(z)\,, \quad (2.2)$$

where $F(z|Y) = \mathbb{P}(Z \leq z|Y)$, $F(z) = \mathbb{P}(Z \leq z)$, $F_r(z) = \mathbb{P}(Z \leq z|Y = y_r)$ and $p_r = \mathbb{P}(Y = y_r)$ for $r = 1, \ldots, R$. It has been verified that $\mathrm{MV}(Z|Y) = 0$ if and only if Y and Z are independent, and thus the MV index characterizes both linear and nonlinear correlations between a categorical random variable Y and a continuous random variable Z.

Let $\{(Y_i, Z_i) : 1 \leq i \leq n\}$ be an i.i.d random sample of size n. Let $\hat{F}(Z)$ and $\hat{F}_r(Z)$ be some sample estimators of $F(Z)$ and $F_r(Z)$. Then, the MV index can be estimated by

$$\widehat{\mathrm{MV}}(Z|Y) =: \mathrm{MV}_n(Z|Y) =: \frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\hat{p}_r[\hat{F}(Z_i) - \hat{F}_r(Z_i)]^2\,, \tag{2.3}$$

where $\hat{p}_r = n^{-1}\sum_{i=1}^{n}\mathrm{I}\{Y_i = y_r\}$ with $\mathrm{I}(\cdot)$ representing the indicator function. Cui, Li and Zhong (2015) used the empirical distributions of Z and $Z|Y$ as their sample estimators in a screening procedure.

3. Maximum mean variance

Based on the MV index, we now introduce the MMV approach to sufficient dimension reduction for high-dimensional classification problems. The idea is to make use of the MV index to find a few linear combinations (or indexes) of the possibly high-dimensional original predictors $\mathbf{X} \in \mathbb{R}^p$ that contribute to classification without any loss of information. These derived low-dimensional indexes can then be utilized for classification.

Recall that $\mathrm{MV}(Z|Y) = 0$ if and only if Z and Y are statistically independent. Thanks to this property, the MV index is used for marginal feature screening in discriminant analysis (Cui, Li and Zhong, 2015). Our novel idea is to abandon this, and on the contrary, we seek a $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ achieves its maximum under some constraints. This is why we named this method maximum mean variance. A sequential algorithm is elaborated as follows: we find the first linear combination of the predictors from

$$\boldsymbol{\beta}_{01} = \arg\max_{\boldsymbol{\beta}_1} \mathrm{MV}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) \;\; \text{s.t} \;\; \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 = 1\,.$$

Then the kth linear combination can be calculated from

$$\boldsymbol{\beta}_{0k} = \arg\max_{\boldsymbol{\beta}_k} \mathrm{MV}(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}|Y) \tag{3.4}$$

$$\text{s.t} \quad \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k = 1, \quad \text{and} \quad [\boldsymbol{\beta}_{01}, \cdots, \boldsymbol{\beta}_{0(k-1)}]^{\mathrm{T}}\boldsymbol{\beta}_k = \mathbf{0}$$

for $k \geq 2$. We continue this process till the MV index reaches 0. This procedure is indeed conducting sufficient dimension reduction for the predictors $\mathbf{X}$ with respect to the response Y, which can be seen clearly from the following theorem. Hereafter, we assume that the linearity condition and the coverage condition hold.

Theorem 1. *Suppose there exists a positive integer $d < p$ such that $\mathrm{MV}(\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}|Y) \geq \mathrm{MV}(\boldsymbol{\beta}_{02}^{\mathrm{T}}\mathbf{X}|Y) \geq \cdots \geq \mathrm{MV}(\boldsymbol{\beta}_{0d}^{\mathrm{T}}\mathbf{X}|Y) > 0 = \mathrm{MV}(\boldsymbol{\beta}_{0(d+1)}^{\mathrm{T}}\mathbf{X}|Y) = \cdots = \mathrm{MV}(\boldsymbol{\beta}_{0p}^{\mathrm{T}}\mathbf{X}|Y)$ where $\boldsymbol{\beta}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_{0i} = 1$ and $\boldsymbol{\beta}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_{0j} = 0$ for $i, j = 1, \ldots, p$ and $i \neq j$. It holds that*

$$\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathrm{span}\{\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0d}\}\,,$$

and for any integer $0 < k < d$, if $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$,

$$\mathrm{span}\{\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0k}\} \not\supseteq S_{Y|\mathbf{X}}\,.$$

Notice that the subscript 0 in $\boldsymbol{\beta}_{0i}$s in Theorem 1 is used to indicate that these parameters are specified in the population level. The existence of the integer $d < p$ is validated in the classical Linear Discriminant Analysis (LDA) and in the index model setting. See more details in Section 4.

Remark 1. Theorem 1 shows that the MMV procedure does not actually find the *central subspace*. Instead, solving the MMV sequentially, some tight upper bound for the *central space* can be obtained. In some specific models, MMV can exactly find the *central subspace*. See Section 4.2 for details.

In practice, the population MV index remains unknown for any given $\boldsymbol{\beta}$, and we use its sample counterpart $\widehat{\mathrm{MV}}$ specified in (2.3) to conduct the sequential optimization procedure. Let $\{(Y_i, \mathbf{X}_i), 1 \leq i \leq n\}$ be an i.i.d sample with $Y_i \in \{y_1, \ldots, y_R\}$ and $\mathbf{X}_i \in \mathbb{R}^p$, and $Z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and $Z_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i$ for a fixed $\boldsymbol{\beta}$. Then, it is natural to estimate $F(z)$ in (2.2) by its empirical counterpart: $\hat{F}(z) = n^{-1}\sum_{i=1}^{n} \mathrm{I}(Z_i \leq z)$, as done in Cui, Li and Zhong (2015). However, the empirical distribution is a step function, which makes the optimization problematic. Hence, here we adopt a kernel estimator

$$\hat{F}(z) := \hat{F}_h(z) = \int_{-\infty}^{z} \hat{f}_h(u)\mathrm{d}u = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{z} K_h(Z_i - u)\mathrm{d}u\,, \tag{3.5}$$

where $\hat{f}_h$ is a kernel density estimator of the density f of F and $K_h(\cdot) =$

$h^{-1}K(\cdot/h)$ with $K(\cdot)$ being a kernel function and $h = h_n \to 0$ the bandwidth. Similarly, $F_r(z)$ in (2.2) can be estimated by

$$\hat{F}_r(z) := \hat{F}_{hr}(z) = \frac{1}{n_r}\sum_{j=1}^{n_r}\int_{-\infty}^{z} K_{h_r}(Z_j - u)\mathrm{d}u \tag{3.6}$$

for $r = 1, 2, \ldots, R$, where n_r is the sample size of the rth category and Z_j, $j = 1, \ldots, n_r$, are the corresponding sample points in this category. Given (3.5) and (3.6), we can use the estimator $\widehat{\mathrm{MV}}$ specified in (2.3) to implement the optimization. The MMV procedure is summarized in Algorithm 1. Suppose $\epsilon > 0$ is a predefined small number, say $\epsilon = 10^{-3}$.

Since the optimization problems in Step 1 and Step 3 of Algorithm 1 are classic constrained nonlinear problems, our suggestion is to use the ready-made function **fmincon** in matlab to solve the two optimizations. In fmincon, **sqp** (sequential quadratic programming) is selected to adapt to possibly high-dimensional problems. Moreover, due to the nonconvexity of the optimizations, the choice of the starting point of **sqp** is of importance. For the optimization in Step 1, since MMV equals LDA as shown in Corollary 2, we suggest using the LDA solution as the starting point. In Step 3, we choose the initial point for $\boldsymbol{\beta}_k$ as the eigenvector of $\mathrm{cov}(\mathbf{X})$ associated with its kth largest eigenvalue. Such choices works fine throughout our numerical experiments.

Algorithm 1: The MMV procedure for sufficient dimension reduction

```
Input: $\{Y_i, \mathbf{X}_i\}_{i=1}^n$, $Y_i \in \{y_1, \ldots, y_R\}$ and $\mathbf{X}_i \in \mathbb{R}^p$
Compute $\widehat{\boldsymbol{\beta}}_1$ as
    $$\widehat{\boldsymbol{\beta}}_1 = \arg\max_{\boldsymbol{\beta}_1} \widehat{\mathrm{MV}}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) \ \ \text{s.t} \ \ \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 = 1\,,$$
    where $\widehat{\mathrm{MV}}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) = n^{-1}\sum_{r=1}^R \sum_{i=1}^n \hat{p}_r[\hat{F}(Z_i) - \hat{F}_r(Z_i)]^2$ with
    $\hat{p}_r = n^{-1}\sum_{i=1}^n \mathrm{I}\{Y_i = y_r\}$, $Z_i = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i$, $\hat{F}(Z_i)$ and $\hat{F}_r(Z_i)$ specified in
    (3.5) and (3.6).
for $k = 2, \cdots, p$ do
    Compute $\widehat{\boldsymbol{\beta}}_k$ as
        $$\widehat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}_k} \widehat{\mathrm{MV}}(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}|Y)$$
        $$\text{s.t} \quad \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k = 1, \quad \text{and} \quad [\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_{k-1}]^{\mathrm{T}}\boldsymbol{\beta}_k = \mathbf{0}\,,$$
        where $\widehat{\mathrm{MV}}(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}|Y)$ is similarly defined as above.
    if $\widehat{\mathrm{MV}}(\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\mathbf{X}|Y) \leq \epsilon$, let $d = k - 1$; break.
end
Output: $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_d$
```

4. MMV in classification

MMV has intimate relationship with several classical classifiers, such as Fisher's LDA, logistic regression and more complex methods.

4.1 Fisher's LDA

Consider a binary classification problem. Suppose we have n labeled i.i.d training samples $(Y_i, \mathbf{X}_i), 1 \leq i \leq n$, where $\mathbf{X}_i$ is a p-dimensional feature vector and $Y_i \in \{-1, 1\}$ is the corresponding class label. Let $p_1 = \mathbb{P}(Y_i = 1)$, $p_{-1} = \mathbb{P}(Y_i = -1)$, and assume

$$\mathbf{X}_i \sim N(Y_i \cdot \boldsymbol{\mu},\ \boldsymbol{\Sigma})\,, \tag{4.1}$$

where $\boldsymbol{\mu}$ is the contrast mean vector between the two classes, and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix. Given a new independent feature vector from the same population, i.e. $\mathbf{X} \sim N(Y \cdot \boldsymbol{\mu},\ \boldsymbol{\Sigma})$, our goal is to train $(Y_i, \mathbf{X}_i)$ to decide whether $Y = -1$ or $Y = 1$. Notice that although we use the contrast mean in model (4.1), the method and result below also apply to a more general model with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$ with no extra difficulty.

Linear discriminant analysis, namely Fisher's LDA, is a well-known method for classification, which is essentially based on a weighted average of the test features $L(\mathbf{X}) = \sum_{j=1}^{p} w_j X_j$ and predicts $Y = \pm 1$ if $L(\mathbf{X}) >< 0$.

Here, $\mathbf{w} = (w_1, \ldots, w_p)^{\mathrm{T}}$ is a preselected weight vector. Fisher showed that the optimal weight vector satisfies

$$\mathbf{w} \propto \boldsymbol{\Omega}\boldsymbol{\mu}\,,$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. In the classical setting where $n \gg p$, $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ can be conveniently estimated and Fisher's LDA is approachable. Unfortunately, in the modern regime where $p \gg n$, Fisher's LDA faces immediate challenges, *i.e.* , the covariance matrix $\boldsymbol{\Sigma}$ is irreversible.

MMV can surprisingly estimate the optimal weight vector $\mathbf{w}$ in LDA without inversing the covariance matrix, and thus bypass the difficulty of estimating $\boldsymbol{\Omega}$ when $p \geq n$, which can be seen clearly form the following corollary.

Corollary 2. *Under model* (4.1)*, $d = 1$ and $\boldsymbol{\beta}_{01} \propto \boldsymbol{\Omega}\boldsymbol{\mu}$, where d is defined in Theorem* 1*.*

Corollary 2 states that for model (4.1), $d = 1$ and $\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}$ contains all the information for classification. This means that the MMV procedure gives exactly the LDA classifier at the population level. The corollary also justifies the efficiency of the LDA for the normal model (4.1) in terms of maximum mean variance. When $p > n$, the LDA needs to estimate the inverse of the covariance matrix, and thus it is unsolvable or requires extra sparsity assumption. MMV is an efficient alternative to circumvent this problem.

In practice, if we have an estimator of $\mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$, say $\widehat{\mathrm{MV}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ as given in (2.3), then we can solve its maximizer with respect to $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}_1$, by routine optimization algorithms. Since $\widehat{\boldsymbol{\beta}}_1$ is an estimator of $\boldsymbol{\beta}_{01}$, by Corollary 2 it is also an estimator of the optimal weight vector $\mathbf{w} \propto \boldsymbol{\Omega}\boldsymbol{\mu}$ in LDA. Therefore, for a new given feature vector $\mathbf{X}$, we classify it as $Y = 1$ if $\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X} > 0$ and $Y = -1$ if $\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X} < 0$. Clearly, MMV provides a new idea for solving the optimal weight vector in LDA.

4.2 Index model

The index model enjoys a lot of popularity in regression and classification. A general index model can be expressed as the following semi-parametric form. Let $Y \in \{y_1, \ldots, y_R\}$ denote the response variable and $\mathbf{X} \in \mathbb{R}^p$ denote the covariates. In a index model, there exist orthogonal p-dimensional vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ with unit norm such that

$$Y = f(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}, \varepsilon) \quad (k < p)\,, \tag{4.2}$$

where f is an arbitrary unknown link function and ε is independent of $\mathbf{X}$. With a slight abuse of notation, the notation k in (4.2) can be seen as a fixed integer indicating the number of the indexes. The column space spanned by $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}$ is defined as the efficient dimension reduction subspace by Li (1991). Under the setting of index model (4.2), we can detail Theorem 1 to some extent. Let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)^{\mathrm{T}}$. Assume

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}\mathbf{X} \,, \tag{4.3}$$

and there exists a p-dimensional vector $\boldsymbol{\gamma}$ such that

$$\mathbf{B}\boldsymbol{\gamma} = \mathbf{0} \,, \quad \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp \mathbf{B}\mathbf{X} \,. \tag{4.4}$$

Then by Lemma 4.3 in Dawid (1979) and Proposition 4.6 in Cook (1998), it holds that $\mathrm{MV}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}|Y) = 0$. This implies that under mild conditions, the MMV method can recover all the information in $\mathbf{X}$ related to classification with $d < p$ indexes specified in Theorem 1. Specifically, when $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can further obtain the following corollary.

Corollary 3. *In model* (4.2)*, assume* $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}\mathbf{X}$ *where* $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)^{\mathrm{T}}$ *with* $k < p$*. For* d *specified in Theorem* 1*, if* $2k \leq p$*, then* $d \leq 2k$*. Specifically, if* $\boldsymbol{\Sigma} = \mathbf{I}$*, then* $d = k$*.*

Corollary 3 indicates that when $\boldsymbol{\Sigma} = \mathbf{I}$, the MMV procedure can exactly recover the efficient dimension reduction subspace with $d = k$ steps under the setting of index model and normal covariates. To be specific, in the

In this special case, if the efficient dimensional reduction subspace equals the *central subspace*, then MMV exactly finds the *central subspace*.

logistic (probit) model with normally distributed covariates, $\boldsymbol{\beta}_{01} \propto \boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1$ denotes the coefficient vector of the logistic (probit) model and $\boldsymbol{\beta}_{01}$ is specified in Theorem 1. This implies that $d = 1$ is enough for the logistic (probit) model. The advantage of our method is that it is a semiparametric method, which does not depend on any specific form of the link functions.

4.3 Other classifiers

Other popular classification methods such as K-Nearest Neighbours (KN-N), neural networks and Support Vector Machine (SVM), can be connected to MMV by a two step procedure, *i.e.* , dimension reduction first and classification next. Such a two-step procedure will improve the accuracy of the classification, because high dimensionality causes problems in the classification algorithms mentioned above. Simulations in Section 6 demonstrate the benefit of such a two-step procedure.

5. Consistency and asymptotic normality

In this section, we establish the consistency and asymptotic normality of the proposed MMV estimator. To simplify the derivation of the proof and the assumptions needed, we consider the case where $d = 1$ with d specified in Theorem 1. For $d > 1$, similar results hold with similar but tediously long conditions.

We first introduce some notations. Recall $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ denotes i.i.d samples, $Y_i \in \{y_1, \ldots, y_R\}$, and n_r denotes the number of samples in the class $Y = y_r$ for $r = 1, \ldots, R$. Let Ω_1 denote the parameter space of $\boldsymbol{\beta}_1$ and $B(\kappa_1) = \{\boldsymbol{\beta}_1 \in \mathbb{R}^p : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \kappa_1\}$ be a ball with center $\boldsymbol{\beta}_{01}$ and radius κ_1, where $\|\cdot\|$ denotes the Euclidean norm of a vector. The boundary of the ball is denoted by $\partial B(\kappa_1)$. Let $\Gamma_1 = \{\boldsymbol{\beta}_1 \in \mathbb{R}^p : \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 = 1\}$. For any $\boldsymbol{\beta}$, we simplify $\mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ and $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ as $\mathrm{MV}(\boldsymbol{\beta})$ and $\mathrm{MV}_n(\boldsymbol{\beta})$, respectively.

The following conditions are required to establish the asymptotic properties of the MMV estimator.

Condition 1. There exist $c_1, c_2 > 0$ such that $c_1/R \leq \min_{1\leq r\leq R} p_r \leq \max_{1\leq r\leq R} p_r \leq c_2/R$, and $R = O(n^\delta)$ for some $\delta \in (0, 1/2]$.

Condition 2. There exists an open subset ω_1 of $\Omega_1 \cap \Gamma_1$ that contains the true parameter $\boldsymbol{\beta}_{01}$ and $\sup_{\boldsymbol{\beta}_1 \in B(\kappa_{01})} \mathrm{MV}(\boldsymbol{\beta}_1) < \infty$ for some constant $\kappa_{01} > 0$. For any $\kappa_1 \in (0, \kappa_{01}]$, $\sup_{\boldsymbol{\beta}_1 \in \partial B(\kappa_1)\cap\Gamma_1} \mathrm{MV}(\boldsymbol{\beta}_1) < \mathrm{MV}(\boldsymbol{\beta}_{01})$.

Condition 3. $\int uK(u)\mathrm{d}u = 0$, $\int u^2K(u)\mathrm{d}u < \infty$, $nh^4 \to 0$ with $h = h_1 = \cdots = h_R$.

Condition 4. For any $\boldsymbol{\beta} \in B(\kappa_{01})$, the cumulative distribution functions of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y = y_r$ $(r = 1, \ldots, R)$ have bounded second derivatives.

Condition 5. For any $\boldsymbol{\beta}_1 \in B(\kappa_{01})$, $\mathrm{MV}^{'}(\boldsymbol{\beta}_1)$ and $\mathrm{MV}^{'}_n(\boldsymbol{\beta}_1)$ exist, and $\sup_{\boldsymbol{\beta}_1 \in B(\kappa_{01})}(\|\mathrm{MV}^{'}(\boldsymbol{\beta}_1)\| + \|\mathrm{MV}^{'}_n(\boldsymbol{\beta}_1)\|) = o_{\mathrm{p}}(\sqrt{p})$.

Condition 1 requires the proportion of each class of the response to be moderate, not too small nor too large. The assumption $R = O(n^{\delta})$ allows the number of the classes of the response to grow with the sample size, which matches the demands of the big data era. A similar condition was also imposed by Cui, Li and Zhong (2015) in a feature screening procedure. Condition 2 is assumed to ensure the existence of a local MMV optimizer. A similar condition was assumed in Chen, Ma and Zhou (2017) for a likelihood function. Condition 3 is widely used in the literature of kernel density estimation. Together with Condition 4, it ensures the uniform convergence of the kernel estimators of the cumulative distribution functions; see Cheng (2017) for reference. For simplicity and without loss of generality, we assume $h = h_1, \ldots, = h_n$ in Condition 3. Condition 5 is a high-level assumption on the population and sample objective functions. Recall that the MV index is defined on the cumulative distribution functions. Hence, for the population objective function, Condition 5 requires that the cumulative distribution functions of $\boldsymbol{\beta}^{\top}\mathbf{X}$ and $\boldsymbol{\beta}^{\top}\mathbf{X}|Y = y_r$ have bounded first derivatives with respect to $\boldsymbol{\beta}$, and the density function of $\boldsymbol{\beta}^{\top}\mathbf{X}$ has bounded first derivative with respect to $\boldsymbol{\beta}$. When p diverges with n, it also requires these derivatives grow not too fast. For the sample objective function, Condition 5 requires $\mathrm{E}(\|\mathbf{X}/\sqrt{p}\|) < \infty$ and $|K|_\infty = \sup_{u\in\mathbb{R}} |h^{-1}K(u/h)| < \infty$.

Proposition 4 (Consistency)**.** *Under Conditions* 1-5*, it holds that*

(i) *when p is fixed, $\widehat{\boldsymbol{\beta}}_1 \to \boldsymbol{\beta}_{01}$ in probability as $n \to \infty$;*

(ii) *when p satisfies $p^{p/2}n^{-\alpha(1-\delta)} = o(1)$ for any $\alpha \in (0, 1/2)$, $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = o_{\mathrm{p}}(1)$.*

Proposition 4 shows that the MMV estimator is consistent for both fixed and diverging p cases under regularity conditions. For the diverging case, we require p grows quite slowly with n. The condition $p^{p/2}n^{-\alpha(1-\delta)} = o(1)$ might be improved. In the simulation, we find that MMV works well even when $p \gg n$. Furthermore, when p is fixed, the $\sqrt{n}$ consistency and asymptotic normality can be further proved. We introduce some new notation.

Let $C(\kappa_1) = \{\boldsymbol{\beta}_1 \in \mathbb{C}^p : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \le \kappa_1\}$ be a complex ball in $\mathbb{C}^p$ with center $\boldsymbol{\beta}_{01}$ and radius κ_1. Denote $\Gamma_{C_1} = \{\boldsymbol{\beta}_1 \in \mathbb{C}^p : \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 = 1\}$. Let Ω_{C_1} be the parameter space of the complex $\boldsymbol{\beta}_1$. For the case $d = 1$, denote

$$L(\boldsymbol{\theta}_1) =: L(\boldsymbol{\beta}_1, \lambda_1) = \mathrm{MV}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) + \lambda_1(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 - 1)\,,$$

$$L_{nh}(\boldsymbol{\theta}_1) =: L_{nh}(\boldsymbol{\beta}_1, \lambda_1) = \mathrm{MV}_n(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) + \lambda_1(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\beta}_1 - 1)\,,$$

where λ_1 denotes the Lagrange multiplier and $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \lambda_1)^{\mathrm{T}}$. Let $\boldsymbol{\theta}_{01} = (\boldsymbol{\beta}_{01}^{\mathrm{T}}, \lambda_{01})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\theta}}_1 = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \hat{\lambda}_1)^{\mathrm{T}}$ be the maximizers of $L(\boldsymbol{\theta}_1)$ and $L_{nh}(\boldsymbol{\theta}_1)$, respectively. The following assumptions are needed.

Condition 6. There exists a positive constant κ_{01}' such that $\mathrm{MV}(\boldsymbol{\beta}_1)$ is an analytic function of each coordinate of $\boldsymbol{\beta}_1$ in $C(\kappa_{01}') \subseteq \Omega_{C_1} \cap \Gamma_{C_1}$ and $\sup_{\boldsymbol{\beta}_1 \in C(\kappa_{01}')} \mathrm{MV}(\boldsymbol{\beta}_1) < \infty$. For any $\boldsymbol{\beta}_1 \in C(\kappa_{01})$, $\mathrm{MV}^{'}(\boldsymbol{\beta}_1)$, $\mathrm{MV}^{''}(\boldsymbol{\beta}_1)$ and $\mathrm{MV}^{'''}(\boldsymbol{\beta}_1)$ exist, and $\sup_{\boldsymbol{\beta}_1 \in C(\kappa_{01}')}(\|\mathrm{MV}^{'}(\boldsymbol{\beta}_1)\| + \|\mathrm{MV}^{''}(\boldsymbol{\beta}_1)\|) < \infty$.

Condition 7. $L^{''}(\boldsymbol{\theta}_{01})$ is nonsingular.

Condition 6 is an extension of Condition 5 to the complex setting. Condition 7 is required to guarantee the root-n consistency of the proposed estimator, which is in the spirit of the Von Mises proposition (Serfling (1980), Section 6.1).

Let $\alpha(\mathbf{X}_i) = 2\sum_{r=1}^{R} p_r\{F(\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}_i) - F_r(\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}_i)\}\{f(\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}_i) - f_r(\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}_i)\}$ and $\boldsymbol{\Sigma} = \mathbb{E}\{\alpha(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\} + 4\lambda_{01}\boldsymbol{\beta}_{01}\mathbb{E}\{\alpha(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\} + 4\lambda_{01}^2\boldsymbol{\beta}_{01}\boldsymbol{\beta}_{01}^{\mathrm{T}}$. We then define $\mathbf{V}_1 = \mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1$ for $\mathbf{A}_1$ specified in (S5.8) in the supplement. We obtain the following result.

Theorem 5 (Asymptotic normality)**.** *Under Conditions* 1-7*,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})$ *is asymptotically normally distributed with mean zero and covariance matrix* $\mathbf{V}_1$.

6. Numerical studies

In practice, the number of the indexes d is unknown. It is usually better to use cross validation to choose the proper number of d, and the empirical bandwidth h. However, the computation cost of doing so is very high, especially in simulation. For simplicity, instead, in the following simulation studies, we use d as the true dimension of the central subspace and $h = 3 \cdot \mathrm{sd}(\tilde{\boldsymbol{\beta}}_1)n^{-1/3}$ where sd stands for standard deviation and $\tilde{\boldsymbol{\beta}}_1$ is a good initial estimate of $\boldsymbol{\beta}_1$.

We use ten-fold cross validation to calculate the classification error in both simulation and real data analysis. We repeat the experiment 400 times, and the average classification error and the corresponding standard deviation (in parentheses) are then calculated. Let $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (1, -1, 1, -1, 0, \ldots, 0)^{\mathrm{T}}$. The calculation for LDA, logistics regression, SVM and KKN is based on the corresponding Matlab (R2015a) packages using default settings.

Although the MMV method can be readily applied under settings of $n < p$, the computation cost for a sequential algorithm like ours is quite high. If the predictors' dimension is ultra-high, our suggestion is to conduct

We find that the results are robust to the choice of the bandwidth, and we report the results for $h = 3 \cdot \mathrm{sd}(\tilde{\boldsymbol{\beta}}_1)n^{-1/3}$ as a representative.

feature screening first to reduce dimensionality p (say, $\exp(O(n^{\xi}))$ for some $\xi > 0$) to a relatively large scale d' (*e.g.* , o(n)) by fast methods such as those of Fan and Fan (2008) and Cui, Li and Zhong (2015). When the size of p is comparable to n, our method is quite fast and effective.

6.1 Fisher's LDA

In this study, we set p be 50 and 200 respectively, with the sample size $n = 80$. We generate $\mathbf{Y} = (1, \ldots, 1, -1, \ldots, -1)^{\mathsf{T}}$ first, and then generate $\mathbf{X}$ as follows. It is an ordinary LDA model which is in fact an inverse model with a one-dimensional central subspace.

Model I

$$\mathbf{X} = \boldsymbol{\beta}_1 Y + \boldsymbol{\Delta}\boldsymbol{\epsilon}\,,$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$ and $\boldsymbol{\Delta} = (\Delta_{ij})$ with $\Delta_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$.

Table 1 shows that MMV+LDA outperforms LDA significantly in both settings ($p = 50$ and $p = 200$). By applying MMV, the classification error is decreased by 50 percent or so. Although MMV+LDA equals LDA in the population level, the former does work better in the finite sample settings. The reason might be that dimension reduction through MMV increases estimation efficiency. Notice that the performance of LDA gets suprisingly better when p goes higher from 50 to 200. That is because the traditional

Table 1: Average classification error (percentage)

Method	MMV+LDA	LDA
	$p = 50$	
Model I	9.95 (3.43)	24.15 (6.21)
	$p = 200$	
Model I	13.87 (4.42)	19.83 (6.33)

LDA fails when the empirical covariance matrix for $p = 200$ is nonsingular, and the regularized LDA is then applied. See the matlab documentation for details on the regularized LDA. Despite the regularization, our MMV+LDA still gains superiority over LDA.

6.2 Logistic regression

In this study, we set p be 20 and 50 respectively, with the sample size $n = 80$. Since logistic regression utilizes likelihood estimation, the sample size is required to be larger than the dimension of the predictors. We generate data using the logistic model as follows.

Model II

$$Y = \mathrm{I}\left(1/\{1 + \exp(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})\} \geq 0.5\right),$$

where I is the indicator function and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Psi})$ with $\mathbf{\Psi} = (\Psi_{ij})$ and $\Psi_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. In this study, the central subspace is spanned by the direction $\boldsymbol{\beta}_1$.

Table 2: Average classification error (percentage)

Method	MMV+Logistic regression	Logistic regression
	$p = 20$	
Model II	9.33 (3.36)	13.31 (4.16)
	$p = 50$	
Model II	14.85 (4.36)	31.71 (6.27)

It can been seen clearly from Table 2 that MMV, as a dimension reduction technique, improves estimation efficiency, and thus reduces classification error remarkably when it is combined with Logistic regression.

6.3 More complex models

We compare our method with more advanced algorithms like SVM and KKN in this study. Models III and IV are multiple index models. We set p be 50 and 200 respectively, while the sample size $n = 160$. In these two models, the central subspace is spanned by the directions $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

Model III

$$Y = \mathrm{I}\left(\boldsymbol{\beta}_1^{\mathrm{T}} X/\{0.5 + (\boldsymbol{\beta}_2^{\mathrm{T}} X + 1.5)^2\} + 0.2\epsilon \geq 0\right),$$

where $\epsilon \sim N(0,1)$, $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ with $\Psi_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$, and $\mathbf{X} \perp\!\!\!\perp \epsilon$.

Model IV

$$Y = \mathrm{I}\left((\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})^2 + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X})^2 + 0.2\epsilon \geq 1\right),$$

where $\epsilon \sim N(0,1)$, $\mathbf{X} = (X_1, \ldots, X_p)^{\mathsf{T}} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ with $\Psi_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$, and $\mathbf{X} \perp\!\!\!\perp \epsilon$.

Table 3: Average classification error (percentage)

Method	MMV+SVM	SVM	MMV+KKN	KKN
		$p = 50$		
Model III	16.65 (3.61)	20.49 (4.01)	17.80 (3.75)	34.52 (4.51)
Model IV	16.66 (4.09)	25.22 (4.46)	18.33 (4.65)	22.75 (5.43)
		$p = 200$		
Model III	25.18 (4.05)	27.08 (4.45)	25.63 (3.99)	42.21 (4.40)
Model IV	16.20 (5.02)	19.35 (5.89)	14.09 (4.06)	22.70 (6.45)

Table 3 indicates that even for complex classification algorithms such as SVM and KNN, employing MMV before classification still enjoys a significant decrease of the classification error and its variance. This further confirms the efficiency and priority of our proposed method.

6.4 Computation cost

We compare the proposed method (MMV) with SIR and one powerful method called MAximum SEparation Subspace (MASES) recently proposed by Zhang, Mai and Zou (2020) for sufficient dimension reduction with categorical response. The **SIR** function in matlab is used for the SIR method, and we obtain the MASES's matlab code in the author's webpage. The comparison are conducted in Model I as an illustration. For each method, we record the average computation time (in seconds), together with the average classification error across 400 repetitions under each (n, p) setting. The results are reported in Figures 1 and 2.

Although the proposed MMV method exhibits slower computation speed compared to the two contrastive methods, it significantly improves the accuracy of classification. While SIR demonstrates the fastest execution, its

The results are a bit different from those in Table 1 of the paper. The reason is that we generate the categorical response as $Y = (1, \ldots, 1, 0, \ldots, 0)^\top$ instead of $Y = (1, \ldots, 1, -1, \ldots, -1)^\top$. Otherwise, the SIR and MASES code would produce an error and fail to generate any results.

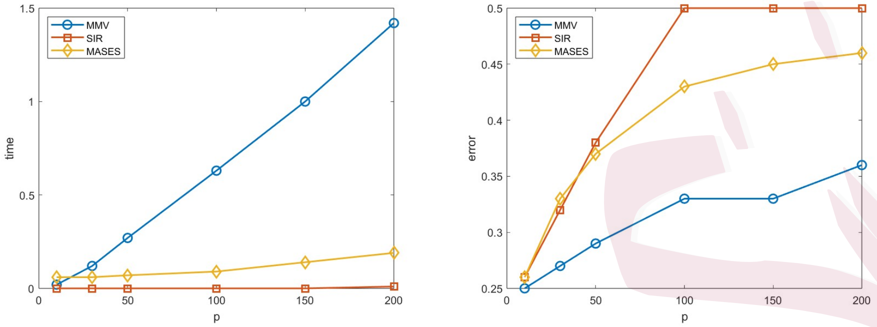

Figure 1: Average computation time (in seconds) and classification error versus the problem dimension with $n = 80$.

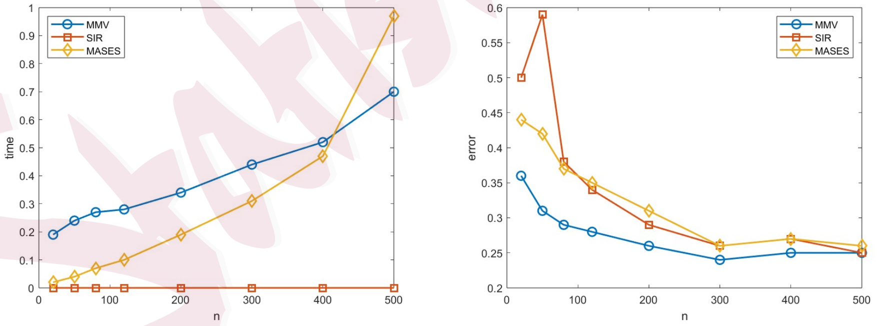

Figure 2: Average computation time (in seconds) and classification error versus the sample size with $p = 50$.

classification error is notably high. In fact, as shown in Figure 1, when the problem dimension p goes to 100, the classification error of SIR reaches 50%, which equals random guess in this binary classification problem. Moreover, compared with MASES, MMV exhibits some robustness against the sample size n, as shown in the left subfigure of Figure 2. In summary, the proposed method can achieve satisfactory estimation and classification within an acceptable timeframe.

6.5 Real data analysis

We apply our method to human colon cancer data with $n = 62$ and $p = 2000$, which is available in R. There are 40 samples from tumors "t", and 22 samples are from normal "n" biopsies. The data was originally collected on microarrays with 6500 probes. 2000 of them were selected apparently, randomly, to be used for demonstrating statistical methods. We first screen the number of the predictors to 100 by the method of Cui, Li and Zhong (2015), and compare our methods with LDA, SVM and KKN. We apply the same bandwidth selection and cross validation methods as those in the simulation study. A few choices of the dimension $d = 1, 2, 3$ are tried, and they have rather similar results. Here we only present the result with fixed $d = 1$. To get a fair comparison, we repeat the permutation 100 times for cross validation results. The table below summarizes the average classification errors and the corresponding standard deviations.

Table 4: Average classification error (percentage)

MMV+LDA	LDA
11.24 (1.16)	16.74 (3.00)
MMV+SVM	SVM
12.40 (1.52)	16.53 (2.41)
MMV+KKN	KKN
14.37 (2.14)	18.73 (1.43)

Table 4 demonstrates that “MMV+.” performs much better than the original classification method. It seems that the performances of the three methods are comparable to each other, with SVM performing a little better, while MMV+LDA performs the best among the three “MMV+.” methods. We conjecture that the relationship between MMV and the original classification algorithms may be more than simple addition.

7. Conclusion

In this paper, we propose a new sufficient dimension reduction approach – Maximal Mean Variance, which is designed deliberately for high-dimensional classification. Our method requires fairly mild restrictions on the predictors and keeps the model-free advantage without estimating the link function. Hence, it can be applied to a wide range of scenarios. The relationships between MMV and several popular classification methods are discussed in detail. The asymptotic properties of the MMV estimator are investigated. Numerical experiments show the potential of the proposed method for $p \gg n$ problems.

While cross validation can be employed to choose a proper dimension d of the central subspace in practice, it would be quite challenging to derive an optimal d theoretically. This is partially because the algorithm of MMV is a stepwise procedure. We leave it for further research. Besides, the MMV method can be readily applied to the ultra high-dimensional setting by conducting a screening procedure first. This two-scale learning framework carrys forward the spirit of Fan and Lv (2008) for sure independence screening.

Supplementary Materials

The supplementary material includes all the theoretical proof of the main paper.

Acknowledgements

Jia Zhang was supported by the National Natural Science Foundation of China (grant nos. 71991472 and 72003150).

References

Bura, E. and Forzani, L. (2015). Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, **110**, 420–434.

Bura, E., Duarte, S. and Forzani, L. (2016). Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, **111**, 1313–1329.

Cheng, F. (2017). Strong uniform consistency rates of kernel estimators of cumulative distribution functions. *Communications in Statistics-Theory and Methods*, **46(14)**, 6803-6807.

Chen, X., Ma, X. and Zhou, W. (2017). Distribution Regression. arXiv preprint arXiv:1712.08781.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177–189.

REFERENCES

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.

Cook, R. D. and Li, L. (2009). Dimension reduction in regressions with exponential family predictors. *Journal of Computational and Graphical Statistics*, **18(3)**, 774-791.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428.

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

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, **110**, 630–641.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, **B**, 1-31.

Fabian, V. (1985). *Introduction to probability and mathematical statistics*. John Wiley and Sons Incorporated.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, **36(6)**, 2605.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, **B**, **70(5)**, 849-911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32(3)**, 928-961.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 867-889.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R.* Chapman and Hall/CRC.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–342.

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(420)**, 1025–1039.

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

Liu, R. and Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, **20(8)**, 661-677.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics (Vol. 162)*. John Wiley and Sons.

Sheng, W. and Yin, X. (2013). Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, **122**, 148–161.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, **103(482)**, 811-821.

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

Sheng, W. and Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics* **25**, 91–104.

Chen, X., Zhang, J., and Zhou, W. (2022). High-dimensional elliptical sliced inverse regression in non-Gaussian distributions. *Journal of Business & Economic Statistics*, **40(3)**, 1204–1215.

Zhang, X., Mai, Q., and Zou, H. (2020). Maximum separation subspace in sufficient dimension reduction with categorical response. *Journal of Machine Learning Research*, **21(29)**, 1–36.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, **101(476)**, 1638-1651.

Department of Statistics & Data Science, Southern University of Science and Technology, Shenzhen, China

E-mail: chenx8@sustech.edu.cn

Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada

E-mail: jinwu@ucalgary.ca

Department of Statistics and Data Science, National University of Singapore, Singapore

E-mail: zhigang.yao@nus.edu.sg

Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China

E-mail: zhangjia@swufe.edu.cn