STAT 578: Report

Sheng Hu

Apr 29, 2020

Contents

Introduction

Sufficient dimension reduction (SDR) is a useful concept to deal with the dimension reduction problem without influence the relationship between predictor and response variable. The most important idea in SDR is Central Subspace (CS), the smallest Dimension Reduction Subspace (DRS). Most of the previous methods used to estimate CS have based on the moment information, such as SIR, SAVE. There are also various assumptions in these methods, such as linear conditional mean (LCM) and constant conditional variance (CCV) assumption. Xin Zhang (2020) proposed a new method called Maximum Separation Subspace (MASES) which uses the distribution information of variables to estimate the Dimension Reduction Subspace on the setting of categorical response. First, the MASES always exists, which is better than CS, because the existence of CS is not always guaranteed. Second, the proposed estimator of MASES is consistent which shows the MASES could be estimated well based on large samples. Finally, the relationship between MASES and CS is close. When CS exists, they are equivalent. Thus, MASES has the nice property of CS.

Proposed method

General design

Suppose we have the univariate response $Y \in \{1, 2, \ldots, C\}$ and the multivariate predictor $X \in R^p$, where C is a constant. Let $f_k(X), k = 1, 2, \ldots, C$ denotes the conditional density function of $X|Y = k$, and $f_k(B^T X), k = 1, 2, \ldots, C$ denotes the conditional density function of $B^T X|Y = k$, where $B \in R^{p \times q}$. Let be $\delta(f_1, f_2)$ a distance of the two (conditional) probability density functions with the following three properties.

$$\delta(f_1, f_2) = \delta(f_2, f_1) \tag{1}$$
$$\delta(f_1, f_2) >= 0, equality\ holds\ only\ when\ f_1, f_2\ are\ almost\ the\ same \tag{2}$$
$$\delta(f_1, f_2) < \delta(f_1, f_3) + \delta(f_3, f_2) \tag{3}$$

Apparently, the further properties of $\delta(f_1, f_2)$ depend on the specific form of $\delta(f_1, f_2)$. For further exploration, we suppose that $\delta(f_1, f_2) = \delta(f_1', f_2')$, where $f_k', k = 1, 2$ denotes the conditional density function of$X - \mathbf{a}|Y = k$, where $\mathbf{a}$ is a vector with constant value.

In this paper, the authors mainly use Squared Hellinger distance as the specific form of distance.

$\delta(f_1, f_2) = \frac{1}{2}\int(\sqrt{f_1(\mathbf{x})} - \sqrt{f_2(\mathbf{x})})^2 d\mathbf{x}$

When $C > 2$, we use $D(X) = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} w_{ij}\delta(f_i(X), f_j(X))$ as metric, where $w_{ij} > 0$ is weight for each distance. Denotes $\mathcal{H}(X)$ as the $D(X)$ when using Squared Hellinger distance.

Core idea

Based on the following five properties of $D(X)$, we could use $D(X)$ to evaluate how much information the subspace of X contains. Moreover, use it to find the subspace which contaion the same conditional imformation as the full space.

For any matrix $A \in R^{p\times r}$ and $B \in R^{p\times q}$.

Boundedness $0 \leq D(A^T X) \leq 1$

Indistinguishability $D(A^T X) = 0$ if and only if all pairs of probability density functions $f_i(A^T X)$ and $f_j(A^T X)$ are identical almost every on $A^T X \in R^r$

Perfect separation $D(A^T X) = 0$ if and only if all pairs of probability density functions $f_i(A^T X)$ and $f_j(A^T X)$ have non-overlapping support on $A^T X \in R^r$ for $i \neq j$

Invariance If $span(A) = span(B)$, then $D(A^T X) = D(B^T X)$

Monotonicity If $span(A) \subseteq span(B)$, then $D(A^T X) \leq D(B^T X)$

Proposition 1 If any of the previous five properties are satisfied for $C = 2$, then they are also true for $C > 2$. And $\mathcal{H}(X)$ satisfies the five properties.

Comment: The five properties give us more information about how $D(X)$ could matric the information contained by the subspace. First, the indistinguishability shows that when $D(X)$ achieve 0, then we could not distinguish Y based on X, because every conditional distribution is the same. When $D(X) = 1$, then every Y are fully distinguished by looking at X, since there is not overlapping support region of X. However, this situation is not always achievable even in the full space of X. From the perspective of Sufficient dimension reduction, we don't need it achieves one, because the invariance property indicates that only the subspace of X is identifiable. We only need the value $D(A^T X)$ equal to $D(X)$, then we know the subspace has the same conditional information. And we have the following corollary which shows the maximum dimension we could conduct in the sense of Sufficient Dimension Reduction.

Corollary 1 There always exists an integer $0 \leq d \leq p$ such that either $0 = D_0 = \cdots = D_d = \cdots = D_p$ or $0 = D_0 \leq \cdots \leq D_d = \cdots = D_p$ Where Let $D_q = max_{A\in R^{p\times q}} D(A^T X), q = 1, 2, \ldots, p$ and $D_0 = 0$

Definition 1 Let $\beta = argmax_{B\in R^{p\times d}} D(B^T X)$. The subspace $span(\beta)$ is called Maximum Seperation Subspace (MASES) based on $D(X)$. Denote the MASES as $D_{Y|X}$.

Comment: MASES is the core concept of this paper. This MASES is based on the preperties of $D(X)$, and try to find the subspace which constains the same conditional information as the full X space. Because the MASES is the minimizer of $\mathcal{H}(B^T X)$, if we consider Squared Hellinger distance, the MASES has several insteresting and useful attributes.

Proposition 2 The MASES always exists. For any non-stochastic full rank matrix $A \in R^{p\times p}$ and vector $\mathbf{a} \in R^p$, the MASES of $Z = AX - \mathbf{a}$ on Y satisies $A^T D_{Y|Z} = D_{Y|X}$

Theorem 1 For any matrix $B \in R^{p\times q}, q < p$, we have the following equivalence.

$$\mathcal{H}(B^T X) = \mathcal{H}(X) \iff X \perp Y | B^T X$$

Theorem 2 If the CS exists, then the MASES is the CS, and is therefore unique and is the smallest DRS.

Comment: Proposition 2 illustrates the existence of MASES which is an advantage of MASES compare to CS. Moreover, MASES has a relationship between the predictor and linear transformed predictor. It's a useful relationship that could be used in the sample level when we need to standardize the sample first. Theorem 1 shows that the equal value of $\mathcal{H}(X)$ in subspace and full space of X is equivalent to the conditional independence which is used to do sufficient dimension reduction. It bridges the SDR and MASES. And then We could find the relationship between CS and MASES.

Remark: In the original paper of MASES, the proof of Theorem 1 has flaw in the final step which try to conduct conditional independence from all Zero value of function. However, it could be revised by just add an indicator set.

Estimation and consistency

With the setting $C = 2$, we have the following

$F_{pop}(B) = 1 - \mathcal{H}(B^T X) = \int \sqrt{f_1(B^T \mathbf{x}) f_2(B^T \mathbf{x})} d\mathbf{x} = E\{\frac{\sqrt{f_1(B^T X) f_2(B^T X)}}{f(X)}\}$

$F(B) = \sum_{i=1}^{n} \frac{\sqrt{\hat{f}_1(B^T X_i) \hat{f}_2(B^T X_i)}}{\hat{p_1} \hat{f}_1(B^T X_i) + \hat{p_2} \hat{f}_2(B^T X_i) + \delta_n} = \sum_{i=1}^{n} F_i(B)$

Where $\hat{p_k} = \frac{n_k}{n}$ is the percetages of the kth catagory. $\hat{f}_k(B^T X_i)$ is the guassian kernal estimator of the conditional ditribution of $B^T X|Y = k$.

The problem $argmax_{B \in R^{p \times d}} \mathcal{H}(B^T X) = argmin_{B \in R^{p \times d}} F_{pop}(B)$

In the sample level, we do $argmin_{B \in R^{p \times d}} F(B)$. This is an optimization problem that has constraint $B^T B = I$ which is called Grassmann manifolds optimization which is easy. Usually, an algorithm designed to solve Grassmmna manifolds optimization problems is easily stuck at a local minimum. So a good initial point for the iterative algorithm is crucial. The authors provided a sequential algorithm that could be used to search an initial point and also find the smallest dimension d we mentioned before.

"Minimization with orthogonality constraints (e.g., $X^T X = I$) and/or spherical constraints (e.g., $||x||_2 = 1$) has wide applications. These problems are difficult because the constraints are not only non-convex but numerically expensive to preserve during iterations." (Zaiwen Wen, 2012)

Based on the explicit forms of $F(B)$ and $dF(B)/dB$, we may use any off-the-shelf optimization methods to obtain the MASES estimator. The author's current implementation adopts the sg_min Matlab package for Stiefel and Grassmann manifolds optimization, which preserves the orthogonality constraint $B^T B = I$.

To repeat the simulation work, I use the R package "GrassmannOptim", which could find the maximum of a function constrained on Grassmann manifold.

Theoretical justification

Advantage

The maximum separation subspace is based on the idea of finding the subspace of X which could separate the conditional distribution given the categorical response Y the most. The MASES could contain more information than the conditional moment relation method, because MASES tries to get the information from the distribution. Moreover, for continuous response Y, To apply the MASES method, we could slice Y like what SIR and SAVE method did.

Limitation

The MASES method uses distribution information to find a dimension reduction subspace. It needs to collect distribution information from the sample level. However, the method will face two problems when using a kernel density estimator to construct a density function. First, the kernel density estimator is not desirable in the high dimension situation. It may lead to sparse distributed sample, and then the kernel estimator will be close to zero at some point which makes the derivative of the density function goes to infinity. To solve this problem, tuning the bandwidth parameter is necessary. Considering the time consumed for finding the maximizer among the Grassmann manifold each time, the tuning process will be hard and time-consuming. Second, because this kernel density function is an un-convex function with Grassmann manifold as parameter space, the optimization algorithm will be stuck at the local optimizer almost every time. Indeed, the author provided a sequential algorithm to find some “better” initial point. However, this local optimizer problem is inevitable. To get a better solution, the relatively high time expense will be another weakness of the MASES method.

Simulation

In the simulation part, we give two example to show the simulation outcome. we consider a binary response $Y \in \{1, 2\}$, and a multivariate predictor vector $X \in R^p$ with $p = 5$. We let $\beta \in R^{p\times 1}$ be a basis for the subspace of interest, i.e. $span(\beta) = S_{Y|X}$. For each simulation setting, a random vector β and its orthogonal completion $B_0 \in R^{p\times(p-1)}$ are randomly simulated such that (β, B_0) is an orthogonal basis for R^p.

Mixture Discriminant Analysis model (Inverse model)

The mixture discriminant analysis model. Since (β, B_0) forms an orthogonal basis for R^p , we generated $\beta^T X$ and $B_0^T X$ separately and then let $X = \beta\beta^T X + B_0 B_0^T X$. The discriminantive component $\beta^T X$ is generated from two different mixture distributions $\beta^T X|(Y = 1) \sim \frac{1}{2}N(-2, 0.1) + \frac{1}{2}N(2, 0.1)$ and $\beta^T X|(Y = 2) \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(5, 1)$. The other components are generated as $B_0^T X = N(0, I_{p-1})$, independent of Y. We generated i.i.d. samples with sample size $n = 100$ for each class.

Table 1: n = 100

	MASES	ENERGY	SIR	SAVE	DR
Mean	9.61	10.24	34.76	47.83	41.47
Standard Error	6.28	12.35	11.69	14.76	13.13

Table 2: n = 200

	MASES	ENERGY	SIR	SAVE	DR
Mean	5.15	5.47	25.73	31.63	27.31
Standard Error	2.54	2.24	10.27	8.12	6.73

Remark The values in the above tables are the angle between the true β and estimated value $\hat{\beta}$.

Analysis In the Mixture Discriminant Analysis (MDA) setting, the response Y is well separated among the $\beta^T X$ direction. The Maximum separation method outperforms the moment method. The reason is that the methods, MASES, and ENERGY, could extract the distribution information from the data when the mean and variance information could not well represent all relationship between X and Y. Moreover, with the sample with more samples, all method performs better. It reflects the consistency of these estimators.

Logistic regression model (Forwaed model)

Logistic regression model with normal predictors $X \sim N(0, I_p)$ and $p(X) = logit(2\beta^T X)$. Where $p(X) = 1 - Pr(Y = 1|X) = Pr(Y = 2|X)$ and $logit(x) = 1/(1 + exp(-x))$. And We generated i.i.d. sample with the total sample size n for two classes.

Table 3: n = 200

	MASES	ENERGY	SIR	SAVE	DR
Mean	12.63	10.43	9.72	14.69	10.68
Standard Error	4.31	3.26	2.98	8.11	3.20

Table 4: n = 400

	MASES	ENERGY	SIR	SAVE	DR
Mean	9.43	7.42	6.33	8.68	7.09
Standard Error	2.97	1.65	2.09	2.16	1.85

Remark The values in the above tables are the angle between the true β and estimated value $\hat{\beta}$.

Analysis In the setting of the Logistic model, the sample is not well separated even though the true direction which used to generate the data. Even though the MASES and ENERGY methods are not the best-performed methods, the performance of them is not bad compared to the moment methods' outcome. It shows that the Maximum separation methods could still enough information in this setting and performs well enough. Moreover, with the sample with more samples, all method performs better.

Conclusion and future work

The maximum separation subspace is based on the idea of finding the subspace of X which could separate the conditional distribution given the categorical response Y the most. It outperforms the moment related method because MASES considered the distribution information. However, MASES is a time-consuming method, and tuning parameter is needed to get a good result. In order to eliminate the weakness, the energy distance could be possibly applied to the MASSES method. A preliminary discussion about applying energy distance is provided.

Energy distance

Population Level

Consider the energy distance in the population level:

If X_1, X_1', X_2, X_2' are independent random vectors in R^d with finite expectations, $X_1 \sim X_1'$ and $X_2 \sim X_2'$, then define

$$E(X_1, X_2) = 2E|X_1 - X_2| - E|X_1 - X_!'| - E|X_2 - X_2'|$$

$E(X_1, X_2) \geq 0$, and equality holds if and only if X_1 and X_2 are identically distributed.

Because of the condition of holding equality, We could use the energy distance to find the Maximum separation subspace.

Consider the $Y \in \{1, 2\}$, $X \in R^p$, and X_j represent the conditional distribution of $X|Y = j$. Then we have the following theorm.

Theorem If $\beta \in R^{p\times d}, \beta^T\beta = I_d$ is a basis matrix of CS. $Var(X) = I_p$, if $P_\beta X \perp Q_\beta X$, where P_β is the projection matrix of β, $Q_\beta = I - P_\beta$. Then For any $\eta \in R^{p\times d}, \eta^T\eta = I_d$, $E(\eta^T X_1, \eta^T X_2) \leq E(\beta^T X_1, \beta^T X_2)$, equality holds when $span(\eta) = span(\beta)$.

Proof Let each column of η projects to space $span(\beta)$, and denoted the project point as β_0, and let $\beta^\perp = \eta - \beta_0$. Then exist a invertable matrix $A \in R^{d\times d}$ such that $\eta = \beta A + \beta^\perp$. $I_d = \eta^T\eta = (\beta A)^T\beta A + \beta^{\perp,T}\beta^\perp = A^T I_d A + \beta^{\perp,T}\beta^\perp$ $det(A^T A) \leq 1 - det(\beta^{\perp,T}\beta^\perp) = 1$,thus $det(A) \leq 1$ so $1 = det(I_d) = det(A^T I_d A + \beta^{\perp,T}\beta^\perp) \geq det(A^T I_d A) + det(\beta^{\perp,T}\beta^\perp)$

$$E(\eta^T X_1, \eta^T X_2) = E((\beta_0 + \beta^\perp)^T X_1, (\beta_0 + \beta^\perp)^T X_2)$$

$$= \frac{1}{C(d,a)} \int_{R^d} \frac{|Ee^{i<\mathbf{t},(\beta_0+\beta^\perp)^T X_1>} - Ee^{i<\mathbf{t},(\beta_0+\beta^\perp)^T X_2>}|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t}$$

$$= \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},(\beta_0+\beta^\perp)^T X>}|Y = 1) - E(e^{i<\mathbf{t},(\beta_0+\beta^\perp)^T X>}|Y = 2)|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t}$$

$$
\begin{aligned}
&= \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},\beta_0^T X>} E(e^{i<\mathbf{t},\beta^{\perp,T} X>}|\beta_0^T X, Y=1)|Y=1) - E(e^{i<\mathbf{t},\beta_0^T X>} E(e^{i<\mathbf{t},\beta^{\perp,T} X>}|\beta_0^T X, Y=2)|Y=2)|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t} \\
&= \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},\beta_0^T X>} E(e^{i<\mathbf{t},\beta^{\perp,T} X>})|Y=1) - E(e^{i<\mathbf{t},\beta_0^T X>} E(e^{i<\mathbf{t},\beta^{\perp,T} X>})|Y=2)|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t} \\
&= \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},\beta_0^T X>}|Y=1) E(e^{i<\mathbf{t},\beta^{\perp,T} X>}) - E(e^{i<\mathbf{t},\beta_0^T X>}|Y=2) E(e^{i<\mathbf{t},\beta^{\perp,T} X>})|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t} \\
&= \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},\beta_0^T X>}|Y=1) - E(e^{i<\mathbf{t},\beta_0^T X>}|Y=2)|^2 |E(e^{i<\mathbf{t},\beta^{\perp,T} X>})|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t} \\
&\leq \frac{1}{C(d,a)} \int_{R^d} \frac{|E(e^{i<\mathbf{t},\beta_0^T X>}|Y=1) - E(e^{i<\mathbf{t},\beta_0^T X>}|Y=2)|^2}{|\mathbf{t}|^{d+a}} d\mathbf{t} \\
&= E(\beta_0^T X_1, \beta_0^T X_2) \leq E(\beta^T X_1, \beta^T X_2)
\end{aligned}
$$

QED

Sample Level

Use the previous theorem, if we could know the exact dimension of CS, then we could use the Energy distance to find the CS by maximizing the Energy distance of subspace.

In the sample level, we could estimate the CS by doing the following optimization. Where $\{a_i\}_{1:n_1}$ is the sample of X given $Y = 1$, $\{b_i\}_{1:n_2}$ is the sample of X given $Y = 2$.

$$
\beta = argmax_{A^T A = I_d, A \in R^{p \times d}} \{ \frac{n_1 n_2}{n_1 + n_2} (\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ||A^T a_i - A^T b_j||
$$

$$
- \frac{1}{n_1 n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} ||A^T a_i - A^T a_j|| - \frac{1}{n_2 n_2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} ||A^T b_i - A^T b_j||) \}
$$

This is still a Grassmann manifold optimization problem, we could use the iterative algorithm to find the maximizer. And the simulation outcome has shown before in the simulation part.

Justification

The energy method is more convenient than the MASES, because we have no need to tuning parameter when we use the observed value of X and Y to find the distribution information.

However, the energy method is still a non-convex optimization problem and hard to solve.

Reference

Adragni, K. P., Cook, R. D., & Wu, S. (2012). Grassmannoptim: An R package for Grassmann manifold optimization. Journal of Statistical Software, 50(5), 1-18.

Chen, X., Yuan, Q., & Yin, X. (2019). Sufficient dimension reduction via distance covariance with multivariate responses. Journal of Nonparametric Statistics, 31(2), 268-288.

Kreyszig, E. (1978). Introductory functional analysis with applications (Vol. 1). New York: wiley. Chapter 3.

Sheng, W., & Yin, X. (2013). Direction estimation in single-index models via distance covariance. Journal of Multivariate Analysis, 122, 148-161. Székely, G. J., & Rizzo, M. L. (2004). Testing for equal distributions in high dimension. InterStat, 5(16.10), 1249-1272.

Szekely, G. J., & Rizzo, M. L. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. Journal of classification, 22(2).

Zhang, X., Mai, Q., & Zou, H. (2020). The Maximum Separation Subspace in Sufficient Dimension Reduction with Categorical Response. Journal of Machine Learning Research, 21(29), 1-36.