Variable Selection based on a Two-stage Projection Pursuit Algorithm

Shu Jiang[1] and Yijun Xie[2]
[1] *Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis, St. Louis, U.S.A.*
[2] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada*

Keywords: Two-stage Projection Pursuit, Variable Selection, Optimization.

Abstract: Dimension reduction methods have gained popularity in modern era due to exponential growth in data collection. Extracting key information and learning from all available data is a crucial step. Principal component analysis (PCA) is a popular dimension reduction technique due to its simplicity and flexibility. We stress that PCA is solely based on maximizing the proportion of total variance of the explanatory variables and do not directly impact the outcome of interest. Variable selection under such unsupervised setting may thus be inefficient. In this note, we propose a novel two-stage projection pursuit based algorithm which simultaneously consider the loss in the outcome variable when doing variable selection. We believe that when one is keen in variable selection in relation to the outcome of interest, the proposed method may be more efficient compared to existing methods.

1 INTRODUCTION

Tremendous amounts of data are being collected in the hopes of finding significant factors that may be associated with, for example, disease progression in clinical studies. With the exponential growth in data collection, a natural question is how to select a smaller subset of meaningful variables from the larger pool. A naturally adopted method in overcoming such burden is by dimensional reduction techniques. Principal component analysis (PCA) has been arguably most commonly adopted technique for such purpose. The mathematical properties as well as highly optimized algorithms for eigen-decomposition make PCA a very appealing and prevalent technique for dimension reduction. More precise descriptions on relevant uses of PCA can be found in Krzanowski (1987), King and Jackson (1999), Cadima and Jolliffe (2001), and Cadima et al. (2004).

We should note that, however, the fundamental purpose of PCA decomposition is to maximize the proportion of total variance of the explanatory variables explained by the principal components, and therefore minimize the variance of residuals. Such unsupervised approach focusing only on the variance decomposition of explanatory variables may well mimic the structure of the variables, but does not impact the relationship between the explanatory variables and the outcome/response under a regression setting. Therefore, such unsupervised method may not be the optimal approach if one's goal lies in dimensional reduction in relation to the outcome of interest.

To overcome such burden, various efforts have been made in developing projection pursuit based methods for selecting the best set of variables. Montanari and Lizzani (2001) discussed a projection pursuit algorithm to identify multivariate variables for classification. Enshaei and Faith (2015) developed an algorithm based on perceptron learning and attraction-repulsion algorithms to find the variable that best separates the data. More relevantly, Hwang et al. (1994) discussed projection pursuit learning algorithm for regression-based problems.

One of the drawbacks of traditional projection pursuit algorithm is that it requires considerable amount of computing power. Such limitation prevented previous efforts in implementing projection pursuit in higher dimensional space. The computational burden has gotten worse in recent years due to exponential increase in the number of variables being collected in the dataset. Therefore we are motivated to introduce a novel dimensional reduction technique, the two-stage projection pursuit method, for variable selection in high dimensional variable space. We believe that the proposed two-stage procedure could lead to the most efficient selection of the set of variables that exert relatively large effects on the outcome

Jiang, S. and Xie, Y.
Variable Selection based on a Two-stage Projection Pursuit Algorithm.
DOI: 10.5220/0009098901880193
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS*, pages 188-193
ISBN: 978-989-758-398-8; ISSN: 2184-4305

of interest without much computational burden.

This paper is organized as follows. We first review the principal component analysis (PCA) and its feature selection techniques in Section 2.1. We then introduce a new dimension reduction framework for high dimensional data based on a two-stage projection pursuit algorithm in Section 2.2. A detailed outline of our purposed algorithm is included in Section 2.2.2. In Section 3 we conduct a small scale simulation study to compare the performance of our proposed algorithm with existing methods including PCA. We present a data example in Section 4 and end with a discussion in Section 5.

2 DIMENSION REDUCTION

2.1 Principal Component Analysis

Let $X = [X_1, \cdots, X_d]'$ be a d-dimensional random vector with zero mean. We let $v_m = [v_{m1}, \cdots, v_{md}]'$ be a vector of length d where the norm is defined as the L^2 norm, *i.e.* $||v_m|| = v_m' v_m = \sqrt{\sum_{j=1}^d v_{mj}^2} = 1$. We further let W_m denote the inner product of v_m and X, which is often referred to as the projection score of X onto v_m in the literature. Specifically,

$$W_m = \langle v_m, X \rangle = \sum_{j=1}^d v_{mj} X_j. \tag{1}$$

The first principal component W_1 can thus be defined as

$$W_1 = \langle v_1, X \rangle,$$

where

$$v_1 = \underset{v_m \in \mathbb{R}^d, ||v_m||=1}{\text{argmax}} \ \text{Var}(W_m)$$

is the unit length vector in an $\mathbb{R}^d$ space that maximizes the variance of the projection scores. We can see from above that the first principal component tries to find a unit length vector in the d-dimensional Euclidean space such that the projection score of a higher dimensional random vector onto this unit length vector has the maximum variance among all projection scores. Such procedure will decompose the random vector X into two parts: the projection that is in the same direction as v_1, and residuals that are orthogonal to v_1. Each subsequent v_k is defined as the direction that will maximize the variance of the residuals after the $(k-1)^{th}$ projection, i.e.

$$v_k = \underset{\substack{v_m \in \mathbb{R}^d, ||v_m||=1, \\ \langle v_m, v_q \rangle = 0 \text{ for } q<k}}{\text{argmax}} \ \text{Var}(W_m).$$

The k^{th} principal component W_k can then be written as

$$W_k = \langle v_k, X \rangle.$$

The estimation of the components $v_1, \ldots, v_k$ involves the covariance matrix Σ of X which is assumed to be full rank. Specifically for v_1, we would need to maximize $v_1' \Sigma v_1$ subject to $v_1' v_1 = 1$ and one possible approach is to use Lagrange multipliers

$$v_1' \Sigma v_1 + \lambda (v_1' v_1 - 1), \tag{2}$$

where λ is a constant. By differentiating (2) with respect to v_1 we would get

$$(\Sigma - \lambda I) v_1 = 0, \tag{3}$$

where I is a $d \times d$ identity matrix. We can see from (3) that λ is an eigenvalue of Σ and v_1 is the corresponding eigenvector where

$$v_1' \Sigma v_1 = v_1' \lambda v_1 = \lambda v_1' v_1 = \lambda.$$

Hence, the maximum of $v_1' \Sigma v_1$ is achieved when $\lambda = \lambda_1$, the largest eigenvalue of Σ with v_1 the eigenvector corresponding to λ_1. Similarly, one can show that v_k is the eigenvector corresponding to the k^{th} largest eigenvalue λ_k. More details can be found in Jolliffe (2011).

2.2 Projection Pursuit Algorithm

2.2.1 Methods and Notations

It is clear that PCA is targeted at maximizing the variance of the projection scores for some high dimensional vector X. However, we stress that such unsupervised approach may not always the optimal choice when the goal for variable selection is associated with the outcome of interest. We are thus motivated to introduce an alternative dimension reduction technique, the projection pursuit algorithm in this subsection (Kruskal, 1972; Friedman and Tukey, 1974).

Similarly to the principal component analysis in the multivariate setting, we want to find the set of d-dimensional vectors $v_k = [v_{k1}, \cdots, v_{kd}]', k = 1, \cdots, d$, such that

$$v_1 = \underset{||v||=1}{\text{argmax}} \, Q(v) \text{ and} \tag{4}$$

$$v_k = \underset{\substack{||v||=1, \\ v_k' v_m = 0 \text{ for } m<k}}{\text{argmax}} \ Q(v) \text{ for } k = 2, 3, \cdots, d, \tag{5}$$

where $Q(v)$ is defined as the projection index. It can be easily seen that if we specify our projection index $Q(v)$ to be the measure of variance of X, the projection pursuit is equivalent to PCA. Under such setting, the optimal directions coincide with the eigenvectors of the sample covariance matrix.

Since the goal is to relate the set the basis functions to the outcome of interest, the $Q(v)$ function should not be solely based on the covariate X. As an example, under a linear regression setting, we let $y_i = f(x_i) + \varepsilon_i$ where $y_i \sim N(f(x_i), \sigma^2)$ for some arbitrary linear function $f(\cdot), i = 1, ..., n$. If we denote the estimate of y_i as $\hat{y}_i^{(v)} = \hat{f}(\langle x_i, v \rangle)$ for some $\|v\| = 1$. Then the projection index may be defined as

$$Q(v|x,y) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{(v)}\right)^2,$$

where $x = (x_1, ..., x_n)'$, $y = (y_1, ..., y_n)'$ and the associated set of basis functions $v_1, \ldots, v_k$ can be estimated from equations (4) and (5).

2.2.2 Two-stage Algorithm

Previous studies try to find the best projection direction by directly searching on a high dimensional unit sphere. The previous approaches, although performs well when the dimension of the unit sphere is low, could fail when working with high dimensional data. The "curse of dimensionality" would prevent one from extracting meaningful information from a sparsely distributed samples.

To address this problem, we propose a two-stage optimization algorithm for variable selection based on projection pursuit. We denote the target unit sphere in a k-dimensional Euclidean space as U^k, and each point v on U^k is a unit length vector with length k. In the first step, we generate N uniformly distributed point on this k-dimensional unit sphere, and denote them as $v_1, \ldots, v_N$. We denote the desired projection index of our data corresponding to v_j as

$$Q_j = Q(v_j),$$

where $Q(v_j) = Q(v_j|x,y)$ and we omit the conditions for simplicity in the algorithm. These Q_j's are then ranked and we pick the largest M of them as $Q_{(1)} \geq \cdots \geq Q_{(M)}$ with their corresponding unit length vector denoted as $v_{(1)}, \ldots, v_{(M)}$. The tuning parameters, N and M are user specified and will be accompanied with larger computational burden as they get larger. However, if the surface of projection index $Q(v)$ is quite smooth on U^k, then $N = 1000$ and $M = 5$ shall be enough. The tuning parameters could also be determined using a elbow-plot.

In our second step, we apply an optimization algorithm in a small neighbor near selected $v_{(1)}, \ldots, v_{(M)}$ using some general-purpose optimization method that allows user-specified searching boundary, such as L-BFGS-B proposed by Byrd et al. (1995). Let $\tilde{v}$ denote the the unit length vector that maximize our projection index, then we can obtain the weight of each variable regarding desired projection index from $\tilde{v}$. Note that Algorithm 2.1 gives a general guideline for approximating the first v. For $v_2, \ldots, v_d$ following the first direction, one can repeat Algorithm 2.1 on unit spheres orthogonal to all previously approximated directions.

This algorithm is close to the coarse-to-fine optimization schemes which is often discussed in the machine learning literature. For more detailed reference, one can find Pedersoli et al. (2015) and Charniak and Johnson (2005) for applications in computer vision and natural language processing.

Algorithm 2.1: Two-Stage Projection Pursuit-based Algorithm for Variable Selection.

```
**Input:** $x_1, \ldots, x_n, y_1, \ldots, y_n$
**Result:** $\tilde{v}$
generate $v_1, \ldots, v_N$;
**for** $j = 1$ **to** $N$ **do**
    denote $Q_j = Q(v_j)$;
**end**
rank $Q_1, \cdots, Q_J$ in decreasing order as
  $Q_{(1)}, \cdots, Q_{(J)}$;
**for** $m = 1$ **to** $M$ **do**
    find $v_{(m)}$ corresponding to $Q_{(m)}$;
    find optimized $\tilde{v}_{(m)}$ in the near neighbor
      that maximize the objective function;
**end**
let $\tilde{v} = \{\tilde{v}_{(m)} : Q(\tilde{v}_{(m)}) = \max_{m=1,\cdots,M} Q(\tilde{v}_{(m)})\}$;
return $\tilde{v}$.
```

For the rest of this paper, we adopt the proposed algorithm and demonstrate that projection pursuit method can lead to an efficient and robust dimension reduction for high dimensional data.

3 SIMULATION STUDY

We will investigate the performance of the proposed method with PCA under this section. We generate 51 centered and normally distributed random variables $X_1, \ldots, X_{51}$, where $X_1 \sim N(0,1)$ and $X_j \sim N(0,2), j = 2, \ldots, 51$. We simulate $i = 1, \ldots, 1000$ individuals in this study. We set the linear model to be

$$y_i = x_{i1}\beta_1 + \sum_{j=2}^{51} x_{ij}\beta_j + \varepsilon_i$$

for $i = 1, \ldots, 1000$, where $\beta_1 = 1$ and $\beta_2, \ldots, \beta_{51} = 0$, and $\varepsilon \sim N(0, 0.1)$. We would like to do variable selection using PCA and our proposed two-stage projection pursuit method, and assess the mis-selection rate. The proposed simulation procedure has been repeated 1000 times, and the comparison between PCA and two-stage projection pursuit is presented in Figure 1. From the histogram we can tell that projection pursuit method select the correct variable every time in the simulation, while PCA is tricked by the high variance of noise terms and can never select the correct variable. While this is an overly simplified example, we can learn from the results that PCA may not be a reliable way of choosing parameters.

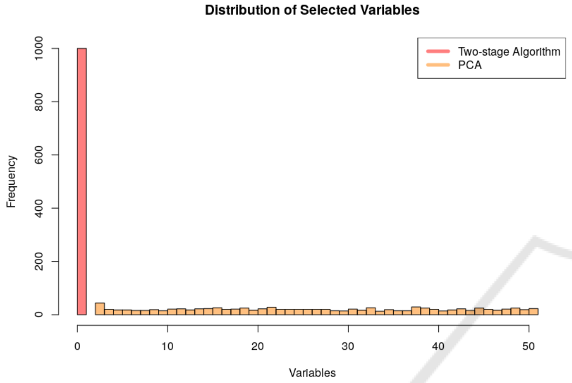

Figure 1: Distribution of selected variables. The red bars denote weights found using two-stage projection pursuit algorithm. The orange bars denote weights found using principal component analysis.

To further emphasize the advantage of our proposed two-stage projection pursuit algorithm, we conduct another simulation study with the same setting as above. In stead of using PCA for dimension reduction, we apply the random projection pursuit by generating 10^3 uniformly distribution random points on the 51-dimensional unit sphere. The frequencies of selected variables are presented in Figure 2. While the random projection pursuit select the correct variable about 60% of the time, we can tell that about $1/3$ of the time it will fail such a simple task due to the high dimension of our variable space. All above results are also summarized in Table 1.

Table 1: Counts of Selected Variables by Two-stage Projection Pursuit, Random Projection Pursuit, and PCA.

Method	X_1	$X_2,\ldots,X_{51}$
Two-stage Projection Pursuit	1000	0
Random Projection Pursuit	692	308
PCA	0	1000

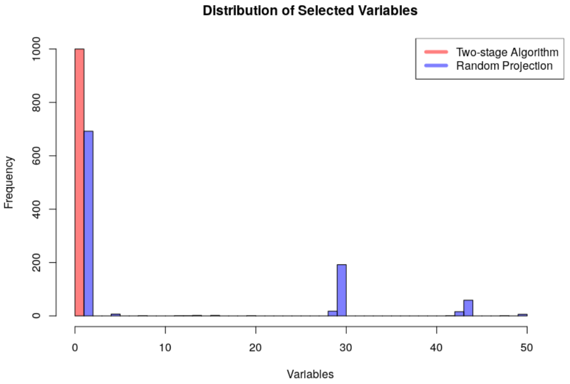

Figure 2: Distribution of selected variables. The red bars denote weights found using two-stage projection pursuit algorithm. The blue bars denote weights found using random projection pursuit method.

4 DATA EXAMPLE

`Bostonhousing` is a popular dataset that was collected by Harrison Jr and Rubinfeld (1978). In this datset there are 13 variables that are potentially related to the housing price in Boston, and they are summarized in Table 2.

Table 2: 13 Explanatory Variables and 1 Response Variable in Boston Housing dataset.[1]

Variable	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $ 10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $ 1000's

In this data example, we estimate the weight of each of these 13 variables when fitting a linear regression model with mean absolute error (MAE). We first compare our two-stage projection pursuit algorithm with the first principal component. While practitioner often use PCA as a technique for feature selection, we

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

can tell from Figure 3 that the results could be very different from projection pursuit which is specialized in finding the optimal direction.

We further compare our results with a random projection search by generating 10^5 uniformly distribution random points on the 13-dimensional unit sphere. The estimated weights are presented in Figure 4. From the plot we can easily tell that there is considerable differences for all variable weights except one. Our explanation is that even though we generate 10^5 uniformly distribution random points on the 13-dimensional unit sphere, they are actually still distributed very sparsely in the space. These random points may not be able to cover the whole space, and hence may very likely to miss the try direction that will maximize our projection index which is defined as MAE in this particular example.

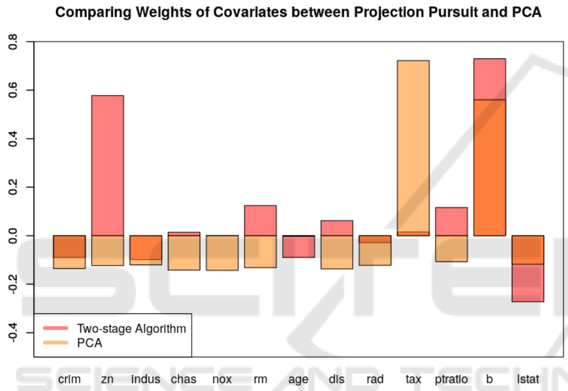

Figure 3: Weight of each of 13 variables in Boston Housing dataset. The red bars denote weights found using two-stage projection pursuit algorithm. The orange bars denote weights found using principal component analysis.

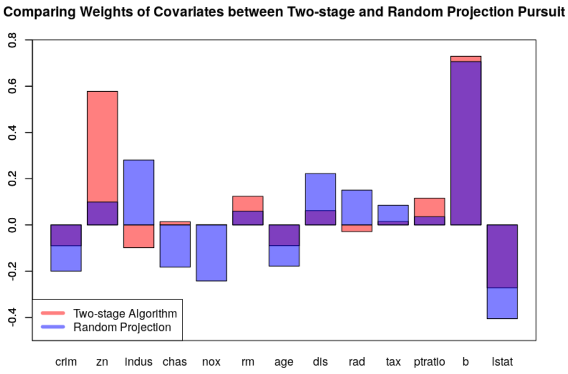

Figure 4: Weight of each of 13 variables in Boston Housing dataset. The red bars denote weights found using two-stage projection pursuit algorithm. The blue bars denote weights found using random projection pursuit method.

5 CONCLUSIONS

In this note we have introduced a new technique, namely the two-stage projection pursuit algorithm in achieving variable selection with high dimensional data. We stress that PCA is based on maximizing the proportion of total variances explained by the principal components which may not be suitable in variable selection under certain scenarios as shown under our simulation studies. Projection pursuit algorithm, on the other hand, can be applied to a more flexible objective function which include PCA as a special case. Previous efforts have been made in optimizing such projection indices only in lower dimensional unit sphere due to computation burden. Our proposed two-stage algorithm overcomes such limitation in the optimization process within a high dimensional variable space. We believe this projection pursuit based method is more flexible and can be more efficient for feature selection. In this paper we used a common dataset in machine learning to illustrate the performance of our projection pursuit based method. Note that the proposed method can be applied to other application settings without much modification. Furthermore, a larger and more intensive simulation study is needed to consolidate our proposed method and will be included in future work.

REFERENCES

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cadima, J., Cerdeira, J. O., and Minhoto, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components. *Computational statistics & data analysis*, 47(2):225–236.

Cadima, J. F. and Jolliffe, I. T. (2001). Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(1):62.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics.

Enshaei, A. and Faith, J. (2015). Feature selection with targeted projection pursuit. *IJ Information Technology and Computer Science*, 7(5):34–39.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890.

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of* *environmental economics and management*, 5(1):81–102.

Hwang, J.-N., Lay, S.-R., Maechler, M., Martin, R. D., and Schimert, J. (1994). Regression modeling in back-propagation and projection pursuit learning. *IEEE Transactions on neural networks*, 5(3):342–353.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

King, J. R. and Jackson, D. A. (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics: The official journal of the International Environmetrics Society*, 10(1):67–77.

Kruskal, J. B. (1972). Linear transformation of multivariate data to reveal clustering. *Multidimensional scaling*, 1:101–115.

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(1):22–33.

Montanari, A. and Lizzani, L. (2001). A projection pursuit approach to variable selection. *Computational statistics & data analysis*, 35(4):463–473.

Pedersoli, M., Vedaldi, A., Gonzalez, J., and Roca, X. (2015). A coarse-to-fine approach for fast deformable object detection. *Pattern Recognition*, 48(5):1844–1853.