# A Brief Review of Sparse Principal Components Analysis and its Generalization*


*Submitted by:*

Arkajyoti Bhattacharjee [†‡]

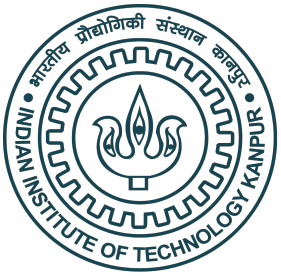
Rachita Mondal [§†]

Ritwik Vashishtha [¶†]

Shubha Sankar Banerjee [‖†]

*Supervised by:*

Dr. Minerva Mukhopadhyay [†]


April 30, 2022

**Abstract**

Principal Component Analysis is a widely studied methodology as it is a useful technique for dimension reduction. In this report, we discuss Sparse Principal Component Analysis (SPCA), which is a modification over PCA. This method is able to resolve the interpretation issue of PCA. Additionally, it provides sparse loadings to the principal components. The main idea of SPCA comes from the relationship between PCA problem and regression analysis. We also discuss GAS-PCA, which is a generalization over SPCA and this method performs better than SPCA, even in finite sample cases. Our report is mainly based on Zou et al. (2006) and its extension Leng and Wang (2009).

[*]This report has been prepared towards the partial fulfillment of the requirements of the course *MTH514A: Multivariate Analysis.*

[†]Department of Mathematics & Statistics, Indian Institute of Kanpur, India.
[‡]201277, M.Sc. Statistics (Final year).
[§]201374, M.Sc. Statistics (Final year).
[¶]201389, M.Sc. Statistics (Final year).
[‖]201416, M.Sc. Statistics (Final year).

# Contents

# 1 Introduction

In the literature of data processing and dimension reduction, Principal Component Analysis (Jolliffe (1986), Jolliffe (2022)) has been extensively studied. It has widespread applications in the fields of biology (Todorov et al. (2018)), engineering (Tzeng and Berns (2005))and in many other areas (Yoo and Shahlaei (2018)). PCA is a methodology to select those linear combinations of original variables which are able to capture maximum variability of the data. The concept of Singular Value Decomposition (SVD) can be incorporated to calculate the PCs.

Let $\mathbf{X}$ be an $n \times p$ data matrix. Here $n$ is the number of observations and $p$ is the number of independent variables. Without loss of generality, it can be assumed that the column means of $\mathbf{X}$ are zero , i.e. the columns of $\mathbf{X}$ are centered. Also, suppose that the SVD of $\mathbf{X}$ is given as:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Note that, $\mathbf{Z} = \mathbf{U}\mathbf{D}$ are the Principal Components and the columns of $\mathbf{V}$ are the corresponding loadings of the PCs. Here, $\mathbf{D}_{ii}^2/n$ is the sample variance of $i$-th PC. If only the first $q(<< p)$ PCs are chosen to represent the data then a significant amount of dimension reduction is possible. PCA is beneficial mainly for the following two properties:

i) It is able to capture the maximum variability among the columns of $\mathbf{X}$ sequentially. Hence, a very minimum amount of information loss is incurred.

ii) PCs are uncorrelated, hence different PCs can be used and interpreted independently of other PCs.

But one of the crucial drawback of PCA is that, each PC is a linear combination of all the $p$ variables and the loadings are non-zero. This creates an interpretation issue for the derived PCs.

Similar interpretation issue is common for multiple linear regression model. There the variable selection steps in. LASSO (Tibshirani (1996)) is a popular methodology for model selection. It produces

a sparse model with greater prediction accuracy. As a generalization of LASSO, Zou and Hastie (2003) has proposed Elastic Net. This report will discuss the a different approach of deriving the PCs with sparse loading, viz. SPCA. SPCA uses the fact that PCA can be written as a regression-type optimization problem.

The rest of the report is organised as follows: In Section (2) we present the popular variable selection methods in regression analysis viz. LASSO and Elastic Net. Section (3) includes a discussion regarding the relationship between regression analysis and PCA. This section provides a method of formulating the SPCA-criterion and also discusses numerical soluton techniques. Section (4) deals with one of the extensions of SPCA, which is called GAS-PCA. All the theories have been implemented in the Section (5) via synthetic data analysis and real data analysis. Finally, we conclude in the Section (6).

## 2 The LASSO and Elastic Net

Let us consider a regression model with $n$ observations and $p$ regressors. $\mathbf{Y} = (y_1, \ldots, y_n)'$ is the reponse vector. $\mathbf{X}$ is the design matrix. Let $\mathbf{X}_j = (x_{1i}, \ldots, x_{nj})$. Assume that, $\mathbf{X}$ are $\mathbf{Y}$ are centered. LASSO uses $L_1$ norm regularization technique and the lasso estimate is given by,

$$\hat{\beta}_L = \arg\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

Here $\lambda$ is the tuning parameter. Due to the nature of $L_1$ penalty, some coefficients shrinks towards zero. This allows us to have a sparse model with higher prediction accuracy. But one of the major disadvantages of LASSO is that, in case of high-dimensional regression $(n << p)$, LASSO fails to select more than $n$ predictors. This drawback has been overcome by the Elastic Net. It combines the

$L_1$ and $L_2$ penalties. For any non-negative $\lambda_1$ and $\lambda_2$ the elastic net estimate is given by,

$$\hat{\beta}_E = (1+\lambda_2)\left\{ \arg\min_{\beta}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)+\lambda_1\sum_{j=1}^{p}\mid\beta_j\mid+\lambda_2\sum_{j=1}^{p}\mid\beta_j\mid^2 \right\}$$

Clearly, for $\lambda_2 = 0$ elastic net reduces to LASSO. For $n < p$ case we can choose $\lambda_2 > 0$, thus it is helpful in high-dimensional regression problems.

# 3 Sparse Principal Component analysis (SPCA)

In both LASSO and Elastic Net sparsity is achieved only through the $L_1$ penalty. Jolliffe et al. (2003) introduced SCoTLASS procedure which obtains sparse loadings by imposing $L_1$ penalty on PCA.It successively maximizes the variance $a_k^T(\mathbf{X}^T\mathbf{X})a_k$ subject to $a_k^T a_k = 1,\ (k \geq 2),\ a_h^T a_k = 0,\ (h < k)$ and $\sum_{j=1}^{p}\mid a_{kj}\mid \leq t$, for some tuning parameter $t$. However, the loadings from SCoTLASS are not sparse enough when high percentage of explained variability is required. In this report a different approach will be shown to obtain the PCs without using the variance maximization.

## 3.1 Direct Sparse Approximation

The main idea of this report lies on the fact that, since each PC can be expressed as a linear combination of the $p$ variables, it is possible to recover the loadings by regressing PC on the $p$ variables.

**Theorem 1.** *For each i denote the i-th PC by* $\mathbf{Z}_i = \mathbf{U}\mathbf{D}_i$. *Consider a positive* $\lambda$ *and the ridge estimate is given by,*

$$\hat{\beta}_R = \arg\min_{\beta}||\mathbf{Z}_i - \mathbf{X}\beta||^2 + \lambda||\beta||^2. \tag{1}$$

*Let* $\hat{v} = \frac{\hat{\beta}_R}{||\hat{\beta}_R||}$, *then* $\hat{v} = V_i$.

*Proof.* Note that, $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Hence we have,

$$\begin{aligned}\hat{\beta}_R &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Z}_i \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{X}V_i) \\ &= V_i\frac{D_{ii}^2}{D_{ii}^2 + \lambda}\end{aligned}$$

Hence we have $\hat{v} = V_i$. □

Theorem (1) establishes the connection between PCA and the regression method. Note that, when $n > p$ and $\mathbf{X}$ is of full column rank, the theorem does not require a positive $\lambda$. But for $n < p$ and $\lambda = 0$, multiple linear regression has no unique solution which is $V_i$. The same happens when $n > p$ and $\mathbf{X}$ is not a full rank matrix. On the other hand PCA gives unique solution for all the cases. This issue is eliminated by imposing positive ridge penalty. Normalization of the estimates makes $\hat{v}$ independent of $\lambda$. Thus it is only used for reconstruction of the PCs and not for penalization.

If we add an $L_1$ penalty to (1) and consider the following optimization problem,

$$\hat{\beta} = \arg\min_{\beta}(\mathbf{Z}_i - \mathbf{X}\beta)^T(\mathbf{Z}_i - \mathbf{X}\beta) + \lambda||\beta||^2 + \lambda_1||\beta||_1. \tag{2}$$

We consider $\hat{V}_i = \frac{\hat{\beta}}{||\hat{\beta}||}$ as an approximation to $V_i$. Consequently, $\mathbf{X}\hat{V}_i$ is the $i$-th approximated PC. Note that, (2) differs from elastic net by a scaling factor $(1+\lambda)$. Zou and Hastie (2003) called it Naive Elastic Net. Since we are normalizing the coefficient, the effect of scaling factor is irrelevant. A large $\lambda_1$ will provide a sparse $\hat{\beta}$ and hence a sparse $V_i$.

## 3.2 SPCA Criterion

Theorem (1) depends on the results of PCA and so it is not an alternative procedure. It can be used in a two-stage exploratory analysis. First, we need to perform PCA, then using (2) it is possible to find

a sparse approximation.

In this section we will discuss a "self-contained" regression type criterion to derive PCs. Let $\mathbf{x}_i$ denote the $i$-th row of $\mathbf{X}$. We first state the theorem to get the leading PC and in the next theorem we will generalize it for first $k$ PCs.

**Theorem 2.** *For any $\lambda > 0$, let*

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} ||\mathbf{x}_i - \alpha\beta^T \mathbf{x}_i||^2 + \lambda||\beta||^2 \tag{3}$$
$$\text{subject to } \ ||\alpha||^2 = 1.$$

*Then $\hat{\beta} \propto V_1$.*

**Theorem 3.** *Suppose we are considering the first k PCs. Let $\mathbf{A}_{p\times k} = [\alpha_1, \ldots \alpha_k]$ and $\mathbf{B}_{p\times k} = [\beta_1, \ldots \beta_k]$. Then for any $\lambda > 0$ let,*

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i||^2 + \lambda \sum_{i=1}^{k} ||\beta_j||^2 \tag{4}$$
$$\text{subject to } \ \mathbf{A}^T\mathbf{A} = \mathbf{I}_{k\times k}.$$

*Then $\hat{\beta}_j \propto V_j$ for $j = 1, 2, \ldots, k$.*

*Proof.* Let us first have the following lemma.

**Lemma 1.** *Consider the ridge regression criterion,*

$$C_\lambda(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||^2$$

*Then if $\hat{\beta} = \arg\min_\beta C_\lambda(\beta)$ we have,*

$$C_\lambda(\hat{\beta}) = \mathbf{y}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$$

*where* $\mathbf{S}_\lambda$ *is the ridge operator and it is given by,*

$$\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$$

*Proof.* Differentiating $C_\lambda$ with respect to $\beta$ we get,

$$-\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta} = 0$$

pre-multiplying both sides by $\hat{\beta}^T$ we get,

$$\lambda||\hat{\beta}||^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})\mathbf{X}\hat{\beta}$$

Using the fact that $\mathbf{X}\hat{\beta} = \mathbf{S}_\lambda\mathbf{y}$ and $||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^T\mathbf{y} - \mathbf{y} - \mathbf{X}\hat{\beta}^T\mathbf{X}\beta$ we get, $C_\lambda(\hat{\beta}) = \mathbf{y}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$. □

Let,

$$C_\lambda(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2 \tag{5}$$

Now using some linear algebra trick it is easy to see that,

$$\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i||^2 = Tr(\mathbf{X}^T\mathbf{X}) + Tr(\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B}) - 2Tr((\mathbf{A}^T\mathbf{X}^T\mathbf{X})\mathbf{B})$$

For fixed $\mathbf{A}$ we have the minimizer of (5) as,

$$\hat{\beta}_j = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\alpha_j, \quad \text{for } j = 1, \ldots, k,$$

or equivalently,

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{A}$$

And thus,

$$\hat{\mathbf{A}} = \arg\max_{\mathbf{A}} Tr\left(\mathbf{A}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{A}\right).$$
$$\text{subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I}_k.$$

Since the eigen-vectors of $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}$ are the columns of $\mathbf{V}$ we get the following,

$$\hat{\alpha}_j = s_j V_j,$$
$$\hat{\beta}_j = s_j \frac{\mathbf{D}_{jj}^2}{\mathbf{D}_{jj}^2 + \lambda} V_j,$$

where $s_j = 1$ or $-1$ and $j = 1, \ldots, k$. □

Theorem (2) and (3) effectively transform the PCA problem to a regression problem. If we restrict $\mathbf{B} = \mathbf{A}$ then the minimizer of $||\mathbf{x}_i - \mathbf{A}\mathbf{A}^T\mathbf{x}_i||^2$ under the orthonormal constraint on $\mathbf{A}$ is exactly the first $k$ loading vectors of ordinary PCA. Theorem (3) shows that we can have exact PCA while relaxing the restriction $\mathbf{B} = \mathbf{A}$. and adding the ridge penalty term.

Adding LASSO penalty to (4) and considering the following optimization problem,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i||^2 + \lambda \sum_{i=1}^{k} ||\beta_j||^2 + \sum_{j=1}^{k} \lambda_{1,j}||\beta_j||_1 \tag{6}$$
$$\text{subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I}.$$

We can carry on the connection between PCA and regression using the LASSO approach to produce sparse loading. Here same $\lambda$ is used for all the $k$ components and different $\lambda'_{1,j}s$ are used for penal-

ization. For $p > n$ a positive $\lambda$ is required to get exact PCA when the sparsity constraint vanishes i.e. when $\lambda_{1,j} = 0, \forall j$. (6) is referred to as the SPCA criterion hereafter.

## 3.3 Numerical Solution

We will follow the alternative minimization algorithm for SPCA criterion proposed by Zou et al. (2006).

From the proof of theorem (3), we get,

$$\begin{aligned}
&\sum_{i=1}^{n} |\mathbf{x} - \mathbf{A}\mathbf{B}^T\mathbf{x}_i|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 + \sum_{j=1}^{p} \lambda_{1,j}|\beta_j|_1 \\
&\quad = Tr(\mathbf{X}^T\mathbf{X}) + \sum_{j=1}^{p} \left(\beta_j^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta_j - 2\alpha_j^T\mathbf{X}^T\mathbf{X}\beta_j + \lambda_{1,j}|\beta_j|_1\right) \qquad (7)
\end{aligned}$$

Clearly, if $\mathbf{A}$ is known, it is equivalent to solving $k$ independent elastic net problems to get $\hat{\beta}_j$ for $j = 1, 2, \ldots, k$. Further, if $\mathbf{B}$ is fixed, we can re-write the SPCA objective function as:

$$\begin{aligned}
&\sum_{i=1}^{n} |\mathbf{x} - \mathbf{A}\mathbf{B}^T\mathbf{x}_i|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 + \sum_{j=1}^{p} \lambda_{1,j}|\beta_j|_1 \\
&\quad = Tr(\mathbf{X}^T\mathbf{X}) - Tr(\mathbf{A}^T(\mathbf{X}^T\mathbf{X})\mathbf{B}) + Tr(\mathbf{B}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{B}) + \sum_{j=1}^{k} \lambda_{1,j}|\beta_j|_1 \qquad (8)
\end{aligned}$$

It is clear that given $\mathbf{B}$, and subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$, to get estimate of $\mathbf{A}$, we need to maximize $Tr(\mathbf{A}^T(\mathbf{X}^T\mathbf{X})\mathbf{B})$. The following theorem provides a solution for the maximization problem.

**Theorem 4.** *Let $\mathbf{A}_{m\times k}$ and $\mathbf{B}_{m\times k}$ be two matrices, with $\mathbf{B}$ being of rank k. Consider the constrained minimization problem:*

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} Tr(\mathbf{A}^T\mathbf{B}) \text{ subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I}_{k\times k}. \qquad (9)$$

*Suppose the SVD of $\mathbf{B}$ is $\mathbf{U}\mathbf{D}\mathbf{V}^T$, then $\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$.*

*Proof.* By our assumption, $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k$ and $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_k$. The constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$ consists of $k(k+1)/2$ constraints, which are of the form:

$$\begin{aligned} \alpha_i^T \alpha_i &= 1, & i &= 1,2,\ldots,k \\ \alpha_i^T \alpha_i &= 0, & j &> i \end{aligned}$$

We thereon use the Lagrangian multipliers method. We define,

$$\mathscr{L} = -\sum_{i=1}^{k} \beta_i^T \alpha_i + \sum_{i=1}^{T} \frac{1}{2}\lambda_{i,i}(\alpha_i\alpha_i - 1) + \sum_{j>1}^{k} \lambda_{i,j}(\alpha_i^T \alpha_i)$$

Clearly, $\partial\mathscr{L}/\partial\alpha_i = 0$ gives $\beta_i = \lambda_{i,i}\hat{\alpha}_i + \lambda_{i,j}\hat{\alpha}_j$; This can be re-written in matrix form as $\mathbf{B} = \hat{\mathbf{A}}\Delta$, where $\Delta_{i,j} = \lambda_{j,i}$. Both $\mathbf{B}$ and $\mathbf{A}$ are of full rank, thus $\Delta$ is invertible and $\mathbf{A} = \mathbf{B}\Delta^{-1}$. Thus we have,

$$\begin{aligned} Tr(\hat{\mathbf{A}}^T\mathbf{B}) &= Tr(\Delta^{-1}\mathbf{B}^T\mathbf{B}) = Tr(\Delta^{-1,T}\mathbf{V}\mathbf{D}^2\mathbf{V}^T) \\ \mathbf{I}_k &= \hat{\mathbf{A}}^T\hat{\mathbf{A}} = \Delta^{-1,T}\mathbf{B}^T\mathbf{B}\Delta^{-1} = \Delta^{-1,T}\mathbf{V}\mathbf{D}^2\mathbf{V}^T\Delta^{-1} \end{aligned}$$

Let us set $P = \mathbf{V}^T\Delta^{-1}\mathbf{V}$, we then observe the following:

$$Tr(\Delta^{-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^T) = Tr(\mathbf{V}^T\Delta^{-1}\mathbf{V}\mathbf{D}^2) = Tr(P^T\mathbf{D}^2) = \sum_{j=1}^{k} P_{jj}\mathbf{D}_{jj}^2$$

$$P^T\mathbf{D}^2P = \mathbf{I}_k$$

Now, $P_{jj}^2\mathbf{D}_{jj}^2 \leq 1$, thus,$\sum_{j=1}^{k} P_{jj}\mathbf{D}_{jj}^2 \leq \sum_{j=1}^{k}\mathbf{D}_{jj}$.

The equality occurs if and only if $P$ is a diagonal matrix and $P_{jj} = \mathbf{D}_{jj}^{-1}$.

Therefore, $\Delta^{-1} = \mathbf{V}P\mathbf{V}^T = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T$. Then, $\mathbf{A} = \mathbf{B}\Delta = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{U}\mathbf{V}^T$. □

Let $Y_j^* = X\alpha_j$. In order to solve (7), we would only require the Gram matrix $\mathbf{X}^T\mathbf{X}$, since,

$$||Y_j^* - \mathbf{X}\beta_j||^2 + \lambda||\beta_j||^2 + \lambda_{1,j}||\beta_j||_1 = (\alpha_j - \beta_j)^T\mathbf{X}^T\mathbf{X}(\alpha_j - \beta_j) + \lambda||\beta_j||^2 + \lambda_{1,j}||\beta_j||_1 \tag{10}$$

The same is also true for (8).

$n^{-1}\mathbf{X}^T\mathbf{X}$ is the sample covariance matrix of $\mathbf{X}$. Thus, we can replace $n^{-1}\mathbf{X}^T\mathbf{X}$ with $\Sigma$, the covariance matrix of $\mathbf{X}$ if $\Sigma$ is known, in (10) and thereby have the population version of SPCA. If $\mathbf{X}$ is standardized beforehand, then we use the sample correlation matrix which is preferable when the scales of variables are different.

It is important to note that replacing $\Sigma$ in place of $\mathbf{X}^T\mathbf{X}$ in (10) is not an elastic net problem. We can however turn it into one. We create an artificial response $Y^{**}$ and $\mathbf{X}^{**}$ as follows:

$$Y^{**} = \Sigma^{1/2}\alpha_j \qquad \mathbf{X}^{**} = \Sigma^{1/2}; \tag{11}$$

then it is easy to show that

$$\hat{\beta}_j = \arg\min_{\beta} ||Y^{**} - \mathbf{X}^{**}\beta||^2 + \lambda||\beta||^2 + \lambda_{1,j}||\beta||_1. \tag{12}$$

The following algorithm summarizes our discussion.

**Algorithm 1** General SPCA Algorithm

1. Let $\mathbf{A}$ start at $\mathbf{V}[,1:k]$, the loadings of the first $k$ ordinary principal components.
2. Given a fixed $\mathbf{A} = [\alpha_1, \ldots, \alpha_k]$, solve the following elastic net problem for $j = 1, \ldots, k$

$$\beta_j = \arg\min_{\beta} (\alpha_j - \beta)^T \mathbf{X}^T \mathbf{X} (\alpha_j - \beta) + \lambda ||\beta||^2 + \lambda_{1,j} ||\beta||_1$$

3. For a fixed $\mathbf{B} = [\beta_1, \ldots, \beta_k]$, compute the SVD of $\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, then update $\mathbf{A} = \mathbf{U}\mathbf{V}^T$.
4. Repeat steps 2-3 until convergence.
5. Normalization: $\hat{V}_j = \beta_j / ||\beta_j||, j = 1, \ldots, k$.

It must be noted that for $n > p$, the default choice of $\lambda$ can be zero. Empirical evidence suggests that the output of the above algorithm does not change much as $\lambda$ is varied. $\lambda$ is however chosen to be small positive number to overcome potential collinearity problems in $\mathbf{X}$.
In principle, since the above algorithm converges quite fast, we can try several combinations of $\{\lambda_{1,j}\}$ to figure out a good choice of the tuning parameters. We choose a $\lambda_{1,j}$ which gives a good compromise between variance and sparsity. However, when we face a variance-sparsity trade-off, we let variance have a higher priority.

## 3.4 Adjusted Total Variance

The ordinary principal components are uncorrelated and their loadings orthogonal. Let $\hat{\Sigma} = \mathbf{X}^T \mathbf{X}$, then $\mathbf{V}^T \mathbf{V} = \mathbf{I}_k$ and $\mathbf{V}^T \hat{\Sigma} \mathbf{V}$ is diagonal. It is clear that only the loadings for ordinary principle components satisfy these conditions. In Jolliffe et al. (2003), the loadings were forced to be orthogonal which came at the sacrifice of the uncorrelated property. SPCA also does not explicitly impose the

uncorrelated components condition either.

Let $\hat{\mathbf{Z}}$ be the modified PCs. Usually the total variance explained by $\hat{\mathbf{Z}}$ is calculated by $Tr(\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})$. This is a reasonable process when $\hat{\mathbf{Z}}$ are uncorrelated. However, if they are indeed correlated, then $Tr(\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})$ is too optimistic for representing the total variance. Suppose $(\hat{\mathbf{Z}}_i,\ i = 1,\ldots,k)$ are the first $k$ modified PCs any method and the $(k+1)^{th}$ modified $\hat{\mathbf{Z}}_{k+1}$ is obtained and we want to compute the total variance explained by the first $k+1$ modified PCs, which should be the sum of the explained variances by the first $k$ modified PCs and the additional variance from $\hat{\mathbf{Z}}_{k+1}$. If $\hat{\mathbf{Z}}_{k+1}$ is correlated with $(\hat{\mathbf{Z}}_i,\ i = 1,\ldots,k)$, then its variance contains contributions from $(\hat{\mathbf{Z}}_i,\ i = 1,\ldots,k)$, which should not be included into the total variance given the presence of $(\hat{\mathbf{Z}}_i,\ i = 1,\ldots,k)$.

Zou et al. (2006) proposes a new formula to compute the total variance explained by $\hat{\mathbf{Z}}$, which takes into account the correlations among $\hat{\mathbf{Z}}$. Using regression projections, we remove the linear dependence between the correlated components. Denote $\hat{\mathbf{Z}}_{j\cdot 1,\ldots,j-1}$ the residuals after adjusting $\hat{\mathbf{Z}}_j$ for $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_{j-1}$, i.e.

$$\hat{\mathbf{Z}}_{j\cdot 1,\ldots,j-1} = \hat{\mathbf{Z}}_j - \mathbf{H}_{1,\ldots,j-1}\hat{\mathbf{Z}}_j, \tag{13}$$

where $\mathbf{H}_{1,\ldots,j-1}$ is the projection matrix on $\hat{\mathbf{Z}}_1^{j-1}$. Then the adjusted variance of $\hat{\mathbf{Z}}_j$ is $||\hat{\mathbf{Z}}_{j\cdot 1,\ldots,j-1}||^2$, and the total total variance is defined $\sum_{j=1}^{k}||\hat{\mathbf{Z}}_{j\cdot 1,\ldots,j-1}||^2$. When the modified PCs $\hat{\mathbf{Z}}$ are uncorrelated, the new formula agrees with $Tr(\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})$.

The above computations do depend on the order of $\hat{\mathbf{Z}}_i$. However, since we have a natural ordering in PCA, the ordering does not create any issue. Using the QR decomposition, it is easy to compute the adjusted variance. Suppose $\hat{\mathbf{Z}} = \mathbf{QR}$, where $\mathbf{Q}$ is orthonormal and $\mathbf{R}$ is upper triangular. Then,

$$||\hat{\mathbf{Z}}_{j\cdot 1,\ldots,j-1}||^2 = \mathbf{R}_{jj}^2. \tag{14}$$

Hence the explained total variance is equal to $\sum_{j=1}^{k}\mathbf{R}_{jj}^2$.

## 3.5 Computational Complexity

PCA is computationally efficient for both $n > p$ and $p \gg n$ data. We discuss the computational cost of general SPCA algorithm.

1. When $n > p$:

   This type of data is where the traditional techniques work quite well. Although the SPCA criterion is defined using $\mathbf{X}$, it only depends on $\mathbf{X}$ via $\mathbf{X}^T\mathbf{X}$. We need to compute $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$ which requires $np^2$ operations. Then the same $\hat{\Sigma}$ is used at each step within the loop. Computing $\mathbf{X}^T\mathbf{X}\beta$ costs $p^2k$ and the SVD of $\mathbf{X}^T\mathbf{X}\beta$ is of order $O(pk^2)$. Each elastic net operation requires at most $O(p^3)$ operations. Since $k \leq p$, the total computation cost is at most $np^2 + mO(p^3)$, where $m$ is the number of iterations before convergence. Thus. SPCA algorithm is able to efficiently handle data with huge $n$, as long as $p$ is small.

2. When $p \gg n$:

   The method of calculating $\hat{\Sigma}$ stated above is no longer applicable, owing it to be a huge matrix of order $p \times p$. The most consuming step is solving the elastic net, whose cost is of the order $O(pnJ + J^3)$ for a positive finite $\lambda$, where $J$ is the number of non-zero coefficients. The overall cost is of the order $mkO(pKn + J^2)$, which can be expensive for large $J$ and $p$.

# 4 GAS-PCA

When the number of responses, $n$, is considerably larger than the dimensionality, $d$ i.e. $n >> d$, Zou (2006) show that the excessive shrinkage equally applied by lasso to each coefficient tend to be problematic, at least in the least-squares setting. So, they proposed *Adaptive LASSO*, wherein different shrinkage are used for different coefficients. This leads to a consistent variable selection with high efficiency because intuitively we are using more shrinkage for the zero coefficients with less shrinkage for the nonzero ones.

Motivated by Adaptive LASSO, Leng and Wang (2009) generalizes SPCA in two ways:

1. The lasso penalty in (6) is replaced by an adaptive lasso penalty;

2. The least-squares problem is expanded to a generalized least-squares problem.

So, based on the above two factors, *Generalized Adaptive Sparse Principal Components Analysis* (GAS-PCA, Leng and Wang (2009)) involves minimizing the following objective function, given a fixed matrix $A$:

$$\sum_{j=1}^{d_0} \{(\alpha_j - \beta_j)' \tilde{\Omega} (\alpha_j - \beta_j) + \sum_{k=1}^{d} \lambda_{jk} |\beta_{jk}|\}, \tag{15}$$

where $\tilde{\Omega}$ is a positive definite matrix with a probabilistic limit $\Omega$, a positive definite matrix, referred to as the *kernel matrix*. Once $\beta_j$ is estimated, $\alpha_j$ can be updated like in Zou et al. (2006). We iterate these two steps until convergence. The BIC criterion associated with this method is given by:

$$BIC_{\lambda j} = (\alpha_j - \beta_j)' \tilde{\Omega} (\alpha_j - \beta_j) + df_{\lambda j} \times \frac{\log n}{n}. \tag{16}$$

Here $df_{\lambda j}$ is the number of nonzero coefficients identified in $\hat{\beta}_{\lambda j}$.

**Choice of Kernel Matrix $\tilde{\Omega}$:**

According to Leng and Wang (2009), the choice of the kernel matrix $\tilde{\Omega}$ plays an important role in the finite sample performance. Wang and Leng (2007) proposed a method of least-squares approximation (LSA), where they found that the estimator produced by minimizing the following least-squares-type objective function:

$$(\hat{\theta} - \theta)' \hat{cov}^{-1}(\hat{\theta})(\hat{\theta} - \theta) + \sum_{k=1}^{d} \lambda_k |\theta_k|, \tag{17}$$

possesses excellent finite sample and asymptotic properties, where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ is some unknown parameter, $\hat{\theta}$ is its unpenalized estimator (e.g. the maximum likelihood estimator) and

$\hat{cov}(\hat{\theta})$ is an estimator of $cov(\hat{\theta})$. The GAS-PCA formulation in (15) is similar to that of LSA in (17). In particular, if we replace the unpenalized estimator $\hat{\theta}$ in (17) by the unpenalized eigenvector estimate $\tilde{\beta}_j$ , we can replace $\tilde{\Omega}$ by $cov^{-1}(\tilde{\beta}_j)$. For $cov^{-1}(\tilde{\beta}_j)$, however, no simple formula exists. So, Leng and Wang (2009) propose a bootstrap approach to estimate it. For given $\tilde{\Sigma}$, they apply PCA to bootstrap samples from $\mathcal{N}(0,\tilde{\Sigma})$ to produce a sufficient number of bootstrap estimates, $\hat{\beta}_J^{boot}$ for $\beta_j$ . A natural estimate of $\hat{cov}(\tilde{\beta}_j)$ is the sample covariance of $\hat{\beta}_j^{boot}$, denoted by $\hat{C}_j$, say. Finally, following Leng and Wang (2009), they fix $\tilde{\Omega} = \hat{C}_j$. This approach is referred to as GAS-PCA. The resulting GAS-PCA estimator is given by: $\hat{B}^*_\lambda = \left(\hat{\beta}^*_{\lambda j},\ldots,\hat{\beta}_{\lambda d_0}\right)^*$.

## 4.1 Asymptotic Properties of GAS-PCA

In this section, we discuss certain consistency results. For that purpose, we first define a few notations. We define the maximum shrinkage, $a_n$, applied to the significant coefficients as:

$$a_n = \{\lambda_{jk} : \beta_{jk} \neq 0 : 1 \leq j \leq d_0, 1 \leq k \leq d\},$$

and the minimum shrinkage, $b_n$, for the insignificant ones as:

$$b_n = \{\lambda_{jk} : \beta_{jk} = 0 : 1 \leq j \leq d_0, 1 \leq k \leq d\}.$$

For the ease of understanding, we fix $\hat{\alpha}_{\lambda j}$ to be fixed at $\bar{\alpha}_j \in \mathbb{R}^d$. We further define

$$\bar{\beta}_{\lambda j} = \underset{\beta_j}{\text{argmin}}\{(\bar{\alpha}_j - \beta_j)'\tilde{\Omega}(\bar{\alpha}_j - \beta_j) + \sum_{k=1}^{d} \lambda_{jk}|\beta_{jk}|\}, \tag{18}$$

where the tuning parameters are assumed to be selected according to BIC in (16). Note that (18) defines one iteration step but with a fixed $\bar{\alpha}_j$ value. For example, for the initial step, $\bar{\alpha}_j$ is usually set to be the $j^{th}$ principal component (i.e., $\tilde{\beta}_j$), and then iteratively updated according to (18). An

examination of this process can be helpful in understanding the asymptotic behavior of the fully iterated estimator. Finally, we define the set of nonzero coefficients in $\beta_j$ as $s_j = \{1 \leq k \leq d : \beta_{jk} \neq 0\}$ and the set of nonzero coefficients identified by $\bar{\beta}_{\lambda j}$ as $\hat{s}_j^{BIC} = \{1 \leq k \leq d : \bar{\beta}_{\lambda jk} \neq 0\}$. The BIC criterion in (16) is used to tune the tuning parameters.

**Theorem 5.** *Assume that $\bar{\alpha}_j - \beta_j = O_p(n^{-1/2})$ and that $\tilde{\Omega}$ converges in probability to some positive definite matrix $\Omega$, $\sqrt{n}a_n \to 0$, and $\sqrt{n}b_n \to \infty$. We have:*

1. $\bar{\beta_{\lambda j}} - \beta_j = O_p(n^{-1/2})$
2. $P(\bar{\beta}_{\lambda jk} = 0) \to 1$ *for every* $\beta_{jk} = 0$.

The above theorem can be interpreted as follows:
As long as $\bar{\alpha}_j$ is $\sqrt{n}$ consistent, $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to 0$, the resulting estimator $\bar{\beta}_{\lambda j}$ is $\sqrt{n}$-consistent, and all sparse loadings can be identified consistently.

**Theorem 6.** *Assume that $\bar{\alpha}_j - \beta_j = O_p(n^{-1/2})$ and that $\tilde{\Omega}$ converges in probability to some positive definite matrix $\Omega$. We have:*

$$P(\hat{s}_j^{BIC} = s_j) \to 1.$$

The above theorem can be interpreted as follows:
As long as the BIC criterion (16) is used in either local or global form, the true model is guaranteed to be identified consistently.
The proofs of these two theorems follow arguments similar to those in Wang and Leng (2007) and so, are omitted here.

## 4.2 Optimal Choice of the Kernel Matrix, $\tilde{\Omega}$

The choice of the kernel matrix does not matter asymptotically as long as it converges to a positive definite matrix, based on Theorems (5) and (6). In this section, we theoretically justify choosing

$\Omega = cov^{-1}(\tilde{\beta}_j)$. We compare the asymptotic efficiency of $\bar{\beta}_{\lambda j}$ by fixing $\alpha_{\lambda j} = \tilde{\beta}_j$ (i.e. the one step estimator). We show that the one-step estimator with $\Omega = cov^{-1}(\tilde{\beta}_j)$ has the "smallest" asymptotic covariance. If the one-step estimator with $\Omega = cov^{-1}(\tilde{\beta}_j)$ is asymptotically efficient, it is reasonable to expect that its fully iterated version performs better than the estimators defined by other kernel matrices.

We first partition $\beta_j$ into two parts as $\beta_j = (\beta'_{j,a}, \beta'_{j,b})'$ with $\beta_{j,a} = \{\beta_{jk} : \beta_{jk} \neq 0\}$ and $\beta_{j,b} = \{\beta_{jk} : \beta_{jk} = 0\}$. Similarly, we partition the kernel matrix $\Omega$ as

$$\Omega = \begin{bmatrix} \Omega_{aa} & \Omega_{ab} \\ \Omega_{ba} & \Omega_{bb} \end{bmatrix},$$

and the $C_j = cov(\tilde{\beta}_j)$ as

$$C_j \begin{bmatrix} C_{j,aa} & C_{j,ab} \\ C_{j,ba} & C_{j,bb} \end{bmatrix}.$$

We denote the estimator that minimizes (15) with a general matrix $\tilde{\Omega}$. Under the conditions that $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to 0$, we know by Theorem (6) that $P(\hat{\beta}_{\lambda j,b} = 0) \to 1$. Further, from Wang and Leng (2007) (equations (A5) and (A6), p. 1047), we know that

$$\hat{\beta}^{\tilde{\Omega}}_{\lambda j,a} - \beta_{j,a} = (\tilde{\beta}_{j,a} - \beta_{j,a}) + \Omega^{-1}_{aa}\Omega_{ab}\tilde{\beta}_{j,b} + o_P(n^{-1/2}). \tag{19}$$

On the other hand, if we fix $\tilde{\Omega} = \hat{C}_j^{-1}$, Wang and Leng (2007) have verified that

$$\hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a} = (\tilde{\beta}_{j,a} - \beta_{j,a}) + C_{j,ab}C^{-1}_{j,bb}\tilde{\beta}_{j,b} + o_P(n^{-1/2}). \tag{20}$$

Some algebraic manipulation yields

$$\begin{aligned} cov(\hat{\beta}^{\tilde{\Omega}}_{\lambda j,a} - \beta_{j,a}) = cov(\hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a}) + cov(\hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) \\ -cov(\hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a}, \hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) \\ -cov(\hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}, \hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a}). \end{aligned} \tag{21}$$

Because $\hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a} = O_P(n^{-1/2})$ and also by (20), we know that

$$\begin{aligned} -cov(\hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a}, \hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) \\ = -cov((\tilde{\beta}_{j,a} - \beta_{j,a}) - C_{j,ab}C^{-1}_{j,bb}\tilde{\beta}_b, \hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) + o_P(n^{-1}). \end{aligned}$$

Similarly, by (19)-(20), we have

$$\begin{aligned} cov((\tilde{\beta}_{j,a} - \beta_{j,a}) - C_{j,ab}C^{-1}_{j,bb}\tilde{\beta}_b, \hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) + o_P(n^{-1}) \\ = cov((\tilde{\beta}_{j,a} - \beta_{j,a}) - C_{j,ab}C^{-1}_{j,bb}\tilde{\beta}_b, (C_{j,ab}C^{-1}_{j,bb} + \Omega^{-1}_{aa}\Omega_{ab})\tilde{\beta}_b) + o_P(n^{-1}) \\ = \Big(cov(\tilde{\beta}_{j,a} - \beta_{j,a}, \tilde{\beta}_b) - C_{j,ab}C^{-1}_{j,bb}cov(\tilde{\beta}_{j,b}, \tilde{\beta}_{j,b})\Big) \\ \times (C_{j,ab}C^{-1}_{j,bb} + \Omega^{-1}_{aa}\Omega_{ab})' + o_P(n^{-1}) \\ = (C_{j,ab} - C_{j,ab}C^{-1}_{j,bb}C_{j,bb})(C_{j,ab}C^{-1}_{j,bb} + \Omega^{-1}_{aa}\Omega_{ab})' + o_P(n^{-1}) = o_P(n^{-1}). \end{aligned}$$

The fourth term in (21) is $o_P(n^{-1})$. Thus, the following theorem holds true.

**Theorem 7.** *Assume that $\tilde{\Omega}$ converges in probability to some positive definite matrix $\Omega$, $\sqrt{n}a_n \to 0$, and $\sqrt{n}b_n \to \infty$. We have $cov(\hat{\beta}^{\tilde{\Omega}} - \beta_{j,a}) = cov(\hat{\beta}^{*}_{\lambda j,a} - \beta_{j,a}) + cov(\hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a}) + o_P(n^{-1})$.*

The matrix $cov(\hat{\beta}^{*}_{\lambda j,a} - \hat{\beta}^{\tilde{\Omega}}_{\lambda j,a})$ is positive semi-definite. Thus, Theorem (7) implies that the optimal asymptotic efficiency of $\hat{\beta}_{\lambda j}$ is achieved when $\tilde{\Omega} = \tilde{C}^{-1}_{j}$. This justifies the use of $\Omega = cov^{-1}(\tilde{\beta}_j) =$

$C_j^{-1}$. Furthermore, by (20), one can verify that

$$cov(\hat{\beta}^*_{\lambda,a} - \beta_{j,a}) = C_{j,aa} - C_{j,ab}C^{-1}_{j,bb}C_{j,ba} + o_P(n^{-1}).$$

Since $C_{j,ab}C^{-1}_{j,bb}C_{j,ba}$ is positive semi-definite, (4.2) implies that the estimator in (15) with the optimal kernel matrix estimates the nonzero coefficients more efficiently than the unpenalized estimator.

# 5 Examples

In this section we compare the performance of SPCA and GAS-SPCA on simulated and real data. We simulate a synthetic data such that there are 3 hidden factors generating the explanatory variables. We wish to see if SPCA and GAS-SPCA can capture these hidden factors using sparse principal components. In real data analysis, we consider two datasets - Pitprops data and teaching data. We use the first data to illustrate the advantage of introducing sparsity in PCs through use of SPCA. We use the second dataset to compare the performance of SPCA and GAS-SPCA.

## 5.1 Synthetic Data Analysis

Following, Zou et al. (2006), we first created three hidden factors

$$V_1 \sim N(0,290), \quad V_2 \sim N(0,300)$$

$$V_3 = -0.3V_1 + 0.925V_2 + \varepsilon, \quad \varepsilon \sim N(0,1)$$

$$V_1, V_2 \quad and \quad \varepsilon \quad are \quad independent.$$

Then 10 observed variables were generated as the follows

$$X_i = V_1 + \varepsilon_i^1, \quad \varepsilon_i^1 \sim N(0,1), \quad i = 1,2,3,4,$$

$$X_i = V_2 + \varepsilon_i^2, \quad \varepsilon_i^2 \sim N(0,1), \quad i = 5,6,7,8,$$

$$X_i = V_3 + \varepsilon_i^3, \quad \varepsilon_i^3 \sim N(0,1), \quad i = 9,10,$$

Note that the variance of the three underlying factors is 290, 300 and 283.8, respectively and the numbers of variables associated with the three factors are 4, 4 and 2. The first three PCs together explain 96% of the total variance. These facts suggest that we only need to consider three PCs with sparse representations. As noted by Zou et al. (2006), ideally, the first derived variable should recover the factor $V_2$ only using $(X_5, X_6, X_7, X_8)$, second derived variable should recover the factor $V_1$ only using $(X_1, X_2, X_3, X_4)$ and third derived variable should recover the factor $V_3$ only using $(X_9, X_{10})$ Table (1) summarises the results. We see that both SPCA and GAS-SPCA are able to identify the the three hidden factors correctly through the first three principal components.

Table 1: Simulation Results

| | PCA | | | SPCA | | | GAS-SPCA | | |
|---|---|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| 1 | -0.351 | 0.356 | -0.002 | 0 | 0.499 | 0 | 0 | 0.500 | 0 |
| 2 | -0.351 | 0.357 | -0.001 | 0 | 0.500 | 0 | 0 | 0.500 | 0 |
| 3 | -0.351 | 0.357 | -0.002 | 0 | 0.500 | 0 | 0 | 0.500 | 0 |
| 4 | -0.351 | 0.356 | -0.001 | 0 | 0.501 | 0 | 0 | 0.500 | 0 |
| 5 | 0.356 | 0.351 | 0.003 | 0.499 | 0 | 0 | 0.500 | 0 | 0 |
| 6 | 0.356 | 0.351 | 0.003 | 0.500 | 0 | 0 | 0.500 | 0 | 0 |
| 7 | 0.356 | 0.351 | 0.003 | 0.500 | 0 | 0 | 0.500 | 0 | 0 |
| 8 | 0.357 | 0.351 | 0.003 | 0.500 | 0 | 0 | 0.500 | 0 | 0 |
| 9 | -0.005 | -0.002 | 0.707 | 0 | 0 | 0.707 | 0 | 0 | 0.707 |
| 10 | -0.005 | -0.001 | 0.707 | 0 | 0 | 0.707 | 0 | 0 | 0.707 |

But it can also be seen that the variance of $V_1$, $V_2$ and $V_3$ is very high and both the methods might be giving optimistic results due to such high variance of the hidden factors. Therefore, we decide to simulate the data again but with very low variance of the first two hidden factors as follows -

$$V_1 \sim N(0,0.2), \quad V_2 \sim N(0,0.3)$$

Table 2: Simulation results under modified setup

| | PCA | | | SPCA | | | GAS-SPCA | | |
|---|---|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| 1 | 0.003 | 0.048 | -0.494 | 0 | 0 | 0 | 0 | 0 | 0.501 |
| 2 | 0.003 | 0.048 | -0.497 | 0 | 0 | 0 | 0 | 0 | 0.503 |
| 3 | 0.005 | 0.047 | -0.498 | 0 | 0 | 0 | 0 | 0 | 0.496 |
| 4 | -0.007 | 0.046 | -0.501 | 0 | 0 | 0 | 0 | 0 | 0.500 |
| 5 | 0.501 | 0.029 | 0.004 | 0 | 0 | 0 | 0 | 0.494 | 0 |
| 6 | 0.497 | 0.035 | -0.001 | 0 | 0 | 0 | 0 | 0.503 | 0 |
| 7 | 0.499 | 0.030 | 0.004 | 0 | 0 | 0 | 0 | 0.498 | 0 |
| 8 | 0.499 | 0.028 | 0.008 | 0 | 0 | 0 | 0 | 0.506 | 0 |
| 9 | -0.045 | 0.702 | 0.069 | 0 | 0 | 0 | 0.706 | 0 | 0 |
| 10 | -0.042 | 0.703 | 0.064 | -1 | 0 | 0 | 0.708 | 0 | 0 |

$$V_3 = -0.3V_1 + 0.925V_2 + \varepsilon, \quad \varepsilon \sim N(0,1)$$

$$V_1, V_2 \quad \textit{and} \quad \varepsilon \quad \textit{are} \quad \textit{independent}.$$

Table (2) summarises the results. We now see that only GAS-SPCA is able to identify the the three hidden factors correctly through the first three principal components. Therefore, it can be observed from our simulation study that both SPCA and GAS-SPCA perform well and fulfil their purpose. However, GAS-SPCA outperforms SPCA when the variance of hidden factors is low.

## 5.2 Real Data Analysis

### 5.2.1 Pitprops Data

The pitprops data first introduced in Jeffers (1967) has 180 observations and 13 measured variables. It is the classic example for showing the difficulty of interpreting principal components. Following Jeffers (1967) we tried to interpret the first 6 PCs and compare it with 6 PCs found using SPCA. Table (3) and Table (4) summarize the results. We see that SPCA makes interpretation of the variable clearer through sparsity. Furthermore, the important variables associated with the 6 PCs do not

overlap, which further makes the interpretations easier.

Table 3: PCA (Pitprops Data)

| Variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| topdiam | -0.404 | -0.218 | 0.207 | -0.091 | 0.083 | 0.120 |
| length | -0.406 | -0.186 | 0.235 | -0.103 | 0.113 | 0.163 |
| moist | -0.124 | -0.541 | -0.141 | 0.078 | -0.350 | -0.276 |
| testsg | -0.173 | -0.456 | -0.352 | 0.055 | -0.356 | -0.054 |
| ovensg | -0.057 | 0.170 | -0.481 | 0.049 | -0.176 | 0.626 |
| ringtop | -0.284 | 0.014 | -0.475 | -0.063 | 0.316 | 0.052 |
| ringbut | -0.400 | 0.190 | -0.253 | -0.065 | 0.215 | 0.003 |
| bowmax | -0.294 | 0.189 | 0.243 | 0.286 | -0.185 | -0.055 |
| bowdist | -0.357 | -0.017 | 0.208 | 0.097 | 0.106 | 0.034 |
| whorls | -0.379 | 0.248 | 0.119 | -0.205 | -0.156 | -0.173 |
| clear | 0.011 | -0.205 | 0.070 | 0.804 | 0.343 | 0.175 |
| knots | 0.115 | -0.343 | -0.092 | -0.301 | 0.600 | -0.170 |
| diaknot | 0.113 | -0.309 | 0.326 | -0.303 | -0.080 | 0.626 |

Table 4: SPCA (Pitprops Data)

| Variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| topdiam | -0.477 | 0 | 0 | 0 | 0 | 0 |
| length | -0.476 | 0 | 0 | 0 | 0 | 0 |
| moist | 0 | 0.785 | 0 | 0 | 0 | 0 |
| testsg | 0 | 0.619 | 0 | 0 | 0 | 0 |
| ovensg | 0.177 | 0 | 0.641 | 0 | 0 | 0 |
| ringtop | 0 | 0 | 0.589 | 0 | 0 | 0 |
| ringbut | -0.250 | 0 | 0.492 | 0 | 0 | 0 |
| bowmax | -0.344 | -0.021 | 0 | 0 | 0 | 0 |
| bowdist | -0.416 | 0 | 0 | 0 | 0 | 0 |
| whorls | -0.400 | 0 | 0 | 0 | 0 | 0 |
| clear | 0 | 0 | 0 | -1 | 0 | 0 |
| knots | 0 | 0.013 | 0 | 0 | -1 | 0 |
| diaknot | 0 | 0 | -0.016 | 0 | 0 | 1 |

### 5.2.2 Teaching Data

Following Leng and Wang (2009), we consider the Teaching dataset which is about the teaching evaluation scores of 251 courses taught in the Guanghua School of Management, Peking University. Each observation corresponds to one course taught during the period from 2002 to 2004, and records the average scores on the students' agreement with the following nine statements: (Q1) I think this is a good course; (Q2) The course improves my knowledge; (Q3) The schedule is reasonable; (Q4) The course is difficult; (Q5) The course pace is too fast; (Q6) The course load is very heavy; (Q7) The text book is good; (Q8) The reference book is helpful; (Q9) Opening this course is necessary.

From table (5), it can be seen that both SPCA and GAS-PCA identified similar interpretable patterns. Specifically, PC2 in SPCA and GAS-SPCA suggests that the items Q4, Q5, and Q6 are related to each other. It can be seen from the description of variables that those three items are the only questions asked in a negative manner. For example, Q4 asks students whether they agree that the course is difficult instead of asking students if they think the course is easy. Similar comments are also applicable to Q5 and Q6. We also notice that Q7 and Q8 are the only two items having negative loadings in PC3. Furthermore, the loading patterns of other items in PC3 are similar to those of PC1 for SPCA and GAS-PCA. This observation suggest that Q1, Q2, Q3, and Q9 are likely to be accounted for by one common hidden factor whereas Q7 and Q8 are due to some other factors. From the description of variables we find that Q7 and Q8 are the only two items about teaching materials, whereas all other items are general questions about the course itself. In conclusion, we observe that both SPCA and GAS-SPCA perform well on the teaching data and give similar interpretable results.

Table 5: Comparison of methods on Teaching Data

| | PCA | | | SPCA | | | GAS-SPCA | | |
|---|---|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| Q 1 | -0.390 | -0.221 | 0.247 | 0.487 | 0 | 0.323 | 0.483 | 0 | 0.320 |
| Q 2 | -0.383 | -0.136 | 0.294 | 0.346 | 0 | 0.338 | 0.376 | 0 | 0.331 |
| Q 3 | -0.339 | -0.221 | 0.301 | 0.347 | 0 | 0.308 | 0.328 | 0 | 0.224 |
| Q 4 | -0.283 | 0.493 | -0.099 | 0 | 0.619 | 0 | 0.110 | 0.643 | 0 |
| Q 5 | -0.268 | 0.503 | -0.053 | 0 | 0.559 | 0 | 0 | 0.515 | 0 |
| Q 6 | -0.258 | 0.515 | -0.012 | 0 | 0.552 | 0 | 0 | 0.567 | 0 |
| Q 7 | -0.322 | -0.242 | -0.686 | 0.502 | 0 | -0.636 | 0.458 | 0 | -0.658 |
| Q 8 | -0.330 | -0.229 | -0.455 | 0.399 | 0 | -0.430 | 0.394 | 0 | -0.468 |
| Q 9 | -0.394 | -0.103 | 0.269 | 0.333 | 0 | 0.311 | 0.375 | 0 | 0.291 |

# 6 Conclusion

All the discussions regarding improvements over PCA have given us clear understanding that, SPCA and GAS-PCA facilitates us to interpret the PCs more efficiently as they include only the relevant variable for the construction of PCs by providing sparse loadings. Also, they appear to be more useful and computationally efficient for high-dimensional cases. In this report, we have also discussed how GAS-PCA can be obtained from SPCA by using adaptive lasso penalty instead on lasso penalty. In the cases when the variability of the hidden factor is low, GAS-PCA performed more accurately than SPCA. Hence, this report presented a comparative study between general PCA and PCA after

inclusion of sparsity, as well as between two different methods of obtaining sparse loadings. For more extensive theoretical developments of SPCA and its applications, we refer the interested readers to Guerra-Urzola et al. (2021), Luss and d'Aspremont (2010), Cai et al. (2013), Hubert et al. (2016).

# 7 Supplementary Material

The interested reader is directed to https://github.com/ArkaB-DS/SPCA, which contains all the R codes and datasets used in this report.

# 8 Acknowledgements

We take this opportunity to heartily thank our supervisor Prof. Minerva Mukhopadhyay for her valuable feedback and constant guidance on this project. Her comments and suggestions during the presentation were enriching and helped us improve our project.

# References

Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

Guerra-Urzola, R., Van Deun, K., Vera, J. C., and Sijtsma, K. (2021). A guide for sparse pca: Model comparison and applications. *psychometrika*, 86(4):893–919.

Hubert, M., Reynkens, T., Schmitt, E., and Verdonck, T. (2016). Sparse pca for high-dimensional data with outliers. *Technometrics*, 58(4):424–434.

Jeffers, J. N. (1967). Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 16(3):225–236.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag, New York.

Jolliffe, I. (2022). A 50-year personal journey through time with principal component analysis. *Journal of Multivariate Analysis*, 188:104820.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Leng, C. and Wang, H. (2009). On general adaptive sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 18(1):201–215.

Luss, R. and d'Aspremont, A. (2010). Clustering and feature selection using sparse principal component analysis. *Optimization and Engineering*, 11(1):145–157.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Todorov, H., Fournier, D., and Gerber, S. (2018). Principal components analysis: theory and application to gene expression data analysis. *Genom. Comput. Biol*, 4(2).

Tzeng, D.-Y. and Berns, R. S. (2005). A review of principal component analysis and its applications to color technology. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(2):84–98.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Yoo, C. and Shahlaei, M. (2018). The applications of pca in qsar studies: A case study on ccr5 antagonists. *Chemical biology & drug design*, 91(1):137–152.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Zou, H. and Hastie, T. J. (2003). Regression shrinkage and selection via the elastic net , with applications to microarrays.