SA Youth Labour Market Participation: Applying sPCA as pre-step before modelling

Siphiwe Bogatsu

Dept. of Statistical Sciences

March 22, 2024

Overview

Background

Let the data $\boldsymbol{X}$ be a matrix $n \times$p, where n and p are number of observations and the number of variables, respectively.

- Task: Projection of covariance matrix of $\boldsymbol{X}$ from $\mathbb{R}^p$ to $\mathbb{R}^q$, $q \leq p$
- PCA is a sequence of projections onto a linear manifold in $\mathbb{R}^q$
- The projections to $\mathbb{R}^q$ should be orthogonal, and explain the maximum amount of data variance as in $\mathbb{R}^p$.

There are two equivalent representations of the problem:

- Minimize the reconstruction error for a centered data matrix (least squares problem) $\rightarrow$ singluar value decomposition (SVD)
- Maximize the variance explained over the linear combination for a given covariance matrix (eigenvalue problem) $\rightarrow$ eigenstructure

Motivation for sparse PCA

Each PC uses ALL the predictors regardless if a predictor is a noise - meaning no true effect in the PC; ALL predictor loadings are non-zero (Zou et al, 2006)

- Result: PCs generated likely contaminated with noise such that the resulting PCs may deviate tremendously from the original data!
- Problematic if p is very large

Sparse PCA merges the benefits of traditional PCA, namely data reduction, with sparsity modeling, which removes ineffective variables from the PCA model by reducing their loadings to zero.

SA Youth Labour Participation

Consider this application:

- First quarter of 2022 the unemployment rate was 63,9% for those aged 15-24 and 42,1% for those aged 25-34 years (StatSA,2024)
- Building statistical models to predict employment outcomes for young people promotes evidence-based decision-making, supports economic empowerment, and fosters career success.
- PROBLEM: SA youth survey datasets present are often of high dimensional in nature.

SA Youth Labour Participation

2022-2023 youth survey data of 4020 South Africans is used for demonstration. Demographic information includes gender (female: 2269, male: 1751);their mean age of 1997. Geographical information include geography (urban:2797, rural: 803, suburb:420), and educational attainments.

Variable	Description	Data Type
Round	The four rounds of 6-month interval survey.	numeric
Status	Labour market status of individual	character
Tenure		numeric
Geography	rural, suburb, or urban ?	character
Province	provincial residence of youth	character
Matric	Whether or not the youth has a matric certificate.	numeric
Degree	Whether or not the youth has a degree	numeric
Diploma	Whether or not the youth has a diploma	numeric
School Quintile	South Africa's education system is divided into five quintiles based on the socio-economic status of the area surrounding the school	numeric
Math	Mathematics results in matric level mark	numeric
Mathlit	Mathematical Literacy in matric level mark	numeric
Additional_lang	Additional language in matric level mark	numeric
Home_lang	Home language in matric level mark	numeric
Science	Physical Science or Life Sciences in matric level mark	numeric
Female	Gender of youth	numeric
Age	Age of youth	numeric

Formulations of sPCA problem

Aims to project data points into $\mathbb{R}^q$ ($q \leq p$) with a goal to preserve variation of original sample points as much as possible.

Theorem (Variance Maximisation Approach (VM))

$$\max_{V_1}(V_1' X' X V_1) + \lambda_1 \| V_1 \|_1 \textit{ subject to } V_1' V_1 = 1$$

where $\| V_1 \|_1 = \sum_{i=1}^{p} |V_{i1}|$. The λ_j, $j = 1, 2, \cdots, k$, is a penalty parameter that controls the shrinkage amount on each PC.
V_1 is the first loading coefficient vector, such that the corresponding PC is given by $\boldsymbol{X} V_1$
The higher the value of λ_j, the greater the amount of shrinkage to zero. To implement an VM algorithm an R package, *pcaPP*, is available (Croux et al , 2013).

Formulations of sPCA problem

Zou, Hastie & Tibshirani (2006) reconstruct the product of the loading coefficient matrix, $V^T V$, into two matrices (A, B) - both being $p \times k$ matrices - and then add an L2-norm penalty on B. To impose sparseness, an L1-norm penalty on B is added to obtain sparse loadings.

Theorem (Reconstruction Error Minimisation Approach)

$$\min_{A,B} \sum_{i=1}^{n} ||x_i - AB'x_i||^2 + \lambda \sum_{j=1}^{k} ||B_j||^2 + \sum_{j=1}^{k} \lambda_j ||B_j||_1 \ \textit{subject to}\ A'A = I_k$$

where $||B_j||^2 = \sum_{t=1}^{p} (\sqrt{B_{tj}})^2$ and λ is the penalty parameter. The j^{th} loading is $V_j = \frac{B_j}{||B_j||}$, $j = 1, 2, \cdots, k$.

Note that there is a common λ used for all PCs, but different λ_j's are used for penalizing the loadings of different PCs. A R package, *elasticnet*, is available to perform the REM (Zou, Hastie & Tibshirani, 2006).

Formulations of sPCA problem

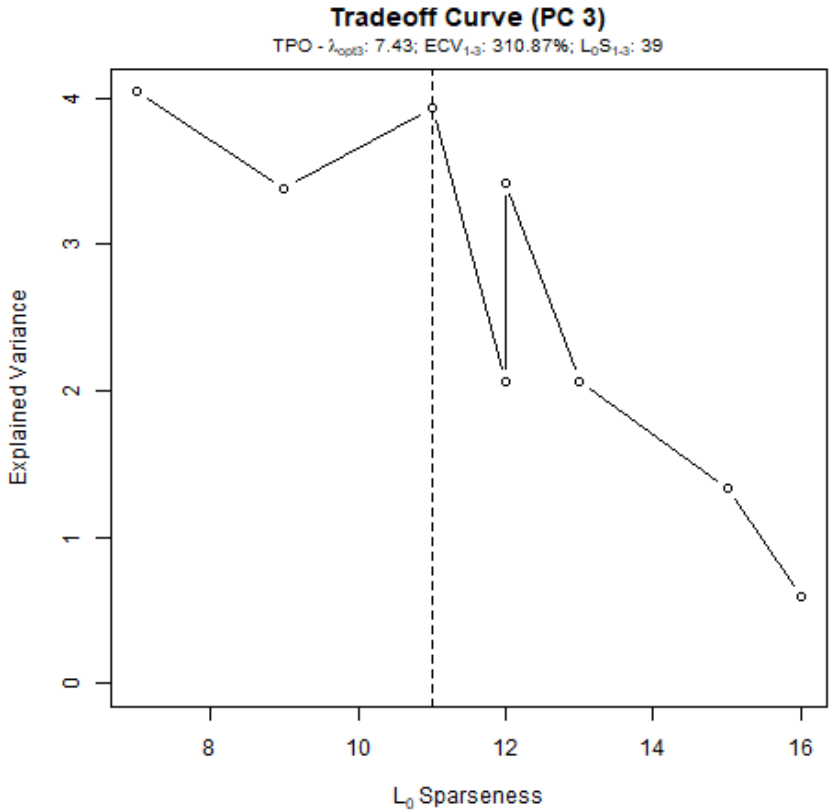
Some other solutions to our issue could be these:

- Sorting simply sorts the diagonal of the covariance matrix and ranks the variables by variance
- Thresholding computes the leading eigenvectors and form sparse vector by thresholding all coefficients below a certain level

I did not consider them in this study.

Illustration: SA Youth Labour Participation

- Use *opt.TPO()* from *pcaPP* package to find the sparseness parameters λ_j's that maximise the trade off curve for a particular component.



- $\lambda_j = (18.27, 3.31, 7.43, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$ and $\lambda = 0$, chosen such that the sparse loadings explain the maximum possible variability as much as an ordinary PC did under PCA.

Illustration: SA Youth Labour Participation

- The loadings and variance information from the PCA constructed using the SVD approach.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Round	-0.03	0.08	0.14	-0.49	-0.24	0.33	-0.20	-0.39	-0.24	-0.12
Status	-0.04	0.09	0.18	-0.53	0.08	0.23	0.09	-0.19	0.44	0.31
Tenure	0.05	0.11	-0.17	-0.09	0.07	-0.09	-0.32	0.22	0.45	0.33
Geography	0.09	0.06	0.30	-0.14	-0.30	-0.51	0.18	0.16	0.13	-0.05
Province	0.07	0.11	0.29	-0.23	-0.39	-0.31	0.19	0.19	-0.13	0.05
Matric	-0.30	-0.45	0.32	-0.06	0.18	-0.04	-0.13	0.09	0.07	0.04
Degree	-0.19	0.10	0.16	-0.10	0.12	-0.04	-0.15	-0.06	0.11	-0.76
Diploma	-0.06	0.02	0.08	-0.30	0.31	-0.04	-0.18	0.34	-0.66	0.25
School quintile	0.13	0.12	0.42	0.13	0.04	0.05	-0.15	-0.04	0.11	0.07
Math	-0.55	0.25	0.13	0.13	0.02	-0.02	0.02	0.02	0.01	0.06
Math lit	0.23	-0.56	0.19	-0.07	0.19	-0.05	-0.02	0.03	0.10	-0.09
Additional_language	-0.42	-0.46	-0.14	-0.14	-0.10	-0.14	-0.02	0.02	0.08	0.01
Home_language	0.16	0.13	0.56	0.29	0.28	0.13	-0.11	0.05	0.02	0.04
Science	-0.52	0.21	0.11	0.14	0.02	-0.03	0.04	0.01	0.04	0.13
Female	-0.01	0.05	-0.01	-0.14	0.49	-0.05	0.74	-0.15	-0.01	-0.01
Birth Month	-0.03	-0.03	0.03	-0.05	-0.20	0.60	0.27	0.68	0.11	-0.18
Age	0.06	0.29	-0.21	-0.32	0.38	-0.24	-0.21	0.29	0.14	-0.25
Variance (%)	14.67	11.10	10.67	7.76	7.07	6.61	6.13	5.64	5.49	4.96
Cumulative Variance (%)	14.67	25.78	36.44	44.21	51.28	57.90	64.03	69.67	75.16	80.12

Table 2: SA youth data: Loadings of the first 10 Principal Components by PCA

Illustration: SA Youth Labour Participation

- VM approach

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Round		0.26	0.38		0.48	-0.21				-0.22
Status		0.27	0.44	0.27	0.21	0.13				
Tenure	-0.16	-0.14	0.20		-0.13	-0.26				0.33
Geography									1.00	
Province	0.11	0.16	0.36	-0.28	0.13	0.13				-0.36
Matric	0.37	0.35	-0.32	0.37		-0.24				
Degree		0.37			-0.12	-0.13				
Diploma		0.18	0.15	0.37	-0.11	-0.27				
School quintile	0.29	0.19	0.26	-0.37	-0.25					0.15
Math	-0.29	0.59	-0.34		-0.11					-
Math lit	0.60	-0.26		0.32		-0.11				
Additional_language								1.00		
Home_language	0.42	0.25	0.13	-0.27	-0.38					0.22
Science							1.00			
Female				0.45	-0.26	0.73				-0.19
Birth Month					0.48	0.34				0.72
Age	-0.34		0.40	0.23	-0.38	-0.16				0.26
Variance (%)	10.50	9.70	7.76	7.2	6.89	6.06	6.06	6.06	6.06	5.79
Cumulative Variance (%)	10.50	20.19	27.96	35.15	42.04	48.10	54.16	60.22	66.27	72.06

Table 3: Table 2. SA Youth Data: Loadings of the First Ten Modified PCs by VM approach. Empty cells have zero loadings.

Illustration: SA Youth Labour Participation

- REM approach

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Round		0.04	0.03			0.07	0.06	-0.12	-0.60	-0.68
Status			0.02	-1.00				0.03	-0.03	-0.07
Tenure	0.03	0.10	-0.15			-0.13	-0.21	0.22	0.33	-0.15
Geography	0.04	0.05	0.27			-0.54	0.37	0.15	0.15	
Province	0.04	0.10	0.24			-0.39	0.47	0.21	-0.13	-0.05
Matric	-0.30	-0.47	0.31			-0.03	-0.18	0.13		-0.04
Degree	-0.21	0.07	0.15			-0.07	-0.23		0.13	-0.51
Diploma			0.06			-0.11	-0.29	0.45	-0.65	0.39
School quintile	0.10	0.12	0.44			0.05	-0.08	-0.07	0.06	-0.05
Math	-0.57	0.24	0.12			0.05		-0.02	0.02	0.07
Math Lit	0.23	-0.57	0.21			-0.06	-0.15	0.06	0.10	-0.08
Additional_language	-0.41	-0.48	-0.17			-0.15	0.07	0.05	0.03	-0.05
Home_language	0.13	0.14	0.63			0.22	-0.25		0.08	0.08
Science	-0.54	0.20	0.11			0.04	0.02	-0.03	0.04	0.10
Female					1.00					
Birth Month		-0.02				0.59	0.36	0.68	0.13	-0.15
Age	0.05	0.26	-0.19			-0.31	-0.45	0.42	0.13	-0.19
Variance (%)	14.6	11.1	10.5	6.0	6.0	6.8	6.7	5.9	5.9	5.4
Cumulative Variance (%)	14.6	25.7	36.2	42.2	48.2	55	61.7	67.6	73.5	78.9

Table 4: Table 2. SA Youth Data: Loadings of the First Ten Modified PCs by REM approach. Empty cells have zero loadings.

Illustration: SA Youth Labour Participation

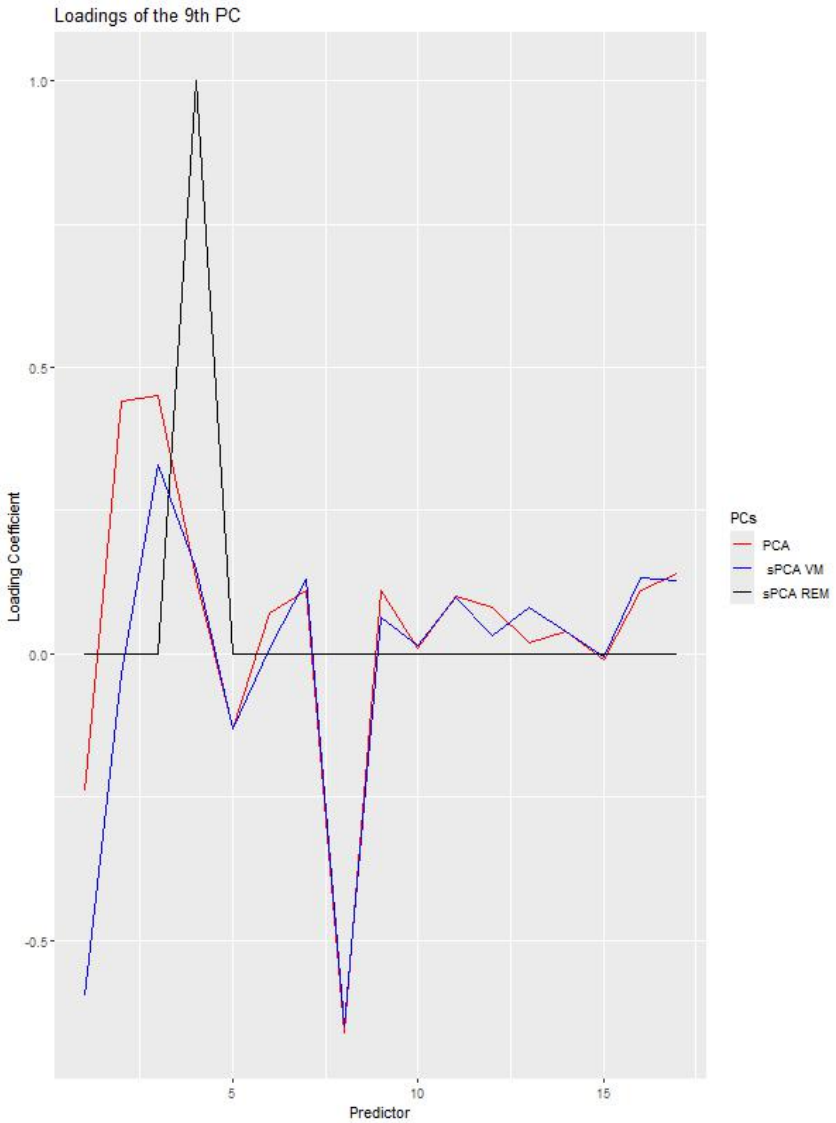


Illustration: SA Youth Labour Participation

- the REM approach (78%) is not far off from the PCA (80%) in terms of variability explained.
- Loading coefficients approaching zero in the traditional PCA are likely to be further reduced to zero by sparse PCAs.
- PCA has *all* loadings non-zero; REM shrinks at total of 40 loadings to zero in the entire structure; and the VM achieves a highest number of loadings shrunk to zero at 99
- *Matric*, *Languages*, *Maths*, *Tenure*, *Province*, *Schoolquintile* are observed to be the important variables in the construction of the first few (sparse) principal components that atleast explain 50% variability in the structure of the SA youth data.

Major Takeaway

As a pre-step before applying statistical modelling techniques, sparse PCA is able to select only the effective variables to construct PCs that can be used to predict the target variable, without the need for us to incorporate variable selection penalty in our models. Thus, statistical modelling techniques that do not consider variable selection can leverage on this sparse PCA's strength to improve their robustness.

References

- Croux, C., Filzmoser, P. and Fritz, H., 2013. Robust sparse principal component analysis. Technometrics, 55(2), pp.202-214.
- Hsu, Y.L., Huang, P.Y. and Chen, D.T., 2014. Sparse principal component analysis in cancer research. Translational cancer research, 3(3), p.182.
- Zou, H., Hastie, T. and Tibshirani, R., 2006. Sparse principal component analysis. Journal of computational and graphical statistics, 15(2), pp.265-286.

The End