Effect of measurement error on sparse principal component analysis

Donghyeok Lee

Tilburg university

Date of submission : 19-06-2022

Email : ldh9509@naver.com

Abstract

Although PCA is one of the widely used methods for dimension reduction, it has two main drawbacks: difficult interpretation and inconsistency in eigen values and eigen vectors in a high dimension low sample size (HDLSS) case. To ease these problems, various sparse PCA methods are proposed. However, there has been little attention on how measurement error affects the performance of sparse PCA. This paper empirically shows the problems of the ordinary PCA with the simulation results in a HDLSS case. In addition, how measurement error affects the performance of sparse PCA is addressed with regard to the four criteria: the expected first eigen vector, the expected PEV (proportion of explained variance), the classification rate and the standard error of the loadings.

Key words : Inconsistency, HDLSS, Sparse loading, Measurement error, sPCA-rSVD

1. Introduction

Principal component analysis (PCA) is one of the widely used methods in the analysis of high dimensional setting, However, one would have a problem of interpretation of each variable in a dataset since each variable is involved in every PCA component; in other words, each PCA component is constructed as a linear combination of all variables. Plus, in a situation where the number of variables is more than observations, another problem arises; estimators from the ordinary PCA are not consistent[Iain M, 1]. In this situation, sparse PCA is frequently used to tackle the two mentioned drawbacks such that each variable is associated with only a few PCA components by rotation technique or low-rank approximation along with lasso penalty term (sPCA-rSVD).

Although the inconsistency and the interpretation problems can be alleviated by sparse PCA, another concern still remains : measurement error. Estimators from the ordinary PCA are often derived from the eigen values and the eigen vectors of a covariance matrix of variables (or

correlation matrix if variables are normalized, which is often the case). By this nature of this PCA calculation, estimators from PCA can be affected by additive Gaussian measurement error since it increases the variance of variables, which as a result dilutes the correlation between variables. The article of Kristoffer Hellton[2] proved that (additive Gaussian) homogeneous measurement error does not cause a bias in loadings, whereas (additive Gaussian) heterogeneous measurement error does.

However, there has been little attention on how measurement error affects the performance of sparse PCA. Therefore, this paper examines how measurement error affects the performance of sparse PCA via the simulation results regarding the four criteria: the expected first eigen vector, the expected PEV, the classification rate and the standard error of the loadings. There have been multiple sparse PCA methods with different mathematical formulations, therefore, the obtained results from each sparse PCA method are significantly different from each other. In case of sparse PCA methods for sparse loadings, these two methodologies are widely used : 'rotation and thresholding' and 'sparse PCA via regularized SVD (sPCA-rSVD)'[Shen and Huang , 3]. It is empirically shown that sPCA-rSVD has the best performance according to the four criteria[Guerra Urzola et al ,4]: the squared relative error (SRE) of the model parameters, the misidentification rate (MR) of zero versus the nonzero status of the sparse coefficients, the proportion of explained variance (PEV), and the cosine similarity. Thus, in this paper sPCA-rSVD is selected in order to investigate the effect of measurement error on sparse PCA.

This paper is organized as follows. Section 2 presents the inconsistency of PCA along with the performance comparison between PCA and sPCA-rSVD in a HDLSS setting. In Section 3, how various measurement error structures affect the performance of sPCA-rSVD is addressed with the simulation results based on the synthetic data which have sparse component loadings. The last chapter, Section 4, includes the conclusion and the limitations of this study. Note that the purpose of this paper is to deliver the conceptual intuitions via the visualizations and the simulation results to potential users of sparse PCA across diverse industries, rather than showing a formal mathematical proof.

2. Drawbacks of PCA in a high dimension low sample size case

As mentioned in Section 1, estimators from the ordinary PCA are not consistent when a sample size is smaller than the number of variables, which is frequently denoted as p >> n case. Section 2.1 and 2.2 firstly demonstrate the inconsistency of PCA with regard to eigen values and eigen vectors via the visualizations and the simulation results. In addition, the comparison of the performance between PCA and sPCA-rSVD in a HDLSS situation (high dimension low sample size) will be shown in Section 2.3.

All the results are based on the synthetic and normally distributed data without or with sparse component loadings, and they are generated from R software. In particular, the population data has the size of 10000 and the samples are n times randomly selected from the population. Through repeating this sampling procedure a 1000 times, the expected estimators and their standard errors are obtained.

2.1 Inconsistency in eigen values

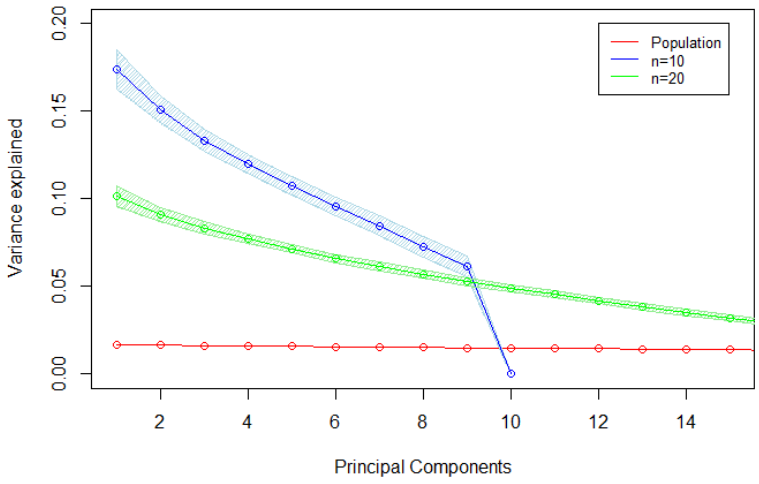


Figure 1 : The points represent the expected PEV (proportion of explained variance) corresponding to each k-th principal component based on the multivariate normal data where n =sample size and p=100. The shaded areas indicate the standard error of the PEV.

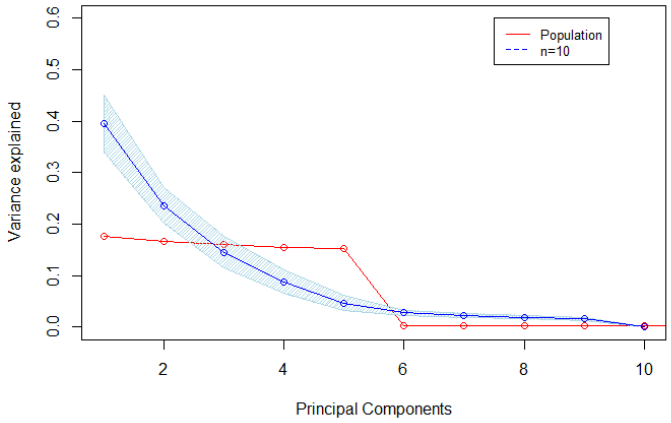


Figure 2 : The expected PEV based on the sparse component loadings structure data where sparsity=30%, p=100 and VAF (variance accounted)=80%.

Figure 1 and Figure 2 show that the eigen values of the first k-th principal components have the significantly large upward bias in both situations: non-structural data (random noise) and the sparse component loadings structure data. The plots also show that the magnitude of the upward bias of the expected PEV and their standard errors are proportional to the decrease of the sample size. It is because in the p >> n case, there exists the n-1 number of eigen values at most in the sample, while the population has the p numer of eigen values. As a result, the first k-th components from the small sample must explain the extra variance that originally belongs to the n-th to p-th components from the population.

2.2 Inconsistency in eigen vectors

The inconsistency in eigen vectors of PCA in a HDLSS case can be shown by a large angle (degree) between the true and an estimated eigen vector. This is accomplished by applying PCA on the datasets which have the same data structure with the different number of variables, while the sample size is fixed. In Algorithm 1, the single component spike covariance datasets with the different number of variables ranging from 10 to 1000 are created, and the angles between the true and the estimated eigen vector are compared, as shown in Figure 3.

Note that the sign of the eigen vector is arbitrary because the data X is expressed as $X = T * P^T$ where T and P are the score and the loading matrix, respectively. It is necessary to define the sign of the eigen vector before calculating the angle, therefore, the sign is defined such that the cosine similarity between $\widehat{v_1}$ (=an estimated eigen vector) and v_1 (=the true eigen vector) is maximized as illustrated in the 7th step in Algorithm 1.

1. Generate X ~ N(0, I_p)

   where X is the matrix of size 10000 x p and I_p is the p x p identity matrix.

2. Obtain U, D, V via SVD(Singular value decomposition) on X.

3. Replace the diagonal elements of D (=eigen values) as follows;

   $\mathrm{d}_1 = \lambda_1$

   $\mathrm{d}_2, \ldots, \mathrm{d}_\mathrm{k} = 1$

4. X ~ $U * D * V^T$

5. Create the sample from randomly selecting the 10 rows of X.

6. Obtain $\widehat{v_1}$ (=the first sample eigen vector) via SVD on the sample.

7. Determine the sign of $\widehat{v_1}$ via the cosine similarity between $\widehat{v_1}$ and v_1 (=the true first eigen vector) as follows;

   Cosine_original = cosine similarity between between $\widehat{v_1}$ and v_1.

   Cosine_flipped = cosine similarity between between -$\widehat{v_1}$ and v_1.

   If Cosine_original < Cosine_flipped,

   $\widehat{v_1}$ = -$\widehat{v_1}$

   Else,

   $\widehat{v_1}$ = $\widehat{v_1}$

8. Calculate the angle between $\widehat{v_1}$ and v_1 via the cosine similarity.

Algorithm 1 : The angle from the single component spike covariance data

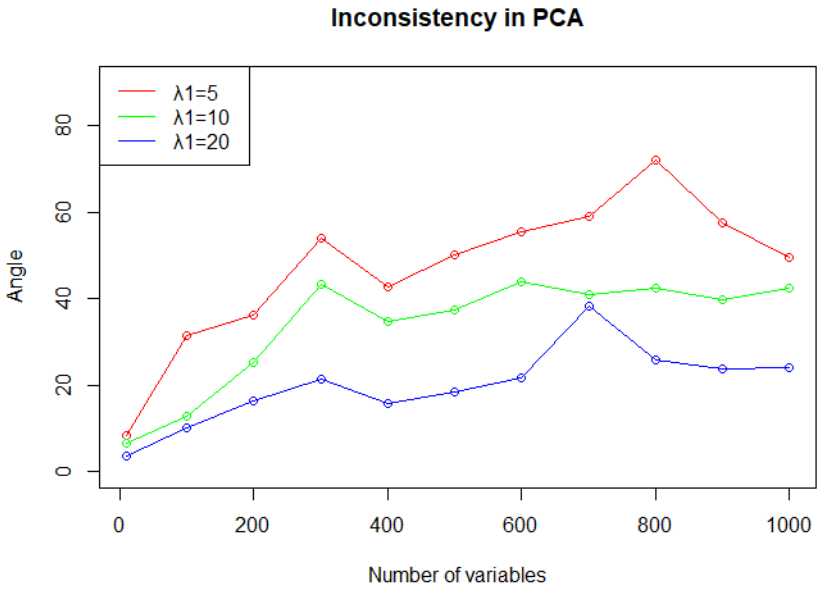


Figure 3 : Angle between the truth and the first eigen vector when the sample size is fixed as 10. It is based on the single component spike covariance data with the different first leading eigen value λ_1, while the remaining eigen values are set to be 1(i.e., $\lambda_2, \ldots, \lambda_k = 1$).

There are two factors to determine the consistency of PCA[Dan Shen et al, 5] : the sample index and the spike index. Note that sparsity of loadings is not related to the consistency of PCA. Each of them is defined as follows;

Definitions

γ (sample index) : n ~ d^{γ} where d is the dimension – the number of variables

α (spike index) : α ~ d^{α}

If γ or α becomes smaller, sample eigen vectors tend to be more different from the truth[Dan Shen et al, 6]. As shown in Figure 3, although there are some fluctuations (such as the data point when x=700 in blue line) occurred by random sampling, the angle tends to become larger as λ_1 decreases or the number of variables increases.

2.3 PCA vs sPCA-rSVD

The simulation results in Section 2.3 are based on the sparse component loadings structure data where sparsity=40%, p=10, n=5, VAF=100%, and the first two true eigen vectors : v1=(0.422, 0.422, 0.422, 0.422, 0, 0, 0, 0, 0.380, 0.380) and v2=(0, 0, 0, 0, 0.489, 0.489, 0.489, 0.489, -0.147, 0.147). The true PEV of the first and the second principal component are around 0.67 and 0.29, respectively. Plus, the population data has the size of 10000 and the samples are n times randomly selected from the population. This procedure is repeated a 1000 times, and the expected estimators along with the standard error of the estimators are obtained. The detailed procedure of data generation is illustrated in Algorithm 2.

1. Generate X ~ N(0, I_p)
2. Obtain U, D, V via SVD(Singular value decomposition) on X.
3. Replace the first two eigen vectors of V by two orthogonal vectors with sparsity as follows;

   v1 = (0.422, 0.422, 0.422, 0.422, 0, 0, 0, 0, 0.380, 0.380)

   v2 = (0 , 0, 0, 0, 0.489, 0.489, 0.489, 0.489, -0.147, 0.147)
4. Replace the diagonal elements of D such that the PEV of the first and second principal component are around 0.67 and 0.29.
5. Obtain V_{new} via QR decomposition on V to have the orthogonal eigen vectors.
6. Obtain cov(X) (=the covariance matrix) via $V_{new} * D * V_{new}^T$
7. X ~ N(0, cov(X))

Algorithm 2 : Sparse component loadings data generation

2.3.1 PCA

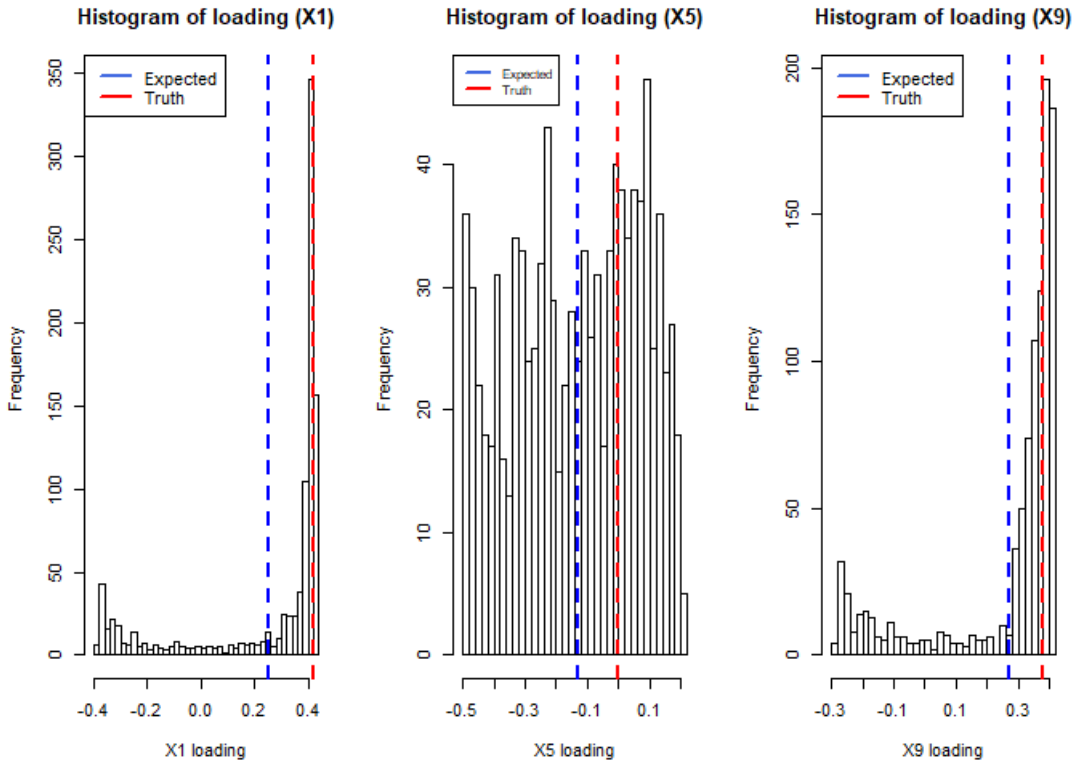


Figure 4 : The estimated loadings of each variable in the first component based on the data with sparsity=40%, p=10, and n=5.

In the p >> n case, not only the expected loadings are different from the truth, but also they have a high variance. Due to the high variance, 'flipped sign problem' or 'non-zero problem' might occur, which refers to the problem that the variable of which the true loading is large (such as X1 and X9) might have the sign that is opposite to its true sign or the variable of which the true loading is zero might have a significantly large loading. In Figure 4, the histogram of X1 and X9 loading shows the flipped sign problem, while the non-zero problem is observed in the histogram of X5 loading.

2.3.2 sPCA-rSVD

According to sPCA-rSVD proposed by Heipeng Shen and Jianhua[7], the regularization penalty is introduced in order to promote sparsity in PC loadings, and three different penalty functions are suggested, namely soft thresholding, hard thresholding and SCAD(smoothly clipped absolute deviation) penalty. From the simulation results of their paper, the performance of each penalty function - the cosine similarity between the true loading vector and the sample loading vector and the rate to correctly identify the true zero loadings - is not significantly different in various conditions (classical multivariate data and HDLSS data). Therefore, the penalty function with the least compuation cost, the hard thresholding, is selected in this paper to ease the computation intensity of the simulation experiment. In addition, this paper assumed that the true degree of sparsity is known when sPCA-rSVD is performed, therefore, an additional technique to decide the sparsity of loadings such as ad hoc approach or CV (K-fold cross validation) is not implemented in this paper.

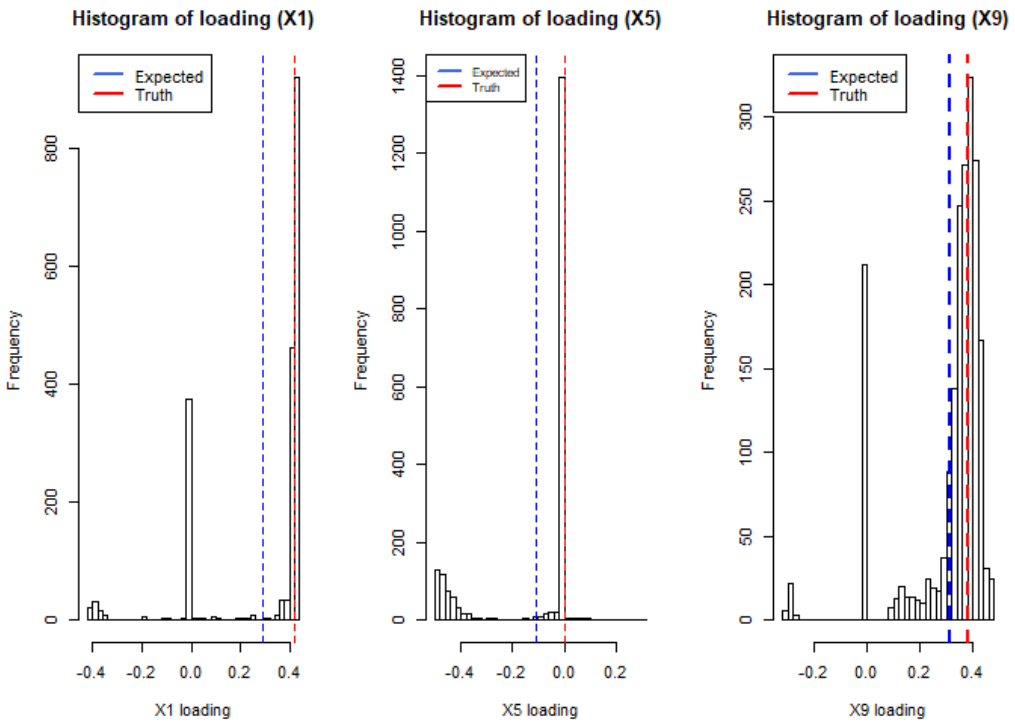


Figure 5 : The estimated loadings of each variable in the first component of sPCA-rSVD based on the data with sparsity=40%, p=10, and n=5.

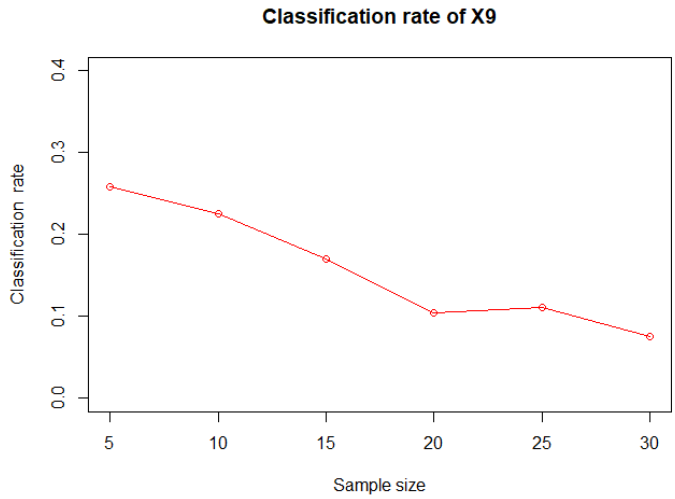


Figure 6 : The classification rate that sPCA-rSVD estimates the loading of X9 as zero.

Figure 5 shows the histogram of the estimated loadings of each variable. Although sPCA-rSVD also has the flipped sign and non-zero problem, they occur less than in PCA. In the case of sPCA-rSVD, the frequency of the estimated loading of X5 is much more concentrated in 0 than in PCA, which leads sPCA-rSVD to perform better when there is sparsity in loadings. However, the smaller a sample size is, the more likely the variable of which the true loading is not zero is identified as zero, as shown in Figure 6 (The true loading of X9 is 0.380).

2.3.3 Comparison between PCA and sPCA-rSVD

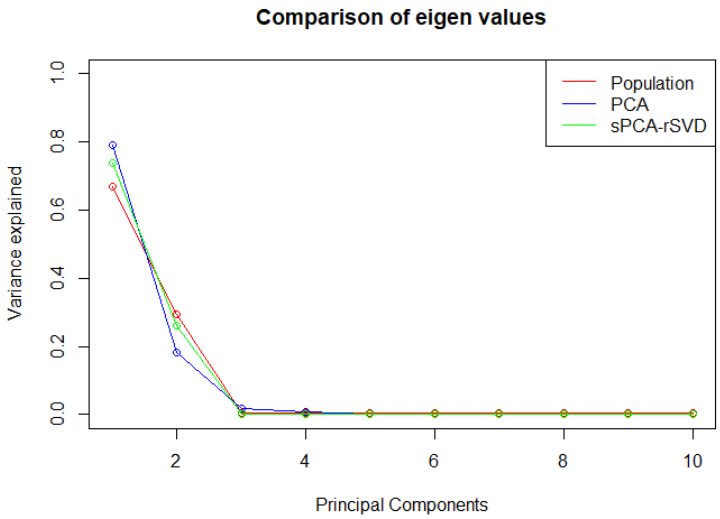


Figure 7 : Comparison of the expected PEV between PCA and sPCA-rSVD

Standard Error of PEV		
	The first PC	The second PC
PCA	0.11	0.11
sPCA-rSVD	0.12	0.12

Table 1 : Comparison of the standard erorr of PEV between PCA and sPCA-rSVD (rounded to 2 decimal places)

In Figure 7, the points indicate the expected PEV corresponding to the k-th principal component. It shows that the expected PEV from sPCA-rSVD is closer to the true PEV than the ordinary PCA, but there is no difference in the standard error of the PEV between them as shown in Table 1.

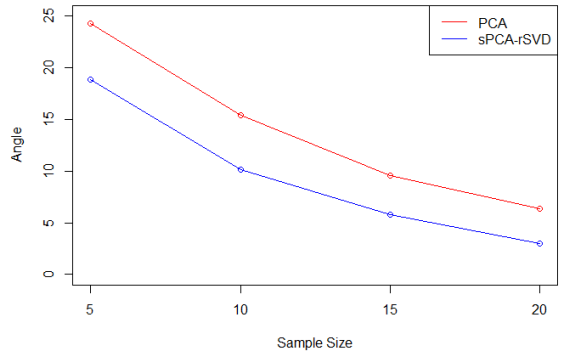


Figure 8 : Median angle between the true eigen vector and the sample eigen vector per each sample size

Angle between v and E[$\hat{v}$]		
	The first PC	The second PC
PCA	1.55	5.2
sPCA-rSVD	1.55	3.75

Table 2 : Comparison of the angle between the truth and the expected sample eigen vector when n=5

From repeating the sampling (n=5) a 1000 times, the angle between the true and the expected eigen vector is derived, along with the median angle of the 1000 number of the angles. As mentioned before in Section 2.2, if the first eigen values are large enough compared to the other remaining eigen values, the corresponding eigen vectors can be consistente even in the p >> n case. The PEV of the first principal component is around 0.67, which is large enough. Consequently, as shown in Table 2, the expected eigen vector in PCA is close to the truth. However, the obtained angles from PCA is larger than sPCA-rSVD in general, which results in the median angles in PCA to be bigger across all the different sample sizes, as shown in Figure 8.

When there is high sparsity in the true eigen vector and the non-zero elements of the true eigen vector have large absolute values — an element of the true eigen vector is either large or zero. — sPCA-rSVD estimates the eigen vector much more precise than PCA in a HDLSS situation. According to the simulation results from Heipeng Shen and Jianhua[8], the median angles, obtained from PCA and sPCA-rSVD when n=50, p=500 and sparsity=98%, are 19.69 and 1.21, respectively. In conclusion, sPCA-rSVD has the better performance than PCA with regard to estimating eigen values and eigen vectors when sparsity in loadings exists.

3. Effect of measurement error on sPCA-rSVD

In this section, how measurement error affects the performance of sPCA-rSVD is examined with regard to the four criteria: the expected first eigen vector, the expected PEV, the classification rate and the standard error of the loadings. The classification rate refers to the rate that sPCA-rSVD correctly identifies the variable of which the true loading is zero. The data with measurement error is generated as Algorithm 3. The population data has the size of 10000 and the samples are n times randomly selected from the population. This procedure is repeated 1000 times, and the expected estimators along with the standard errors are obtained. Note that every result is related to the first principal components of the samples. In addition, the variance of each variable without error is as follows; $Var(X_i)$ = (47.41 59.13 46.01 40.83 32.01 28.90 42.60 41.79 46.90 35.79) where i=1, ..., 10.

1. X ~ $TP^T + E$ where

   TP^T is from Algorithm 2

   The error matrix E ~ N(0, $d * I_p$)

2. If E is homogeneous error,

   d is a constant value to determine the variance of the error.

   If E is heterogeneous error, error is only in one variable.

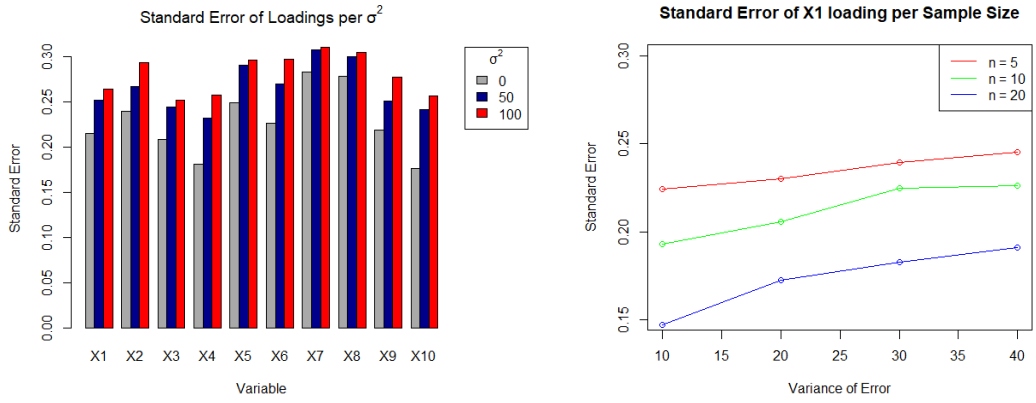
   d is the vector with the length p of which the elements are zero except the selected one.

Algorithm 3 : Data with measurement error

3.1 Homogeneous error

Sample size	Variance of error(σ^2)	Angle(degree)	Classification rate
n = 30	0	0.7	0.9
n = 5	0	1.55	0.63
	10	1.61	0.60
	20	1.47	0.58
	30	1.50	0.54

Table 3 : The angle (between the truth and the expected first eigen vector) and the average classification rate of the variables with the true zero loading per the different variance of the error (σ^2).

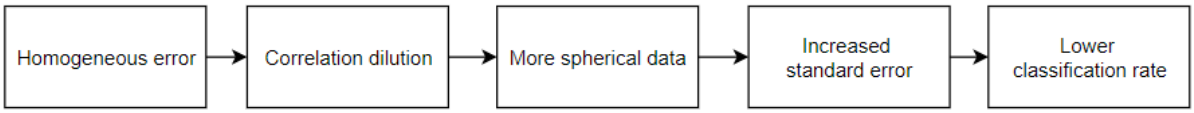


(a) Standard error of the loadings (n=5) (b) Standard error of X1 loading

Figure 9 : Effect of the variance of the error and the sample size on the standard error of the loadings

The homogeneous error equally increases the variance of each variable, in other words, the data structure remains the same. Thus, there is no change in the expected eigen vector, which is shown as no significant change in the angles in Table 3. Although the data structure remains the same, the correlation between the variables decrease, which makes the spiked covariance data more spherical. This results in the increase of the standard error of the loadings and the decrease of the average classification rate as illustrated in Table 3 and Figure 9. In detail, Figure 9b shows that the standard error of the loadings depend on the sample size and the variance of error. The homogeneous error proportionally increases the standard error of the loadings of all the variables, while the sample size is inversely proportional to the standard error.

The effect of the homogeneous error is summarized as follows;



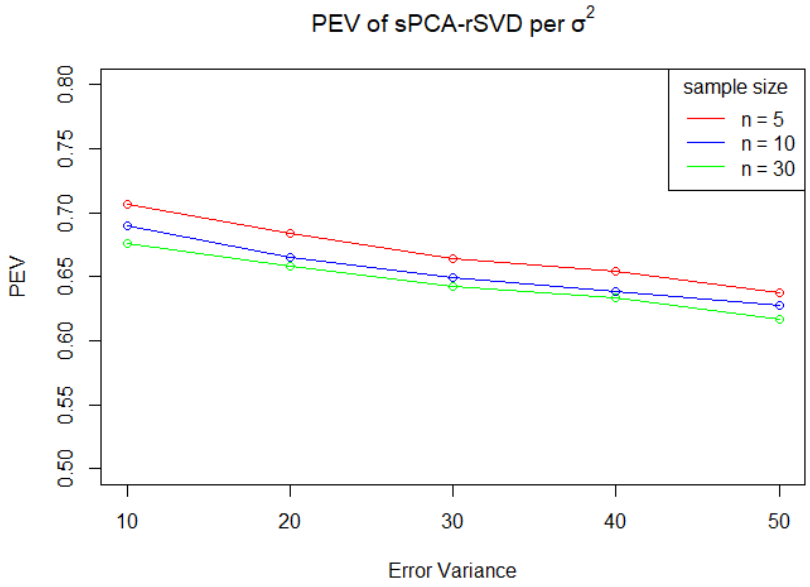


Figure 10 : The effect of the homogeneous error on the expected PEV of the first PC (principal component) in sPCA-rSVD per the different sample size.

According to the article of Kristoffer Hellton[9], additive Gaussian homogeneous measurement error equally increases the eigen values of each principal component of PCA by the variance of error, which can be denoted by $E(\nabla\lambda_i|X) = \sigma^2$ where i indicates the i-th PC. Using this fact, this paper defines the PEV with the homogeneous error as follows;

$$\text{PEV} = \frac{\lambda_1+\sigma^2}{\lambda_1+\cdots+\lambda_n+n*\sigma^2}$$

Therefore, the PEV of the first PC decreases by the homogeneous error. This phenomenon also holds for sPCA-rSVD. Figure 7 in Section 2.3.3 shows that sPCA-rSVD estimates the PEV better than PCA in the p >> n case, which means that the eigen values added to each components by the homogeneous error tend to be more equal to each other than PCA in the p>>n case (note that there exists a bias in the eigen values of PCA in a HDLSS situation.). Figure 10 illustrates that the PEV of the first PC decreases as the variance of error increases, regardless of the sample size.

3.2 Heterogeneous error

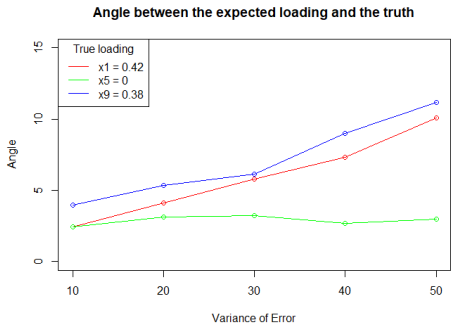
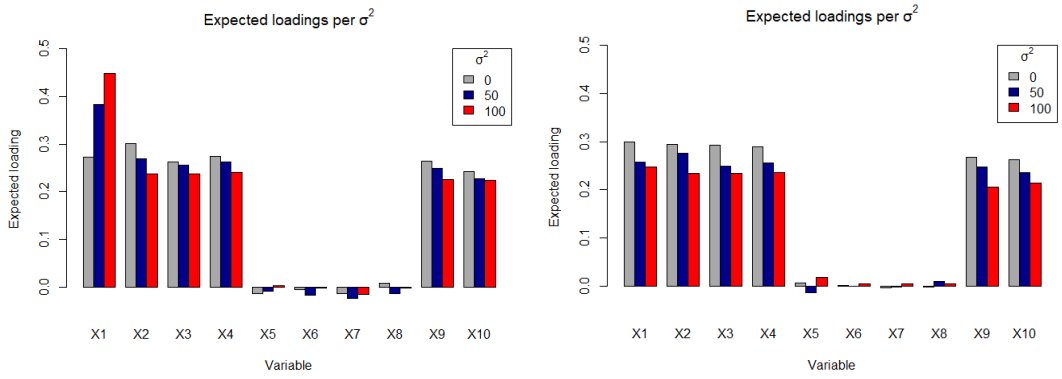
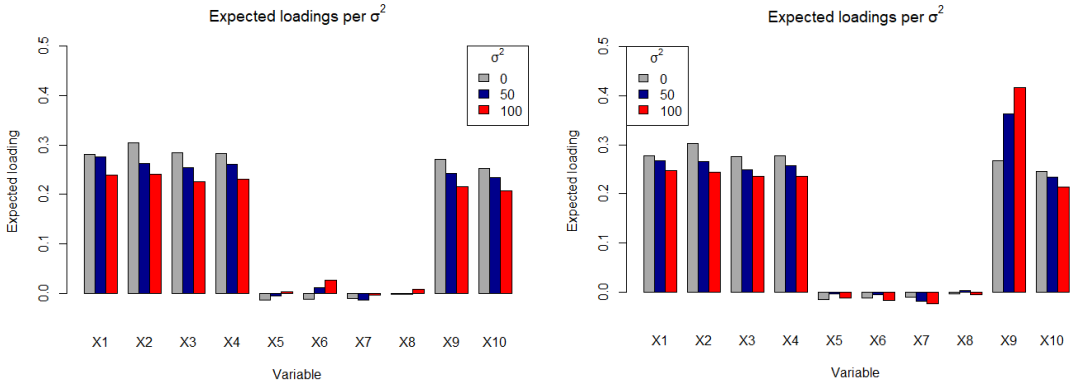


Figure 11 : The angle between the expected first eigen vector and the truth when the error is in one variable. Blue line : error in X9, Red line : error in X1 and Green line : error in X5. The true loading of X1, X5 and X9 are presented in the legend.



(a) : error in X1

(b) : error in X5

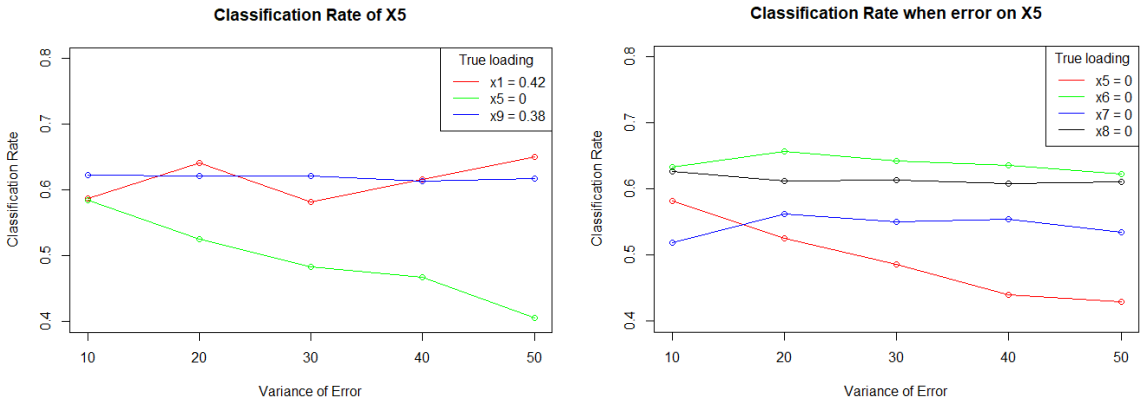


(c) : error in X6

(d) : error in X9

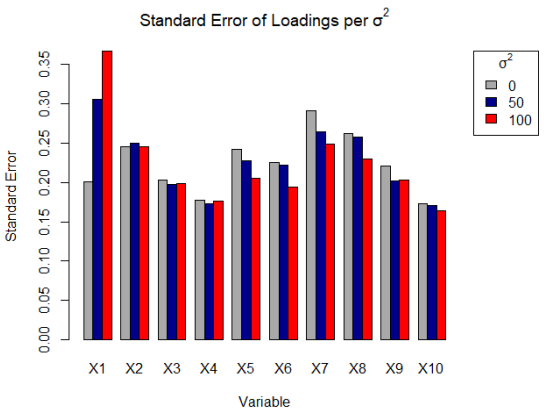
Figure 12 : The expected first eigen vector per the different variance of the error when the error is in one variable.

Unlike the homogeneous error, the heterogeneous error changes the data structure, which causes the change in the expected eigen vectors. Figure 11 and Figure 12 show the effect of the heterogeneous erorr on the expected first eigen vector. From Figure 11, the angle significantly increases when the error is in the variable of which the true loading is large (such as X1 and X9), while there is only a slight change in the angle when the error is in X5. The reason is that the expected loadings of the variables, of which the true loadings are zero, are rather insensitive to the error; these variables are not associated with the direction of the biggest variance in data, therefore, increasing the variance of these variables by the error does not significantly affect the first eigen vector. This insensitiveness is empirically shown in Figure 12; the variables with the true zero loading (X5 and X6) are insensitive to the error in both cases; the error in themselves or the error in the other variable. Additionally in Figure 12, if error is in a variable (denoted by X), the expected loading of X tends to be larger as the error increases the variance of X, while the expected loadings of other variables shrink as the direction of the biggest variance moves towards the X-axis.

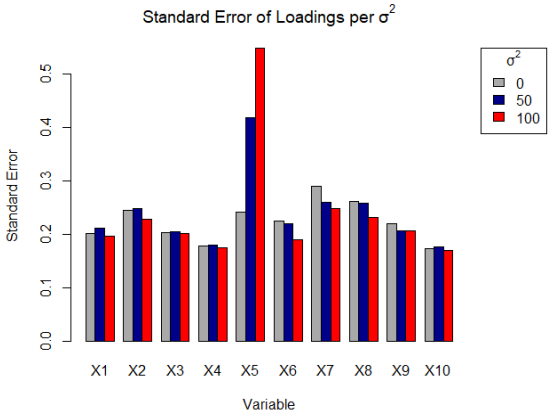


(a) error in one variable (b) error in X5

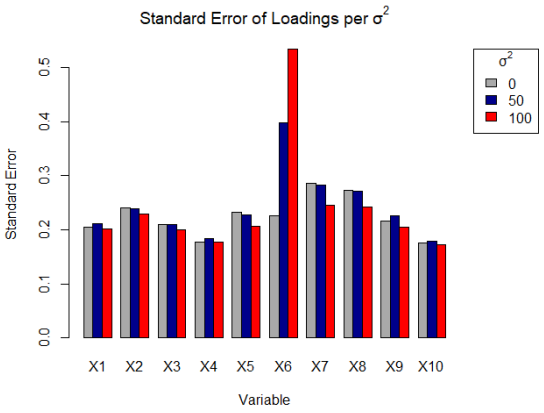
Figure 13 : The left plot shows the change of the classification rate of X5 when error is in one variable. The right plot shows the change of the classification rate of the variable with the true zero loading when error is in X5.



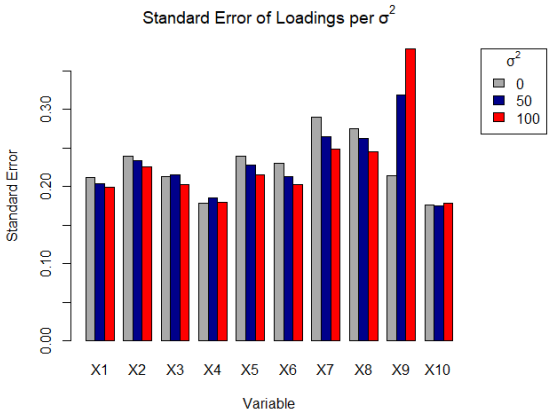
(a) : error in X1



(b) : error in X5



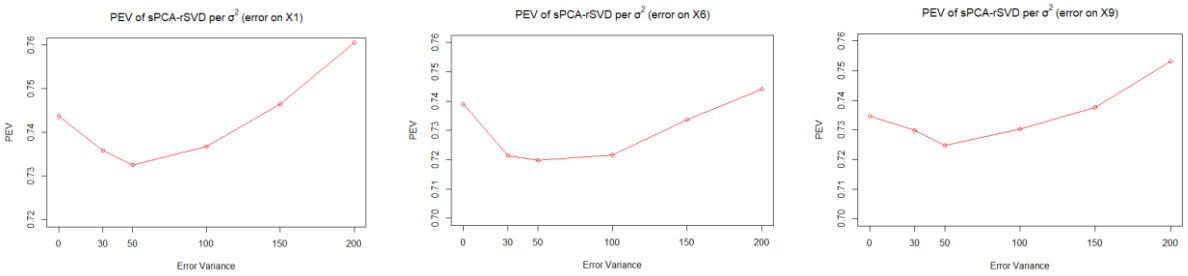
(c) : error in X6



(d) : error in X9

Figure 14 : The effect of the variance of the error on the standard error of the loadings when the error is in one variable.

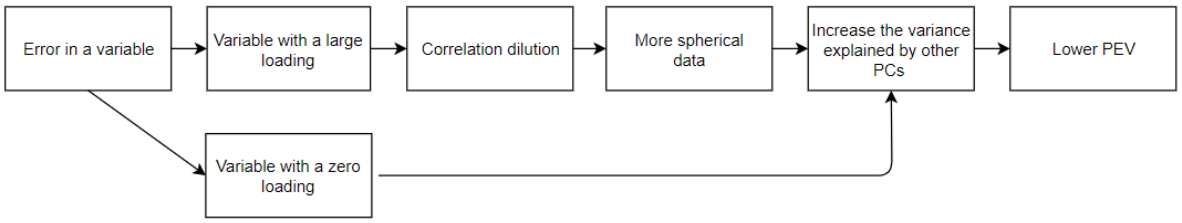
In Section 3.1, how the error increases the standard error of the loading is discussed with regard to correlation dilution. The same mechanism also holds for the heterogeneous error. When there is an error in one variable (denoted by X), the correlation between X and another variable is diluted, which makes data 'less spiked' (=more spherical) towards the X-axis and as a result, increases the standard error of the loading of X as shown in Figure 14. By the fact that the error in X does not significantly affect the standard error of the loadings of the other variables, the error in X only affects the classification rate of X as illustrated in Figure 13. The classification rate of X5 only significantly decreases when the error is in X5.



(a) : error in X1 (b) error in X6 (c) : error in X9

Figure 15 : The effect of the heterogeneous error on the expected PEV of the first PC in sPCA-rSVD when the error is in one variable.

While the homogeneous error consistently decreases the expected PEV of the first principal component, the heterogeneous error reduces it until a specific point as seen in Figure 15. When the error is in a variable of which a true loading is large (such as X1 or X9), the data becomes less spiked towards the axis of that corresponding variable, which increases the variance explained by the second or another PC. As a result, it decreases the PEV of the first PC. On the other hand, when the error is in a variable of which a true loading is zero (such as X5 or X6), this directly increases the variance explained by the second or another PC as that variable might be associated with the second or another PC, but not the first PC. In short, this mechanism can be summarized as follows;



However, as the error in a variable continuously increases beyond a certain point, the direction of the largest variance in data becomes closer to the axis of that corresponding variable, and the variance explained by the first PC becomes larger. Consequently, the PEV of the first PC becomes larger in the end.

4. Conclusion and Discussion

In terms of the eigen value and the eigen vector, this study discussed three topics via the simulation results: the inconsistency of PCA, the comparison of performance between PCA and sPCA-rSVD, and the effect of measurement error on sPCA-rSVD. Section 2 showed that, when p is fixed, the inconsistency of the eigen value depends on the sample size, while the inconsistency of the eigen vector is determined by two factors: the sample size and the spike index. To overcome these drawbacks of PCA, sPCA-rSVD is proposed[Shen and Huang , 3]. Section 2.3 showed that the loadings from sPCA-rSVD are not only more interpretable, but also more consistent than PCA in a HDLSS situation, especially when there is high sparsity in the loadings. Finally, in Section 3, how the homogeneous or heterogeneous error affects the performance of sPCA-rSVD is addressed regarding the four criteria: the expected first eigen vector, the expected PEV of the first PC, the classification rate and the standard error of the loadings.

Yet, there are some limitations of the simulation experiments in Section 3. Only the first principal component is addressed, and multiple datasets should have been created to examine whether another factor such as the sparsity or the spike index is related to the effect of measurement error on sPCA-rSVD. Plus, the critical limitation is that the unrealistic assumption - the sparsity of the loadings are known. – is made for the simulation experiments in this paper. In practice, other techniques – e.g., ad hoc approach or CV (K-fold cross validation) - to determine the sparsity of the loadings need to be additionally implemented. Thus, there should be further research on the effect of measurement error on another PC and techniques to determine the sparsity of the loadings, based on multiple datasets with different factors and levels.

References

[1] Iain M. Johnstone, Arthur Yu Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. J Am Stat Assoc. (2009), 3-4.

[2] KRISTOFFER HERLAND HELLTON and MAGNE THORESEN. The Impact of Measurement Error on Principal Component Analysis. Scandinavian Journal of Statistics Vol. 41, No. 4(2015), 7-8.

[3] Shen and Huang. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis Volume 99, Issue 6(2008).

[4] Guerra Urzola, Rosember; Van Deun, Katrijn; Vera, J. C.; Sijtsma, K. A GUIDE FOR SPARSE PCA: MODEL COMPARISON AND APPLICATIONS. psychometrika—vol. 86, no. 4, 893–919(2021), 25.

[5] Dan Shen, Haipeng Shen and J. S. Marron. A GENERAL FRAMEWORK FOR CONSISTENCY OF PRINCIPAL COMPONENT ANALYSIS. JMLR 17(150):1－34(2016), 3-4.

[6] Dan Shen, Haipeng Shen and J. S. Marron. A GENERAL FRAMEWORK FOR CONSISTENCY OF PRINCIPAL COMPONENT ANALYSIS. JMLR 17(150):1－34(2016), 3-4.

[7] Shen and Huang. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis Volume 99, Issue 6(2008) 3-5.

[8] Shen and Huang. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis Volume 99, Issue 6(2008) 11.

[9] KRISTOFFER HERLAND HELLTON and MAGNE THORESEN. The Impact of Measurement Error on Principal Component Analysis. Scandinavian Journal of Statistics Vol. 41, No. 4(2015), 6.