DOI : https://dx.doi.org/10.26808/rs.st.i7v6.08

Studying Of Gamma Principal Components: Analysis of Molding Noise

Zakiah I. Kalantan[a], Samia A. Adham[b], Ameerah O. Bahashwan[c]

[a,b,c]Department of Statistics, King Abdul-Aziz University, Jeddah, Saudi Arabia

[a]Email: zkalanten@kau.edu.sa

[b]Email: sayadham@hotmail.com

[c]Email:Abahashwan003@stu.kau.edu.sa

Abstract

Big data is affected by the presence of noise which is unavoidable problem. To overcome such a problem, it is a favorable to reduce the high dimensions while retaining the most important and useful data. The dimension reduction method such as principal component analysis is achieved by projecting the input data via a subset of principal components that describes the most variance of the data. PCA based on Gaussian noise model is sensitive to the noise of large magnitude. In this paper, we propose Gamma Principal Component Analysis instead of Gaussian PCA .We will utilize Gamma distribution to model noise where they consider effective PCA method to noise. The effectiveness of Gamma PCA model is studied using simulated data. In addition, the comparison with other noise models is discussed.

Keywords:

Principal component analysis, Gamma PCA, location-scale family PCA.

1. Introduction

Principal Components Analysis (PCA) is linear technique for dimensionality reduction[7,8]. The goal of PCA is to reduce the number of interesting variables into a smaller set of *components.* It seeks for a linear data pattern where the variance of the data in the low-dimensional representation is maximized [1]. PCA based on Gaussian noise model is sensitive to noise of large magnitude.

Various studies discussed the weak of Gaussian noise and suggested another distribution to model noise. Such researches proposed Student-t distribution and Laplace distribution because they have heavy tailed compared to Gaussian noise model which make them suitable to robust PCA. Peel and McLachlan (2000) [4] replaced Gaussian distribution by the Student-t distribution to increas the robust of PCA. While Ke and Kanade (2005) propsed Laplace noise assumption[2].The Student-t distribution and Laplace distribution are suitable for modeling spiky noise with large magnitude as result to their heavy tailed. However, but both methods suffer of some problems. Khan and Dellaert (2004) and Archambeau et al. (2006) tried to modify PCA by Student-t distribution [4]. The Student-t distribution was very similar behave to the Gaussian PCA[3, 4] but works much worse on large noises than Laplace PCA. In other hand, Laplace PCA cannot deal with dense noise. Pengtao and Eric (2014) discussed the capable of Cauchy noise assumption model and the effectiveness of Cauchy PCA[5].

2. Concepts Of Principal Component Analysis

2.1 PCA

Principal component analysis is one of the most important and powerful methods that interested with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. PCA is a procedure that uses an orthogonal transformation to convert a number of correlated variables into a small number of independent linear combinations of those variables called principal components. PCA seeks the linear combinations of the original variables such that the derived variables capture maximal variance [7,8].

2.2 Finding PCAs

Given the D-dimensional random variable $X = (X_1, \cdots, X_D)^T$, the PCA find a lower dimensional representation of it, $S = (S_1, \cdots, S_D)^T$ with $d \leq D$, that captures the content in the original data, according to some criterion. The components of S are called the hidden components. Linear techniques result in each of the $d \leq D$ components of the new variable being a linear combination of the original variables:

$$S_i = w_{i,1}x_1 + \cdots + w_{i,D}x_D \quad i = 1,\cdots,d \tag{1}$$

$$S = WX,$$

Where$W_{d\times D}$ is the linear transformation weight matrix. Expressing the same relationship as:

$$X = AS$$

with$A_{D\times d}$, note that the new variables S are also called the hidden or the latent variable[6].

Let $X = (X_1,\cdots,X_D)^T$ be a random vector has a probability density function $f(x)$ from multivariate Gaussian distribution with mean and variance denoted by μ and $\sum$, respectively. Assume that a sample of size N is drawn from the random vector X, yielding data $Z = \{x_1,\cdots x_N\}\, \varepsilon\, R^D$, which are N independent and identically distributed (iid) observations. The matrix Z has the following structure:

$$Z = \begin{bmatrix} x_{11}\; x_{12} \cdots x_{1j} \cdots x_{1D} \\ x_{21}\; x_{22} \cdots x_{2j} \cdots x_{2D} \\ x_{31}\; x_{32} \cdots x_{3j} \cdots x_{3D} \\ \vdots \quad \vdots \;\cdots\; \vdots \;\cdots\; \vdots \\ x_{N1} x_{N2} \cdots x_{Nj} \cdots x_{ND} \end{bmatrix}$$

this can be rewritten as:

$$Z = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix},$$

where D is the number of variables and N is the number of observations [7]. Now, the joint density function of $X_1, X_2,\cdots,X_n$ is given by

$$f(x_1,\cdots,x_n) = \prod_{j=1}^{n}\left\{\frac{1}{(2\pi)^{\frac{p}{2}}|\sum|^{\frac{1}{2}}} e^{\frac{-(x_j-\mu)^t\Sigma^{-1}(x_j-\mu)}{2}}\right\},$$

$$= (2\pi)^{\frac{-np}{2}}|\sum|^{\frac{-n}{2}} e^{\frac{-\sum_{j=1}^{n}\left[(x_j-\mu)^t\Sigma^{-1}(x_j-\mu)\right]}{2}}.$$

Now, the maximum likelihood estimators of μ and $\sum$ are $\hat{\mu}$and $\widehat{\sum}_{ML}$, respectively.

Where

$$\hat{\mu} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_D \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_{i1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} x_{iD} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_2 \end{pmatrix},$$

and

$$\widehat{\sum}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T.$$

and the sample variance matrix is given by

$$\widehat{\sum}_{sample} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T = \frac{N}{N-1}\widehat{\sum}_{ML}.$$

Note that $\hat{\mu}$ is an unbiased estimator ofμ and $\widehat{\sum}_{sample}$ is an unbiased estimator of$\sum$. An important characteristic of PCA is the decomposition of the variance of X. The illustration of decomposing the variance is presented below:

For the j^{th}eigenvector γ_j of $\sum$, one has

$$\sum\gamma_j = \lambda_j\gamma_j, \quad \mathrm{j} = 1,\cdots,\mathrm{D},$$

where $\sum$ is variance-covariance matrix and γ_j is one of the orthogonal eigenvectors$\gamma_1,\cdots,\gamma_D$ of $\sum$, and λ is one of the D eigenvalues of $\sum$ such that$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D > 0$.

Therefore, the eigenvectors is chosen corresponding to the largest eigenvalueλ_j, which can be written as

$$\sum(\gamma_1,\cdots,\gamma_D) = (\gamma_1,\cdots,\gamma_D)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{pmatrix}.$$

Then,

$$\sum\Gamma = \Gamma\Lambda \tag{2}$$

$$\sum = \Gamma\Lambda\Gamma^{-1} = \Gamma\Lambda\Gamma^T,$$

where, Γ transpose equals Γ inverse, meaning the columns of Γ are orthonormal.$\Lambda = diag(\lambda_1,\cdots,\lambda_D)$is a diagonal matrix containing the ordered eigenvalues of $\sum$, and Γ is an orthogonal matrix. The columns of$\Gamma = \gamma_1,\cdots,\gamma_D$ are the eigenvectors of $\sum$ and are called principal component loadings.

This decomposition is called the eigen-decomposition of $\sum$ have

$$\lambda_j = var(\gamma_j^T X), \qquad j = 1,\cdots,D.$$

Which means thatλ_j provides some decomposition of variance, and, from Eq. (2), their sum is

$$\lambda_1 + \cdots + \lambda_D = Tr(\Lambda).$$

Thus,

DOI : https://dx.doi.org/10.26808/rs.st.i7v6.08

$$\lambda_1 + \cdots + \lambda_D = Tr(\Gamma^T \sum \Gamma) = Tr(\Gamma^T \Gamma \sum)$$

The trace of the variance matrix is called the total variance. Therefore,

$$\frac{\lambda_j}{\lambda_1 + \cdots + \lambda_D} = \frac{var\left(\gamma_j^T X\right)}{TV(X)},$$

is the proportion of total variance explained by thej^{th} principal component. A software R package illustrates this decomposition by plotting λ_j versus j using the scree-plot tool.

Assume that PCA has been carried out on a data set Zyieldingμ, $\sum$,$\gamma_1, \cdots, \gamma_D$; $\lambda_1, \cdots, \lambda_D$. Then to compress the data Z to a smaller dimension $d \le D$ means to project all data points N onto the d-dimensional subspace spanned by the d largest principal components:

$$\phi: R^D \to R^D, x_i \to (\gamma_1, \cdots, \gamma_D)^T (x_i - \mu), \quad i = 1, \cdots, N.$$

The $\phi(x_i) \equiv P_i$are called scores. It is obvious that the original data will not be reconstructed exactly unless$d = D$.[7].
The application of principal component analysis postulates implicitly some form of linearity (Einbeck, Evers, and Bailer-Jones, 2008)[8]. The first Principal Component (PC) is the linear combination with the largest variance. The second PC is the linear combination with the second largest variance and orthogonal to the first PC and so on.
In essence, PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations (the PCs) of the original variables with the largest variance. For many datasets, the first several PCs explain most of the variance, so that the rest can be disregarded with minimal loss of information[6].

2.3 Determine the Number Of Components

There are many approaches of deciding how many principal components to use. These ways can detect a different number of components for the same data. The amount of accuracy of the work depended on how many principal components should use. In such cases, choose the most interpretable and logical solution for your data.

Percentage of Variation Explained: Retain components that cumulatively explain a certain percentage of variation. Some research needs 50% of the variance, rather than 40% or 60%. The Strategy is adding component until achieve the acceptable level of explained variance. Often, should chose component explain 70 to 80% of the variance at least percentage of variation explained

Kaiser Method: Retain components with eigenvalues greater than 1. This approach called Kaiser Methodor eigenvalue one criterion

Scree Plot: It is a graphical method; retain those components in the steep curve before the first point that starts the flat line trend. A scree plot displays the eigenvalues associated with a component in descending order versus the number of the component. Scree plots use in principal components analysis for visually estimate which components explain most of the variability in the data.

3. Gamma PCA Modeling Noise

This section illustrates location-scale family distributions. Some intuition of choosing Gamma distribution is presented to model noise by comparing its density curve with other distributions.

3.1 Location-Scale Family PCA

Location-scale family is a class of distributions parameterized by a location parameter and a scale parameter. The most important property of this family is that distributions are closed under linear transform. If X is a random variable drawn from this family, then $a\mathrm{X} + b$ is also from this family. This property provides convenience to model additive noise. In PCA setting, assume each entry of noise matrix E is from iid location-scale family distribution

$$E_{ij} \sim p\left(E_{ij} \,|0, s\right) \tag{3}$$

with location parameter zero and scale parameter s. According to the closure under linear transformation property and additive noise assumption $\mathrm{M} = \mathrm{L} + \mathrm{E}$, observation matrix M can be modeled as

$$M_{ij} \sim p\left(M_{ij} \,|L_{ij}, s\right). \tag{4}$$

with shifted location parameter L_{ij}. L can be estimated by maximizing the likelihood of observations (or minimizing the negative log likelihood) with low rank constraint

$$max_L \prod_{rank\ (L) \le k} p\left(M_{ij} \,|L_{ij}, s\right).$$

Gaussian PCA , Laplace PCA and Cauchy PCA (Candes et al., 2009; Ke & Kanade, 2005; Xie & Xing; 2014) are special cases of the general framework by specifying the distribution in Eq.(3) to Gaussian , Laplace ,Cauchy,and Gamma distribution respectively.

3.2 Vision on Gamma PCA with the Previous Studies

PCA technique assumes that the noise has a Gaussian model that is sensitive to noise of large magnitude. There are number of improvements have been proposed, such as Student-t distribution that has heavy tails that can reasonably explain data far away from the mean. Thus, they are suitable for modeling spiky noises with large magnitude. Gaussian PCA is limited to small noise and Laplace PCA is only suitable for sparse noise. Once noise is both large and dense, neither of the two PCA methods will be sufficed. In practice, the Student-t methods have very similar performance as the Gaussian PCA [5].

On the other hand, in many problems, noise patterns are mixed. Gaussian PCA and Laplace PCA are not applicable in this case since neither of them are able to deal with the two types of noise simultaneously. Therefore, will study an alternative probabilistic PCA method called Gamma PCA to discover its effectiveness to model noise and derive Gamma PCA under maximum likelihood estimation.

Given Gamma distribution with location parameter zero to model noise **E**

$$\mathrm{p}\left(\mathrm{E_{ij}}\right) = \frac{\mathrm{x}^{\mathrm{k}-1}}{\theta^{\mathrm{k}}\,\Gamma k}\,\mathrm{e}^{-\frac{\mathrm{x}}{\theta}} \tag{5}$$

where k is a shape parameter and θ is the scale parameter. The likelihood function of Gamma distribution is:

$$L(x) = \theta^{-nk}\left(\Gamma k\right)^{-n}\left[\prod_{i=1}^{n} x_i\right]^{k-1} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} \tag{6}$$

The maximum likelihood is $\hat{\theta} = \frac{\bar{x}}{\alpha}$.

Now will introduce the comparative study between the Gamma PCA, Gaussian PCA, Laplace PCA, Logistic PCA and Cauchy PCA to show their resistance to noise.

Figure 1(a) shows the density curves of univariate Gaussian, Laplace, Logistic, Cauchy and Gamma distributions. To enable a clear comparison, density curves are aligned to the same location and peak. The motivation of aligning their peaks is to inspect an interesting phenomenon: if we put the same amount of probability on the mode of each distribution, how much probability will each distribution allocates for other values? This can give a good sense of heavy-tail-ness. As data points get far away from the center, Gaussian probability drops quickly to zero while Gamma has a heavy tail than Gaussian and Logistic as shown in Figure 1(b), thus means Gamma consider best from Gaussian probability. In addition, Laplace and Cauchy probabilities remain a certain amount. In other words, Laplace and Cauchy density curves have longer tails than others.

In terms of noise modeling under a probabilistic framework, large noises can be reasonably explained by heavy tail distribution. Thereby, Laplace PCA and Cauchy PCA naturally possess the ability of dealing with large noise due to their heavy-tail-ness. At the same location (Figure 1(c)), Laplace distribution is not differentiable. The non-smoothness property induces sparsity, which makes Laplace distribution unsuitable to model dense noise. Gamma distribution has smooth property that makes it best than Laplace in dealing to model dense noise. It is suitable to model dense noise. Gamma distribution and Logistic distribution highly resembles Gaussian distribution in shape except a slightly heavier tail. Therefore, its behavior in modeling noise should be very similar to Gaussian distribution. Among the five, Cauchy distribution owns two appealing advantages. First, it is a smooth then it is suitable for modeling dense noise. Second, it has a much heavier tail than others, thus it is highly capable of modeling large noise.

4. Experimental Results

A simulation study is performed to evaluate the Gamma PCA to the noise with two scenarios.

Firstly:

A data matrix of 100 observations are generated from Gamma ditribution with paramters $(\alpha = 9\ ,\ \beta = .5)$ which is a symmetric Gamma curve. To study the effecting of Gamma noise, we corrupt the data with noise rate 10% of sample size. Next, the Gamma PCA is performed to both the original and new data.

Figure 2 displays the scree plots of Gaussian PCA implementation for both two cases of the data. The scree plots illustrate the effect of noise data on the variation of each component after adding the Gamma noise data. Table 1 presents the cumulative proportion for components before and after adding noise data. It is obvious that components 1 and 2 explain 88% of the total variation before existing any noise in data, as displayed in Table 1 [b1]. While after adding the noise points, Table 1 [b2] presents that the cumulative proportion of the components 1 and 2 explain 82% of the total variation. Therefore, one can notice there is a little bit difference between the two cases but also still have a reasonable sense of interpretation of the data.

Secondly:

A Gaussian data matrix $(m \times m)$ is generated with parameters $(\mu = 3.5\ , \sigma = 1.25)$ with sample size 100 observations. Then corrupt them with noise that has rate 10%. The Gamma PCA is implemented for both data cases, before and after adding Gamma noise.

Hence, the scree plots in Figure 6 present the PCA components. We can recognize that components 1 and 2 explain 87% of the total variation before existing any noise in data, as displayed in Table 2 [b1]. While 72% of the variation is explained by both components 1 and 2 after adding Gamma noise data .

To sum up, both scenarios provide that Gamma PCA technique has a suitable behaved on the data with noise points. One can conclude that Gamma PCA technique able to skip the noise in the data efficiently. Which means that Gamma PCA technique can able to reduce the dimensions actively and exclude the data that cause the noise. While in a simulation of Gaussian PCA, it is sensitively appeared to the noise compared to Gamma distribution.

DOI : https://dx.doi.org/10.26808/rs.st.i7v6.08

5. Conclusion

The aim of this paper was to provide a new approach to model noise via Gamma distribution. It is known that the Gaussian model is sensitive to noise of large magnitude. Our approach aims to overcome that issue and deal with dense noise. Obviously, Gamma distribution with $(\alpha = 9\,, \beta = .5)$ appears longer tail than Gaussian. From the experimental results, one can conclude that Gamma PCA technique able to skip the noise in the data efficiently. Especially when noise data with 10% of data is considered as a big rate. Therefore, this means Gamma PCA technique can able to reduce the dimensions well and exclude the data that cause the noise. Gamma PCA technique has acceptable behave with data has noise and in some cases, it could resistant noising in data efficiently

.

6. Reference

[1] Jolliffe I. T. (1986) Principal Component Analysis and Factor Analysis, In *Principal component analysis*. ed: Springer. pp. 115-128.

[2] Ke Q. and Kanade T. (2005) Robust L/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming. In *Computer Vision and Pattern Recognition. CVPR 2005. IEEE Computer Society Conference on*. pp. 739-746.

[3] Khan Z. and F. Dellaert F. (2004) Robust generative subspace modeling: The subspace t distribution. Georgia Institute of Technology2004.

[4] Archambeau C., Delannay N. and Verleysen M.(2006) Robust probabilistic projections. In *Proceedings of the 23rd International conference on machine learning*. pp. 33-40.

[5] Xie P. and Xing E. (2014) Cauchy Principal Component Analysis. *arXiv preprint arXiv:1412.6506*.

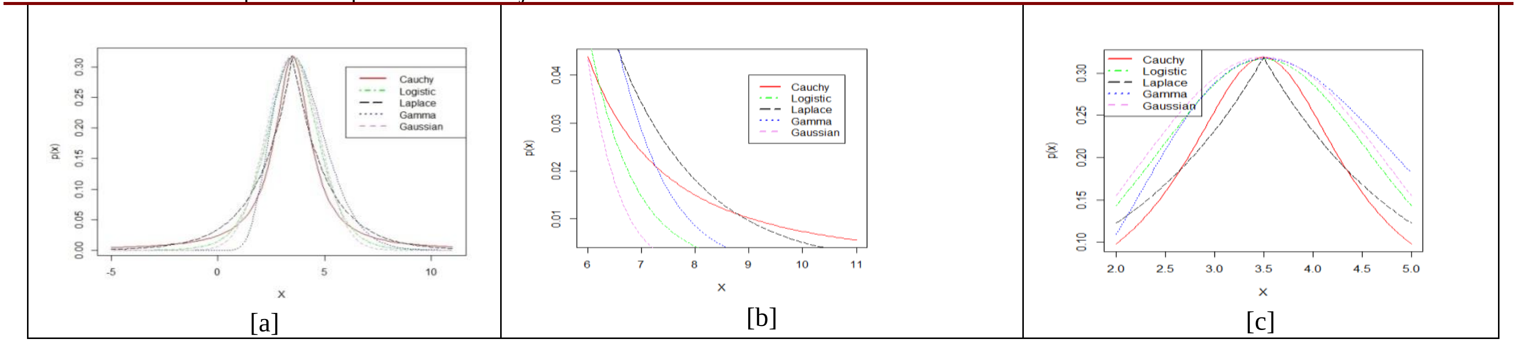
[6] Fodor I. K. (2002) A survey of dimension reduction techniques. Lawrence Livermore National Lab. CA (US).

[7] Kalantan Z., (2014) Methods for Estimation of Intrinsic Dimensionality. Ph. D. Thesis in Statistics. Durham University. UK.

[8] Kalantan Z., (2017) implementing Correlation Dimension: K means clustering via correlation dimension. In proceeding : The 3rd international Conference on computing, mathematics and Statistics (ICM2017) and the5thejoint international conference on new challenges forstatistical software: the use of r in official statistics(uros2017)- asia pacific, Lanckawi, Malaysia

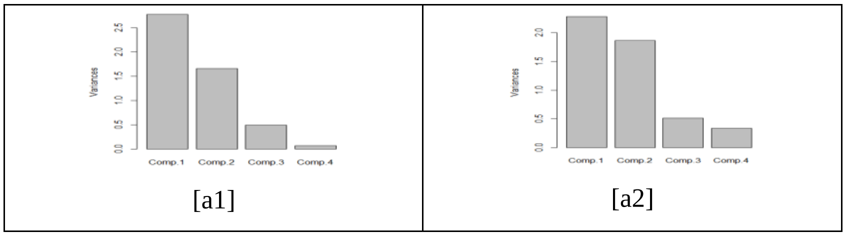
[9] Einbeck, J., Evers, L. and Bailer-Jones, C. (2008). Representing complex data using localized principal components with application to astronomical data. *Principal Manifolds for Data Visualization and Dimension Reduction*, In: *Lecture Notes in Computational Science and Engineering*, **58**, pp 180–204.

DOI : https://dx.doi.org/10.26808/rs.st.i7v6.08



[a] [b] [c]

Fig 1: Density curves of Gaussian, Laplace, Logistic, Cauchy and Gamma distribution over different ranges of random variable X.



[a1] [a2]

Fig 2: Gamma PCA at N=100; [a1] Scree plot for data before noising , [a2] Scree plot when 10% noising done.

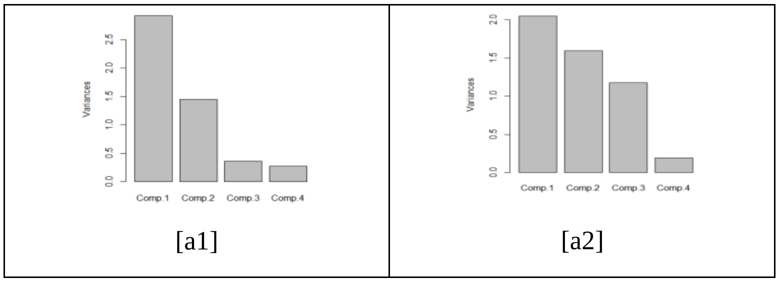
```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation       1.665661  1.28938   0.7040613  0.2595484
Proportion of Variance   0.55488   0.33250   0.0991404  0.0134730
Cumulative Proportion    0.55488   0.8873    0.9865269  1.000000
```

[b1]

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation       1.51070  1.365398 0.717915   0.5814121
Proportion of Variance   0.45644  0.372862  0.1030804  0.0676080
Cumulative Proportion    0.45644  0.829311  0.932392   1.0000000
```

[b2]

Table 1: Summary table of the implementation of Gamma PCA at N=100; [b1] before nosing, [b2] when noising done.



[a1] [a2]

Fig 3: Gaussian PCA at N=100; [a1] scree plot for data before noising [a2] scree plot when 10% gamma noising

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation        1.70817   1.20357   0.5988074  0.524392
Proportion of Variance 0.58357    0.28971   0.0717140  0.054997
Cumulative Proportion   0.58357   0.87328   0.9450026 1.0000000
```

[b1]

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation        1.4304    1.26121   1.08293   0.43630500
Proportion of Variance 0.4092     0.31813   0.23454   0.03807241
Cumulative Proportion   0.4092    0.727379   0.96192   1.00000000
```

[b2]

Table 2: Summary table of the implementation of Gamma PCA on Gaussian data (N=100); [b1] before nosing, [b2] when noising done.