Quantile Kurtosis in ICA and Integrated Feature Extraction for Classification

Md Shamim Reza and Jinwen Ma(✉)

Department of Information Science, School of Mathematical Sciences and LMAM,
Peking University, Beijing 100871, China
shamim@pku.edu.cn, jwma@math.pku.edu.cn

Abstract. As an effective statistic in independent component analysis (ICA), kurtosis can provide valuable information for testing normality, determining features shape and ordering independent components of feature extraction in classification analysis. However, it may lead to the poor performance in certain situations so that the quantile kurtosis has been developed. In this paper, we propose a robust quantile measure of kurtosis in ICA for feature extraction. Moreover, we also present a feature extraction method which integrates the extracted features of principal component analysis (PCA), linear discriminant analysis (LDA), ICA and random forest algorithm (RFA) together. For the ICA based feature extraction, independent components are sorted according to the proposed quantile kurtosis. The experimental results show that our integrated feature extraction method, especially with the help of the proposed quantile kurtosis, outperforms the others.

Keywords: Quantile · Kurtosis · Normality · ICA · PCA · LDA · RFA

1 Introduction

Over the last few decades, Independent Component Analysis (ICA) has been found to be a very convenient and effective

The fundamental restriction of ICA is that independent components (IC's) must be non-Gaussian, and we cannot determine the order of the obtained independent components [13]. This two ambiguities of ICA are the main obstacle for extracting representative feature in classification analysis. Clearly, the Principal Components (PC's) can be sorted according to the related eigenvalues, but there is no reasonable measure to order the independent components (IC's) [12, 13]. However, the past studies have shown that non-Gaussian IC's are sometimes significant to classification and kurtosis statistic can be considered as a measure of non-gaussianity as well as for sorting the IC's [14, 15].

Given this emerging concentration of kurtosis in ICA, all of the previous work concerning kurtosis in ICA has used the classical measures of kurtosis [14, 15]. Usually, classical measures of kurtosis are based on the sample average and very sensitive to outliers. In order to overcome this problem, Moors [20] proposed a quantile kurtosis alternatively, but this quantile kurtosis is not so robust to ordering independent components. In this paper, we propose an improved quantile measure of kurtosis to sort the independent components and compare their performances with four other kurtosis measures that are found in the recent statistics literature.

In the classification problem, there may happen irrelevant features that affect the learning process and thus lead to an unsatisfactory result [16]. Additionally, as the dimension of feature space becomes very large, the classification method requires many attributes to find out the association of the features, which triggers slow training and testing in both of supervised and unsupervised learning algorithm. Some of the feature extraction techniques such as ICA, PCA, LDA, random forest algorithm (RFA) and wrapper method can be directly used for feature extraction, but they cannot guarantee to generate the useful information individually [11, 12, 17]. In our earlier work, we proposed two integrated feature extraction methods only based on ICA and PCA, which outperformed the others adoptive methods [12]. In this paper, we further integrate with LDA, PCA, ICA, and RFA feature based on some statistical criterions to generate a more representative feature for the purpose of improving the classification performance.

The remaining of this paper is organized as follows. Section 2 gives a review of classical and quantile measures of kurtosis. Section 3 presents our proposed robust method of quantile kurtosis and the experimental results to demonstrate its performance. In Sect. 4, we propose our integrated feature extraction method for classification. Section 5 summarizes the experimental results and comparisons. Finally, we conclude briefly in Sect. 6.

2 Review of Kurtosis Measures

Pearson [18] originally introduced kurtosis as a measure of how flat the top of a symmetric distribution is in comparison with a normal distribution of the same variance. This conventional measure can be formally defined as the standardized fourth population moment about the mean.

$$K_1 = \frac{E(x-\mu)^4}{\left(E(x-\mu)^2\right)^2} - 3 = \frac{\mu_4}{\sigma^4} - 3 \tag{1}$$

Since the conventional measures of kurtosis are essentially based on sample averages, they are sensitive to outliers. Moreover, the impact of outliers is greatly amplified in the conventional measures of kurtosis due to the fact that they are raised to the third and fourth powers [19].

To overcome of conventional measure of kurtosis, Moors [20] proposed a quantile kurtosis alternative to K_1. The quantity of Moors kurtosis is

$$K_2 = \frac{(E_7 - E_2) + (E_3 - E_1)}{(E_6 - E_2)} \tag{2}$$

where E_i is i-th octile; that is $E_i = F^{-1}(i/8)$. For Gaussian independent components, Moor's quantile kurtosis is equal to 1.23. One advantage of the quantile measures of kurtosis is that it doesn't depend on the first moment and second moment. So the measure is not affected by outliers.

While investigating how to test light-tailed distributions against heavy-tailed distributions, Hogg [21] found that the following measure of kurtosis performs better than the traditional measure in detecting heavy-tailed distributions:

$$K_3 = \frac{U_\alpha - L_\alpha}{U_\beta - L_\beta} \tag{3}$$

Where U_α (L_α) is the average of the upper (lower) α quantile defined as:

$$U_\alpha = \frac{1}{\alpha}\int_{1-\alpha}^{1} F^{-1}(y)dy, \text{ and } L_\alpha = \frac{1}{\alpha}\int_{0}^{\alpha} F^{-1}(y)dy$$

for $\alpha\epsilon$ (0,1). According to Hoggs simulation experiments, $\alpha = 0.05$ and $\beta = 0.5$ gave the most satisfactory results. For the normal distribution, Hogg kurtosis value is 2.54. Hence, the centered Hogg coefficient is given by:

$$K_3 = \frac{U_{0.05} - L_{0.05}}{U_{0.5} - L_{0.5}} - 2.54 \tag{4}$$

Another interesting measure based on quantiles has been used in Crow and Siddiqui [22], which is given by

$$K_4 = \frac{F^{-1}(1-\alpha) + F^{-1}(\alpha)}{F^{-1}(1-\beta) + F^{-1}(\beta)} \tag{5}$$

where $\alpha, \beta\epsilon$ (0, 1). Their choices for α and β are 0.025 and 0.25 respectively. For these values, we obtain $F^{-1}(0.975) = -F^{-1}(0.025) = 1.96$ and $F^{-1}(0.75) = -F^{-1}(0.25) = -0.68$ for N (0,1) and the coefficient is 2.91.

3 Robust Quantile Kurtosis

In this study, we take an attempt to propose a robust measure of kurtosis which based on quantile and modification of moor's kurtosis estimator. In statistics, quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. It generalization of the idea of the median, where median is the value which splits data into two equal parts. Similarly, a quantile partitions the data into other proportions. Dividing ordered data into essentially equal-sized data subsets is the motivation for q-quantiles. We have used specialized 16-quantiles are called hexadeciles where ordered data are divided into 16 equal sizes. The proposed measure is given by

$$K_5 = \frac{\left(E_{15} - E_9\right) + \left(E_7 - E_1\right)}{\left(E_{15} - E_1\right)} \tag{6}$$

Where E_i is i-th hexadeciles; that is $E_i = F^{-1}(i/16)$. We easily to calculate that $E_1 = -E_{15} = -1.53$, $E_7 = -E_9 = -0.16$ for N (0,1) and therefore the coefficient of kurtosis is 0.8975. Hence, the centered coefficient is given by:

$$K_5 = \frac{\left(E_{15} - E_9\right) + \left(E_7 - E_1\right)}{\left(E_{15} - E_1\right)} - 0.8975 \tag{7}$$

Since our measure based on hexadeciles, it covers a wide range of data and doesn't depend on sample mean and variance. So it's less affected by outlier and more robust than the classical measure of kurtosis. In the next section, we will discuss the robustness of an estimator.

3.1 Qualitative Robust Index

Qualitative robustness, influence function, and breakdown point are three main concepts to judge an estimator from the viewpoint of robust estimation. Nasser et al. [23] have proposed a definition of finite-version qualitative robustness, their estimator with finite breakdown point equal to zero should have empirically lower QRI whereas estimators with high breakdown point should have higher QRI. They proposed two versions of SQRI (SQRI-1 and SQRI-2):

$$SQRI - I = \frac{1}{1 + max_j\left|\widehat{\theta} - \widehat{\theta}(j)\right|} \tag{8}$$

$$SQRI - II = \frac{1}{1 + max_{i \neq j}\left|\widehat{\theta}(i) - \widehat{\theta}(j)\right|} \tag{9}$$

It is easy to prove (i) It's maximum value is 1. (ii) It's minimum value is zero or above zero. The more SQRI the estimator is more qualitative robust.

3.2 Datasets and Results

To find out qualitative robust measure of kurtosis by using SQRI-1 and SQRI-2 method, we have used USGS (United States Geological Survey) data, counting earthquake by yearly contains 112 observations (1900-2011) and considered magnitude range of earthquake 7.0 to 9.9. In our experiment, we have taken four datasets, where three from UCGS, earthquake data and one set of melanoma skin cancer data. Sample data structure are given below:

- Data-1: 60 sample have drowned from 112 observations of UCGS data.
- Data-2: 60 sample have drowned from 112 observations and 5 samples have drawn from the Cauchy distribution with parameter 2.
- Data-3: 60 sample have drowned from 112 observations and 5 samples have drowned from the student-t distribution with degrees of freedom 2.
- Data-4: 37 sample have taken from melanoma skin cancer incidence data during the year (1936 – 1972). Source: R data package (lattice).

In descriptive statistics, the box plot is a convenient way of graphically displaying variation in samples of a statistical population without making any assumptions about the underlying statistical distribution. Figure 1, we have taken 60 samples in 1000 times from earthquake data and calculate each of kurtosis estimators. The graph shows that propose kurtosis estimator (K_5) distribution is more consistent than others. To check

Table 1. Results comparison for SQRI-1 of six kurtosis estimators.

Kurtosis estimators	Data-1	Data-2	Data-3	Data-4
K_1	0.725	0.825	0.874	0.952
K_2	0.891	0.852	0.837	0.944
K_3	0.849	0.935	0.918	0.946
K_4	0.698	0.782	0.807	0.955
K_5 (proposed)	**0.925**	**0.942**	**0.931**	**0.968**

numerically of a robust estimator, we have used two SQRI for all kurtosis estimators which result are displayed in Tables 1 and 2.

Table 2. Results comparison for SQRI-2 for six kurtosis estimators.

Kurtosis estimators	Data-1	Data-2	Data-3	Data-4
K_1	0.575	0.726	0.810	0.905
K_2	0.853	0.830	0.767	0.902
K_3	0.812	0.879	0.882	0.904
K_4	0.602	0.662	0.735	0.917
K_5 (proposed)	**0.905**	**0.922**	**0.894**	**0.926**

Notes and Comments: In our experiments, we consider bootstrap techniques to find θ, where the number of replication N = 1000. The results show that the proposed method successfully chooses the best robust estimator (K_5) because it's all value closest to 1. According to SQRI index, the more SQRI value conveys high break down point and indicating as the more robust estimator. So this robust kurtosis estimator can provide useful information in ICA as well as to sorts of independent components and extract representative features in a classification problem, in details refer to Sect. 4.

4 Integrated Feature Extraction Paradigm

Although, the performances of PCA, ICA, and LDA are powerful in the field of data visualization and blind source separation. For classification problem, feature extraction technique of LDA performances is good if certain assumptions are hold in data [24], but PCA and ICA are not as good as expected [11, 25]. To overcome the problem, we propose a feature extraction method, which integrates with LDA, PCA, ICA, and a feature selection technique random forest algorithm (RFA) to represent significant feature sets for classification problem.

The idea of the proposed feature extraction is very simple. In the proposed approach, LDA, ICA, PCA, and feature selection algorithm have applied to the original data individually, we then retain those PC's that can explain at least 80% of the total variation, most sub-Gaussian IC's (kurtosis < 0) are ordered by using propose quantile measure of kurtosis, (class-1) number of LD components and 20% most important original features which selected by random forest algorithm based on features weight. This proposed approach is named as integrated feature extraction. Figure 2 shows the flow chart of implementing on the four databases.

In the proposed approach, the procedure has extracted features from PCA which are uncorrelated and gives maximum variation (>80%) of the data. In ICA, the selected features are not only uncorrelated but also independent and we have chosen sub-Gaussian independent components that can play a vital role in the classification problem. LDA is powerful feature extraction in supervised classification because of the extracted components (class-1) are best characterize or separate between the classes of data. Feature selection algorithm FS-RFA also selects best original features based on feature

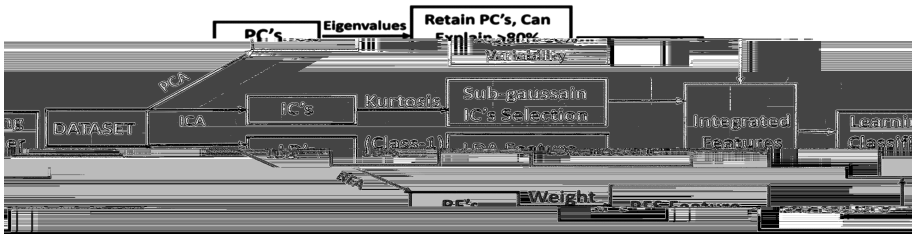

Fig. 2. Flow chart for implementing integrated feature extraction.

weights of class information. Finally, we have been integrated a significant features sets to learning classifiers.

5 Experimental Framework and Results

In this section, our proposed feature extraction approach is tested on a simulated dataset, and three real datasets from UCI database [26], namely Satellite, Ionosphere, and Sonar datasets, respectively.

In order to test the efficiency of the proposed feature extraction methods, we select the most significant number of original attributes by using random forest algorithm (FS-RFA), which is available in R package, FSelector [27]. In random forest algorithm FS-RFA, first, employs a weight function to generate weights for each feature. To select significance of weight, the algorithm use mean decrease accuracy, besides it selects an optimum number of subset feature through the statistical function chi-square and information gain. Finally, the procedure sorts a top most dominant original subset of features. To apply ICA for feature extraction, training data was transformed to zero mean and unit variance by using fastICA algorithm [13], PCA and LDA were applied directly over data.

In classifier system, we have used multi-layer perception (MLP), support vector machine (SVM), decision tree (C5.0), and naive Bayes classifier. In our experiment, we have driven 10-fold cross validation in getting the performance as follows: The observations have been divided randomly into 10 disjoint fold or sets. For each experiment, 9 of these fold is used as training data, while 10th set observation is reserved for testing. The experiment is repeated 10 times in such a way that every fold appears once as a part of a test set.

To show the effectiveness of our method, we have compared the performances of the proposed methods with PCA, LDA, ICA, IPCA, IC-PC and FS-RFA. Feature extraction techniques, IPCA and IC-PC we have already been proposed our earlier work [12]. All the experiments conduct here are implemented in R-studio software with different R-CRAN (The Comprehensive R Archive Network) packages of machine learning [27], and run on a 4 core GPU system.

5.1 On Simulation Dataset

In the simulation study, we have simulated 200 observations from each of the four well-known probability distributions, standard normal ($n = 200$, $\mu = 0$, $\sigma^2 = 1$), student's t ($n = 200$, degrees of freedom, $v = 1$), chi-square (χ^2) ($n = 200$, d. f. $= 1$), and standard uniform distribution ($n = 200$).

To analyze the synthetic data in the classification problem, we have divided each 200 observations into four columns in such a way that each column has 50 observations. An additional column has also been inserted to input class label. As for example of the Gaussian distribution (mean = 0, variance = 1), the generated 200 observations have been divided into four columns, where each column contains 50 observations, then each of the first 50 observations has been labeled by 1 in the additional column. Similarly, for chi-square (χ^2) distribution, 200 generated observations were divided into four columns and inserted the class label 2 and so on. Finally, we have combined the observations to obtain a data frame that contains 5 features including one class label attribute each with 200 observations.

In simulated data, first 3 PC's can explain 84.23% of the total variation, then we have applied ICA algorithm on 3 PC's to construct IPCA feature. In IC-PC, we have united first 3 PC's and one sub-Gaussian (kurtosis < 0) IC's. To make propose integrated feature, we have combined first 3 PC's with sub-Gaussian IC's (kurtosis < 0), one LD component and one most important original feature which obtained by using random forest algorithm. Table 3 shows the classification performances of different classifiers. The classification accuracy is obtained by using 10-fold cross-validation and we found that our integrated feature extraction methods improvement in all the classification performances in a certain degree.

Table 3. Classification accuracy (%) for simulated data (parentheses are the number of PC's & IC's respectively)

Features	SVM	Naïve Bayes	C5.0	MLP
Original	62.5	66.5	69.0	51.5
FS-RFA	64.0	69.0	65.0	56.0

PCA546Dσ.237(t)-7.4(a TDσ.re)7656.3(0)-4486.29980468σ6.606(F)-10.5(A)3(0)-4488(d306Jσ/.re)7

5.2 On Satellite Dataset

The original Landsat data for this database was generated from data purchased from NASA by the Australian Centre for Remote Sensing and used for research at the Centre for Remote Sensing, University of New South Wales, Australia. These data have been taken from the UCI Repository of Machine Learning Databases [26].

Data frame with 36 inputs, one target on 6435 observations. The database consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values.

In satellite data, only first 4 PC's (out of 36) can explain 92.02% of the total variation. All of the classifiers performs well when we integrate with first 7 PC's, one most sub-Gaussian IC's that ordered by our proposed quantile kurtosis, (6-1) class i.e. 5 LD components and most two important original variables that obtained from random forest algorithm (RFA). Table 4 shows that the performances of integrated and IC-PC feature extraction method have achieved 100% classification accuracy of the decision tree (C50) classifier. The IC-PC (7,2) method integrated only 7 PC's and 2 most sub-Gaussian IC's features out of 36 features and achieved 100% accuracy.

Table 4. Classification accuracy (%) for satellite data (parentheses are the number of PC's & IC's respectively)

Features	SVM	Naïve Bayes	C5.0	MLP
Original	80.34	79.12	84.61	83.31
FS-RFA	85.12	72.69	83.12	81.18
PCA	87.88 (7)	82.18 (7)	85.85 (7)	83.02 (7)
LDA	87.13	84.55	84.97	84.04
ICA	86.20	80.26	79.37	83.68
IPCA	87.44 (5)	82.79 (6)	85.68 (7)	83.80 (7)
IC-PC	87.56 (5,1)	82.64 (7,1)	**100 (7,2)**	83.73 (7,1)
Integrated	**89.29**	**85.75**	**100**	**84.16**

5.3 On Ionosphere Dataset

These data have been taken from the UCI Repository of Machine Learning Databases [26]. This radar data was collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. "good" radar returns are those showing evidence of some type of structure in the ionosphere. "bad" returns are those that do not; their signals pass through the ionosphere. Data frame with 351 observations on 35 independent variables, some numerical and 2 nominal, and one last defining the class. This dataset is often used to test and compare the performances of various classification algorithms.

In Ionosphere data, first 11 PC's can explain 80% of the total variation, while original feature number is 35. The classification accuracy of the four classifiers are displayed in Table 5. It can be seen that propose integrated features (11 PC's, 1 IC's, 1 LD component, and 1 original attribute) perform better than others. The cross validation classification

accuracy of SVM classifier exceeds than others in the past work on this dataset. In Ionosphere data, the naive Bayes classifier also performs better than others because of our proposed feature extraction method produced uncorrelated and independent features which coincide the assumptions of naive Bayes classification techniques.

Table 5. Classification accuracy (%) for ionosphere data (parentheses are the number of PC's & IC's respectively)

Features	SVM	Naïve Bayes	C5.0	MLP
Original	94.87	78.89	88.91	89.77
FS-RFA	94.87	91.73	89.75	90.31
PCA	96.01 (11)	90.32 (11)	89.73 (11)	87.46 (11)
LDA	89.45	88.89	88.89	89.45
ICA	93.73	89.74	84.33	87.22
IPCA	95.72 (11)	92.89 (11)	87.76 (11)	87.17 (11)
IC-PC	95.43 (11,3)	90.60 (11,2)	90.59 (11,1)	87.18 (11,2)
Integrated	**96.29**	**95.73**	**91.46**	**91.73**

5.4 On Sonar Dataset

This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network. The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock. These data have been taken from the UCI Repository of Machine Learning Databases [26], and data frame with 208 observations on 61 variables, all numerical and one (the class) nominal.

In sonar data, only first 14 PC's (out of 60 PC's) can explain 81.19% of the total variation. We have then compared proposed feature extraction approach with PCA, LDA, ICA, IC-PC and IPCA. In Table 6 show that most of the cases, LDA and our extracted feature (15 PC's, 2 IC's, one LD components, and 3 original variables) outperforms the others.

Table 6. Classification Accuracy (%) for Sonar data (Parentheses are the number of PC's & IC's respectively)

Features	SVM	Naïve Bayes	C5.0	MLP
Original	84.59	66.38	73.09	83.17
FS-RFA	83.62	73.07	81.73	80.78
PCA	85.02 (16)	75.97 (9)	76.50 (11)	82.19 (15)
LDA	88.46	85.22	**89.90**	**90.86**
ICA	79.83	60.64	59.66	75.02
IPCA	85.02 (16)	69.26 (16)	75.50 (15)	80.71 (15)
IC-PC	86.93 (16,3)	75.48 (16,2)	76.50 (16,2)	84.09 (15,1)
Integrated	**90.83**	**89.45**	86.50	87.45

13. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **4–5**(13), 411–430 (2000)
14. Reza, M.S., Nasser, M., Shahjaman, M.: An improved version of kurtosis measure and their application in ICA. Int. J. Wirel. Commun. Inf. Syst. **I**(1), 6–11 (2011)
15. Scholz, M., Gibon, Y., Stitt, M., Selbig, J.: Independent component analysis of starch-deficient pgm mutants. In: Proceedings of the German Conference on Bioinformatics, pp. 95–104 (2004)
16. Cios, K.J., Pedrycz, W., Swiniarski, R.W.: Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, Boston (1998). Chap. 9
17. Fan, L., Poh, K.L., Zhou, P.: Partition-conditional ICA for Bayesian classification of microarray data. Export Syst. Appl. **37**, 8188–8192 (2010)
18. Pearson, K.: Skew variation, a rejoinder. Biometrika **4**, 169–212 (1905)
19. Kim, T.H., White, H.: On more robust estimation of skewness and kurtosis: simulation and application to the S\&P500 index, Department of Economics, UCSD (2003)
20. Moors, J.J.A.: A quantile alternative for kurtosis. The Stat. **37**, 25–32 (1988)
21. Hogg, R.V.: More light on the kurtosis and related statistics. J. Am. Stat. Assoc.