Communications for Statistical Applications and Methods
2019, Vol. 26, No. 1, 57–67

https://doi.org/10.29220/CSAM.2019.26.1.057
Print ISSN 2287-7843 / Online ISSN 2383-4757

Nonparametric analysis of income distributions among different regions based on energy distance with applications to China Health and Nutrition Survey data

Zhihua Ma[a], Yishu Xue[a], Guanyu Hu[1,a]

[a]Department of Statistics, University of Connecticut, USA

Abstract

Income distribution is a major concern in economic theory. In regional economics, it is often of interest to compare income distributions in different regions. Traditional methods often compare the income inequality of different regions by assuming parametric forms of the income distributions, or using summary statistics like the Gini coefficient. In this paper, we propose a nonparametric procedure to test for heterogeneity in income distributions among different regions, and a K-means clustering procedure for clustering income distributions based on energy distance. In simulation studies, it is shown that the energy distance based method has competitive results with other common methods in hypothesis testing, and the energy distance based clustering method performs well in the clustering problem. The proposed approaches are applied in analyzing data from China Health and Nutrition Survey 2011. The results indicate that there are significant differences among income distributions of the 12 provinces in the dataset. After applying a 4-means clustering algorithm, we obtained the clustering results of the income distributions in the 12 provinces.

Keywords: income distribution, nonparametric test, energy distance, K-means clustering

1. Introduction

The income distribution in economic theory describes how a region's total wealth is distributed amongst its population (Sullivan, 2003). Back to classical economists time, income distribution was a key concern of economic theory, as it plays an important role in measuring the health of a region's economy. For modern economics, more attention is paid to inequalities in certain measurements of income distribution, an example of which is the Gini coefficient (Yitzhaki, 1979), now universally used by many international organizations such as the United Nations and the World Bank. The Gini coefficient, however, summarizes information about a distribution in one number, and loses some information such as the intrinsic structure of the distribution. Other measurements such as the Lorenz curve (Lorenz, 1905), also have such kind of disadvantage.

The probability density of the income distribution can often be estimated when the sample size is large enough. From the 1890s, many parametric distributions were introduced to model income distribution. In 1895, Pareto (1964) (originally published in 1895) first proposed the Pareto density function to model the income distribution. Gibrat (1931) suggested the usage of a two-parameter log-normal distribution. Salem and Mount (1974) proposed a two-parameter Gamma density to approximate the distribution of personal income. Bartels and Van Metelen (1975) suggested another

[1] Corresponding author: Department of Statistics, University of Connecticut, Room 323, Philip E. Austin Building, 215 Glenbrook Rd. U-4120, Storrs, CT 06269, USA. E-mail: guanyu.hu@uconn.edu

Published 31 January 2019 / journal homepage: http://csam.or.kr
© 2019 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

two-parameter density, the Weibull distribution, to model personal income. McDonald (1984) considered two generalized Beta distributions as models for the distribution of income. Furthermore, McDonald and Xu (1995) introduced a five-parameter beta distribution which nests the generalized Beta and Gamma distributions to model income distribution. In summary, these methods used heavy tail distributions to model personal income. When testing the equality of income distributions, people often compared parameters in the parametric models. This method leans on the assumption that the parametric models sufficiently describe both income distributions. Still, some information loss would occur in such comparisons. In addition, it is often not reasonable to assume the same parametric model form for different regions, and such comparison would be impossible when the regional models are parameterized differently.

The energy distance (Székely and Rizzo, 2004) is a measure that is often used to test for equality of distributions. Rizzo and Székely (2010) provided a nonparametric version of the analysis of variance (ANOVA) called the distance components (DISCO), which partitions the total dispersion in the dataset into components that are analogies of ANOVA's variance components. In addition, the energy distance can also be used in feature selection and generalizations of clustering algorithms (Rizzo and Székely, 2015). Li and Rizzo (2017) extended the usage of the energy distance and proposed k-groups, a generalization of the K-means clustering algorithm. K-groups aims to put similar samples in the same cluster so that the dispersion between the k groups are maximized. The maximization is achieved upon computation of energy distances, instead of Euclidean distances.

In order to address the disadvantages of the Gini coefficient and parametric models, we use the energy distance between different regions to test the equality of income distributions, which is a parametric-free method. Based on the energy distance, we propose a clustering method which clusters not individual observations, but income distributions, in different regions. The main contribution of this work is that all proposed methods are nonparametric - we do not require any parametric assumptions for income distribution.

The remainder of this paper is organized as follows. In Section 2, we give a brief introduction to the energy distance, and propose our testing and clustering procedures. In Section 3, we present the performance of the proposed illustrate our proposed procedures using extensive simulation results. In Section 4, we apply our methods on China Health and Nutrition Survey data. We conclude this paper with a brief discussion in Section 5.

2. Methodology

2.1. Energy distance

The energy distance is a statistical distance defined for two probability distributions. Following Székely and Rizzo (2004), we consider two distributions denoted by F_X and F_Y. The energy distribution between these two distributions is defined as:

$$D_e(F_X, F_Y) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\left\|X - X^{'}\right\| - \mathbb{E}\left\|Y - Y^{'}\right\|, \tag{2.1}$$

where $X, X^{'}, Y, Y^{'}$ are independent random vectors in $\mathcal{R}^d$ with finite expectations, and $\|\cdot\|$ denotes the Euclidean norm. The distance $D_e(F_X, F_Y)$ equals 0 if and only if $F_X \equiv F_Y$, while a large energy distance corresponds to very different distributions. Hence it provides a parametric free similarity measure of two distributions.

Now consider two independent samples from F_X and F_Y: $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. From Székely and Rizzo (2004), the energy distance associated with the independent random samples

can be written as:

$$d_{n,m} = \frac{nm}{n+m}\left(\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|y_i - y_j\|\right), \tag{2.2}$$

where $\|\cdot\|$ is the Euclidean distance. Based on (2.2), we can calculate the distance between two income distributions.

2.2. Hypothesis testing

For the two-sample problem, suppose we have two samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. The null hypothesis of two-sample is:

$$H_0: \quad F_1 = F_2, \tag{2.3}$$

where F_1 and F_2 are the respective distributions of $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. The two-sample test statistic is defined using (2.2):

$$T = \frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|X_i - Y_j\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|X_i - X_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|Y_i - Y_j\|\right).$$

Let $n = n_1 + n_2$ be the total sample size of the sample. Under the null hypothesis, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ follow the same distribution. Therefore, the random permutation without replacement from $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ is $W_1^\pi, \ldots, W_{n_1}^\pi, \ldots, W_n^\pi$. Then we can calculate the test statistics T_π based on $W_1^\pi, \ldots, W_{n_1}^\pi, \ldots, W_n^\pi$. For a fixed number of permutations, under the null hypothesis, the empirical distribution of the test statistic can be obtained. Based on the empirical distribution of the test statistic, the critical value for the statistic, c_α, can be chosen to satisfy:

$$\lim_{n\to\infty} P(T_n > c_\alpha) = \alpha, \tag{2.4}$$

where $\alpha \in (0, 1)$, and T_n is the full permutation test statistic. Proofs of the limiting distribution of test statistics T, the existence of c_α and consistency results are given in Székely and Rizzo (2004). From this limiting distribution, we can reject null hypothesis if $T > c_\alpha$.

It is easy to extend the two-sample testing problem to k-sample problem. Suppose $X_1, \ldots, X_k$ are k independent samples of random vectors with respective distributions $F_1, \ldots, F_k$. The null hypothesis is:

$$H_0 : F_1 = \cdots = F_k, \tag{2.5}$$

and the alternative is $F_i \neq F_j$ for some $i \neq j$.

For X_i and X_j, where $i \neq j$, we can calculate the energy distance T_{ij} between X_i and X_j using (2.4). The k-sample test statistic is:

$$T^k = \sum_{1\le i<j\le k} T_{ij}. \tag{2.6}$$

Similar to in the two-sample case, we can get the critical value from the random permutations from $X_1, \ldots, X_k$.

Algorithm 1 Multidimensional scaling

1: Assign points to arbitrary coordinates in p-dimensional space
2: Compute the Euclidean distances among all pairs of points, to form the $\hat{D}$ matrix
3: Compare the $\hat{D}$ matrix with the input D matrix by evaluating the stress function. The smaller the value, the greater the correspondence between the two
4: Adjust coordinates of each point in the direction that best minimize the stress
5: Repeat step 2 through 4 until stress won't get any lower

Algorithm 2 K-means clustering

1: Set K-means $\{m^{(k)}\}$ to random values
2: Each data point $x^{(n)}$ is assigned to the nearest mean. We denote our guess of the cluster $k^{(n)}$ that the point $x^{(n)}$ belongs to by
$$\hat{k}^{(n)} = \arg\min_k \left(d\left(m^k, x^{(n)}\right)\right).$$
An alternative, equivalent representation of this assignment of points to cluster i given by "responsibilities", which are indicator variables: $R_k^n = 1$ if $\hat{k}^{(n)} = k$, $R_k^n = 0$ otherwise.
3: The model parameters, the means, are adjusted to mach the sample means of the data points that they are responsible for
$$m^{(k)} = \frac{\sum_n r_k^{(n)} x^{(n)}}{R^{(k)}},$$
where $R^{(k)}$ is the total responsibility of mean k,
$$R^{(k)} = \sum_n r_k^{(n)}.$$
4: Repeat the step 2 and step 3 until the assignments do not change

2.3. K-means clustering

According to the energy distance calculated by (2.2), we can perform clustering for different samples. Before clustering, we need to describe the relationship between the different samples based on the energy distance. The technique we use is the multidimensional scaling (MDS) (Kruskal, 1964).

The purpose of the MDS is to provide a visual representation for the pattern of proximities among a set of objects, when only a table of the distances between them is given. When the objects are income distributions, the MDS is used to reduce the matrix of distance to two dimensions, which can be plotted on a two-dimensional coordinate system. Similar functions are placed near each other on the map, while very different functions are placed far from each other. See Algorithm 1 for a brief description of the MDS. Based on the MDS, following the steps of K-means clustering in Algorithm 2, we are able to cluster income distributions from different regions using their energy distance matrix.

3. Simulation study

In this section, we set up different scenarios to assess the accuracy of hypothesis testing and K-means clustering using the methods described in Section 2. For hypothesis testing, we test cases of two-sample and multi-sample under the null and alternative hypotheses respectively. Different sample sizes are considered to compare the error rates under different circumstances. For K-means clustering,

Table 1: Error rates under null hypothesis: two-sample ($\alpha = 0.05$)

	Sample size k	20	40	60	80	100	150	200	250
$X_1 \sim$ Gamma$(2, 0.2)$ $X_2 \sim$ Gamma$(2, 0.2)$	Energy test	0.0524	0.0532	0.0518	0.0516	0.0494	0.0490	0.0486	0.0466
	Wilcoxon rank sum test	0.0528	0.0520	0.0518	0.0499	0.0516	0.0484	0.0502	0.0486
	Kolmogorov-Smirnov test	0.0464	0.0371	0.0354	0.0330	0.0244	0.0436	0.0402	0.0394
	Anderson-Darling test	0.0589	0.0531	0.0513	0.0502	0.0514	0.0494	0.0494	0.0462
	Jonckheere-Terpstra test	0.0522	0.0504	0.0488	0.0499	0.0536	0.0542	0.0518	0.0484
	Rank Score test	0.0560	0.0545	0.0494	0.0502	0.0482	0.0494	0.0470	0.0484
$X_1 \sim$ Log-Normal$(0, 1)$ $X_2 \sim$ Log-Normal$(0, 1)$	Energy test	0.0502	0.0488	0.0510	0.0506	0.0504	0.0492	0.0472	0.0474
	Wilcoxon rank sum test	0.0554	0.0472	0.0508	0.0500	0.0520	0.0482	0.0502	0.0478
	Kolmogorov-Smirnov test	0.0380	0.0334	0.0490	0.0330	0.0400	0.0422	0.0410	0.0386
	Anderson-Darling test	0.0546	0.0498	0.0530	0.0494	0.0550	0.0506	0.0504	0.0458
	Jonckheere-Terpstra test	0.0516	0.0430	0.0498	0.0454	0.0516	0.0514	0.0466	0.0452
	Rank Score test	0.0532	0.0444	0.0490	0.0510	0.0526	0.0482	0.0482	0.0492

we consider 2-means clustering procedure within different sample sizes.

3.1. Simulation under the null hypothesis

3.1.1. Two-sample test

We first consider the two-sample tests under two different circumstances. One case is that two groups of data $X_1 = (x_{11}, \ldots, x_{1k})^\top$ and $X_2 = (x_{21}, \ldots, x_{2k})^\top$ are generated from Gamma$(2, 0.2)$, and another case is that X_1 and X_2 are generated from Log-Normal$(0, 1)$. The test statistic based on energy distance (2.4) is then calculated along with other commonly used test statistics. We considered $k \in \{20, 40, \ldots, 250\}$ to illustrate the performances of hypothesis testing of different statistics under different sample sizes with a significance level of 0.05. For each k, a total of 5,000 replicates are performed. We report the error rates, i.e., empirical probabilities of falsely rejecting the null hypothesis, in Table 1. Generally, under these two circumstances, at small sample sizes, the error rates of the proposed test statistics are slightly above 0.05, and as the sample size increases, the error rates decrease and fluctuate around the nominal value 0.05. These results indicate that the proposed test holds its size under H_0 in the two-sample case. By comparing the results with those of the other traditional testing procedures, we can see that the Energy test outperforms the K-S test under both circumstances, and behaves similarly to the other four testing procedures. the results in Table 1, we can find that the Energy test has similar performance with other traditional testing procedure.

3.1.2. Multi-sample test

For the multi-sample scenario, the same simulation setting is used as the two-sample test in Section 3.1.1 except for the number of groups and the values of k. Here we generated five groups of data $X_1, \ldots, X_5$ and calculated the test statistics with $k \in \{10, 20, \ldots, 55\}$. The error rates under different testing procedures are presented in Table 2. Under both circumstances, the error rates of the Energy test fluctuate around 0.05 slightly. When the samples are generated from Gamma$(2, 0.2)$, the performance of the Energy test is better than the other tests when the sample sizes are larger than 40. Under the case of Log-Normal$(0, 1)$, the Energy test behaves similarly to the other testing procedures.

3.2. Simulation under alternative hypothesis

3.2.1. Two-sample test

In this section, we consider two-sample tests under the alternative hypothesis. We generated two groups of data, X_1 and X_2, with different sample sizes $k \in \{20, 40, \ldots, 250\}$. Two different circum-

Table 2: Error rates under null hypothesis: multi-sample

	Sample size k	10	20	30	40	45	50	55
$X_i \sim$ Gamma$(2, 0.2)$ $(i = 1, \ldots, 5)$	Energy test	0.0502	0.0497	0.0510	0.0476	0.0478	0.0488	0.0484
	Wilcoxon rank sum test	0.0466	0.0523	0.0492	0.0570	0.0468	0.0480	0.0464
	Anderson-Darling test	0.0510	0.0486	0.0504	0.0460	0.0476	0.0440	0.0440
	Jonckheere-Terpstra test	0.0492	0.0542	0.0502	0.0514	0.0466	0.0460	0.0476
	Rank Score test	0.0498	0.0484	0.0514	0.0470	0.0464	0.0424	0.0448
$X_i \sim$ Log-Normal$(0, 1)$ $(i = 1, \ldots, 5)$	Energy test	0.0488	0.0548	0.0462	0.0482	0.0492	0.0494	0.0484
	Wilcoxon rank sum test	0.0468	0.0540	0.0436	0.0506	0.0500	0.0560	0.0430
	Anderson-Darling test	0.0488	0.0566	0.0440	0.0524	0.0500	0.0536	0.0476
	Jonckheere-Terpstra test	0.0538	0.0494	0.0462	0.0498	0.0494	0.0534	0.0452
	Rank Score test	0.0508	0.0546	0.0436	0.0536	0.0514	0.0532	0.0504

Table 3: Error rates under alternative hypothesis: 2 groups

	Sample size k	20	40	60	80	100	150	200	250
$X_1 \sim$ Gamma$(2, 0.2)$ $X_2 \sim$ Gamma$(2, 0.3)$	Energy test	0.6308	0.3428	0.1746	0.0798	0.0354	0.0054	0.0004	0.0000
	Wilcoxon rank sum test	0.6460	0.3778	0.2038	0.0994	0.0474	0.0064	0.0010	0.0000
	Kolmogorov-Smirnov test	0.7612	0.5598	0.2944	0.2060	0.1212	0.0202	0.0046	0.0004
	Anderson-Darling test	0.6528	0.3786	0.1940	0.0936	0.0438	0.0070	0.0008	0.0001
	Jonckheere-Terpstra test	0.9992	0.9942	0.9814	0.9781	0.9583	0.9016	0.8875	0.8666
	Rank Score test	0.6360	0.3638	0.1860	0.0886	0.0384	0.0048	0.0006	0.0000
$X_1 \sim$ Log-Normal$(0, 1)$ $X_2 \sim$ Log-Normal$(0.5, 1)$	Energy test	0.7088	0.4920	0.3040	0.1870	0.1086	0.0240	0.0048	0.0010
	Wilcoxon rank sum test	0.6800	0.4476	0.2552	0.1590	0.0802	0.0184	0.0038	0.0006
	Kolmogorov-Smirnov test	0.7868	0.6398	0.3750	0.3004	0.1944	0.0482	0.0146	0.0036
	Anderson-Darling test	0.6836	0.4622	0.2654	0.1700	0.0876	0.0174	0.0040	0.0010
	Jonckheere-Terpstra test	0.5650	0.3214	0.1670	0.0906	0.0428	0.0060	0.0014	0.0000
	Rank Score test	0.6744	0.4324	0.2374	0.1418	0.0700	0.0126	0.0028	0.0004

stances are considered. The first one is that $X_1 \sim$ Gamma$(2, 0.2)$ and $X_2 \sim$ Gamma$(2, 0.3)$, and second case is that $X_1 \sim$ Log-Normal$(0, 1)$ and $X_2 \sim$ Log-Normal$(0.5, 1)$. The other simulation settings remain the same as Section 3.1.1. The Energy test and the other test procedures are used for hypothesis testing. Under the alternative hypothesis, the error rate is the empirical probability of failing to reject a false null hypothesis, which is equal to 1 minus the power of a test statistics. The error rates under different sample sizes are reported in Table 3. Under both cases, for all the testing procedures, the error rates are larger with smaller sample sizes, and decrease sharply when the sample sizes increase.

Under the first case, the Energy test behaves the best compared to the other testing procedures since it has the greatest power with different values of k. And under the second case, the Energy test still behaves better than the K-S test. Under these two circumstances, the power of the Energy test is greater than 0.95 when k reaches 100 and 150, respectively.

3.2.2. Multi-sample test

For multiple-sample testing, we generated five groups of data $X_1, \ldots, X_5$ with $k \in \{10, 20, \ldots, 55\}$ and kept the other simulation settings same as Section 3.2.1. For the first case, $X_1 \sim$ Gamma$(2, 0.2)$, $X_2 \sim$ Gamma$(2, 0.25)$, $X_3 \sim$ Gamma$(2, 0.3)$, $X_4 \sim$ Gamma$(2, 0.35)$, and $X_5 \sim$ Gamma$(2, 0.4)$. For the second case, $X_1 \sim$ Log-Normal$(0, 1)$, $X_2 \sim$ Log-Normal$(0.1, 1)$, $X_3 \sim$ Log-Normal$(0.3, 1)$, $X_4 \sim$ Log-Normal$(0.5, 1)$, and $X_5 \sim$ Log-Normal$(0.7, 1)$. Hypothesis testings are performed with $\alpha = 0.05$ and the corresponding error rates are reported in Table 4. Similarly, the error rates decrease as the sample sizes increase. Under the first case, the Energy test behaves the best under different values of the sample size, and reaches a power larger than 0.95 when the sample size in each group exceeds 40.

Table 4: Error rates for the proposed nonparametric test under the alternative hypothesis: multi-sample

	Sample size k	10	20	30	40	45	50	55
$X_i \sim \text{Gamma}(2, a_i)$ $a = (0.2, 0.25, 0.3, 0.35, 0.4)$	Energy test	0.6144	0.2646	0.0946	0.0220	0.0154	0.0074	0.0026
	Wilcoxon rank sum test	0.9968	0.9978	0.9992	0.9994	0.9998	0.9998	0.9998
	Anderson-Darling test	0.6420	0.3054	0.1194	0.0306	0.0202	0.0116	0.0036
	Jonckheere-Terpstra test	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Rank Score test	0.6216	0.2896	0.1048	0.0282	0.0176	0.0086	0.0038
$X_i \sim \text{Lognormal}(b_i, 1)$ $b = (0, 0.1, 0.3, 0.5, 0.7)$	Energy test	0.8118	0.6004	0.3972	0.2692	0.1940	0.1534	0.1138
	Wilcoxon rank sum test	0.6994	0.4516	0.2612	0.1564	0.1122	0.0794	0.0622
	Anderson-Darling test	0.7888	0.5540	0.3430	0.2034	0.1548	0.1076	0.0790
	Jonckheere-Terpstra test	0.4784	0.2128	0.0818	0.0332	0.0208	0.0118	0.0060
	Rank Score test	0.7780	0.5254	0.3068	0.1734	0.1272	0.0826	0.0616

Table 5: Error rate for the proposed nonparametric clustering procedure under the alternative hypothesis: multi-sample

Density 1	Gamma(2, 0.2)	Gamma(2, 0.2)	Gamma(2, 0.2)	Gamma(2, 0.2)	Gamma(2, 0.2)
Density 2	Gamma(2, 0.25)	Gamma(2, 0.3)	Gamma(2, 0.35)	Gamma(2, 0.4)	Gamma(2, 0.45)
$k = 500$	0.854	0.361	0.090	0.035	0.008
$k = 600$	0.823	0.303	0.073	0.012	0.001
$k = 700$	0.767	0.159	0.020	0.003	0.001
$k = 800$	0.721	0.133	0.015	0.001	0.001
$k = 900$	0.635	0.066	0.011	0.001	0.000
$k = 1000$	0.624	0.064	0.006	0.001	0.000
$k = 1500$	0.408	0.015	0.000	0.000	0.000

Under the second case, even though the Energy test cannot outperform the other testing procedures, it still has the power larger than 0.88 when the sample size reaches 55.

3.3. Simulation: K-means clustering

In this case, we test the accuracy of K-means clustering using energy distance through simulation. We considered the case where in each replicate, 60% of the data are generated from Gamma(2, 0.2), and the other 40% are generated from another Gamma distribution with different rates as displayed in Table 5. For each sample size $k \in \{500, 600, 700, 800, 900, 1000, 1500\}$, 1,000 replicates of simulations are performed. The error rate in clustering procedure is the proportion of observations that are incorrectly clustered. They are reported in Table 5. It can be seen that the accuracy is related to the sample size as well as the extent how different the two distributions are. With the same alternative distribution, the accuracy of clustering increases with larger sample size. With the same sample size, as the difference between the distributions gets larger, the accuracy of clustering increases.

4. Illustration: China Health and Nutrition Survey data

We apply the proposed methods to the analysis of a China Health and Nutrition Survey (CHNS) dataset for year 2011. The dataset contains information of 4,346 households in 12 provinces. Based on our simulation results, this sample size is big enough to have powerful testing results and clustering results. A brief description of the dataset by province is given in Table 6.

From the Gini coefficient calculated in Table 6, Beijing has the lowest Gini coefficient, which means that Beijing has the most balanced income distribution. On the contrary, Henan has the highest Gini coefficient, which means the income distribution in Henan is the least balanced. Most provinces have similar Gini coefficients, and it is difficult to set a threshold to say whether there are statistically

Table 6: Description of China Health and Nutrition Survey data

Province	Sample size	Average household income in thousands (SD)	Gini coefficient
Beijing	416	74.98 (49.9)	0.3236
Liaoling	395	49.43 (47.9)	0.3953
Heilongjiang	398	46.11 (44.4)	0.4307
Shanghai	424	87.46 (68.7)	0.3700
Jiangsu	414	60.93 (43.5)	0.3712
Shandong	399	41.00 (40.9)	0.4031
Henan	299	36.57 (42.7)	0.5027
Hubei	339	50.04 (57.3)	0.4142
Hunan	245	47.95 (43.5)	0.3986
Guangxi	362	36.81 (33.4)	0.3800
Guizhou	339	45.39 (52.7)	0.4172
Chongqing	329	40.98 (39.9)	0.4361

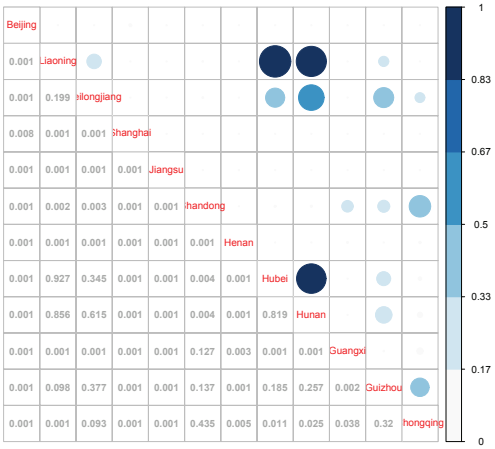

Figure 1: *A visualization of pairwise p-values.*

significant differences among these provinces.

4.1. Hypothesis testing results

We calculated the pairwise test statistic in (2.4) for the 12 provinces, giving rise to 66 statistic values. The resulted p-values are visualized in Figure 1. While most p-values are small, indicating a significant difference of the two distributions they correspond to, some p-values suggest that the two income distributions are not significantly different, such as Liaoning and Hubei, Liaoning with Hunan, and Hunan with Hubei.

Also, we do the multi-sample test for all the provinces using (2.6). The energy statistic of this data is calculated to be 87,944,000. Based on the permutation test procedure, the p-value is 0.000999. We therefore reject the null hypothesis, and conclude that not all income distributions are the same in these 12 provinces.

4.2. Clustering results

From the test results of previous analysis, we learned that the income distributions among 12 provinces are different, but smaller subsets of provinces have similar poverty pattern. Therefore, it is of interest

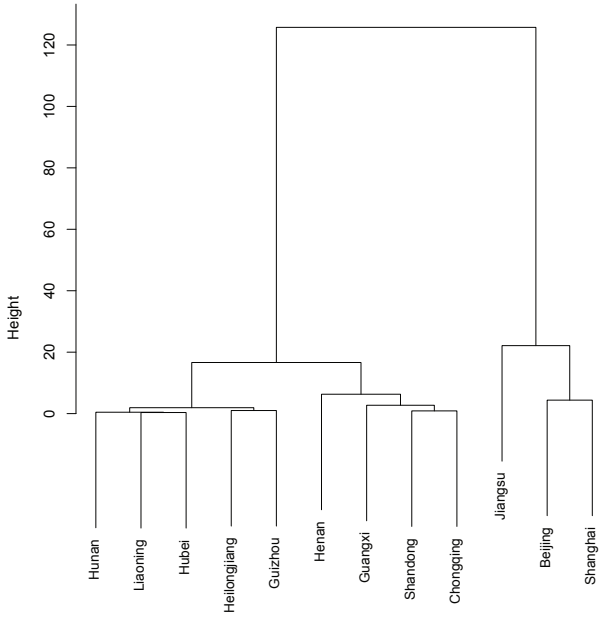

Figure 2: *Hierarchical clustering results.*

to further cluster the 12 provinces. Some exploratory data analysis is performed before clustering. We used the hierarchical clustering methods (Bibby *et al.*, 1979; Johnson and Wichern, 2007) based on the energy distances calculated using (2.1) to see the clustering pattern of different provinces. The result is plotted in Figure 2.

From Figure 2, we can easily find that there are four major clusters based on the energy distance. Therefore, we chose $k = 4$ for the clustering method described in Section 2.3, and the result is visualized in Figure 3.

From Figure 3, it is found that Beijing and Shanghai are in cluster 1. Jiangsu province is the only one province in cluster 2. Hunan, Hubei, Chongqing, Guizhou, Heilongjiang, and Liaoning are in cluster 3, and Henan, Guangxi, and Shandong are in cluster 4. These clustering results are consistent with explanatory analysis based on the Gini coefficient calculation and testing results. Also, the clustering results reflects actual development of economy in these provinces. Beijing, Shanghai, and Jiangsu have relatively better economic development, while Henan, Guangxi, and Shandong have worse economic development.

5. Discussion

We developed nonparametric test and clustering procedure for income distribution in different geographical regions. The test was inspired by the energy distance between distributions (Székely and Rizzo, 2004), and the MDS (Kruskal, 1964) for reducing the dimension of data based on a distance matrix so that an appropriate clustering algorithm can be implemented. In simulation studies, the nonparametric test has size close to the nominal 0.05 level. When the samples are generated from different distributions, or the same distribution with different parameters, the proposed test has strong power in detecting the violation of the null hypothesis, and the power is even stronger than other frequently

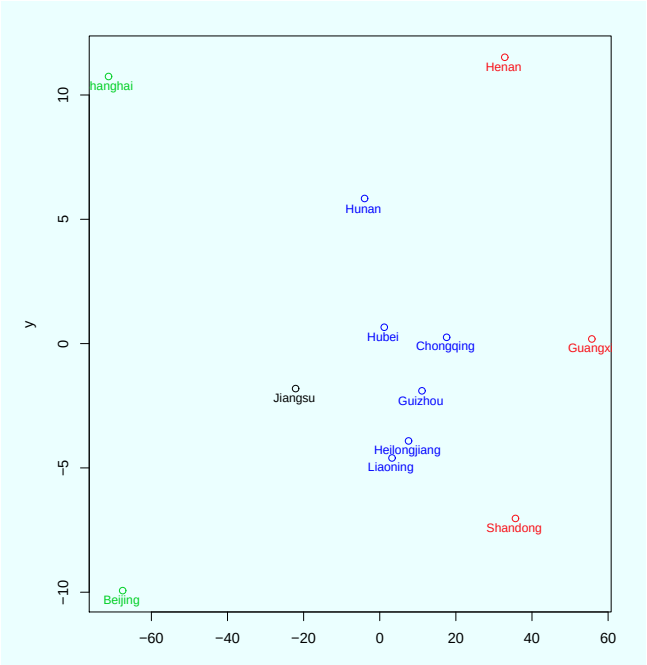

Figure 3: *4-means clustering results.*

used testing procedures. In the application to CHNS data, the proposed methods rejected the null hypothesis, and identified four clusters of income distribution, which is robust against choice of initial centroids and consistent with the benchmark hierarchical clustering results. In this work, there are no significant differences between the results of hierarchical clustering and K-means clustering because of the relative small number of provinces we analyzed. If we compare the county level data, however, it is very difficult for us to interpret results form the dendrograms. K-means clustering method is more robust than the HC.

In this work, we are only concerned with using household incomes to calculate the energy distance between different regions. There could be other information sources, such as the value of property, amount of mortgage, and consumptions, that can be used to provide more accurate description of the distribution of income, as the energy distribution provides good measurement of high-dimensional cumulative distribution functions. Geographical information, in addition, can be taken into account in comparing income distributions.

Acknowledgement

Dr. Hu's research was supported by Dean's office of College of Liberal Arts and Sciences in University of Connecticut.

References

Bartels CP and Van Metelen H (1975). *Alternative probability density functions of income: A comparison of the lognormal-, Gamma-and Weibull-distribution with Dutch data*, Vrije Universiteit, Economische Faculteit.

Bibby J, Kent J, and Mardia K (1979). *Multivariate Analysis*, Academic Press, London.

Gibrat R (1931). *Les inégalités économiques*, Recueil Sirey.

Johnson R and Wichern D (2007). Discrimination and classification, *Applied Multivariate Statistical Analysis*, **4**.

Kruskal JB (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1–27.

Li S and Rizzo ML (2017). K-groups: a generalization of K-means clustering, *ArXiv e-prints*.

Lorenz MO (1905). Methods of measuring the concentration of wealth, *Publications of the American Statistical Association*, **9**, 209–219.

McDonald JB (1984). Some generalized functions for the size distribution of income, *Econometrica: Journal of the Econometric Society*, **52**, 647–665.

McDonald JB and Xu YJ (1995). A generalization of the beta distribution with applications, *Journal of Econometrics*, **66**, 133–152.

Pareto V (1964). *Cours d'économie politique*, volume 1, Librairie Droz.

Rizzo ML and Székely GJ (2010). Disco analysis: a nonparametric extension of analysis of variance, *The Annals of Applied Statistics*, **4**, 1034–1055.

Rizzo ML and Székely GJ (2015). Energy distance, *Wiley Interdisciplinary Reviews: Computational Statistics*, **8**, 27–38.

Salem ABZ and Mount TD (1974). A convenient descriptive model of income distribution: the gamma density, *Econometrica: Journal of the Econometric Society*, **42**, 1115–1127.

Sullivan A (2003). Economics: Principles in action.

Székely GJ and Rizzo ML (2004). Testing for equal distributions in high dimension, *InterStat*, **5**, 2004.

Yitzhaki S (1979). Relative deprivation and the Gini coefficient, *The Quarterly Journal of Economics*, **93**, 321–324.

Received October 23, 2018; Revised December 4, 2018; Accepted December 5, 2018