Jour. Ind. Soc. Ag. Statistics
53(1), 2000 : 1–11

The Common Principal Components Approach for Clustering under Multiple Sampling

B.S. Kulkarni and G. Nageswara Rao
Acharya N.G. Ranga Agricultural University, Hyderabad 500030
(Received : August, 1998)

SUMMARY

The suitability of using Common Principal Components (CPC) for investigating the crucial assumption of homogeneity of covariance matrices (which is not satisfied in general) is examined in the context of clustering under the multi-sample situation. The approach mitigates the crucial assumption by providing a "common" (pooled) estimate for the principal components of the objects instead of a pooled estimate for the covariance matrices of the objects. An illustration of the approach with regard to the rainfall based classification of the districts of Andhra Pradesh State is described.

Key words : Cluster analysis, Homogeneity of covariance matrices, Principal components, Common principal components.

1. Introduction

Homogeneity of covariance matrices has been a crucial assumption in several multivariate statistical analysis procedures. Attempts at providing a solution to the situation arising out of heterogeneity of covariance matrices are very few.

The Common Principal Component (CPC) analysis (Krzanowski [3]) can be considered as a way to deal with the problem of heterogeneity of covariance matrices. The CPCs provide a pooled estimate for the principal components of the covariance matrices which are heterogeneous. An application of the procedure to the discriminant analysis (when the covariance matrices are heterogeneous) has been attempted by Darghai-Noubary [1] and Owen [4]. These researchers observed that for the large samples, the performance of discrimination procedure based on the CPCs tends to be better than the conventional procedures such as the Quadratic or Linear discriminant functions.

In the present study, the applicability of common principal components in clustering is examined for the multi-sample case.

2. *Material and Methods*

The Multi-Sample Case: Let there be m "objects" which are to be clustered into g (< m) homogeneous groups. Suppose that the j-th object has N_j observations (j = 1, ..., m) which are recorded by drawing a random sample of size N_j from it. Let X be the random vector consisting of k variables, then X_{ij} represents the i-th observation vector on the j-th object (i = 1, ..., N_j, j = 1, ..., m). Thus on the basis of the observation vector X_{ij} the m objects are to be classified into g (<m) distinct groups. The objects can be conveniently clustered by applying the canonical approach.

The Canonical Approach: The approach involves extracting first p (<k) latent vectors α_j (j = 1, ..., p) of the matrix product $(W^{-1}B)$; where the matrices B and W form the following identity as a result of a one-way MANOVA model assumption

$$T = B + W$$

where

T = Matrix of Total Sum of Squares and Products (S.S.P.)

B = Matrix of "Between Samples" S.S.P., and

W = Matrix of "Within Samples" S.S.P.

Note that W is the Pooled Matrix of S.S.P. of the multi-samples.

The latent vector α_j is also referred to as canonical variable, $Y_j = \alpha_j' X$ (j = 1, ... , p). Each of the i-th object can be represented by the p-dimensional co-ordinates $\alpha_1'X_i$, $\alpha_2'X_i$ ..., $\alpha_p'X_i$; (i = 1, ..., m). For the purpose of clustering, the canonical means (scores) of the objects are obtained on the basis of the means of the objects ($\overline{X}_i$), i.e., $\alpha_1'\overline{X}_i$, $\alpha_2'\overline{X}_i$ ..., $\alpha_p'\overline{X}_i$; (i = 1, ..., m). The objects are then grouped into clusters on the basis of similarities in the co-ordinates of the objects in the p-dimension. For $p \leq 3$, it can be conveniently carried out by plotting the co-ordinates.

The canonical approach thus involves use of the common (pooled) covariance matrix (W/df) for determining the clusters. However, the use of pooled covariance matrix is reasonable only when the covariance matrices (S.S.P.) of the samples satisfy the homogeneity condition. Thus when the covariance matrices of the objects are heterogeneous, as an alternative, let us consider the applicability of Common Principal Components approach for obtaining the clusters.

The Common Principal Components: This approach involves determining a vector subspace which represents the vector subspaces of all the objects as closely as possible. Several developments have taken place in this field (Flury [2]). The present study considers the procedure developed by Krzanowski [3]. Suppose that principal component analysis has been carried out for each of the m objects. Furthermore, the first p (< k) principal components are adequate for summarising the total variance of each of the covariance matrices. Let L_t (p × k) be the matrix of these vectors corresponding to the t-th object (t = 1, ..., m) whose rows are the eigen vectors of the p-principal components. Let $H(k \times k) = \Sigma L_t' L_t$. Then the first q (< k) principal components of H represent the "common principal components". Following inferences were drawn about these CPCs (Krzanowski [3]):

1. Let b be an arbitrary vector in the original k-dimensional data space and δ_t, the angle between b and the vector most nearly parallel to it in the space generated by the first p-components of the t-th object. Then the value of b that minimizes $V = \Sigma \cos^2\delta_t$ is given by the latent vector b_1 corresponding to the largest latent root $\cap_1$ of H.
2. Obviously, corresponding to b_1 the value of V (= V_1) is given by $\cap_1$, which is the largest latent root of H.
3. V represents the measure of closeness of the components of H to all the p-dimensional subspaces of the objects. The average component that agrees most closely with all the m subspaces (i.e., the principal components) of the objects is given by b_1. Thus the measure of discrepancy between b_1 and the subspace of the t-th object is given by $\cos \delta_t = \sqrt{[b_1' L_t' L_t b_1]}$
4. Completing the latent root and vector analysis of H leads to a subspace of dimension q (< k) that represents all the m subspaces of the objects as closely as possible.

The basis for CPCs is the closeness of the estimated plane of H with the subspaces of the objects. Since the covariance matrices of the objects are non-homogeneous, it is obvious that their subspaces would not be comparable (similar) with all the components of H, but with only the first few components, q (< k).

These components can be readily identified on the basis of the V_1 values and also the angular separation (δ_t) between the latent vector b_1 and the subspace of the t-th object (i = 1, ..., k, t = 1, ..., m). These q identified components thus provide a common estimate for the principal components of the m objects.

The q CPCs, as identified above, can be used for obtaining the component scores for the objects (based on their mean values) in the q-dimensional plane of H. Clustering can then be carried out by grouping the objects with similar scores in the q-dimensional plane.

The approach based on CPC thus duly considers the sampling variations within the objects by estimating the individual component subspaces and then a common subspace for clustering.

Application

The approach outlined above is applied for obtaining the clustering of the districts of Andhra Pradesh (A.P.) State on the basis of rainfall of South-West monsoon season. The observation vector is the monthly rainfall from June to September. Rainfall data of 30 years from 1961-62 to 1990-91 has been used for the purpose of this clustering. The clustering is restricted to the 20 districts (instead of 23 districts) due to non-availability of the complete data on the 3 newly formed districts (Vizianagaram, Prakasam and Ranga Reddy). Thus there are 20 objects (i.e., the districts) each having 30 sample observations corresponding to the 30 years.

For comparison, clustering was also obtained by applying the canonical approach, as described earlier.

The relevant rainfall data were collected from the Season and Crop Reports and Statistical Abstracts of A.P. State.

3. Results and Discussion

Andhra Pradesh State is classified into 3 administrative regions, viz. Coastal Andhra (7 districts), Rayalseema (4 districts) and Telangana (9 districts). South West monsoon is the main rainy season of the state. There is a considerable disparity in the average rainfall received in the districts as well as its consistency over time (Table 1).

Clustering with Common Principal Components: It was found that the covariance matrices of all the 20 districts were significantly different, implying heterogeneity among them (Chi-square Statistic = 587.20, df = 190, significant at 0.01 level of probability).

For obtaining the CPCs, each of the covariance matrices of the districts were subjected to principal component analysis. It was observed that the first 3 PCs accounted for atleast 86 percent of the sample variance and so adequately summarized the total variance of the 4 rainfall variates in all the 20 districts. Hence the matrix L_t ($t = 1, ..., 20$) was defined on the basis of the first 3 components.

Table 1 : Average monthly rainfall of Andhra Pradesh South-West monsoon (1961–62 to 1990–91)

	District	June	July	August	September
	I Coastal Andhra Region				
1.	Srikakulam (SRK)				
	Mean (mm)	145.96	178.54	182.79	199.08
	C.V. (%)	45.55	33.34	34.44	44.84
2.	Visakhapatnam (VZG)				
	Mean (mm)	123.67	156.12	173.83	168.41
	C.V. (%)	43.60	32.58	41.12	36.48
3.	East-Godavari (EGD)				
	Mean (mm)	128.67	199.33	201.17	165.29
	C.V. (%)	50.44	48.08	47.70	39.82
4.	West-Godavari (WGD)				
	Mean (mm)	130.00	212.54	219.75	160.58
	C.V. (%)	65.71	49.99	51.95	38.25
5.	Krishna (KRSN)				
	Mean (mm)	110.46	199.88	189.83	156.75
	C.V. (%)	42.01	49.72	49.32	48.18
6.	Guntur (GNTR)				
	Mean (mm)	84.12	151.54	142.83	141.79
	C.V. (%)	43.01	54.60	55.88	50.67
7.	Nellore (NLR)				
	Mean (mm)	39.08	92.63	88.21	109.54
	C.V. (%)	42.58	48.02	76.25	59.43
	II Rayalseem Region				
8.	Kurnool (KRNL)				
	Mean (mm)	72.29	111.58	122.83	142.38
	C.V. (%)	28.59	51.72	73.80	43.84
9.	Anantapur (ATP)				
	Mean (mm)	47.42	67.29	69.42	143.25
	C.V. (%)	45.32	89.73	78.50	49.83
10.	Cuddapah (CDP)				
	Mean (mm)	58.00	105.21	101.54	135.67
	C.V. (%)	38.80	67.07	74.84	53.21
11.	Chittoor (CHTR)				
	Mean (mm)	54.13	106.04	95.83	141.42
	C.V. (%)	38.43	46.34	60.27	42.99

Table 1 contd.....

	District	June	July	August	September
	III Telangana Region				
12.	Hyderabad (HYD)				
	Mean (mm)	121.58	181.00	175.17	159.21
	C.V. (%)	31.18	53.56	58.85	60.45
13.	Nizamabad (NZB)				
	Mean (mm)	172.63	298.25	315.33	180.54
	C.V. (%)	56.44	49.89	59.50	76.86
14.	Medak (MDK)				
	Mean (mm)	139.46	237.25	228.58	169.25
	C.V. (%)	38.38	60.47	51.57	59.24
15.	Mehaboobnagar (MBNR)				
	Mean (mm)	84.00	144.17	151.21	147.46
	C.V. (%)	25.23	47.01	55.42	43.72
16.	Nalagonda (NLG)				
	Mean (mm)	101.92	152.00	141.21	140.12
	C.V. (%)	41.67	56.67	44.25	59.51
17.	Warangal (WGL)				
	Mean (mm)	151.54	272.33	229.25	152.25
	C.V. (%)	48.51	53.13	44.46	59.83
18.	Khammam (KHM)				
	Mean (mm)	149.54	291.75	253.50	170.63
	C.V. (%)	43.94	41.10	45.02	40.95
19.	Karimnagar (KRMN)				
	Mean (mm)	160.46	262.83	250.38	152.17
	C.V. (%)	40.36	49.84	52.05	50.11
20.	Adilabad (ADB)				
	Mean (mm)	180.04	306.13	306.92	154.75
	C.V. (%)	44.93	57.08	50.87	63.46

The results of the principal component analysis of the matrix $H\ (= \Sigma L'_t L_t)$ which gives the common principal components are presented in Tables 2a and 2b. It can be observed that the latent roots of H, which represent the measure of similarity between the CPC and vector subspaces of all the 20 districts, are almost similar corresponding of the first 3 components, ($\cap_1 = 19.58$, $\cap_2 = 19.31$ and $\cap_3 = 18.31$, Table 2a) whereas it is considerably low in the fourth component ($\cap_4 = 2.85$). The results thus indicate that all

the districts are close together along the first 3 CPCs (i.e., the first three components of H).

An inspection of the angular separation of the subspaces of the districts with the subspaces of CPCs reveal that these angles are considerably low

Table 2 a : Common Principal Components for A.P. districts

Variable	Common Principal Components			
	1	2	3	4
	(Vector Coefficients)			
June	0.1123	0.1832	0.0505	0.9753
July	–0.9380	0.2980	–0.1663	0.0606
August	0.1444	0.8003	0.5483	–0.1953
September	–0.2944	–0.4870	0.8180	0.0830
Latent Root	19.5840	19.3079	18.3091	2.8461

Table 2 b : Angular separation of the districts with the CPCs

	District	Common Principal Components			
		1	2	3	4
		(Angles in Degrees)			
1.	Srikakulam	8.62	38.65	14.42	46.46
2.	Visakhapatanam	26.19	0.81	1.15	63.66
3.	East Godavari	13.44	7.43	0	74.57
4.	West Godavari	8.89	34.88	45.81	18.27
5.	Krishna	2.81	8.70	13.88	73.75
6.	Guntur	4.80	5.06	4.58	81.66
7.	Nellore	6.59	6.28	6.12	79.12
8.	Kurnool	0.81	14.96	4.05	74.45
9.	Ananatapur	0.81	16.68	2.56	73.13
10.	Cuddapah	4.44	9.60	9.07	76.01
11.	Chittoor	0	10.95	6.07	77.41
12.	Hyderabad	5.00	3.71	0	83.38
13.	Nizamabad	10.26	16.68	19.37	61.74
14.	Medak	3.34	3.71	7.30	81.11
15.	Mehaboobnagar	10.20	10.70	3.34	74.78
16.	Nalagonda	3.97	6.97	13.90	73.89
17.	Warangal	0	0	55.54	65.84
18.	Khammam	1.98	2.56	20.90	68.85
19.	Karimnagar	7.95	11.62	0.81	75.85
20.	Adilabad	1.62	10.98	3.14	78.43

corresponding to the first three CPCs (Table 2b) with the exception of the only district of Visakhapatanam (26.19°)[1] corresponding to the first CPC, the 2 districts of Srikakulam (38.65°) and West Godavari (34.38°) corresponding to the second CPC and the three districts of West Godavari (45.81°), Warangal (55.54°) and Khammam (20.90°) corresponding to the third CPC. These separations further increase corresponding to the fourth CPC (Table 2b).

The vector subspaces of the CPCs (Table 2a) indicate that the vector subspace of the first CPC is heavily loaded on July rainfall (Vector Coeff. = – 0.9380); while rainfall of August (Vector Coeff. = 0.8003) and September (Vector Coeff. = 0.8180) are dominant respectively in second and third CPCs. This behaviour has been suitably exhibited by all the districts in their three-dimensional subspaces (i.e., the 3 principal components), as follow:

July rainfall is dominant among the 8 districts in their first component vector subspaces, the other 5 districts in their second components (vector subspaces) and the remaining 7 districts in their third component subspaces. Similar is the case with rainfall of August and September which were dominant in respectively the second and third CPCs.

Contrary to the above behaviour, the rainfall of June which is dominant in the fourth CPC is represented by only 2 districts in their individual component subspaces.

These results indicate that only the first three CPCs can be considered common to all the vector subspaces of the districts. These three components also reveal the common cause for variation in the rainfall of the districts viz. rainfall of July, August and September.

The three CPCs identified above formed the basis for clustering (being the common or pooled estimates of principal components of the 20 districts). The component scores corresponding to the districts were then obtained on the basis of the mean vectors of the districts (in the three dimensional plane of the CPCs). For the purpose of clustering, the districts with similar scores in the three-dimensional plane of the CPCs were then grouped. This provided a classification of the districts into 11 clusters consisting of 4 single district clusters (Table 3). The basis of classification (as identified in the CPCs) i.e., rainfall of July and August can be readily observed in the clusters. For instance, the cluster formed with the districts of Mehaboobnagar, Nalagonda and Guntur can be seen to be almost closer to each other with regard to the rainfall of

1 Angles as low as 18° have been considered to be within the 95 per cent Monte-Carlo limits (Krzanowski 1979).

Table 3 : Clustering of A.P. districts

Cluster No.	Clustering with	
	Common Principal Components Approach	Canonical Approach
1.	ATP	–
2.	CDP, CHTR, KRNL	CDP, CHTR, KRNL
3.	MBNR, NLG, GNTR	MBNR, NLG, GNTR
4.	SRK	MDK
5.	NLR	NLR, ATP
6.	VZG, HYD	VZG, HYD
7.	EGD, KRSN	EGD, KRSN, SRK
8.	WGD, MDK	WGD
9.	KRMN, WGL	KRMN, WGL, KHM
10.	NZB, ADB	NZB, ADB
11.	KHM	–

July and August (Table 1). Similar observations can be had from the other clusters.

These clusters can be compared with the agro-climatic zones formed by the National Agricultural Research Project (NARP), A.N.G.R. Agricultural University, Hyderabad. Andhra Pradesh State is classified into 7 agro-climatic zones based on the climate, soil types, irrigation and cropping pattern (NARP Status Report, [5]).

Table 4 gives the agroclimatic zones and the districts covered under each zone as well as the clusters formed with the CPC approach alongwith the districts covered under each cluster.

It can be observed from Table 4 that there are only two deviations in classification under the CPC approach vis-a-vis the agro-climatic classification: Medak (MDK) has been classified along with West Godavari (WGO) in Cluster No. 8; whereas, West Godavari is classified in the Krishna Godavari Zone and Medak in the Northern Telangana Zone. Also, Hyderabad (HYD) and Visakhapatnam (VZG) are classified in Cluster No. 6. However, these districts belong to respectively the Southern Telangana and the North Coastal Zones under the agro-climatic classification. The agro-climatic zones are classified not only on the basis of the climate, soil types, irrigation and cropping pattern but also on the basis of the geographical contiguity for administrative convenience.

Table 4 : Agro-climatic classification of Andhra Pradesh State vis-a-vis clusters with CPC approach

S. No.	Classification with Agro-climatic base		Classification with CPC Approach	
	Agro-climatic Zones	Districts Covered	Cluster No**	Districts Covered
1.	Krishna-Godavari Zone	WGD, KRSN, GNTR, EGD, NLG, PRKM*	Cluster No. 7 Cluster No. 3 Cluster No. 8	EGD, KRSN NLP, GNTR, MBNR WGD, MDK
2.	North-Coastal Zone	Most Parts of SRK, VZNGM*, VZG, Uplands of EGO	Cluster No. 4	SRK
3.	Southern Zone	NLR, CHTR, CDP, Southern parts of PRKM*, Eastern parts of ATP	Cluster No. 5 Cluster No. 2	NLR CHTR, COP, KRNL
4.	Northern Telangana Zone	ADB, KRMN, NZB, MDK, WGL, NLG, KHM	Cluster No. 9 Cluster No. 10 Cluster No. 11	KRMN, WGL NZB , ADB KHM
5.	Southern Telangana Zone	HYD, RR*, NBNR, NLG	Cluster No. 6	HYD, YZG
6.	Scarce Rainfall Zone	KRNL, ATP	Cluster No. 1	ATP
7.	High Altitude and Tribal Zone	High Altitudes of SRK, VZNGM*, VZG, EGD, KHM	–	–

** Table 3

* The districts of Vizianagaram (VZNGM), Prakasan (PRKM) and RR (Ranga Reddy) are excluded in the analysis due to incomplete data.

The clusters formed with the CPC approach do not consider such contiguity. Barring these two deviations, most of the clusters are in agreement with the agro-climatic classification.

Clustering with the Canonical Approach: The canonical analysis revealed that about 93 per cent of the variation in $W^{-1}B$ is accounted by the first canonical root (Table 5). Hence clustering was carried out on the basis of the canonical scores corresponding to only the first component. The vector coefficients of the first component indicate that July rainfall is the main cause of variability (Vector Coeff. = 0.71), followed by rainfall of June

Table 5 : Canonical analysis

Component	Canonical Cum		Vector Coefficients			
	Root	(%)	June	July	August	Sept.
Component I	51.01	93.21	0.55	0.71	0.40	0.11
Component II	2.89	98.49	0.45	–0.35	–0.22	0.79

(Vector Coeff. = 0.55). The clusterings obtained with this approach were entirely different from those based on the CPC approach. The deviation is obviously due to taking a pooled estimate W of the covariance matrices of the objects (assuming that these are "homogeneous"). However, there are instances of similarities with regard to the clusters of (Cuddapah, Chittoor, Kurnool), (Mehaboobnagar, Nalagonda, Guntur), (Nizamabad, Adilabad) and (Visakhapatanam, Hyderabad). It can be mentioned, however, that the clustering obtained with this approach are not meaningful as the covariance matrices of the districts are heterogeneous.

In conclusion, it can be said that the common principal components provide a useful solution for dealing with the heterogeneous covariance matrices and in particular for obtaining the clustering under the multi-sample situation when the covariance matrices of the objects are heterogeneous.

REFERENCES

[1] Dargahi-Noubary, G.R., 1981. An application of discrimination when covariance matrices are proportional. *Australian J. Stat.*, **23**, 38-44.

[2] Flurry, B., 1988. *Common Principal Components and Related Multivariate Models*. John Wiley and Sons, New York.

[3] Krzanowski, W.J., 1979. Between groups comparison of principal components. *J. Amer. Stat. Assoc.*, **74**, 703-707.

[4] Owen, A., 1984. A neighbourhood-based classifier for LANDSAT data. *The Canadian J. Stat.*, **12**, 191-200.

[5] Status Report, 1989. National Agricultural Research Project - Southern Zone Volume 1, Regional Agricultural Research Station, Andhra Pradesh Agricultural University, Tirupati.