- تحلیل مولفه‌های اصلی

تحلیل مولفه های اصلی یکی از تکنیک های کاهش داده ها می باشد که مجموعه متغیرهای اصلی را به مجموعه کوچکتری تبدیل می کند به طوریکه این مجموعه کوچک علت بیشتر واریانس موجود در داده هاست. هدف از تحلیل مولفه های اصلی آن است که واریانس موجود در داده های چندین متغیر را به مولفه هایی تجزیه می کند که اولین مولفه تا آن جا که ممکن است علت بیشترین واریانس موجود در داده ها باشد. دومین مولفه علت بیشترین واریانس ممکن بعد از مولفه اول باشد و الی آخر (فرشادفر ، ۱۳۸۴). به علاوه در این روش هر مولفه مستقل از مولفه ی دیگر است و بین مولفه ها همبستگی وجود ندارد. بنابراین اگر متغیرهای X_1، $X2$ و Xp را مورد بررسی قرار دهیم، تابع خطی:

$PC_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{P1}X_P$	

را اولین مولفه ی اصلی گویند. به همین ترتیب دومین مولفه ی اصلی به صورت زیر خواهد بود :

$PC_2 = a_{12}X_1 + a_{22}X_2 + ... + a_{P2}X_P$	

و الی آخر به طوری که به تعداد متغیرها مولفه خواهیم داشت. با این تفاوت که همان چند مولفه اول بیشتر تغییرات را توجیه و تبیین خواهند کرد. در این معادلات a_{ij} مثل ضرایب رگرسیون ثابت هستند. بنابرین در ارتباط با توجیه واریانس متغیرها توسط مولفه ها ترکیب زیر را می توان نوشت:

$Var(PC_1) \geq Var(PC_2) \geq ... \geq Var(PC_P)$	

لازم به توضیح است که اگر بین متغیرهای اولیه همبستگی وجود نداشته باشد ، این تجزیه مطلقا بی ارزش خواهد بود. مراحل اساسی در تحلیل مولفه های اصلی :

- مرتب نمودن ماتریس داده های خام شامل بارش و مشخصات آن
- تهیه ماتریس همبستگی(R) از روی داده های خام
- محاسبه مولفه های اصلی (PC)

راه حل پایانی (چرخش یافته) با توجه به نتایج حاصل از راه حل اولیه گام ۳ بدست می آید. مولفه های حاصل از تحلیل اولیه معمولا چرخش داده میشوند تا مطلوبترین طرح که به آسانترین شیوه قابل تفسیر باشد به دست آید. چرخش محورها ساختار مولفه ها را ساده می کند. مولفه های چرخش یافته که از راه های متفاوت به دست می آید ، ممکن است الگوی معنا داری از متغیر به دست دهد. توجه داشته باشیم که یک تحلیل مولفه اصلی به گونه نرمال شامل دست کم دو راه حل متمایز (هرچند مرتبط با یکدیگر) اولیه (چرخش نایافته) و پایانی (چرخش یافته) است. متغیر ها (داده های خام) می بایست با یکدیگر همبستگی داشته باشند تا تحلیل مولفه های اصلی با موفقیت انجام شود.

کارگاه برنامه نویسی دانشگاه خوارزمی- دوستکامیان- جلسه هشت و نه

دستور اول

[pc, score, latent] = princomp(X) Var=cumsum (latent)./sum(latent);	
ضرایب بار عاملی	Pc
بار عاملی	Score
مقدار ویژه	Latent
درصد توجیه واریانس	Var

دستور دوم

علاوه بر دستور که گفته شد می توات با اجرای دستور زیر اطلاعات کامل تری نسبت pc ها بدست آورد:

[Coeff, score, latent, tsquared, explained, mu] = pca(x);	
توضیحات	فراسنج
ضرایب عامل ها	Coeff
بارهای عاملی	score
مقدار ویژه	latent
مجموع مربعات نمرات استاندارد شده برای هر مشاهده	tsquared
درصد از واریانس کل توضیح داده شده توسط هر جزء اصلی	explained
میانگین متغیرهای اصلی	mu

دستور سومی

[coeff, score, pcvar, mu, v] = ppca(x,k);	
توضیحات	**فراسنج**
متغیرهای ورودی	
اسم ماتریس ورودی مورد مطالعه	x
تعداد عامل‌های، که باید کمتر یا برابر با تعداد متغیرها تعیین شود	k
متغیرهای خروجی	
ضرایب عامل ها	Coeff
بارهای عاملی	score
مقدار ویژه	pcvar
میانگین متغیرهای اصلی	mu
واریانس باقیمانده همسانگرد، به عنوان یک ارزش عددی	v

مثال- در جدول زیر داده‌های ۱۰ متغیر اقلیمی در ایستگاه زنجان را نشان می‌دهد. در این مثال قصد داریم با

استفاده از تحلیل مولفه‌های اصلی مهمترین عامل را مشخص کنیم. البته باید توجه کرد تعداد متغیرها در انتخاب نتیجه بهینه مناسبا است. برای مثال بعضی از آمار دانان معتقدند که هرچه که تعداد متغیرها کمتر باشد اجرای تحلیل چند متغیره نایج خیلی مناسبی ارایه نمی دهد دلیل ان هم این است که این روش‌ها و تکنیک ها برای متغیرهای بسیار زیاد نتیجه مناب تری می‌دهند. بنابراین دقت کنید که هرچه که متغیرها مورد مطالعه بیشتر باشد نتایج کار مناسب تر و دقیق تر می‌باشد.

جدول--: متغیر های مورد مطالعه

Rain	tem	t- max	t-min	hum	h- max	h- min	wind	mix	dew
222.9	10.3	17.8	2.8	52	73	32	3.6	4.72	-1.2
339.7	9.9	16.8	3	55	77	34	3.8	4.82	-0.7
309.3	8.4	15	1.8	57	80	36	3.5	4.78	-1.3
360.2	9.7	16.4	3.1	57	79	35	4.6	4.85	-0.7
415.5	10.6	17.4	3.9	57	81	34	4.7	5.28	0.6
242.8	11	18.5	3.5	53	78	30	3.8	5.01	-0.4
295.6	10.9	17.6	4.3	54	76	32	4.8	4.84	-0.5
260.6	10.7	17.9	3.6	54	77	32	6.7	4.9	-0.5
257.7	11.4	18.4	4.5	55	76	34	6.5	5.22	0.4
187.5	11.9	19.2	4.7	51	74	30	6.6	4.86	-0.3
309.7	11.3	18.1	4.4	53	75	33	6.5	4.96	-0.1
201.1	12.2	19.3	5.2	50	72	30	5.9	4.93	-0.1
255.4	11.4	18.4	4.4	54	74	34	5.9	5.13	0.3
354.9	11.5	18.2	4.8	57	79	36	6	5.62	1.5
323.8	11.5	18.5	4.5	58	79	37	5.7	5.65	1.6
251.3	11.3	18.6	4.1	59	80	37	4.9	5.92	1.9
328	11.7	18.6	4.7	57	78	37	5.3	5.87	1.8
381.6	11	17.9	4.1	59	81	37	5.2	5.86	1.6
167.6	11.4	18.6	4.1	55	76	35	5.3	5.69	1
306.3	11.5	18.3	4.6	57	79	36	5.2	5.73	1.7
259.8	13.3	20.7	5.8	52	73	31	4.8	5.54	1.5

حل

از انجای که واحد داده‌ها یکسان نیستد با استفاده از تابع zscore داده‌ها را به نمره استاندارد تبدیل می‌کنیم که برای این کار از مطابق زیر عمل خواهیم کرد:

```
>> x=zscore (data);
```

وقتی که داده‌ها را به نمره z تبدیل کرده‌اید با استفاده از دستور زیر تحلیل مولفه‌ها را انجام دهید:

```
>> [pc, score, latent] = princomp (x);
```

دستور فوق به طور خودکار مولفه‌ها بر اساس ضریب همبستگی محاسبه می‌کند. بنابراین درصورتی که بخواهیم از ماتریس کواریانس برای استخراج مولفه ها استفاده کنیم آنوقت دستور محاسبه تحلیل مولفه‌ها به شکل زیر خواهد بود:

```
>> [COEFF, latent, explained] = pcacov(V)
```

یا می توان همان دستور princomp را به شکل زیر نوشت:

```
>> [pc, score, latent] = princomp (cov(x));
```

نتایج حاصل از دستورات بالا به شرح زیر است. در این دستور نمرات بار عاملی را درpc و واریانس را در latent ایجاد می‌کند:

>> pc

	عامل اول	عامل دوم	عامل سوم	عامل چهارم	عامل پنجم	عامل ششم	عامل هفتم	عامل هشتم	عامل نهم	عامل دهم
Rain	-0.310	0.183	0.371	0.753	-0.235	-0.317	-0.079	0.063	-0.018	-0.016
tem	0.411	0.236	-0.062	0.180	0.007	-0.005	0.129	-0.153	-0.245	0.799
t- max	0.423	0.185	-0.201	0.108	0.171	-0.513	0.404	-0.163	-0.182	-0.469
t-min	0.371	0.276	0.154	0.262	-0.131	0.728	0.042	0.078	-0.134	-0.347
hum	-0.344	0.337	0.013	-0.131	0.060	0.091	0.711	0.453	0.119	0.117
h- max	-0.374	0.247	0.059	0.080	0.746	0.178	-0.090	-0.372	-0.232	-0.028
h- min	-0.295	0.347	0.006	-0.352	-0.564	-0.007	0.048	-0.474	-0.344	-0.071
wind	0.256	0.156	0.832	-0.392	0.134	-0.192	-0.070	0.060	0.043	0.008
mix	0.015	0.487	-0.277	-0.140	0.046	-0.170	-0.520	0.541	-0.267	-0.037
dew	0.085	0.494	-0.151	0.006	-0.013	-0.014	-0.146	-0.274	0.793	0.011

همانطور که اشاره شد با استفاده از ماتریس latent می‌توان مقدار ویژه هریک از عامل ها را مشخص کرد اما باید عملیات ساده زیر را ابتدا انجام داد و با استفاده از نتایج این عملیات درصد توجیه هریک از واریانس ها را تبین کرد:

```
latent =
   4.8224
   4.5099
   0.7942
   0.5654
   0.2202
   0.0460
   0.0312
   0.0076
   0.0028
   0.0003
   0.0000
```

برای درصد توجیه واریانس از دستور زیر استفاده می کنیم:

```
>> VAR=cumsum (latent)./ sum (latent);
```

یا می توانید دستورات زیر را اجرا کنید. نتایج هر دو محاسبات یکی می‌باشد.

```
>> Ss=sum (latent);
>> Vv= (latent./ss);
```

```
>> Cu=cumsum (Vv);
```

درصد توجیه واریانس هریک از عامل ها با توجه با انجام محاسبات فوق در زیر نمایش داده شده است:

```
>> VAR=cumsum(latent)./sum(latent)
VAR =
   0.4512
   0.8392
   0.9141
   0.9700
   0.9920
   0.9965
   0.9991
   0.9997
   1.0000
   1.0000
```

```
>> vv
vv =
   45.1228
   38.8001
    7.4869
    5.5891
    2.1985
    0.4486
    0.2594
    0.0648
    0.0265
    0.0033
```

در انتخاب تعداد عامل ها بهتر است تعداد عامل های را انتخاب کنید که حداقل بالای ۷۰ درصد واریانس را توجیه کرده باشد. برای مثال اگر دقت کنید سه عامل اول ۹۱/۴۱ درصد از واریانس متغیرها را توجیه می‌کند بنابراین از بین ۱۰ عامل تنها سه عامل را انتخاب می کنیم که عامل های اول تا سوم می‌باشد:

جدول— **ماتریس همبستگی بین متغیرها و عامل های استخراج شده**

متغیرها	عامل اول	عامل دوم	عامل سوم
Rain	-0.310	0.183	0.371
tem	**0.411**	0.236	-0.062
t- max	**0.423**	0.185	-0.201
t-min	**0.371**	0.276	0.154
hum	**-0.344**	0.337	0.013
h- max	**-0.374**	0.247	0.059
h- min	-0.295	0.347	0.006
wind	0.256	0.156	0.832
mix	0.015	0.487	-0.277
dew	0.085	0.494	-0.151

بار عاملی همبستگی بین عامل‌های حاصل از تحلیل مولفه‌ها و متغیرهای اصلی را که برای ساختن عامل‌ها مورد استفاده قرار می‌گیرند را توضیح می‌دهند. بار عامل‌ها را همانند بالا می‌توان به شکل یک ماتریسنشان داد که در آن اعداد مربوط به هر ستون همبستگی بین عامل بخصوص با متغیرهای اصلی می‌باشد. کاربرد اولیه این ماتریس شناسایی آن متغیرهای است که دارای همبستگی بالای با یک عامل معیین هستند یعنی بار بالای دارند. با این تفاسیر وقتی که تعدا عامل ها را استخراج کردیم نیاز است بدانیم روی هر عامل کدام یک بیشترین وزن را دارد. اگر دت کنید مشاهده می-شود که بر روی عامل اول حانواده دما بیشترین تاثیر را داشته است. بنابراین می توان گفت که عامل اول عامل دمایی می‌باشد. این در حالی می‌باشد که نسبت آمیختگی و نقطه شبنم نماینده مناسبی

برای عامل دوم به حساب می‌آید.

مثال کاربردی

در این مثال ابتدا داده‌ها طوری مرتب شده‌اند که ردیف‌ها مربوط به ۳۱۹ ایستگاه سنوپتیکی و ستون‌ها برای ۱۹ متغیر اقلیمی می‌باشند. با اجرای تحلیل مولفه‌ها بر روی این متغیرها ۳ عامل که در مجموع نزدیک ۸۰ درصد واریانس داده‌ها را توجیه می‌کند قابل تقلیل می‌باشد:

```
>> [coeff, score, pcvar, mu, v] = ppca(x);
```

جدول ۱ : درصد و مقدار ویژه هریک از عامل ها (pcvar)

عامل‌ها	مقدار ویژه	درصد واریانس	درصد تجمعی واریانس
عامل دمایی	7.8	41.3	41.09
عامل رطوبتی	5.9	31.4	72.5
عامل بارشی	1.1	6.4	80.3

جدول— ماتریس همبستگی بین متغیرها و عامل های استخراج شده(coeff)

	عامل دمایی	**عامل رطوبتی**	**عامل بارشی**
حداکثر دما	**0.287**	0.153	0.172
حداقل دما	**0.288**	0.208	-0.026
حداکثر مطلق	**0.265**	0.054	0.193
حداقل مطلق	**0.239**	0.193	-0.162
حداکثر بارندگی	-0.020	0.276	**0.485**
روزهای با بارندگی ۱ میلیمتر	**-0.305**	0.179	0.096
روزهای با بارندگی ۱۰ میلیمتر	-0.201	0.252	**0.417**
روزهای برفی	**-0.275**	-0.101	-0.149
گردوغبار	**0.238**	0.040	0.033
طوفان تندری	-0.153	0.067	**-0.216**
فشارQFA	0.167	**0.331**	-0.124
نقطه شبنم	0.132	**0.348**	-0.232
فشار بخارآب	0.135	**0.336**	-0.232
آسمان صاف	**0.287**	-0.176	0.178
روزهای ابری	-0.227	**0.261**	0.022
ساعات آفتابی	0.212	**-0.268**	0.191
میانگین ماهانه	**0.321**	0.146	0.057
مجموع بارندگی	-0.215	0.262	**0.379**
رطوبت نسبی	-0.163	**0.317**	-0.281

یکی دیگر از کاربردهای تحلیل مولفه‌های اصلی را می‌توان در تعیین روز نماینده می‌باشد. فرض کنید که شما در یک دوره ۵۰ ساله در محدوده اقلیمی ایران فشار سطح ۵۰۰ هکتوپاسکال تعداد روزهای که بارش فوق سنگین

داشته استخراج کرده اید و با استفاده از تحلیل مولفه های اصلی می خواهیم بفهمیم کدام روز (حالت جوی) بیشترین بار عاملی را در وقوع بارش های سنگین داشته است. از این رو ماتریس متغیرها به شکل زیر خواهد بود:

ایستگاه	روز اول	روز دوم	روز سوم	n
A				
B				
C				
D				
n				

با تشکر دوستکامیان