
بسم الله الرحمن الرحيم
إنه من سليمان وإنه

ادامه فصل۳:

معرفی سایر روش های رج بندی
آنالیز مولفه های اصلی (PCA)
ویژگی های مولفه های تولیدی
بردارها و مقادیر ویژه
مراحل روش PCA
تمرکز داده ها
رج بندی های R و Q
روش دو پلاتی (Biplot)

معرفی سایر روش های رج بندی:

آنالیز مولفه های اصلی (PCA)
روش آنالیز مولفه های اصلی نخستین بار توسط کارل پیرسون(۱۹۰۱)

مطرح شده است. روش محاسباتی آن نیز توسط هاتلینگ (۱۹۳۳)

ارایه شده است.

گودال (۱۹۵٤) برای نخستین بار از روش آنالیز مولفه های اصلی به

عنوان آنالیز عاملی استفاده کرده است.

روش PCA پس از ارایه مقاله اورلوسی (۱۹٦٦) کاملا شناخته شد و استفاده از آن متداول گردید.

از روش PCA به طور عمده برای سنتز داده های محیطی و تولید رج بندی واحد های نمونه برداری (یا قاب ها) بر اساس متغیرهای محیطی استفاده می شود و این روش برای رج بندی داده های فلوریستیکی (ترکیب گونه ای) چندان مناسب نیست و توصیه نمی شود، هر چند که به این منظور از آن استفاده زیادی شده است.

روش PCA نخستین روشی است که محورهای رج بندی در آن بر اساس داده های جدول ماتریس محاسبه می شود و نیازی به وزن دادن داده ها و یا استفاده از نقاط انتهایی و یا سایر اطلاعات ذهنی در فرآیند محاسباتی این روش نمی باشد.

در مقایسه با روش ریاضی(جبر ماتریس)، تشریح هندسی PCA برای دانشجویان مناسب تر و راحت تر است.

در روش PCA سطر های جدول ماتریس(گونه ها) کاهش می یابد و گونه های جدید ساخته و تولید می شود.

به طوری که ماتریس که دارای n واحد نمونه برداری (یا قاب) و m گونه (یا عامل محیطی) است به n واحد نمونه یا قاب و تعدادی مولفه یا عامل کاهش می یابد.

این مولفه های ساختگی به عنوان متغیرهای سوپر شناخته می شوند و معرف ترکیب کاملاً همبسته ای از گونه ها یا عامل های محیطی هستند.

ویژگی های مولفه های تولیدی:

۱- مولفه های جدید بر خلاف کلیه گونه ها یا عامل های محیطی که در قبل به طور کامل با یکدیگر همبستگی داشتند، غیر همبسته هستند و به آنها مولفه های متعامد (Orthogonal)گفته می شود.

۲- همانگونه که در جدول ماتریس داده های اولیه برای گونه یا متغیرهای محیطی مقادیر عددی اختصاص یافته است در جدول ماتریس جدید نیز برای هر یک از مولفه ها هم مقادیر عددی در نظر گرفته می شود که به آن نمرات مولفه گفته می شود.

۳- هسته مرکزی روش PCA را شاخص های بردارهای ویژه (eigenvectors) و مقادیر ویژه (eigenvalues) تشکیل می دهد.

بردارها و مقایر ویژه:

الف- بردارهای ویژه Eigenvectors

هر مولفه تولیدی (ساختگی) شامل مجموعه نمراتی است که هر یک از این نمرات وزن هر یک از گونه ها یا متغیرهای اولیه را بر روی مولفه را بیان می کند.

نمرات بردار ویژه مانند ضرایب همبستگی بین دو حد صفر و ۱+ و صفر تا ۱- تغییر می کنند.
هر چه نمرات بردار ویژه (مربوط به گونه ها یا متغیرها) از صفر دورتر و به یک نزدیکتر باشد از اهمیت بیشتری گونه یا متغیر برخوردار بوده و وزن بیشتری به مولفه داده می شود.
در این صورت مولفه تولیدی به طور عمده بر حسب این گونه ها و یا متغیرها شناخته و تفسیر می شود.

به عنوان مثال PCA از جدول ماتریس داده هایی که از n واحد نمونه برداری و ٦ گونه (A تا F) تشکیل یافته است، دو مولفه با نمرات بردارهای ویژه به شرح زیر ساخته است.

در بعضی از آنالیزها از بردارهای ویژه و نمرات آنها تحت عنوان مولفه یا بارهای عاملی (Factor loading) نام برده می شود.

ب- مقادیر ویژه (Eigenvalue)

مقادیر ویژه سهم نسبی هر مولفه در تبیین کل داده ها را ارائه می کند.

اندازه مقدار ویژه برای یک مولفه اهمیت آن مولفه را در تشریح کل تغییرات در داخل مجموعه داده ها به طور مستقیم نشان می دهد

(به عبارتی مقدار ویژه واریانس هر محور را بیان می کند).

مراحل روش PCA:

گام ۱- استاندارد کردن داده های ماتریس

برای استاندارد کردن داده ها بر مبنای روش استاندارد سازی میانگین صفر و واریانس واحد، از رابطه زیر استفاده می شود.

$$SSi = \frac{Si - \bar{S}}{\delta S}$$

که در این رابطه:

SSi : نمرات استاندارد شده گونه یا متغیر S در واحد نمونه برداری(قاب)

Si: نمرات اولیه گونه یا متغیر S در واحد نمونه برداری(قاب i)

$\bar{S}$: میانگین نمره گونه یا متغیرSدرکلیه واحدهای نمونه برداری(قاب ها)

δS : انحراف معیار گونه یا متغیر S در کلیه واحد های نمونه برداری (یا قاب ها) می باشد.

انحراف معیار را می توان از رابطه زیر محاسبه نمود.

$$\delta S = \sqrt{\sum_{j=1}^{n}(Sij - \ \bar{S}i)^2}$$

به عنوان مثال برای گونه یا متغیر A استاندارد سازی به شرح زیر است:

۵۵ ÷ ۱۰۰ = ۵ /۵

$\bar{S}$ = ۵/۵

انحراف معیار برابر ۳/۰۳ است و نمره اولیه واحد نمونه برداری (قاب) برای گونه A برابر ۱ است که پس از استاندارد سازی ۱/٤۹ می شود.

(۱–۵/۵) ÷ ۳/۰۳ =۱/٤۹

برای واحد نمونه برداری ٤ که نمره گونه A برابر ۹ است، نمره پس از استانداردسازی برابر ۱/۱٦ خواهد بود.

(۹ –۵/۵) ÷ ۳/۰۳ =۱/۱٦

بدین صورت کلیه نمرات در جدول ماتریس داده های اولیه استاندارد می شود.
اگر داده ها استاندارد نگردد، در این صورت آنالیز در جهت گونه ها

یا متغیرهایی که دارای بیشترین واریانس هستند اریبی پیدا می کند.

اگر واحد های نمونه ای در سطر ها قرار گیرند به همین صورت برای

آنها استاندارد سازی می تواند انجام شود.

اگر از ضرایب همبستگی به عنوان معیار تشابه برای تشکیل ماتریس

تشابه استفاده شود، در این صورت به طور خودکار استانداردسازی

داده ها هم انجام می گردد.

گونه‌ها یا متغیرها		قابها									
		1	2	3	4	5	6	7	8	9	10
	A	1	5	7	9	10	8	6	4	3	2
	B	8	10	7	9	6	5	4	3	1	2
	C	3	6	7	9	10	8	5	4	2	1
	D	4	8	7	10	9	6	5	3	2	1
	E	10	9	7	8	6	5	4	3	1	2
	F	2	6	7	9	10	8	5	4	3	1
	G	1	6	8	9	10	5	7	4	3	2

گام ۲- محاسبه ماتریس تشابه(یا ضریب همبستگی)

برای تشکیل این ماتریس، اغلب همبستگی بین گونه ها و یا متغیرهای محیطی با استفاده از ضرایب همبستگی پیرسون محاسبه می شود.

مانند جدول ماتریس تشابه زیر:

A	1.00						
B	0.35	1.00					
C	0.95	0.58	1.00				
D	0.83	0.79	0.93	1.00			
E	0.20	0.96	0.47	0.67	1.00		
F	0.98	0.49	0.99	0.90	0.36	1.00	
G	0.93	0.42	0.88	0.85	0.26	0.90	1.00
	A	B	C	D	E	F	G

گام ۳- محاسبه شاخص تشابه بین گونه و واحد های نمونه برداری

الف - برای این منظور ابتدا از روی ماتریس اولیه Xmax ، ماتریس جدید As×n ساخته می شود.
برای ساخت ماتریس جدید A از رابطه زیر استفاده می شود:

$$aij = \frac{Xij}{\sqrt{ri \,.\, cj}}$$

که در این رابطه:
aij : داده های ماتریس جدید A
ri : مجموع داده های هر سطر در ماتریس اولیه
cj: مجموع داده های هرستون در ماتریس اولیه است.

به عنوان مثال برای ماتریس X مقادیر ri و cj برای گونه A به ترتیب ۵۵ و ۲۹ می باشد.

بر این اساس مقدار aij در ماتریس جدید به جای عدد ۱، ۰/۲۵ محاسبه می شود.

$$aij = \frac{1}{\sqrt{55 \times 29}}$$

ب) در ادامه ماتریس جدید $A_{s\times n}$ وارونه می شود و جای ردیف ها

و ستون ها تعویض می شود و ماتریس $A_{n\times s}$ تولید می گردد.

ماتریس ضریب همبستگی گونه ها (یا تشابه)(R) و ماتریس ضریب همبستگی (یا تشابه) واحدهای نمونه برداری(Q) ازضرب ماتریس جدید $A_{s\times n}$ در ماتریس جدید وارونه $A^{t}_{n\times s}$ به شرح زیر بدست می آید:

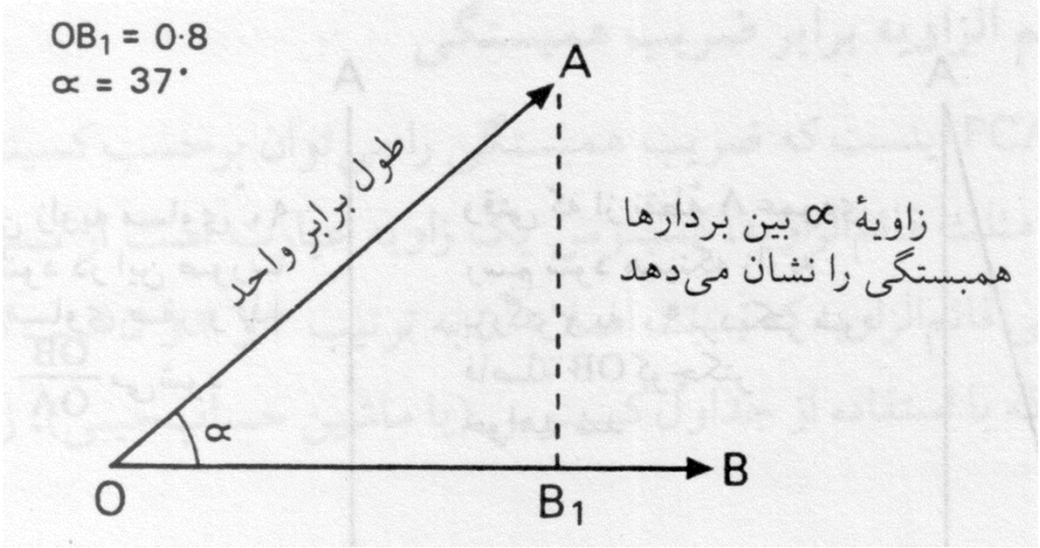
$$R\ s.s = A\ s.n\ \times\ A^{t}_{n.s}$$

$$Q\ n.n = A\ s.n\ \times\ A^{t}_{n.s}$$

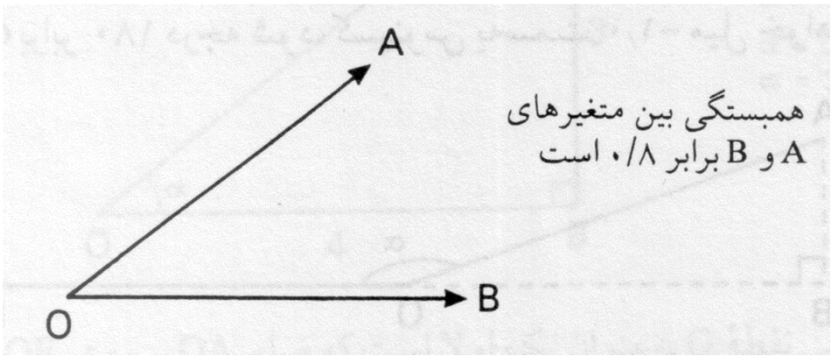
$$\begin{pmatrix} 3 & 2 & 1 & 0 \\ 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 3 & 2 & 0 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 14 & 9 & 3 \\ 9 & 7 & 4 \\ 3 & 4 & 6 \end{pmatrix}$$

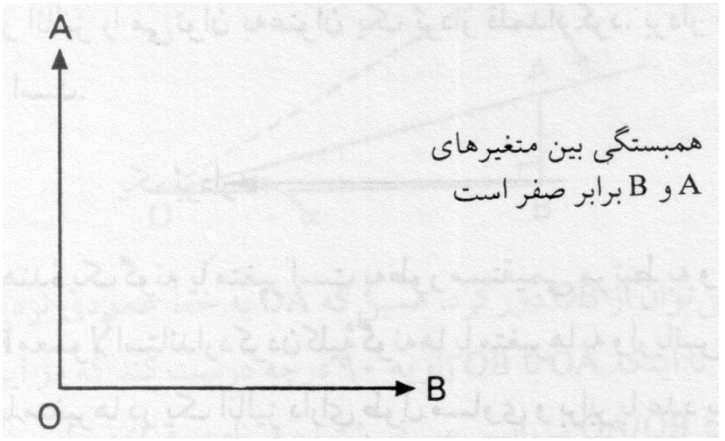
گام٤- محاسبه بردار و مقایر ویژه ماتریس R

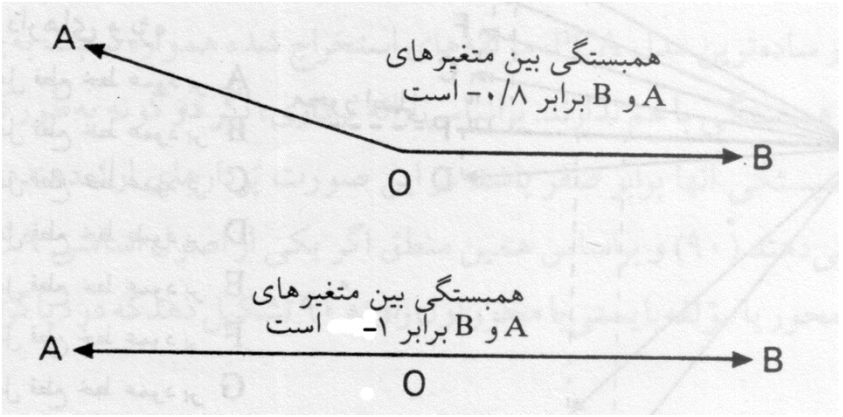
در مثلث قائم الزاویه OBA کسینوس زاویه آلفا (نسبت ضلع مجاور به وتر) برابر ضریب همبستگی است.



هر گونه یا متغیر را در آنالیز رج بندی می توان یک بردار محسوب کرد. چنین بردارهایی دارای اندازه طول و جهت هستند(⟶).


A
B
O
همبستگی بین متغیرهای
A و B برابر ۰/۸ است


A
B
O
همبستگی بین متغیرهای
A و B برابر صفر است


A
B
O
همبستگی بین متغیرهای
A و B برابر ۰/۸- است
A
B
O
همبستگی بین متغیرهای
A و B برابر ۱-
است

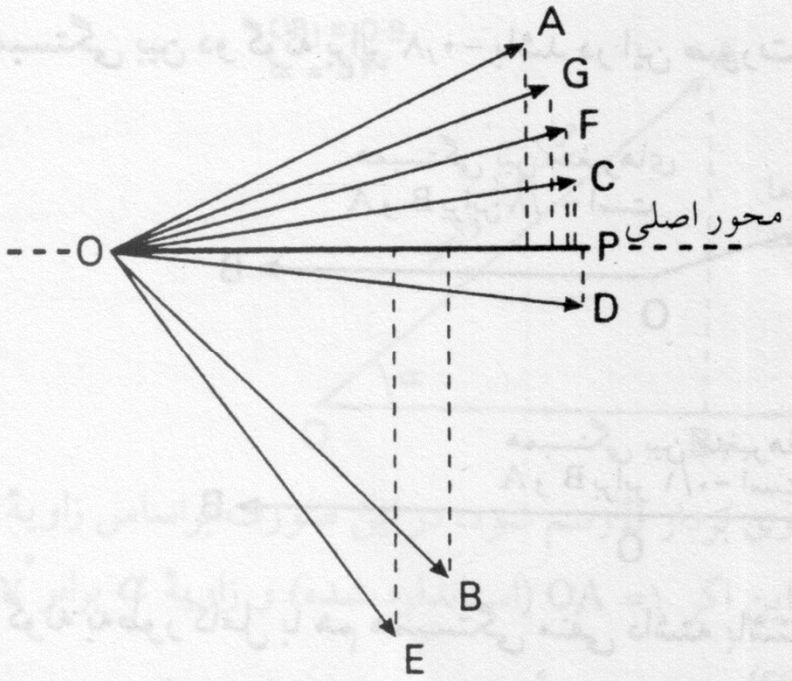
هدف روش آنالیز مولفه اصلی (PCA) پیدا کردن جهت کلی فرود بردارها از طریق عبور یک خط یا محور از مبدا مشترک است به طوری

که هر برداری که یک گونه را ارائه می دهد(معرف یک گونه است) با تصویر آن بردار بر روی این محور زاویه ۹۰ درجه تشکیل دهد.
این خط یا محور به عنوان محور اصلی شناخته می شود.

از آنجا که طول کلیه بردارها برابر یک واحد است، تصویر آنها بر روی محور اصلی برابر کسینوس زوایای بین بردارها و محور اصلی است.

محور اصلی به آهستگی حول مبدا O می چرخد و در هر لحظه طول هایی بر روی محور OP از مبدا O تا نقاطی که از نوک بردارها بر خط OP عمود شده اند، اندازه گیری می شود(اشکال تا).

محور اصلی(محور اول):



بردارهای ویژه

A از O مبدأ تا محل قطع خط عمود بر = ۰/۹۰

B از O مبدأ تا محل قطع خط عمود بر = -۰/۷۱

C از O مبدأ تا محل قطع خط عمود بر = ۰/۹۷

D از O مبدأ تا محل قطع خط عمود بر = -۰/۹۹

E از O مبدأ تا محل قطع خط عمود بر = -۰/۵۹

F از O مبدأ تا محل قطع خط عمود بر = ۰/۹۵

G از O مبدأ تا محل قطع خط عمود بر = ۰/۹۲

Op (محور اصلی) = ۱/۰

بردار های ویژه (eigenvectors):

تصویر بردارهای گونه ها بر روی محور اصلی، بردار ویژه (یا بار گونه) نامیده می شود.

بردار ویژه (یا بار گونه ها) در واقع درجه همبستگی هر گونه با محور اصلی را بیان می دارد.

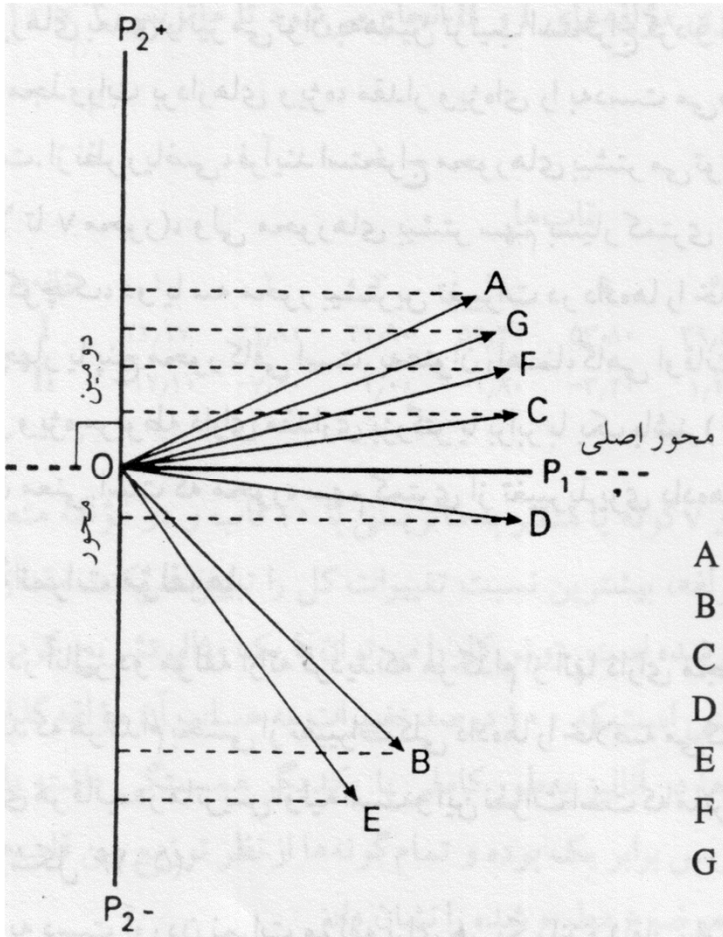
متغیر / گونه	بردار ویژه
A	۰/۹
B	۰/۷۱
C	۰/۹۷
D	۰/۹۹
E	۰/۵۹
F	۰/۹۵
G	۰/۹۲

مقادیر ویژه (eigenvalue):

مقادیر ویژه از مجموع مجذورات همبستگی بین گونه ها، با محور اصلی محاسبه می شود.

محور دوم:

همانطوری که در قبل اشاره شد در مدل PCA مولفه های تولیدی نسبت به یکدیگر متعامد هستند.
بنابراین با توجه به زاویه ۹۰ درجه بین آنها، همبستگی دو مولفه عمود بر هم صفر می باشد.
بر این اساس محور دوم محوری است که عمود بر محور اول(اصلی) است و زاویه ۹۰ درجه با آن تشکیل می دهد.



بُردارهای ویژه

A	فاصلهٔ O از محل نقطهٔ عمود بر	=۰٫۴۳
B	فاصلهٔ O از محل نقطهٔ عمود بر	=−۰٫۷۰
C	فاصلهٔ O از محل نقطهٔ عمود بر	=۰٫۲۴
D	فاصلهٔ O از محل نقطهٔ عمود بر	=−۰٫۱۴
E	فاصلهٔ O از محل نقطهٔ عمود بر	=−۰٫۸۱
F	فاصلهٔ O از محل نقطهٔ عمود بر	=۰٫۳۱
G	فاصلهٔ O از محل نقطهٔ عمود بر	=۰٫۳۹

Op_2 دومین محور = $\pm$۱٫۰

محاسبه بردار و مقادیر ویژه برای محور دوم:

متغیر/ گونه	بردار ویژه
A	۰/٤۳
B	-۰/۷
C	۰/۲٤
D	-۰/۱٤
E	-۰/۸۱
F	۰/۳۱
G	۰/۳۹

$$\text{eigenvalues} = (0.43)^2 + \ldots\ldots + (0.39)^2 = 1.66$$

ملاحظه می شود که مقادیر ویژه برای محور دوم (۱/٦٦) خیلی کمتر از مقادیر ویژه برای محور اول (۵/۳۳) است. این روند از ویژگی های PCA است.

به همین ترتیب مقادیر ویژه برای محورهای بعدی کمتر خواهد شد.

محور اول در آنالیز PCA از اهمیت بیشتری برخوردار است و واریانس (یا سهم تغییرات مجموعه داده ها) را بیشتر توجیه می کند.

فرآیند استخراج محورها تا تعداد گونه ها (که در این مثال ۷ گونه است) می تواند ادامه یابد.

اغلب برای مجموعه داده های کوچک ۲ تا ۳ محور و برای مجموعه داده های بزرگتر ٤ تا ۵ محور بخش عمده واریانس را توجیه می کنند.

گام۵- محاسبه نمرات مولفه ها

از ضرب مقادیر اولیه گونه ها (یا متغیرها) در یک واحد نمونه برداری در نمرات مولفه مربوطه (بردارهای ویژه) و سپس جمع مقادیر حاصل، نمره مولفه برای آن واحد نمونه محاسبه می شود.

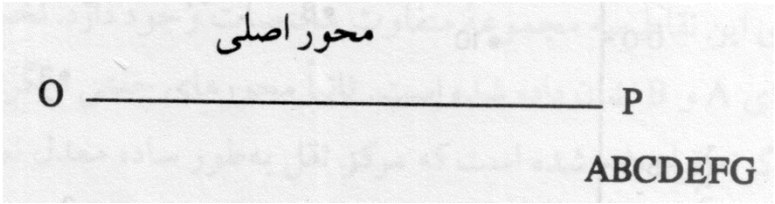
$$+ (۰/۹۹ \times ۴) + (۰/۹۷ \times ۳) + (۰/۷۱ \times ۸) + (۰/۹ \times ۱)$$

$$\underline{۲۲/۱۷} = (۰/۹۳ \times ۱) + (۰/۹۵ \times ۲) + (۰/۵۹ \times ۱۰) +$$

گرنه‌ها	متغیرها	قابها 1	2	3	4	5	6	7	8	9	10	بردارهای ویژه مؤلفه I	مؤلفه II
	A	1	5	7	9	10	8	6	4	3	2	0.90	0.43
	B	8	10	7	9	6	5	4	3	1	2	0.71	−0.70
	C	3	6	7	9	10	8	5	4	2	1	0.97	0.24
	D	4	8	7	10	9	6	5	3	2	1	0.99	−0.14
	E	10	9	7	8	6	5	4	3	1	2	0.59	−0.81
	F	2	6	7	9	10	8	5	4	3	1	0.95	0.31
	G	1	6	8	9	10	5	7	4	3	2	0.92	0.39

قاب‌ها

مؤلفه‌ها	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
I	۲۲٫۱۷	۴۱٫۸۰	۴۲٫۹۰	۵۴٫۴۰	۵۳٫۸۰	۳۹٫۵۰	۳۱٫۴۰	۲۱٫۷۰	۱۳٫۴۰	۹٫۱۰
II	-۱۲٫۱۰	-۷٫۶۰	-۲٫۰۰	-۱٫۸۰	-۳٫۴۰	۱٫۴۰	۱٫۳۰	۰٫۵۰	۲٫۱۰	-۰٫۹۷



ملاحظه می شود که ماتریس اولیه $X_{7\times10}$ به ماتریس $X'_{2\times10}$ کاهش یافت و ۷ گونه در قالب ۲ مولفه متعامد خود را نشان می دهند.

در صورتی که کلیه گونه ها با یکدیگر همبستگی کاملی داشته باشند، یک مولفه PCA می تواند ۱۰۰ درصد واریانس(یا تغییرات مجموعه داده ها) را توجیه نماید.

در این صورت مقادیر همبستگی گونه ها در جدول ماتریس همبستگی برابر ۱ خواهد بود واز نظر هندسی بردارها بر روی یکدیگر منطبق می شوند.

در چنین شرایطی بردار ویژه برای هر گونه ۱ خواهد بود و مقدار ویژه نیز برابر تعداد گونه ها(در اینجا برابر ۷) می باشد.

$$\text{eigenvalues} = (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 = 7$$

در شرایط متعارف چون یک محور PCA، ۱۰۰ درصد تغییرات را معرف نیست، با استفاده از مقادیر ویژه محاسبه شده برای محورها و تعداد گونه های ماتریس، می توان سهم هر یک از مولفه ها را از کل واریانس محاسبه کرد.

به طور مثال:

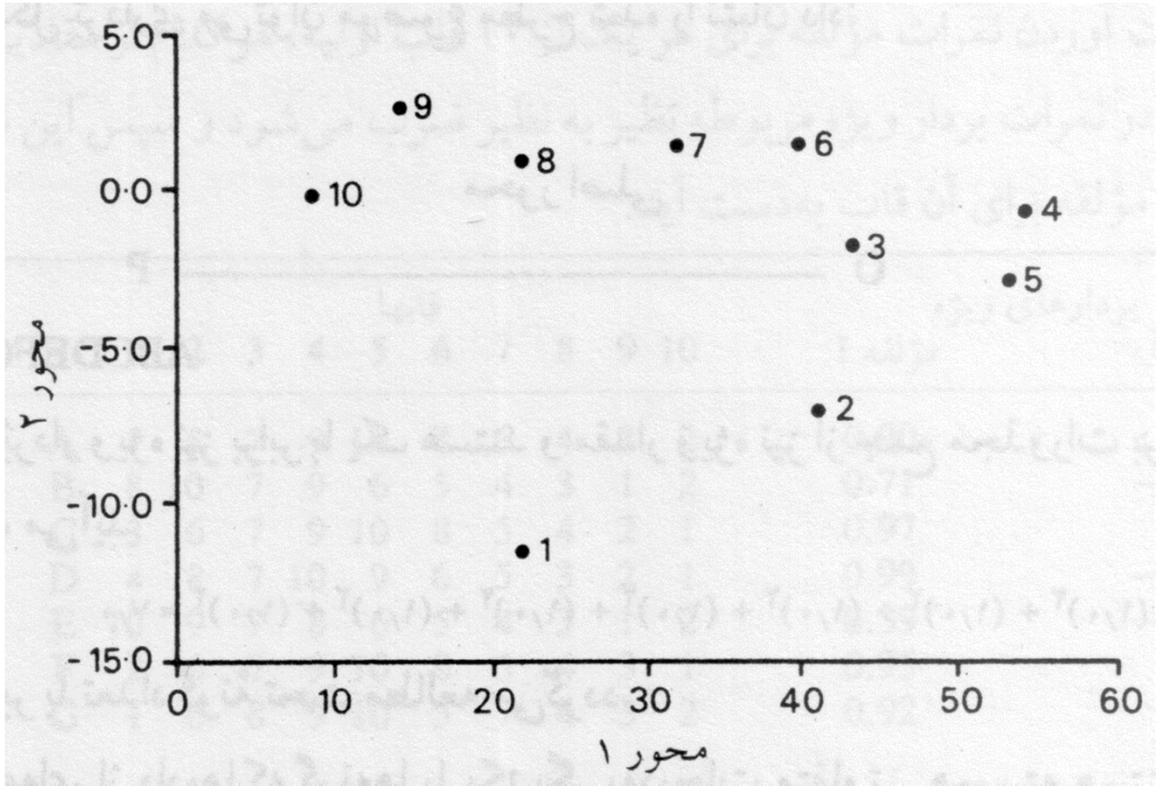
(۵/۳۳ ÷ ۷) × ۱۰۰ = ٪ ۷۶/۲

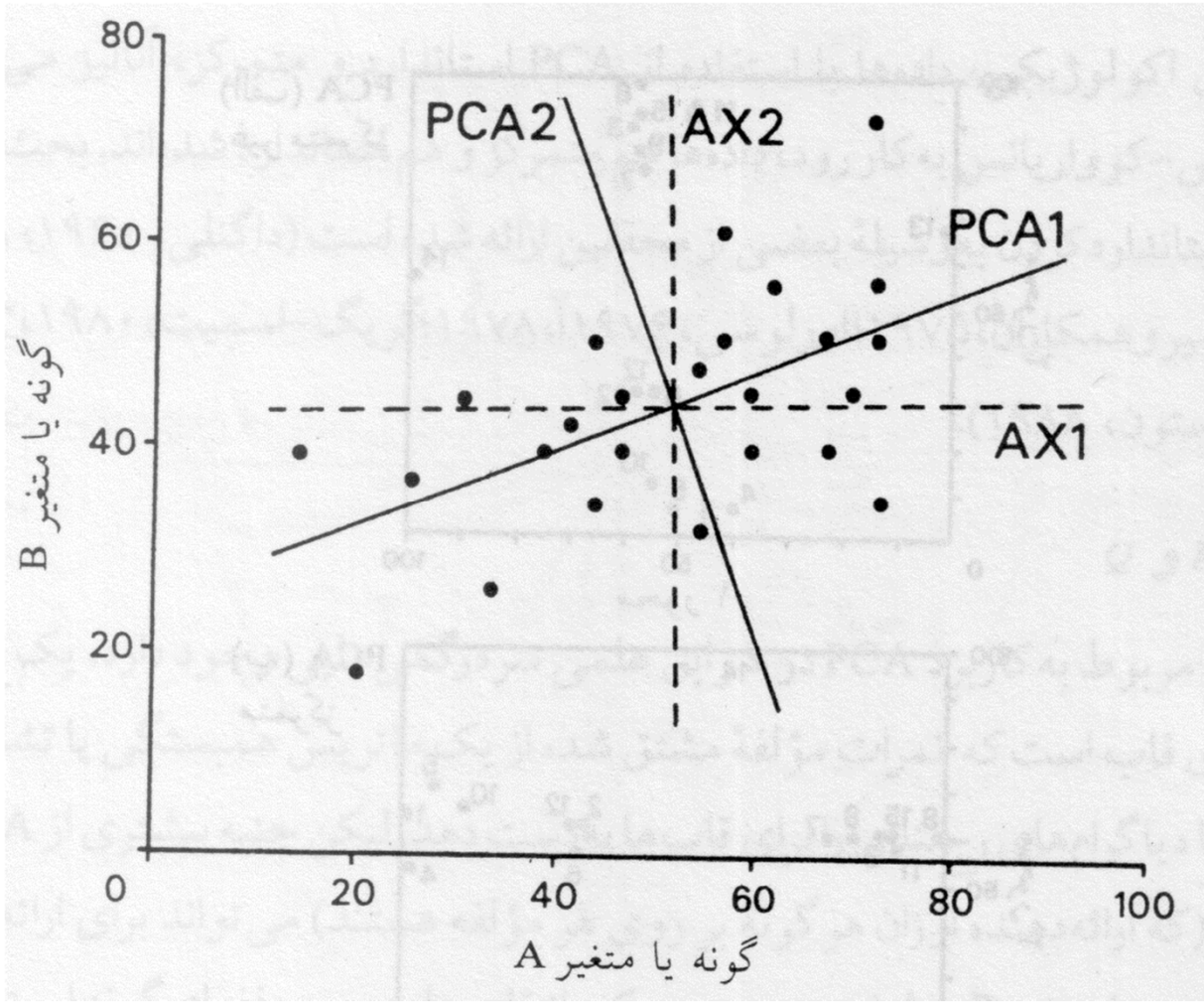
(۱/۶۶ ÷ ۷) × ۱۰۰ = ٪ ۲۳/۶۵

۷۶/۲ + ۲۳/۶۵ = ٪ ۹۹/۸۵

دو مولفه I و II، ۹۹/۸۵ درصد از مجموع واریانس داده ها را توجیه می کنند و ۰/۱۵ درصد نیز سهم سایر مولفه ها (۵ مولفه بعدی) است.

گام٦- رسم نمودار نهایی رج بندی(واحد نمونه برداری یا قاب)



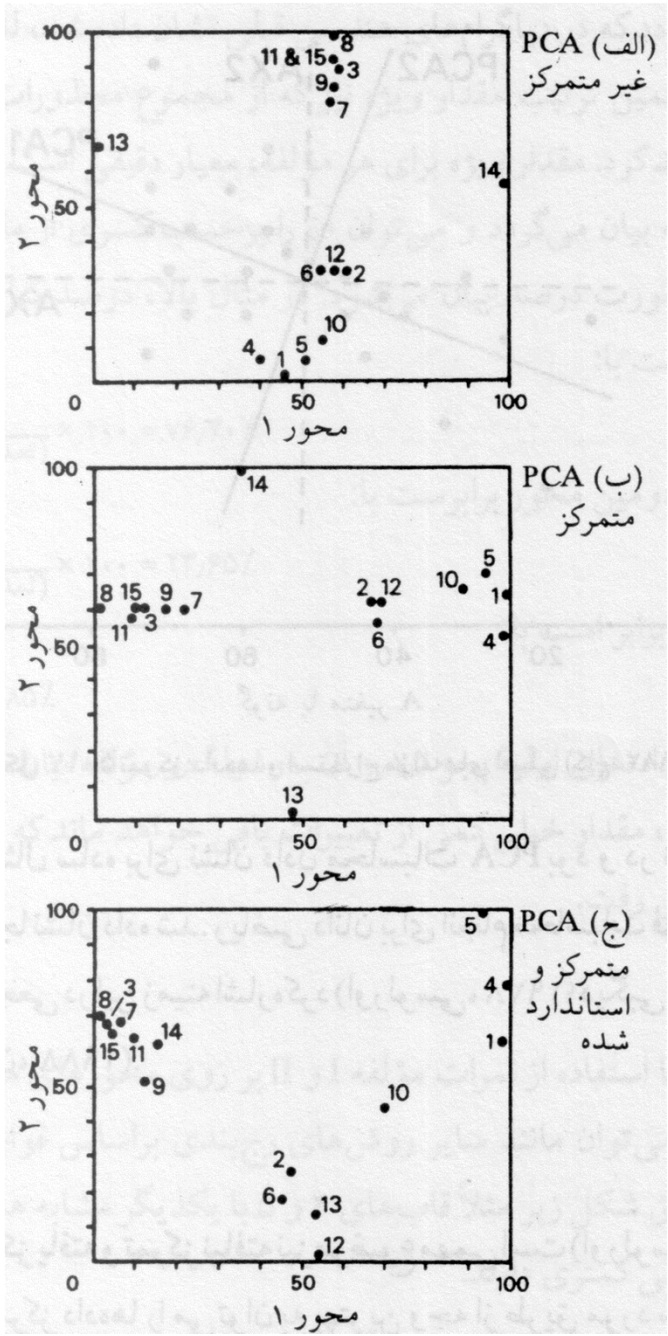

80
60
40
20
0
20
40
60
80
100
PCA2
AX2
PCA1
AX1
گونه یا متغیر B
گونه یا متغیر A

تمرکز داده ها:

آنالیز PCA به روشهای ثقلی(Centroid) و غیر ثقلی می تواند انجام گیرد.
وقتی داده ها پیوسته است نتیجه آنالیز ثقلی و غیر ثقلی تفاوت چندانی ندارد ولی در صورتی که داده ها ناپیوسته باشد، موقعیت محور اول مقداری تغییر می کند و در نتایج رج بندی تفاوت مشاهده می شود.

اگر داده ها استاندارد شوند خروجی رج بندی تفاوت زیادی در مقایسه با داده های استاندارد نشده پیدا می کند.

به عبارتی، استانداردسازی داده ها اثر معنی داری بر روی رج بندی داده ها دارد.


(الف) PCA
غیر متمرکز
(ب) PCA
متمرکز
(ج) PCA
متمرکز و
استاندارد
شده
محور ۲
محور ۱
0
50
100

رج بندی های R و Q

در آنالیز PCA، رج بندی R به معنی رج بندی واحدهای نمونه برداری

(یا قابها) بر مبنای ماتریس تشابه یا همبستگی گونه ها است.
بنابراین بر روی نمودار رج بندی موقعیت واحد های نمونه ای (یا قابها) آشکار و مشخص می گردد.

رج بندی Q نیز به منزله رج بندی گونه ها بر مبنای ماتریس تشابه یا همبستگی واحد های نمونه ای (یا قابها) است که موقعیت گونه ها (یا متغیرهای محیطی) را بر روی نمودار رج بندی نشان می دهد.

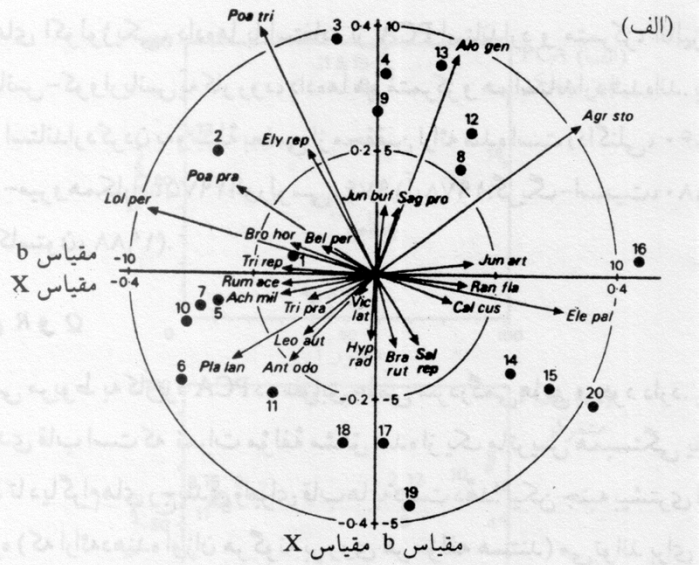
اگر استانداردسازی و تمرکز یکسان بر روی داده ها بکار رود، نتایج حاصل از آنالیز رج بندی R و آنالیز رج بندی Q یکسان می باشد.

روش دو پلاتی:

به طور همزمان نتایج آنالیز رج بندی R و Q را می توان بر روی یک نمودار رج بندی حاصل از PCA نشان داد که به آن اصطلاحاً نمودار دو پلاتی (Biplot) گفته می شود.

البته مقیاس بکار رفته برای هر یک از آنالیزها R و Q متفاوت می باشد.

(الف)

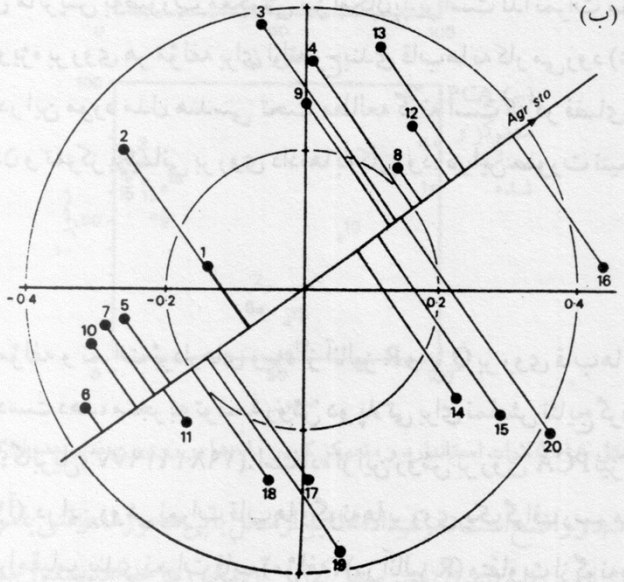


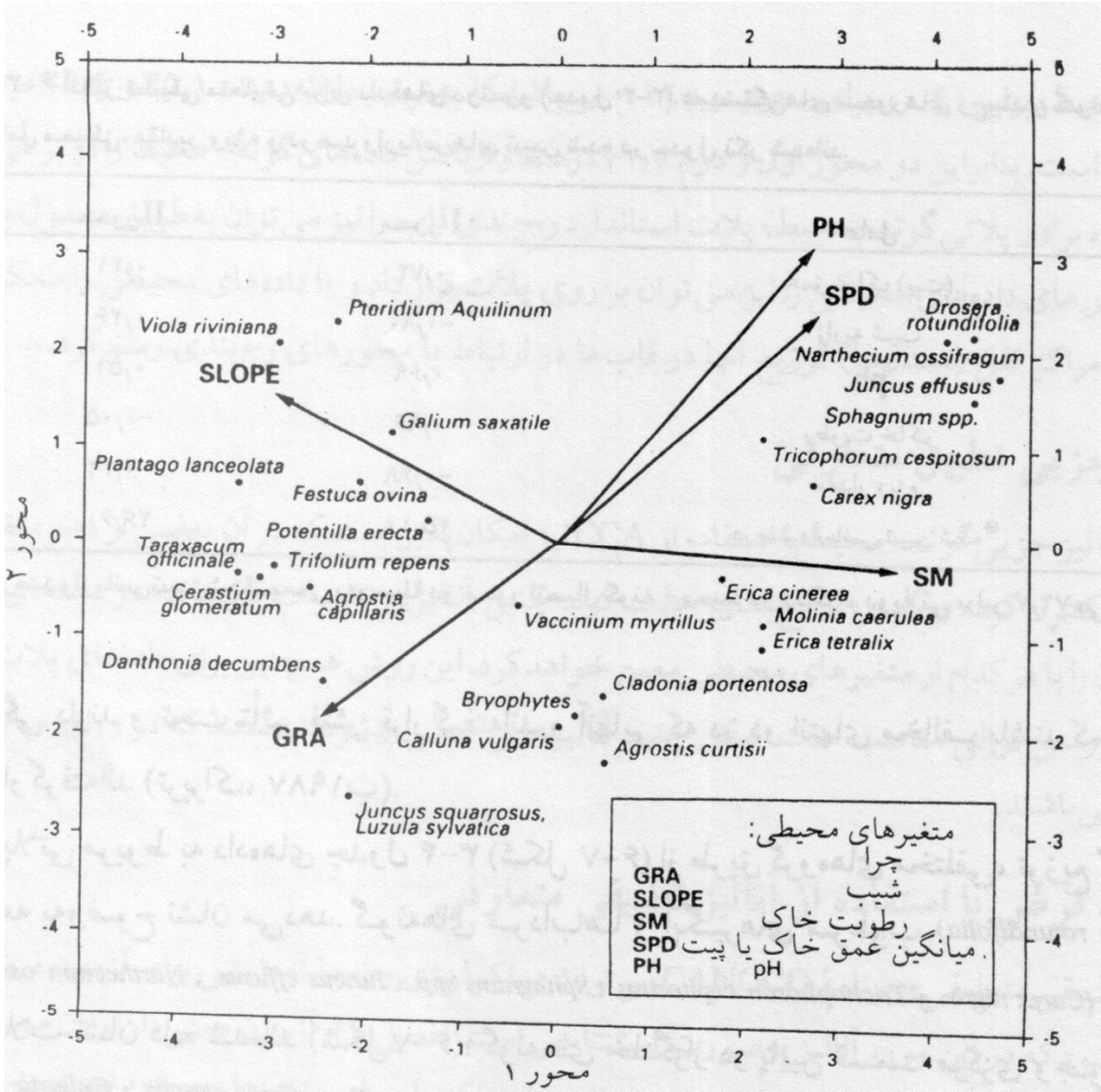
مقیاس b

مقیاس X

مقیاس b مقیاس X

(ب)




PH
SPD
SLOPE
SM
GRA
Pteridium Aquilinum
Viola riviniana
Drosera rotundifolia
Narthecium ossifragum
Juncus effusus
Sphagnum spp.
Galium saxatile
Tricophorum cespitosum
Plantago lanceolata
Carex nigra
Festuca ovina
Potentilla erecta
Taraxacum officinale
Trifolium repens
Erica cinerea
Cerastium glomeratum
Agrostia capillaris
Molinia caerulea
Vaccinium myrtillus
Erica tetralix
Danthonia decumbens
Cladonia portentosa
Bryophytes
Calluna vulgaris
Agrostis curtisii
Juncus squarrosus, Luzula sylvatica
متغیرهای محیطی:
GRA چرا
SLOPE شیب
SM رطوبت خاک
SPD میانگین عمق خاک یا پیت
PH pH
محور ۲
محور ۱

پایان