تحلیل مولفه‌های اصلی (قسمت ۱)

Principal Component Analysis (Part 1)

سجاد طلایی
Talaei.s@arc-ordc.ir
کارشناس ارشد اصلاح نباتات، مرکز تحقیقات کاربردی و تولید بذر، شرکت توسعه کشت دانه‌های روغنی

به طور کلی به مجموعه روش‌هایی که به کاهش ابعاد داده‌ها کمک می‌کند تحلیل عاملی می‌گویند. یکی از این روش‌ها تحلیل مولفه‌های اصلی است که اولین بار توسط پیرسون در سال ۱۹۰۱ بکار گرفته شد و در سال ۱۹۳۳ توسط هتلینگ گسترش یافت (فرشادفر، ۱۳۸۹). کاربرد این روش چند متغیره آماری در کشاورزی بسیار زیاد است. وقتی در آزمایشات زیستی با حجم بالایی از متغیرها و اندازه‌گیری‌ها مواجه هستیم این روش کاربرد زیادی دارد. با افزایش تعداد متغیرها، تعداد ضرایب همبستگی نیز زیاد می شود. از روش تجزیه به مولفه‌های اصلی در جهت ابعاد اطلاعات موجود استفاده می‌شود. تجزیه به مولفه‌های اصلی واریانس بین داده ها را تبدیل به مولفه‌هایی می‌کند که این مولفه‌ها کاملا از هم مستقل هستند. نحوه کار این روش به این صورت است که سعی می شود تا آنجا که ممکن است تغییرات داده ها در مولفه اول توجیح شوند. مولفه دوم بعد از مولفه اول بقیه تغییرات و تنوع را توجیه می کند و به همین صورت تا مولفه آخر همه تغییرات تبین می شوند. تعداد کل مولفه‌ها با تعداد کل متغیرها برابر است. اما معمولا به چند مولفه اول که تغییرات بیشتر را تبیین می‌کنند بسنده می شود.

محاسبات:

برای انجام محاسبات تجزیه به مولفه‌های اصلی از روش ریاضی تجزیه به مقادیر منفرد به کمک ماتریس‌ها صورت می‌گیرد.

فرض کنید ماتریس A میانگین تکرار داده‌های اندازه‌گیری شده می‌باشد.

ارتفاع بوته	وزن هزار دانه	عملکرد دانه	ژنوتیپ
۱۰۵	۵	۲۵	۱
۱۱۰	۷	۲۹	۲
۹۰	۱۰	۳۵	۳

فرض کنید ماتریس داده ها (A) را برابر با ماتریس L باشد.

$$A = L$$

اکنون هر دو طرف معادله را در یک ماتریس دیگری به نام V ضرب می‌شود. (البته به شرط اینکه ماتریس V صفر نباشد). لذا فرمول مذکور بصورت زیر در می‌آید:

$$A\,V = L\,V$$

$$A=\begin{pmatrix} ۲۵ & ۵ & ۱۰۵ \\ ۲۹ & ۷ & ۱۱۰ \\ ۳۵ & ۱۰ & ۹۰ \end{pmatrix}$$

دو مقدار مجهول L و V بدست می‌آید که باید مقادیر آنها محاسبه گردد.

$$L\,V - A\,V = 0$$

از ماتریس V فاکتور گرفته می شود: $V(L-A)=0$

برای حل معادله فوق از دترمینان استفاده می‌شود: $Det(V(L-A))=0$ که بصورت زیر در می‌آید:

$$Det\,(L-A)I = 0$$

ماتریس I یک ماتریس با قطر ۱ می باشد که ماتریس همانی گفته می شود.

$$Det = \begin{pmatrix} 25-L_1 & 5-0 & 105-0 \\ 29-0 & 7-L_2 & 110-0 \\ 35-0 & 10-0 & 90-L_3 \end{pmatrix} = 0$$

اگر دترمینان حل شود مقادیر L بدست می‌آیند. این مقادیر L مقادیر ویژه نامیده می‌شوند. با توجه به اینکه سه متغیر در نظر گرفته شده بود، پس سه مقدار ویژه نیز برای آن وجود دارد.

با استفاده از دستور زیر در نرم افزار SAS می توان مقادیرL یا بردار ویژه را بدست آورد (Little *et al.*, 2006).

```
PROC IML;
A={105  5   25,
   110  7   29,
    90  10  35 };
 call svd  (LEFT, L, RIGHT ,A);
 print L;      run;
```

خروجی نرم افزار بصورت زیر ارائه شده است. مقادیر بدست آمده همان مقادیر ویژه (eigen value) می‌باشند.

L_1 184.29
L_2 11.22
L_3 0.40

همانطور که ملاحظه می‌شود مقدار عددی مقادیر ویژه کاهشی است.

محاسبه تعداد مولفه‌های اصلی:

برای پاسخ به این سوال که چند مولفه باید محاسبه گردد، می‌توان بصورت زیر عمل کرد:

184.29+ 11.22+ 0.40 = 195.91 ⟵ جمع کل مولفه‌ها
(184.29÷195.91)*100= 94% ⟵ سهم مولفه اول
(11.22 ÷ 195.91)*100 = 5.73% ⟵ سهم مولفه دوم
(0.4 ÷ 195.91)*100 = 0.2 % ⟵ سهم مولفه سوم

سهم مولفه‌های اول (۹۴ درصد) و دوم (۵/۷۳ درصد) بدست آمد. محققین معمولا سهم مولفه‌ها را حساب می‌کنند. معمولا وقتی سهم مولفه‌های ابتدایی از ۷۰ درصد بیشتر می‌شود سهم بقیه مولفه‌ها در نظر گرفته نمی‌شود زیرا دارای اطلاعات غیرضروری است. در مثال ارائه شده مولفه اول به تنهایی ۹۴ درصد از واریانس تغییرات را توجیه می نماید. بنابراین استفاده از همین یک مولفه اطلاعات زیادی را توجیه می نماید. در این زمینه بین صاحب نظران اختلاف نظر وجود دارد. آزمون‌های آماری برای بررسی کفایت تعداد مولفه با استفاده از شبیه سازی بصورت پارامتریک و غیرپارامتریک انجام شده است که با استفاده از آن شبیه سازی‌ها می‌توان بصورت دقیق‌تر تعداد مولفه را مشخص کرد (Forkman and Piehpho, 2014). اگر همبستگی بین متغیر بالا باشد با تعداد مولفه کمتری سهم بیشتری از تغییرات تبیین می‌گردد. ولی اگر همبستگی متغیرها پایین باشد تعداد مولفه‌های بیشتری باید محاسبه شوند. وقتی مولفه‌های اول نتوانند سهم زیادی از تنوع موجود در داده‌ها را تبیین کنند تفسیر نتایج مشکل می‌شود. در این صورت بهتر است از تجزیه خوشه‌ای استفاده گردد.

منابع:

فرشادفر، ع (۱۳۸۹). اصول و روش های تجزیه و تحلیل های آماری چند متغیره. انتشارات دانشگاه رازی کرمانشاه. ۷۳۴ص.

Forkman, J., and Piepho, H.P. (2014). Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. Biometrics, 70(3), 639-647.

Littell, R.C., Milliken, G.A., Stroup, W., Wolfinger, R.D., and Oliver, S. (2006). SAS for Mixed Models. SAS Institute. 828 P.