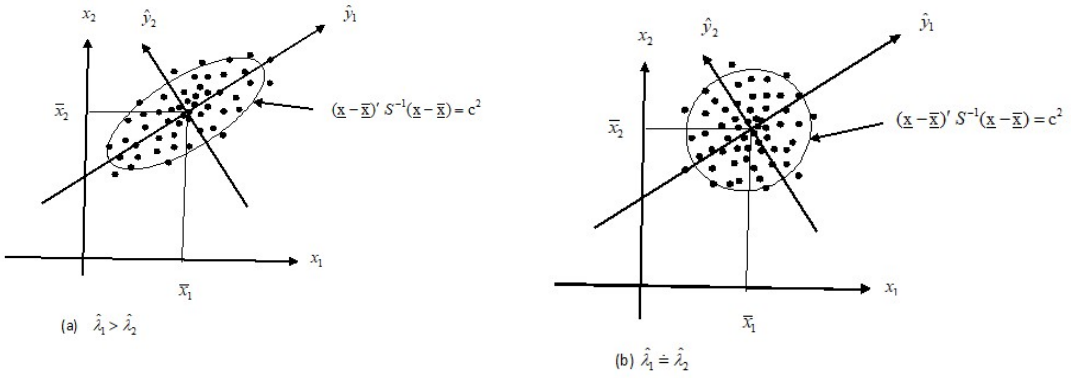
4. HAFTA

Temel Bileşenlerin Geomitrik Yorumu

Geometriksel olarak veriler *p* boyutlu uzayda, *n* tane nokta biçiminde gösterilebilir. Eğer **S** pozitif tanımlı ise, $(\underline{x}-\underline{\mu})'\,\Sigma^{-1}(\underline{x}-\underline{\mu})=c^2$ eşitliğini sağlayan bütün $px1$ tipindeki $\underline{x}$ vektörleri, eksenleri $\mathbf{S}^{-1}$ veya **S** nin özvektörleri ile verilen merkezi $\underline{x}$ olan bir hiper elipsoidi verirler. Bu eksenlerin uzunlukları $\sqrt{\hat{\lambda}_i}$; $i=1,2,...,p$ ' ye orantılıdır, burada $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ ' lar **S** 'nin özdeğerleridir. *i* inci temel bileşenin mutlak değeri, $|\hat{y}_i| = |\underline{\hat{e}}'_i(\underline{x}-\overline{\underline{x}})|$, $\underline{\hat{e}}_i$ birim özvektörü üzerine $(\underline{x}-\overline{\underline{x}})$ izdüşümünün uzunluğunu verir. Örneklem temel bileşenlerinin bu geometriksel yorumu *p*=2 için aşağıda verilmiştir.



Şekil a'da $\hat{\lambda}_1 > \hat{\lambda}_2$ olduğunda, merkezi $\overline{\underline{x}}$ olan sabit yoğunluk elipsidir. Örneklem temel bileşenleri iyi bir şekilde belirlenmiştir. Toplam örneklem varyansının büyük bir oranı birinci temel bileşen tarafından açıklanmaktadır. Bu temel bileşen elipsin büyük ekseni boyunca uzanmıştır. Şekil b'de $\hat{\lambda}_1 \doteq \hat{\lambda}_2$ olduğunda, merkezi $\overline{\underline{x}}$ olan sabit yoğunluk elipsi bir dairedir. Bu durumda eksenler tek biçimde belirlenememektedir. Toplam örneklem varyansı, her iki temel bileşen tarafından aynı oranda açıklanmaktadır. Bu durumda örneklem değişkenliği tüm yönler için aynıdır ve veriler *p* boyuttan daha az boyuta indirgenemez.

$\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$, normal bir kitleden alınan n birimlik örnekleme ilişkin gözlem değerleri ise, $\hat{y}_i = \underline{\hat{e}}'_i(\underline{x}-\overline{\underline{x}})$ örneklem temel bileşenleri, $Y_i = \underline{e}'_i(\underline{X}-\underline{\mu})$ kitle temel bileşenlerinin gerçekleşmesidir. Bu durumda kitle temel bileşenlerinin dağılımı $N_p(\underline{0},\Lambda)$ dır, burada $\Lambda = Diag\{\lambda_i\}$; $i=1,2,...,p$ olan diagonal bir matris ve $(\lambda_i, \underline{e}_i)$ ' ler **Σ** 'nın özdeğer ve birim

özvektör çiftleridir. Ayrıca örneklem verilerden elde edilen sabit yoğunluk elipsoidleri, $(\underline{x}-\underline{\mu})'\,\Sigma^{-1}(\underline{x}-\underline{\mu})=c^2$ sabit yoğunluk elipsoidlerinin tahminleridir. Normallik varsayımı temel bileşenlerin elde edilmesi için gerekli değildir, ancak sonuç çıkarımı için aranan bir koşuldur.

Genelde örneklem temel bileşenleri kitlede olduğu gibi değişkenlerin ölçü birimlerinden etkilenir. Değişkenlerin ölçü birimleri farklı ise standartlaştırılmış değişkenlere göre temel bileşenlerin elde edilmesi faydalı olacaktır. Örneklem için standartlaştırılmış *j* inci gözlem

$$\underline{z}_j = D^{-1/2}(\underline{z}_j-\underline{\bar{x}}) = \begin{bmatrix} \frac{x_{1j}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{2j}-\bar{x}_2}{\sqrt{s_{22}}} \\ . \\ . \\ . \\ \frac{x_{pj}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \begin{bmatrix} z_{1j} \\ z_{2j} \\ . \\ . \\ . \\ z_{pj} \end{bmatrix} ;\ j=1,2,...,n$$

biçimindedir ve standartlaştırılmış *pxn* tipindeki veri matrisi

$$Z = \begin{bmatrix} z_{11} & z_{12} & . & . & . & z_{1n} \\ z_{21} & z_{22} & . & . & . & z_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ z_{p1} & z_{p2} & . & . & . & z_{pn} \end{bmatrix}$$

olarak elde edilir. Buradan standartlaştırılmış gözlemlerin örneklem ortalama vektörü

$$\underline{\bar{z}} = \frac{1}{n} Z\underline{1} = \underline{0}$$

ve örneklem varyans kovaryans matrisi

$$\mathbf{S}_z = \frac{1}{n-1}(Z-\frac{1}{n}Z\underline{1}\,\underline{1}')(Z-\frac{1}{n}Z\underline{1}\,\underline{1}')' = \mathbf{R}$$

dır.

Böylece örneklem varyans kovaryans matrisi **S**′ nin yerine örneklem korelasyon matrisi **R** ′nin alınmasıyla, standartlaştırılmış gözlemlere ilişkin temel bileşenler elde edilmiş olur. Buradan standartlaştırılmış gözlemlerden elde edilen *i* inci örneklem temel bileşeni

$$\begin{aligned}\hat{y}_i &= \underline{\hat{e}}_i' \underline{z} \\ &= \hat{e}_{1i}z_1 + \hat{e}_{2i}z_2 + ... + \hat{e}_{pi}z_p \quad ; \; i = 1,2,...,p\end{aligned}$$

biçimindedir, burada $(\hat{\lambda}_i, \underline{\hat{e}}_i)$, **R**′ nin i inci özdeğer ve ilişkili birim özvektör çiftidir ve $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ ' dir. Buradan,

$\hat{y}_i$ nin örneklem varyans, $s^2_{\hat{y}_i} = \hat{\lambda}_i \quad ; \quad i = 1,2,...,p$

$(\hat{y}_i, \hat{y}_k)$ arasındaki örneklem kovaryansı, $s_{\hat{y}_i,\hat{y}_k} = 0 \quad ; \quad i,k = 1,2,...,p, \quad i \neq k$

dır. Ayrıca, Toplam örneklem varyansı;

$$\begin{aligned}tr(\mathbf{R}) &= p \\ &= \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p\end{aligned}$$

ve

$$r_{\hat{y}_i,z_k} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i} \quad ; \quad i,k = 1,2,...,p, \quad i \neq k$$

dir.

Ayrıca *i* incı örneklem temel bileşenine göre toplam varyansın açıklanan oranı

$$\frac{\lambda_i}{p}; \quad i = 1,2,...,p$$

dir.

ÖDEV 1 : Beş hisse senedi seçilerek haftalık kar oranları incelenmiştir. Kar oranı

$$\text{Bir hisse senedinin haftalık kar oranı} = \frac{\left(\begin{array}{l}\text{söz konusu haftanın}\\ \text{cuma günki kapanış fiyatı}\end{array}\right) - \left(\begin{array}{l}\text{bir önceki haftanın}\\ \text{cuma günki kapanış fiyatı}\end{array}\right)}{\left(\begin{array}{l}\text{bir önceki haftanın}\\ \text{cuma günki kapanış fiyatı}\end{array}\right)}$$

formülü ile hesaplanmaktadır. Birbirini izleyen 100 hafta için gözlemler bağımsız dağılımlı görünmektedir. Ancak hisselerin karşılıklı haftalık kar oranları ilişkilidir.

x_1 (Allied Chemical) , x_2 (Du Pont), x_3 (Union Clarbide), x_4 (Exxon) ve x_5 (Texaco) beş hisse senedi için haftalık kar oranlarına ilişkin gözlemleri göstersin. Örneklem ortalama vektörü,

$$\underline{\bar{x}}' = [0,0054 \quad 0,0048 \quad 0,0057 \quad 0,0063 \quad 0,0037]$$

ve standartlaştırılmış gözlemlerin örneklem varyans- kovaryans matrisi

$$R = \begin{bmatrix} 1 & 0,577 & 0,509 & 0,387 & 0,462 \\ & 1 & 0,599 & 0,389 & 0,322 \\ & & 1 & 0,436 & 0,426 \\ & & & 1 & 0,523 \\ & & & & 1 \end{bmatrix}$$

olmak üzere verilenlere göre temel bileşenleri elde ediniz. Temel bileşenlerin toplam örneklem varyansını açıklama oranılarını bulunuz. Sonuçları yorumlayınız.

Küresellik Testi

Bu test ile temel bileşenler analizinin gerekliliğine karar verilmektedir. İlgili hipotezler

$$H_0 : \boldsymbol{\rho} = \mathbf{I}$$
$$H_1 : \boldsymbol{\rho} \neq \mathbf{I}$$

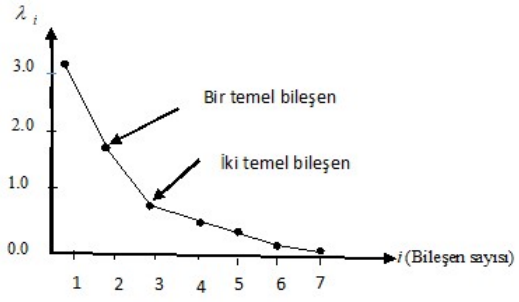
biçimindedir ve test istatistiği

$$-(n-1-\frac{2p+5}{6})\ln(|\mathbf{R}|) \sim \chi^2_{\frac{1}{2}p(p-1)}$$

dir. Eğer test istatistiğinin değeri, $\chi^2_{\frac{1}{2}p(p-1)}(\alpha)$ kritik değerinden büyükse H_0 hipotezi red edilir ve bu durumda temel bileşenler analizi yapmak anlamlıdır sonucuna karar verilir.

Uygun Temel Bileşen Sayısının Belirlenmesi

1. En basit karar verme süreçlerinden biri Varyans –kovaryans matrisinden temel bileşenler elde ediliyorsa $\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \geq \frac{2}{3}$ sonucunu veya korelasyon matrisinden temel bileşenler elde ediliyorsa $\frac{\sum_{j=1}^{m}\lambda_j}{p} \geq \frac{2}{3}$ sonucunu sağlayan en küçük *m* değeri uygun temel bileşen sayısıdır.
2. Korelasyon matrisinden temel bileşenler elde ediliyorsa, 1 den büyük özdeğer sayısı kadar temel bileşen alınabilir.
3. Yamaç eğrisi grafiğinde (Scree plot), temel bileşen sayısının, temel bileşenlerin varyanslarına (özdeğerlere) karşılık grafiği çizilir. Aşağıda verilen örnek yamaç grafiğinde iki temel bileşenin yeterli olduğu söylenebilir. Grafiğin düzleştiği noktadan önceki değer uygun temel bileşen sayısı olarak alınabilir.



4. Bartlet test yaklaşımına göre de uygun temel bileşen sayısı belirlenebilir. İlgili hipotezler

$$H_0 : \left|\boldsymbol{\rho}_g\right| = 0 \quad \text{veya} \quad H_0 : \lambda_{m+1} = \ldots = \lambda_{m+g} = 0$$
$$H_1 : \left|\boldsymbol{\rho}_g\right| \neq 0 \qquad H_1 : \textit{En az biri sıfırdan farklı} \qquad , \quad m+g=p$$

(Burada ilk *m* tane özdeğerin sıfırdan farklı, geriye kalan *p-m* tane özdeğirin sıfıra eşit olduğu iddia edilmektedir.)

biçimindedir ve test istatistiği

$$U_g = \frac{\left|R\right| g^g}{\left(\prod_{i=1}^{m} \hat{\lambda}_i\right)\left(p - \sum_{i=1}^{m} \hat{\lambda}_i\right)^g}$$

olmak üzere

$$-((n-1) - \frac{2p+5}{6} - \frac{2}{3}m)\ln(\left|U_g\right|) \sim \chi^2_{\frac{g}{2}(g+1)-1}$$

dir. Eğer test istatistiğinin değeri, $\chi^2_{\frac{g}{2}(g+1)-1}(\alpha)$ kritik değerinden büyükse H_0 hipotezi red edilir ve m değeri bir artırılarak hipotez kabul edilinceye kadar devam edilir.

Örnek 6 : *n*=11 ve *p*=5 olan bir örnekleme ilişkin örneklem varyans-kovaryans matrisi

$$S = \begin{bmatrix} 1974 & -4.473 & 726.9 & -2218 & -52.01 \\ & 1.488 & -26.62 & 197.5 & 1.577 \\ & & 1224 & -6203 & -56.44 \\ & & & 48104 & 328.9 \\ & & & & 4.087 \end{bmatrix}$$

olarak verilmiştir (Tatlıdil 1996).

a) Temel bileşen analizinin gerekli olup olmadığını $\alpha = 0.05$ anlam düzeyinde test ediniz.

b) Temel bileşen sayısının ne olacağına hem pratik yolla hem de testle karar veriniz.

Çözüm 6 :

a) Veriye temel bileşen analizi yapılıp yapılmayacağı Bartlett küresellik testi ile belirlenir. Küresellik testi özünde değişkenlere ilişkin korelasyon matrisinin (değişkenler arasında ilişki yoktur varsayımına dayanan) birim matrise karşı test edilme ilkesine dayanır. Bu nedenle Bartlett testi aynı zamanda korelasyon matrisinin anlamlılığının bir testidir. Hipotez

$$\begin{matrix} H_0 : \rho = I \\ H_1 : \rho \neq I \end{matrix} \quad , \quad \alpha = 0.05$$

şeklinde kurulur ve test istatistiği ise

$$-\left[n - 1 - \frac{2p+5}{6}\right] \ln\left|R\right| \sim \chi^2_{\frac{1}{2}p(p-1)}$$

biçimindedir. Eğer $\chi^2_h > \chi^2_{\frac{1}{2}p(p-1)}(\alpha)$ ise H_0 hipotezi reddedilir.

Şimdi test istatistiğinin değerini hesaplayalım:

$n = 11$ örneklem genişliği

$p = 5$ değişken sayısı

R : Korelasyon matrisi

$$R = \begin{bmatrix} 1 & -0.087 & 0.491 & -0.239 & -0.607 \\ & 1 & -0.624 & 0.758 & 0.640 \\ & & 1 & -0.808 & -0.798 \\ & & & 1 & 0.742 \\ & & & & 1 \end{bmatrix}$$

$$\ln\left|R\right| = \ln\left|0.0213224\right|$$

$$\chi^2_{hesap} = -\left(11 - 1 - \frac{(2).(5)+5}{6}\right)\ln\left|0.0213224\right| = 28.86$$

$$\chi^2_{tablo} = \chi^2_{\frac{1}{2}5(5-1)}(0.05) = 18.307$$

olup $\chi^2_h = 28.86 > \chi^2_{tablo} = 18.307$ olduğundan H_0 hipotezi reddedilir. Yani örneklemin alındığı kitleye ilişkin değişkenler arasındaki kitle korelasyon matrisi birim matristen farklıdır. Yani bazı değişkenler arasındaki korelasyonlar anlamlıdır. Temel bileşen analizi yapmaya gerek vardır.

(Korelasyon matrisi incelendiğinde bazı değişkenler arasında yüksek korelasyon olduğu da zaten görülmektedir.)

b) ***Pratik Yolla Tespit :***

$\frac{\sum_{i=1}^{k}\hat{\lambda}_i}{p} > \frac{2}{3}$ koşulunu sağlayan ilk *k* değeri uygun temel bileşen sayısıdır.

Korelasyon matrisi yardımıyla elde edilen özdeğerler ve herbir özdeğerin toplam varyansı açıklama oranına katkıları $\left(\frac{\hat{\lambda}_i}{p}\right)$ aşağıdaki gibi elde edilmiştir:

$\hat{\lambda}_i$	3.399	1.016	0.295	0.161	0.130
i. temel bileşenin toplam değişimi açıklama oranı	67.98	20.32	5.90	3.22	2.60

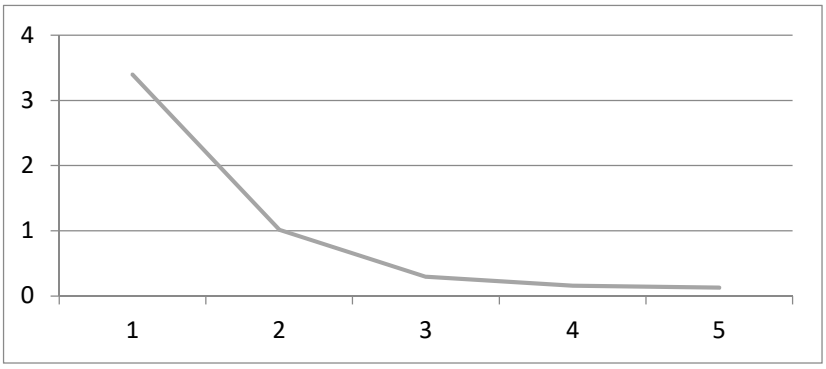
Not: Burada S varyans-kovaryans matrisi kullanılırsa toplam değişimi açıklama oranı

$$\left(\frac{\hat{\lambda}_i}{\sum\hat{\lambda}_i}\right)$$

formülü ile hesaplanır.

$\frac{\hat{\lambda}_1}{p} = 0.6798 > \frac{2}{3}$ olduğundan 1 temel bileşen yeterlidir : $\hat{Y}_1 = \hat{e}_1' Z$. Burada korelasyon matrisinden temel bileşenler yapıldığı için, orijinal değişkenler yerine standartlaştırılmış değişkenler dikkate alınmıştır.

Bununla birlikte aşağıdaki Scree Plot'a göre de uygun temel bileşen sayısına karar verilebilir. Bu grafikte eğri yatay eksene paralel olmaya başladığı bileşen sayısı, uygun temel bileşen sayısı olarak alınmaktadır.



<u>Test ile karar verme:</u>

m : uygun temel bileşen sayısı

g : önemsiz temel bileşen sayısı

p-g=m

$H_0 : |\rho_g| = 0$ veya $H_0 : \lambda_{m+1} = \lambda_{m+2} = ... = \lambda_p = 0$

$H_1 : |\rho_g| \neq 0$ $H_1 : \lambda_{m+i} \neq \lambda_{m+k} \quad , \quad i \neq k$

Test istatistiği

$$U_g = \frac{|R| g^g}{\left(\prod_{i=1}^{m} \hat{\lambda}_i\right)\left(p - \sum_{i=1}^{m} \hat{\lambda}_i\right)^g}$$ olmak üzere

$$-\left[(n-1) - \frac{2p+5}{6} - \frac{2}{3}m\right] \ln\left(U_g\right) \sim \chi^2_{\left[\frac{g}{2}(g+1)\right]-1}$$

biçimindedir. İlk olarak *g*=3 alalım. Bu durumda *m*=*p*-*g*=5-3=2 dir.

$$H_0 : \lambda_3 = \lambda_4 = \lambda_5 = 0$$
$$H_1 : \lambda_{m+i} \neq \lambda_{m+k} \quad , \quad i \neq k$$

$$U_g = \frac{\left|R\right| g^g}{\left(\prod_{i=1}^{m} \hat{\lambda}_i\right)\left(p - \sum_{i=1}^{m} \hat{\lambda}_i\right)^g} = \frac{(0.0213224)(3^3)}{\left(3.399\right)(1.016)\left(5 - (3.399 + 1.016)\right)^3} = 0.832$$

$$\chi^2_{hesap} = -\left[(11-1) - \frac{(2)(5)+5}{6} - \frac{(2)(2)}{3}\right]\ln\left(0.832\right) = 1.134$$

$$\chi^2_{tablo} = \chi^2_{\left[\frac{g}{2}(g+1)\right]-1}(\alpha) = \chi^2_{\left[\frac{3}{2}(3+1)\right]-1}(0.05) = 11.07$$

$\chi^2_{hesap} < \chi^2_{tablo}$ olduğundan H_0 hipotezi red edilemez. Yani $\alpha = 0.05$ hata ile $\lambda_3 = \lambda_4 = \lambda_5 = 0$ dır. Yani 3.,4. ve 5. temel bileşeni almaya gerek yoktur. Bu durumda aynı adımlar tekrarlanır: Bu defa *g*=2 alalım. *m*=*p*-*g*=3 olur.

$$H_0 : \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0$$
$$H_1 : \lambda_{m+i} \neq \lambda_{m+k} \quad , \quad i \neq k$$

$$U_g = \frac{\left|R\right| g^g}{\left(\prod_{i=1}^{m} \hat{\lambda}_i\right)\left(p - \sum_{i=1}^{m} \hat{\lambda}_i\right)^g} = \frac{(0.0213224)(4^4)}{\left(3.399\right)\left(5 - 3.399\right)^4} = 0.244$$

$$\chi^2_{hesap} = -\left[(11-1) - \frac{(2)(5)+5}{6} - \frac{(2)(1)}{3}\right]\ln\left(0.244\right) = 9.63$$

$$\chi^2_{tablo} = \chi^2_{\left[\frac{g}{2}(g+1)\right]-1}(\alpha) = \chi^2_{\left[\frac{4}{2}(4+1)\right]-1}(0.05) = 16.92$$

$\chi^2_{hesap} < \chi^2_{tablo}$ olduğundan H_0 hipotezi red edilemez. Yani $\alpha = 0.05$ hata ile $\lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0$ dır. Yani 2.,3.,4. ve 5. temel bileşeni almaya gerek yoktur. 1. Temel bileşeni almak yeterlidir.