Temel Bileşen Analizi (PCA)

Yusuf Talha ARABACI

Aralık 2022

İçindekiler

1 PCA'ya Genel Bakış

Bu bölümde Temel Bileşen Analizi(PCA) konusuna bir giriş yapacağız.

1.1 Temel Bileşen Analizi Nedir?

Temel Bileşen Analizi çok sayıda birbiri ile ilişkili değişkenler içeren veri setinin boyutlarını veri içerisinde var olan değişimlerin mümkün olduğunca korunarak indirgenmesini sağlayan bir dönüşüm tekniğidir. Verilerin paternini yüksek boyutlu veri setlerinde bulmak zor olabileceğinden, Temel Bileşenler Analizi(PCA), bizlere verileri analiz etmek için güçlü bir araç sunmaktadır. Kısacası, Temel Bileşenler Analizi;

- Veri kümelerinin boyutunu indirgemek
- Yorumlanabilirliği artırmak
- Bilgi kaybını en aza indirmek

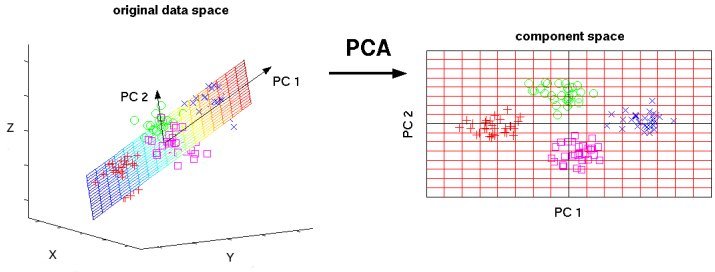
için kullanılan bir tekniktir.Temel Bileşenler Analizinin arkasında yatan temel mantık çok boyutlu bir veriyi, verideki temel özellikleri yakalayarak daha az sayıda değişkenle göstermektir.

1.2 Temel Bileşen Analizinin Çalışma Mantığı

Temel Bileşenler Analizi'nde p sayıda başlangıç değişkenine karşılık elde edilen p sayıda temel bileşenin her biri, orijinal değişkenlerin doğrusal bir bileşimidir. Dolayısıyla, her bir temel bileşen bünyesinde tüm değişkenlerden belirli oranda bilgiyi barındırır. Bu özelliği sayesinde Temel Bileşenler Analizi, p boyutlu veri kümesi yerine, ilk m önemli temel bileşenin kullanılması yoluyla boyut indirgemesi sağlayabilmektedir. İlk m temel bileşen toplam varyansın büyük kısmını açıklıyorsa, geriye kalan p-m temel bileşen ihmal edilebilir.

Veriye Temel Bileşenler Analizi uygulamadan önce mutlaka standardizasyon yapılmalıdır. Farklı ölçeklerdeki veriler yanıltıcı bileşenlere sebep olacaktır. Ayrıca analiz tekniği, aykırı gözlemlerden(outlier) fazlaca etkilenir. Analizden önce veriler aykırı gözlemlerden ayrılmalıdır

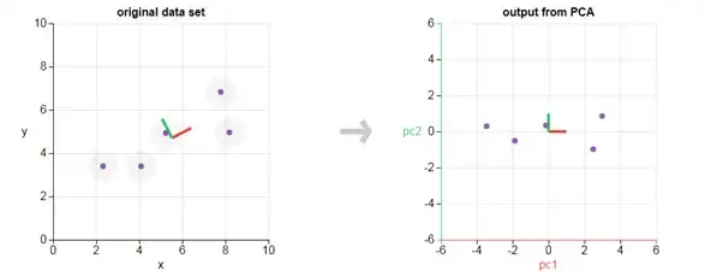
Uygulamada, ister sınıflandırma ister regresyon problemi olsun, bilgi içerdiğini düşündüğümüz gözlem verileri girdi olarak alınır. Ardından:



Şekil 1: Temel Bileşenler Analizi

1. Her boyut için ortalama vektör hesaplanır.
2. Kovaryans matrisleri hesaplanır.
3. Her boyut için özvektörleri ve karşılık gelen özdeğerleri hesaplanır.
4. Her bir özdeğerin özdeğerler toplamına bölünmesi ile temel bileşenlerin toplam
5. Her bir özdeğerin özdeğerler toplamına bölünmesi ile temel bileşenlerin toplam varyansı açıklama yüzdeleri elde edilir.

Bu varyans değerlerine göre yön hesaplanır. PCA ile yapmaya çalıştığımız öyle bir yön bulmak ki, veri noktalarının tamamının o yöne izdüşümü yapılınca varyansı en büyük olsun. Değişkenler arasındaki ilişkiyi en üst düzeye çıkarmak için bileşenler döndürülür.



Şekil 2: Bileşenlerin döndürülmesi

Burada ek olarak bahsetmek istediğim bir kavram olan Kovaryans ise iki değişkenin birlikte ne kadar değiştiklerinin ölçüsüdür. Bu matrisin özvektörleri ve özdeğerleri, PCA'nın temelini oluşturur.

1.3 Temel Bileşen Analizinin Kullanım Alanları

Temel Bileşen Analizinin kullanıldığı alanların bazıları; Yüz Tanıma, Resim Sıkıştırma ve Örrüntü Tanımadır. Bu algoritmanın sağladıkları; Verilerin boyutunu azaltma, Tahminleme yapma ve Veriyi görüntülemedir.

Temel Bileşen Analizinin kullanımına örnek olarak bir Hisse Senedi Tehminleme modeli verilebilir.

Bir hisse senedi performansının tahmine dayalı bir modelini oluşturmanızın istendiğini varsayalım. Doğru modelleme için stok performansıyla ilgili ölçümlerin tüm boyutlarını göz önünde bulundurmanız gerekir. Bu değişkenler, kapanış fiyatı, açılış fiyatı, gün içi yüksek, gün içi düşük, alfa, beta ve bireysel hisse senedi temellerinin yanı sıra finansal oranlar, işgücü piyasası, konut değişkenleri, duygu ölçümleri, GSYİH ölçümleri, enflasyon ve işsizlik gibi farklı kategorilerden oluşabilir. . Hisse seçimi için büyük bir aracı kurum tarafından sağlanan değişkenler, analiz için 30'dan fazla değişken sağlar.

Değişkenlerden bazıları birbiriyle ilişkili olabilir ve bazıları analiz için herhangi bir ek bilgi eklemeyebilir. Tüm değişkenler arasındaki korelasyonları sezgisel olarak değerlendirmek finansal okuryazar bir kişi için bile açık olmayabilir. PCA, bu büyük değişken kümesine uygulanabilir ve orijinal verilerdeki varyansın tamamına yakınını temsil eden azaltılmış bir alternatif değişkenler kümesi çıkarabilir. Daha küçük küme daha sonra stok performansı için farklı modelleri değerlendirmek için kullanılabilir.

2 PCA Hesaplama

2.1 Matematiksel Arka Plan

Temel Bileşen Analizi, orijinal verilerin aynı sayıda veya daha az boyutta bir projeksiyonunu hesaplamak için lineer cebir ve istatistiksel bazı basit matris işlemlerini kullanır. Lineer cebir kullanarak temel bileşenleri türetmek, kovaryans matrisine dayanır. PCA'nın genel bir versiyonu, tekil değer ayrıştırma (SVD) yöntemi kullanılarak türetilebilir. Bu bölümde PCA hesaplamada kullanacağımız matematiksel terimler verilmiştir.

2.1.1 Veri Temsili

X, tüm veri ve değişkenlerin gözlemlerini temsil etsin. O zaman X, X_{mn} matrisi olarak temsil edilebilir; burada m, gözlem sayısını ve n, ölçülen değişkenlerin sayısını temsil eder:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

Şekil 3: X_{mn} matrisi

2.1.2 Kovaryans Matrisi

Değişkenler arasındaki göreli ilişkinin hesaplanması, kovaryans hesaplaması kullanılarak yapılabilir. Değişkenler arasında ilişki yoksa kovaryans 0, değişkenler arasında yüksek korelasyon varsa, kovaryans değeri pozitif korelasyonla 1'e, negatif korelasyon için -1'e yakın olacaktır. Y boyutları (m, n) olan bir matris olsun ve sonra Y için kovaryans matrisi Denklem ile verilir:

$$E = \frac{1}{m}(YY^t) \tag{1}$$

Y^t, Y'nin transpozudur.

2.1.3 Diyagonal Matris

Köşegen elemanlar hariç, matrisin tüm elemanları 0 ise, bu matris bir diyagonal(köşegen) matristir. Önceki denklemdeki E matrisi köşegen bir matristir, o zaman Y'nin elemanları birbiriyle ilişkili değildir ve tüm gözlemlerin varyansını açıklar.

2.1.4 Özdeğer ve Özvektör

Özvektör, bir kare matris M'ye uygulandığında sadece skaler kısmını değiştiren ve yönünü değiştirmeden bırakan bir vektördür.

$$MV = aV \tag{2}$$

V bir vektördür, aynı zamanda M'nin özvektörüdür
a bir skalerdir, ayrıca V'ye karşılık gelen özdeğerdir.

2.1.5 Simetrik Matris

Simetrik bir matris, Denklem ile temsil edildiği gibi, devrik orijinal matrise eşit olma özelliğine sahiptir.

$$M = M^t \tag{3}$$

2.1.6 Temel Bileşenleri Türetme

Belirli bir X matrisi için, Denklem olacak şekilde bir P matrisi bulabilirseniz. 12.4a karşılanır, ardından P satırları X'in ana bileşenlerini temsil eder:

$$Y = PX \tag{4}$$

(YY^t) diyagonal bir matristir

Simetrik bir M matrisi için, aşağıdaki denklem karşılanır:

$$M = EDE^t \tag{5}$$

D bir diyagonal matris. E, M'nin bir özvektörü.

Sonucu gerçekten türetmeden, Denklem 4 ve 5'e karşılık gelen teoremleri uygulayarak, $Y = PX$ denkleminde X'in özvektörü olarak P seçilirse, YY^t'nin köşegen bir matris olduğu görülebilir. Dolayısıyla, X'in temel bileşenleri, P'nin sıraları olan XX^T'nin özvektörleridir.

2.1.7 Tekil Değer Ayrışımı(SVD)

Tekil değer ayrışımı, herhangi bir matris türü için yukarıdaki tekniğin genelleştirilmesidir.

Bir matrisin tekil değer ayrıştırması, aşağıdaki denklemde gösterildiği gibi, bir kovaryans matrisi ortonormal vektörler U, V ve diyagonal matris D'nin pozitif gerçek sayılarla çarpımı olacak şekilde matris X'in faktörlerini bulmaktır. V vektörünün sütunlarının bu ayrıştırmadaki ana bileşenleri temsil ettiği gösterilebilir:

$$X = UDV^T \tag{6}$$

Denklemin sezgisel bir yorumu, X'in m-boyutlu uzayı tarafından temsil edilen veri noktalarının, $n < m$ olduğu V'nin n-boyutlu uzayı tarafından dönüştürülmesi ve n boyut düzleminde sonuçtaki veri noktalarının ortalama karesel hatasını en aza indirmek için D ile ölçeklenmesidir.

2.2 İstatiksel Temel Bileşen Analizi

Temel Bileşenler Analizinin matematiksel tanımı olarak; "bir veri setinin en küçük karesel ortalama hata ile daha küçük boyutlu bir alt uzaya izdüşümünü sağlayan dönüşüm matrisini bulmamıza yarayan bir analiz yöntemidir" verilebilir.

2.2.1 Veri Özellikleri

PCA kullanılarak türetilen temel bileşenler, daha az sayıda değişkene sahip bilgisayarlı PC'lerin varyansındaki orijinal veri setinin genel varyansını hesaba katar. Ortalama ve varyansın, gözlem altındaki değişkenlerin tüm dağılımını tanımladığına dair üstü kapalı bir varsayım vardır. Varyansla tanımlanabilen bu sıfır ortalama dağılım, yalnızca Gauss veya normal dağılımda doğrudur.

Gözlenen değişkendeki büyük varyans, hesaplanan temel bileşenin genel varyansına en çok katkıda bulunacaktır. Değişkenlerin gözlemlenen değerleri çok farklı aralıklara sahipse, temel bileşen hesaplamaları için korelasyon analizi uygulanırken verilerin değişkenler genelinde normalleştirilmesi gerekir.

2.2.2 Veri Ön İşleme

Mesafe ve zaman olarak ölçülen iki değişken örneğini ele alalım. Mesafe santimetre cinsinden ve zaman saat cinsinden ölçülürse, ortaya çıkan ana bileşen

mesafe eksenine doğru eğridir. Bununla birlikte, mesafe kilometre cinsinden ve zaman saniye cinsinden ölçülürse, hesaplanan ana bileşen mesafe yerine zaman eksenine doğru eğilir. Değişken değerinin ölçeğinden kaynaklanan çarpıklık sorununun üstesinden gelmek için, veriler üzerinde ön işleme adımı olarak standardizasyon yoluyla Özellik Ölçeklendirmesi gerçekleştirilir. Z-ortalama normalleştirme olarak da adlandırılan standardizasyon, verileri ortalama 0'a ve standart sapma 1'e dönüştürür. Aşağıdaki denklem, z-normalleştirme yöntemini gösterir:

$$X_{ij} = \frac{X_{ij} - \mu}{\sigma} \tag{7}$$

X_{ij}, i indeksindeki veri noktasıdır
μ, j sütununa karşılık gelen X değişkeni için veri kümesinin ortalamasıdır
σ, j sütununa karşılık gelen X değişkeni için veri kümesinin standart sapmasıdır

2.2.3 Temel Bileşenleri Seçme

Temel bileşenler, bir kovaryans matrisinin hesaplanması ve elde edilen matris üzerinde Tekil Değer Ayrışımı(SVD) gerçekleştirilmesiyle hesaplanır.

$$\sigma = \frac{1}{m}(YY^T) \tag{8}$$

$$UDV^T = SVD(\sigma) \tag{9}$$

$$Pirincible\ Components = U^T Y \tag{10}$$

U matrisi, ana bileşenleri hesaplamak için kullanılan Y matrisinin özvektörlerini temsil eder.

U matrisinin k sütununun seçilmesi, k temel bileşeni sağlayacaktır. k değeri, temel bileşenlerden yakalanmak istenen varyans miktarına göre seçilebilir.

U_k ile temsil edilen U matrisinden k ana bileşenin seçildiğini varsayalım. Orijinal boyut verileri, Denklem 11'e göre indirgenmiş ana bileşenlerden hesaplanabilir. Orijinal boyut verilerini yeniden oluşturma yeteneği, Denklem 12'de verildiği gibi, veri azaltma nedeniyle ortaya çıkan hatayı hesaplamanıza da izin verebilir. Denklem 13 ve 14, k ana bileşen tarafından yakalanan varyanstaki hatayı ölçmek için bir yol sağlar:

$$Y = U_k \cdot PC \tag{11}$$

Y, çıktı veya sonuç matrisidir
U_k, özvektör U'dan seçilen k ana bileşenli matristir
PC, hesaplanan temel bileşenlerin katsayılarının matrisidir

$$E_p = \frac{1}{m} * \sum_{i=1}^{m} (y_i - y_{project\ i})^2 \tag{12}$$

E_p, öngörülen varyansta hatadır
y_i, y'nin orijinal değeridir
$y_{project\ i}$, seçilen k temel bileşenle hesaplanan/yeniden oluşturulan değerdir
m, y'nin gözlem sayısıdır

$$v_t = \frac{1}{m} * \sum_{i=1}^{m} (y_i)^2 \tag{13}$$

v_t, y'nin varyansıdır
y_i, y'nin orijinal değeridir
m, y'nin gözlem sayısıdır

$$E_t = \frac{v_t}{E_p} \tag{14}$$

E_t, k temel bileşen tarafından yakalanan varyansta hatadır
E_p, öngörülen varyansta hatadır
v_t, y'nin varyansıdır

Tipik olarak, varyanstaki hatanın 0,01'den (yani %1) daha az olarak yakalanmasını isteriz ve bileşenler, temel bileşenin veri kümesindeki orijinal varyansın %99'unu yakalamasını sağlayacak şekilde seçilir.

Denklem 9'dan, diyagonal matris D, temel bileşenlerin her birine atfedilen varyansı azalan düzende yakalar. Bu nedenle, hata, hesaplanan toplam varyansın oranını hesaplayarak basit bir şekilde hesaplanabilir:

$$Error = 1 - \frac{\sum_i^k D_{ii}}{\sum_n^1 D_{ii}} \tag{15}$$

k, kullanılan ana bileşenlerin sayısıdır
n, hesaplanan ana bileşenlerin toplam sayısıdır
D, SVD ayrıştırmasında dşyagonal matristir
i, matristeki hücrenin indeksidir

2.3 R ile implementasyon

Bu bölümde, R ortamında bu işlemin nasıl gerçekleştirileceği adım adım açıklanacaktır.

2.3.1 Verileri Yükleme

öncelikle, verileri görselleştirmek ve değiştirmek için çeşitli fonksiyonlar içeren Tidedverse paketini yükleyeceğiz:

```
library(tidyverse)
```

Ardından kullanacağımız verisetini R ortamına yükleyeceğiz. Bu örnekte USArrests datasetini kullanacağız.Aşağıdaki kod, veri kümesinin ilk birkaç satırının nasıl yüklenip görüntüleneceğini gösterir:

```
#load data
data("USArrests")

#view first six rows of data
head(USArrests)
```

```
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

2.3.2 Temel Bileşenleri Hesaplama

Verileri yükledikten sonra, veri kümesinin temel bileşenlerini hesaplamak için R'ın hazır fonksiyonu olan prcomp()'u kullanabiliriz.

Temel bileşenleri hesaplamadan önce veri kümesindeki değişkenlerin her birinin ortalama 0 ve standart sapma 1 olacak şekilde ölçeklenmesi için scale= TRUE belirttiğinizden emin olun.

Ayrıca R'deki özvektörlerin varsayılan olarak negatif yönü gösterdiğine dikkat edin, bu nedenle işaretleri tersine çevirmek için -1 ile çarpacağız.

```
#calculate principal components
results <- prcomp(USArrests, scale = TRUE)

#reverse the signs
results$rotation <- -1*results$rotation

#display principal components
results$rotation
```

```
               PC1        PC2        PC3         PC4
Murder   0.5358995 -0.4181809  0.3412327 -0.64922780
Assault  0.5831836 -0.1879856  0.2681484  0.74340748
UrbanPop 0.2781909  0.8728062  0.3780158 -0.13387773
Rape     0.5434321  0.1673186 -0.8177779 -0.08902432
```

İlk temel bileşenin (PC1) Cinayet, Saldırı ve Tecavüz için yüksek değerlere sahip olduğunu görebiliriz, bu da bu temel bileşenin bu değişkenlerdeki en fazla varyasyonu açıkladığını gösterir.

Ayrıca, ikinci temel bileşenin (PC2) UrbanPop için yüksek bir değere sahip olduğunu görebiliriz, bu da bu ilke bileşeninin en çok kentsel nüfusa ağırlık verdiğini gösterir.

Her durum için temel bileşen puanlarının result$x'de saklandığını unutmayın. Ayrıca işaretleri tersine çevirmek için bu puanları -1 ile çarpacağız:

```
#reverse the signs of the scores
results$x <- -1*results$x

#display the first six scores
head(results$x)
```

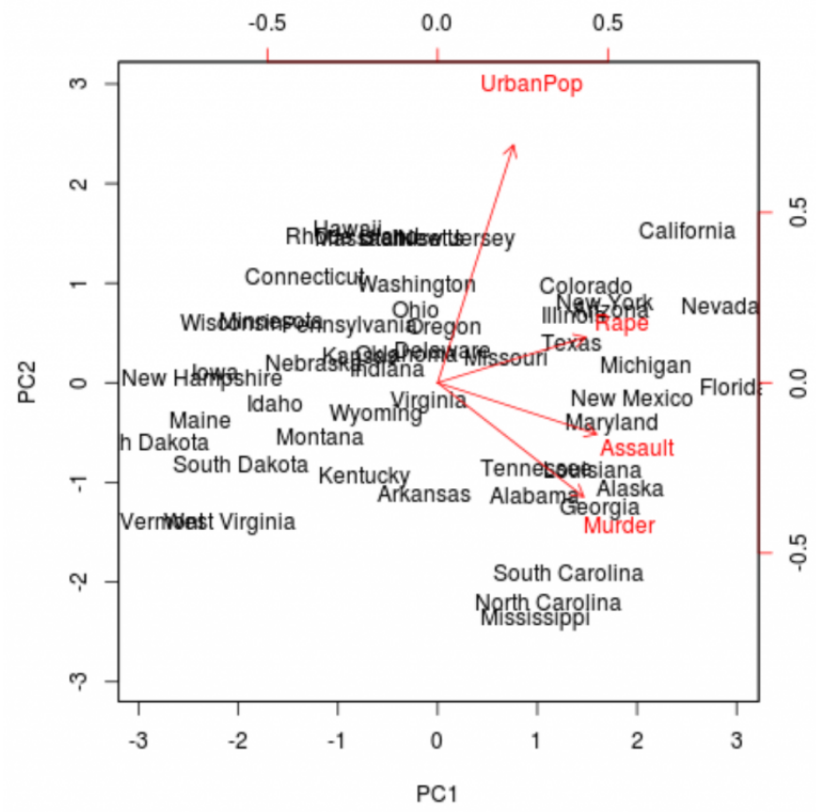
```
                 PC1        PC2         PC3          PC4
Alabama     0.9756604 -1.1220012  0.43980366 -0.154696581
Alaska      1.9305379 -1.0624269 -2.01950027  0.434175454
Arizona     1.7454429  0.7384595 -0.05423025  0.826264240
Arkansas   -0.1399989 -1.1085423 -0.11342217  0.180973554
California  2.4986128  1.5274267 -0.59254100  0.338559240
Colorado    1.4993407  0.9776297 -1.08400162 -0.001450164
```

2.3.3 Sonuçları Görselleştirme

Ardından, veri kümesindeki gözlemlerin her birini eksen olarak birinci ve ikinci ana bileşenleri kullanan bir dağılım grafiğine yansıtan bir çizim olan bir biplot oluşturabiliriz:

scale = 0'ın çizimdeki okların yüklemeleri temsil edecek şekilde ölçeklenmesini sağladığını unutmayın.

```
biplot(results, scale = 0)
```



Grafikten, basit bir iki boyutlu uzayda temsil edilen 50 durumun her birini görebiliriz. Grafik üzerinde birbirine yakın olan durumlar, orijinal veri setindeki değişkenler açısından benzer veri modellerine sahiptir.

Belirli eyaletlerin belirli suçlarla diğerlerinden daha fazla ilişkili olduğunu da görebiliriz. Örneğin, Georgia, olay örgüsünde Cinayet değişkenine en yakın eyalettir. Orijinal veri setinde cinayet oranlarının en yüksek olduğu eyaletlere göz atacak olursak, aslında listenin başında Georgia'nın yer aldığını görebiliriz:

```
#display states with highest murder rates in original dataset
head(USArrests[order(-USArrests$Murder),])
```

```
               Murder Assault UrbanPop Rape
Georgia          17.4     211       60 25.8
Mississippi      16.1     259       44 17.1
Florida          15.4     335       80 31.9
Louisiana        15.4     249       66 22.2
South Carolina   14.4     279       48 22.5
Alabama          13.2     236       58 21.2
```

2.3.4 Temel Bileşenler Tarafından Açıklanan Varyansı Bulma

Her bir temel bileşen tarafından açıklanan orijinal veri kümesindeki toplam varyansı hesaplamak için aşağıdaki kodu kullanabiliriz:

```
#calculate total variance explained by each principal component
results$sdev^2 / sum(results$sdev^2)
```

```
[1] 0.62006039 0.24744129 0.08914080 0.04335752
```

Sonuçlardan şunları gözlemleyebiliriz:

- İlk temel bileşen, veri kümesindeki toplam varyansın
- İkinci temel bileşen, veri setindeki toplam varyansın
- Üçüncü temel bileşen, veri setindeki toplam varyansın
- Dördüncü temel bileşen, veri setindeki toplam varyansın

Böylece, ilk iki temel bileşenin, verilerdeki toplam varyansın büyük bir bölümünü açıkladığı sonucuna varırız.

Bu iyi bir işarettir, çünkü önceki ikili grafik, orijinal verilerdeki gözlemlerin her birini yalnızca ilk iki ana bileşeni hesaba katan bir dağılım grafiğine yansıtmıştır. Bu nedenle, birbirine benzer durumları belirlemek için biplottaki kalıplara bakmak geçerlidir.

3 PCA Uygulamaları

3.1 Görüntü Sıkıştırma

Görüntü sıkıştırma, boyut küçültme tekniği kullanan PCA'nın yaygın bir uygulamasıdır. Şekil 4, 1200 x 795 piksel çözünürlükte çekilmiş bir ay resmini göstermektedir.

Şekil 4: orjinal ay görseli

```
moon <-- readJPEG(\moon.jpg")
moon_pca <- prcomp(moon, center¼FALSE)
percent_variance ¼ sum((moon_pca$sdev[1:k])^2)/sum((moon_pca$sdev^2))
```

PCA sonuçları, temel bileşenlerle ilişkili standart sapmaları içeren moon_pca nesnesinde depolanıyor. Kodun son satırındaki k, dikkate alınan temel bileşenlerin sayısını gösterir ve dolayısıyla varyans yüzdesini etkiler. Şekildeki tablo, k değerine bağlı olarak temel bileşenler tarafından yakalanan yüzde varyansı gösterir (yani, dikkate alınan temel bileşenlerin sayısı).

Number of PCs	Percent variance
5	96.3
10	98.1
20	98.9
100	99.9

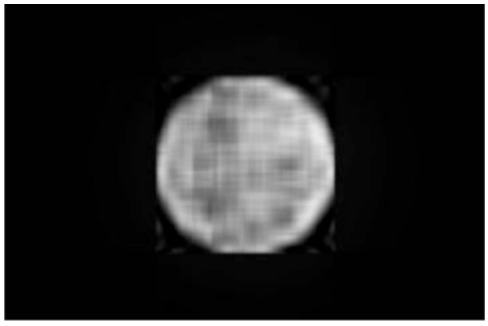
Şekil 5: yüzde varyansları

Bu sonuçlara göre, yalnızca 10 Temel Bileşen seçerek görüntüyü %98 doğrulukla yeniden oluşturabilirsiniz.Bu da, 10/795'lik bir sıkıştırmadır.

Orijinal veriler, aşağıdaki kod alıntısı kullanılarak PC'lerden çoğaltılabilir.

```
moon_compressed ← moon_pca$x[,1:k] * transpose(moon_pca$rotation[,1:k])
```

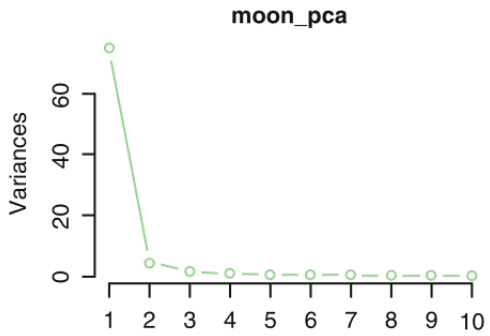
5 temel bileşen ve 20 temel bileşen ile yeniden yapılandırılmış görseller şekil 6 ve şekil 7'de verilmiştir.



Şekil 6: 5 PC

Şekil 7: 20 PC

Şekil 8, temel bileşenlerin sayısıyla ilişkili varyansları temsil eder.



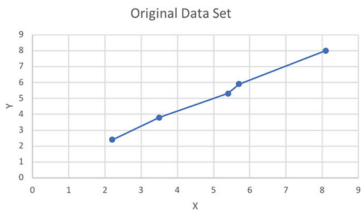
Şekil 8: Temel bileşenlerin sayısıyla oluşturulan varyans grafiği

3.2 Veri Görselleştirme

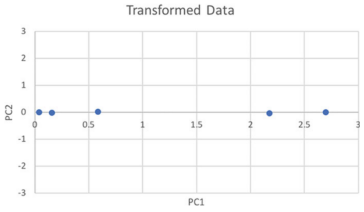
Çok sayıda değişkene sahip veri kümeleri, genellikle değişkenler tarafından temsil edilen varyansın çoğunu yakalayan daha küçük bir değişken alt kümesine sahiptir.

Bir düzlemde çizilen verilerle binaların tabanı ve yüksekliği gibi iki boyutlu bir veri kümesi düşünün. PCA uygulaması, bileşenlerdeki varyansı yakalayan yeni boyut uzayını ortaya çıkaracaktır.

Orijinal boyut verileri Şekil 9'ta çizilmiştir ve alternatif boyut Şekil 10'da çizilmiştir. Orijinal x ve y boyutları, Temel Bileşen1(PC1) ve Temel Bileşen 2(PC2)'nin alternatif boyutlarına dönüştürülür.



Şekil 9: orjinal veri



Şekil 10: dönüştürülmüş veri

Yukarıdaki örnekte dönüşüm, x, y koordinatlarındaki veri noktalarının bir koordinat sistemine döndürülerek eşlenmesinin basit boyutsal dönüşümü olarak yorumlanabilir. Koordinatların dönüşü, yeni eksenle eşleşen PC1 ve PC2 ile ana koordinatı elde etmek için çizginin eğimine eşittir.

Varyansın çoğunun PC1 tarafından dönüştürülmüş boyutta yakalandığını görebilirsiniz. Dönüştürülen veriler, PC1'de x ve y arasındaki doğrusal ilişkiyi yakalar ve bu doğrusal ilişkideki tüm farklılıklar PC2'de yakalanır. PC1 değişkeni, binanın genişliği veya yüksekliği gibi fiziksel bir anlam ifade etmeyebilir; ancak, ilgili boyutların sayısı büyükse önemli faydaları olan verilerin görselleştirilmesinde ve keşfedilmesinde yararlıdır.

4 PCA Uygulamasının Zorlukları

4.1 Overfitting

Özellik sayısı çok olduğunda ve yeterli veri olmadığında modelde aşırı uydurma meydana gelir. Öznitelik sayısını azaltmak için, bazen öznitelik sayısının azalmasının aşırı uydurmayı önleyeceği varsayımı altında PCA kullanılır. Aşırı uydurmayı önlemek için PCA kullanmak kötü sonuçlara yol açabilir. Aşırı uydurma sorunu, düzenlileştirme(regularization) teknikleri ve daha iyi veri kümeleri kullanılarak ele alınabilir.

4.2 Model Üretimi

Model oluşturmak için Temel Bileşenlerin kullanılması, yeni girdiler için verilerin tahmininde büyük hatalara neden olabilir. Bir alternatif, verilerden orijinal faktörleri çıkarmaktır. Ortak faktör analizi denilen bu yöntem, gözlenen değişkenlerin varyansını maksimize edebilen orijinal değişkenlerin kombinasyonunu tanımlar. Bu, orijinal değişkenlerin toplam varyansını maksimize eden PCA'dan farklıdır. Bu değişkenlere, gizli faktörler veya gözlemlenen değişkenlerin birleşik varyansını kapsayan gizli değişkenler de denir. Bu hesaplama faktörleri yöntemi, korelasyon matrisinin yalnızca köşegende değil, aynı zamanda matrisin diğer kısımlarında da yüksek değerlere sahip olduğu durumlarda kullanışlıdır. Gizli faktörler, modellemede Temel Bileşenlerden daha iyi performans gösterme eğilimindedir. Orijinal bileşenlere dayalı modelleme tercih edilmeli ve yalnızca boyutların ölçeğiyle ilgili sorunlar ortaya çıkarsa, Temel Bileşenlere dayalı model oluşturmayı düşünmelisiniz.

4.3 Model Yorumu

Temel bileşen yükleri, değişkenler için korelasyon faktörünü temsil eder; yüklemenin boyutuna ve işaretine göre değişkenin pozitif ilişkili veya negatif ilişkili olduğu sonucuna varabilirsiniz. Ancak, 1.Temel bileşende düşük yükleme değerine sahip bir değişken, diğer Temel bileşenlerde daha yüksek değerlere sahip olabileceğinden, yükleme boyutu ve yönü hakkında fazla yorum yapamazsınız. Modeli yorumlamaya çalışırken ortak faktör (CF) analizinin kullanılması daha iyi bir yorum sağlayacaktır.

5 Özet

Özetle, Temel Bileşen Analizi(Principal Component Analysis - PCA), veri kümesinde var olan değişkenliği mümkün olduğunca korurken, çok sayıda birbiriyle ilişkili değişkenden oluşan bir veri kümesinin boyutsallığını azaltmak için uygulanan çok güçlü bir teknikt. PCA'nın amacı, veri setlerindeki çok sayıda değişkeni, her biri orijinal değişkenlerin doğrusal bir fonksiyonu olan önemli miktarda daha az sayıda bileşen ile temsil etmektir. Bazı değişkenler arasındaki ilişkiyi en üst düzeye çıkarmak için bileşenler döndürülür. Bir Varimax rotasyonundan sonra, her bir orijinal değişken bir (veya az sayıda) bileşenle ilişkili olma eğilimindedir. Sonuç olarak, daha az sayıda temel bileşen, daha büyük orijinal değişkenler grubuyla aynı sonuca ulaşmayı sağlar. PCA çevresel verilerin değerlendirilmesinde yararlı bir teknik olarak gösterilmektedir. Temel bileşenler analizi gruplar arasındaki farklılıkları incelemek ve aralarındaki faktörleri belirlemek amacı ile kullanılmıştır.

Bu raporda, Temel Bileşen Analizi tekniğinin Matematiksel modeli, R ile implemente edilmesi ve uygulamalarına değinilmiştir. Çalışma mantığı ve kullanım alanları anlatılmıştır.

Kaynaklar

[1] Principal Component Analysis Networks and Algorithms - Springer

[2] An Introduction to Machine Learning - Springer

[3] https://aylablgn.medium.com/temel-bileşen-analizi-principal-component-analysis-pca-makine-öğrenmesi-4-be5dd634463b

[4] https://medium.com/machine-learning-türkiye/temel-bileşen-analizi-pca-c58c99718d3

[5] https://medium.com/@gulcanogundur/pca-principal-component-analysis-temel-bileşenler-analizi-bf9098751c62

[6] https://en.wikipedia.org/wiki/Principal_component_analysis

[7] https://youtu.be/fqegK5IyfqU

[8] https://www.statology.org/principal-components-analysis-in-r/

[9] https://www.datacamp.com/tutorial/pca-analysis-r

[10] https://tr.theastrologypage.com/principal-component-analysis