1. **Proje Konusu**

Temel Bileşen Analizi (Principal Component Analysis), Bağımsız Bileşen Analizi (Independent Component Analysis), Negatif Olmayan Matris Faktorizasyonu (Nonnegative Matrix Factorization)

1. **Projenin Amacı**

Temel Bileşenler Analizi, Bağımsız Bileşenler Analizi ve Negatif Olmayan Matris Faktorizasyonu tekniklerini araştırmak, algoritmalarını öğrenmek. Bu tekniklerin günlük hayattaki kullanımlarını ve aynı zamanda teorikteki uygulamalarını öğrenip kendi problemlerimizi çözmek.

1. **Konu Tanımı**

**1.TEMEL BİLEŞENLER ANALİZİ (PCA)**

Çok değişkenli bir veri seti içerisindeki bilgiyi daha az değişkenle ve minimum bilgi kaybıyla açıklamanın bir matematiksel tekniğidir. Temel amacımız daha az değişkenle minimum bilgi kaybı maksimum başarı ve performanstır. Temel Bileşenler Analizini değişken sayısını azaltmak ve çoklu doğrusallığı (multicollinearity) önlemek için veya gözlem sayısına göre çok fazla tahminciniz olduğunda kullanılır.

Proje de yüksek boyutlu bir veri setine uygulanmıştır.

Temel Bileşen Analizinin üç temel amacı vardır:

* Verilerin boyutunu azaltma
* Tahminleme yapma
* Veri setini bazı analizler için görselleştirmek

PCA uygulandığında p boyutlu uzayın gerçek boyutu belirlenir. Bu gerçek boyuta temel bileşenler adı verilir. Temel bileşenlerin üç özelliği vardır:

* Kolerasyonsuzlardır.
* Birinci temel bileşen toplam değişkenliği en çok açıklayan değişkendir.
* Bir sonraki temel bileşen kalan değişkenliği en çok açıklayan değişkendir.

Temel Bileşen Analizi bu soruna şu şekilde yaklaşmaktadır:

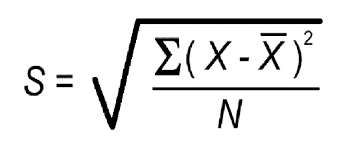
* Çok boyutlu verilere doğru açıdan bakarak genellikle verideki ilişkiler açıklanabilir.
* PCA’nın amacı bu “doğru açıyı” bulmaktadır.

**MATEMATİKSEL TEMEL**

Bu bölümde PCA’nın anlaşılabilmesi için gerekli olan bazı matematiksel ifadelere değinilecektir.

* **Standart Sapma**

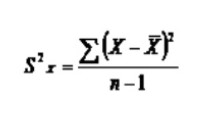
Standart sapma, verilerin nasıl yayıldığına (saçıldığına) dair ölçümsel olarak bilgi verir. Veri değerlerinin yayılımının özetlenmesi için kullanılan bir ölçüdür. Standart sapma varyansın kareköküdür.



X çizgi, X dizisinin ortalamasını ifade etmektedir.

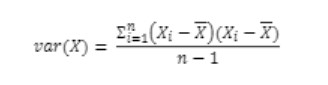
* **Varyans**

Varyans verinin yayılımı ile ilgili bir başka ölçüm bilgisi veren kavramdır. Genellikle değişimi ölçmek için kullanılır. Varyans, standart sapmanın karesidir.

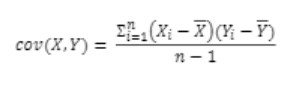


* **Kovaryans**

Olasılık teorisi ve istatistikte, kovaryans iki değişkenin birlikte ne kadar değiştiklerinin ölçüsüdür. Kovaryans, iki rasgele değişkenin beraber değişimlerini inceleyen bir istatistiktir. Standart sapma ve varyans tek boyutlu veriler için kullanılmaktadır. Ancak çoğu zaman veri setleri birden fazla boyuta sahiptir. Kovaryans her zaman iki boyut arasında ölçüm yapmak için kullanılmaktadır.



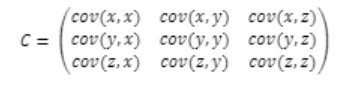
Yukarıdaki formüle benzer şekilde X değişkeni ve Y değişkeni arasındaki ölçüme bakılmak istendiğinde aşağıdaki formül yardımıyla bu iki değişken arasındaki ilişkiye bakılabilir.



Kovaryans değeri, pozitif ise her iki değişkenin birlikte arttığı; negatif ise biri artarken diğerinin azaldığı; sıfır ise bu iki değişkenin bağımsız olduğu yorumu yapılabilir.

* **Kovaryans Matris**

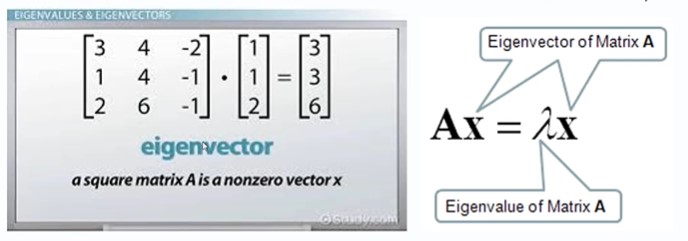
İkiden fazla değişkene bakıldığında kovaryans matris kullanılır. Kovaryans matristeki diyagonal değerler değişkenlerin varyans değerlerine eşittir.Kovaryans matris simetrik bir yapı sergilemektedir.



* **Özvektörler ve Özdeğerler**

Özvektör (Eigen vector), doğrusal bir dönüşüm uygulandığında yönü değişmeden kalan vektördür.

Özdeğer (Eigen value),verilerde ne kadar değişkenlik olduğunu gösteren sayıdır.



**TEMEL BİLEŞEN ANALİZİNİN UYGULAMA ADIMLARI**

**Temel Bileşen Analizi genel olarak 5 temel adımdan oluşur:**

1. Verileri hazırlama

Veriyi Merkezle: Burada her değişkenden ortalaması çıkartılarak veri merkezlenir. Böylece ortalaması 0 olan bir veri seti elde edilir.

Veriyi Ölçekle: Şayet orijinal verisetinin varyansları (aralıkları) birbirinden çok farklı ise verileri bilinen teknikle 0 ortalama ve 1 varyanslı hale dönüştürmek iyi bir fikirdir. Değişkenler karşılaştırılabilir hale gelir.

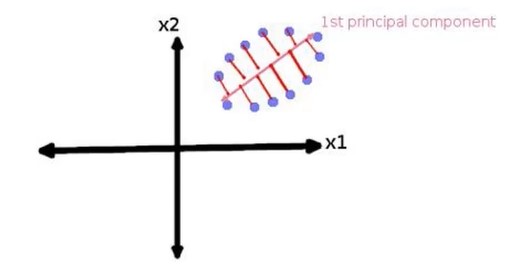
1. Kovaryans/Korelasyon matrisini oluşturma
2. Kovaryans/Korelasyon matrisinin öz değerlerini ve öz vektörlerini hesaplama
3. Temel bileşenleri (Principal Components) seçme
4. Yeni veri setini hesaplama

Öz vektörleri satır olarak transpoze et. Satırları değişkenler ve sütunları vakalar olan veri setini transpoze et.

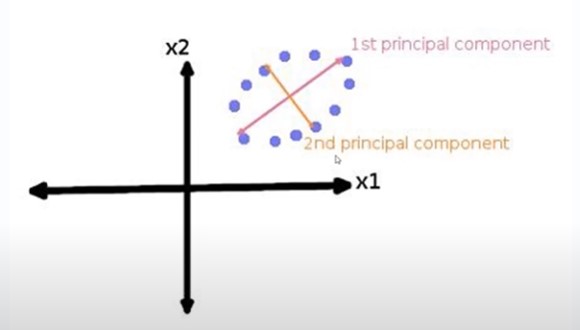
Yeni.Veriseti = (Transpoze.Özvektörler) X (Transpoze .Ayarlanmış Veri seti)

**ADIM ADIM TEMEL BİLEŞENLER ANALİZİ**

PCA veri kümesini daha düşük boyutlu bir alt alana, yani yeni bir koordinat sistemine dönüştürür. Yeni koordinat sisteminde, ilk eksen birinci ana bileşene karşılık gelir. Verilerde en büyük varyans miktarını açıklayan bileşendir.



İkinci Temel Bileşen, birinci ana bileşene dik gelecek şekilde seçilir. Böylece veri kümesinin ilk iki ana bileşene yansıltıldığı gibi görünür.



Örneğin: PCA boyutsallığı azaltabilir, ancak verilerinizdeki değişken sayısını azaltmaz. Bunun anlamı, 1000 özellik veri kümenizdeki varyansın % 99’unu yalnızca 3 temel bileşen kullanarak açıklayabileceğimizdir. Ancak yine de bu 3 temel bileşeni oluşturmak için bu 1000 özelliğe ihtiyacımız vardır.   
Bu aynı zamanda tahmin durumunda da gelecekteki verilerde, karşılık gelen ana bileşenleri oluşturmak için yeni gözlemlerinizde hala aynı 1000 özelliğe ihtiyacımız vardır.

Temel Bileşenler Analizi uyguladığımızda daha az değişkenle verinin anlaşılır hale geldiğini gördük. Veride 6 bileşenle varyansın %88.8’ini açıklayabildik. Boyut indirgeme sayesinde görselleştirme işlemi daha da kolaylaştı.

**KULLANIM ALANLARI**

Temel Bileşenler Analizi ağırlıklı olarak yüz tanıma, bilgisayar görme ve görüntü sıkıştırma gibi alanlarda boyutsallık azaltma tekniği olarak kullanılmaktadır. Aynı zamanda finans, veri madenciliği, biyoinformatik, psikoloji vb. Alanlardaki yüksek boyutlu verilerdeki kalıpları bulmak için kullanılır.

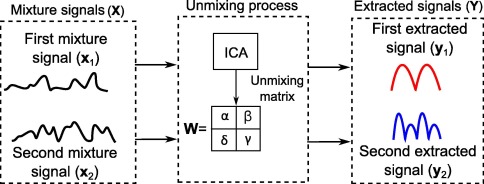
Projemizde yüz tanıma ile ilgili uygulamada da gördüğümüz gibi boyutsallığı azaltarak görüntülerin daha az alan kaplamasını sağlamaktadır.

1. **BAĞIMSIZ BİLEŞEN ANALİZİ (ICA)**

Bağımsız Bileşen Analizi (ICA), bağımsız kaynakları karışık bir sinyalden ayırmak için kullanılan bir makine öğrenme tekniğidir. Tipik olarak büyük ve çok değişkenli veri setleri için üretken bir model tanımlar. Veri noktalarının varyansını en üst düzeye çıkarmaya odaklanan temel bileşen analizinin aksine, bağımsız bileşen analizi bağımsızlığa, yani bağımsız bileşenlere odaklanır. Rastgele değişkenler, ölçümler veya sinyal kümelerinin altında yatan gizli faktörleri ortaya çıkarmak için kullanılır. Yaygın olarak kullanılan bir kör kaynak ayırma tekniğidir.  
Bağımsız Bileşen Analizi (ICA), tek bir kaynağı silebildiğinde veya koruyabildiğinde, boyutsallık azaltma algoritması olarak da kabul edilir . Buna, bazı sinyallerin filtrelenebildiği veya çıkarılabildiği filtreleme işlemi de denir.

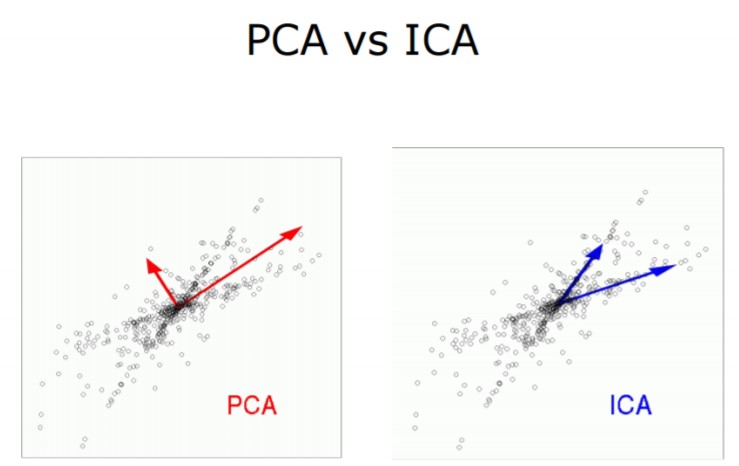
Modelde, veri değişkenlerinin bilinmeyen bazı gizli değişkenlerin doğrusal karışımları olduğu varsayılmaktadır ve karıştırma sistemi de bilinmemektedir. Gizli değişkenlerin normal olmadığı ve karşılıklı bağımsız olduğu varsayılır ve bunlara gözlemlenen verilerin bağımsız bileşenleri denir. Kaynaklar veya faktörler olarak da adlandırılan bu bağımsız bileşenler ICA tarafından bulunabilir.

Sinyal çıkarma işleminin açıklayıcı bir örneğini görüyoruz. İki kaynak sinyali (y1 ve y2) iki karışım sinyalinden (x1 ve x2) karıştırma matrisini kullanarak (W) birbirinden ayrılıyor.



Eğer A matrisi biliniyorsa, bu sorunun çözümü önemsizdir. A'nın basit bir matris işlemi ardından M ile çarpılması cevabı verecektir. Ancak gerçekte A matrisi genellikle bilinmemektedir. Elimizdeki tek bilgi karıştırma işleminin çıktısıdır.

**TEMEL BİLEŞEN ANALİZİ ve BAĞIMSIZ BİLEŞEN ANALİZİ ARASINDAKİ FARKLAR NELERDİR?**



|  |  |
| --- | --- |
| **TEMEL BİLEŞEN ANALİZİ** | **BAĞIMSIZ BİLEŞEN ANALİZİ** |
| PCA'da amaç, verilerinizin değişkenliğini en iyi açıklayan boyuta indirgemedir. PCA ‘nın ilk vektörü, verilerinizin değişkenliğini (maksimum varyans) en iyi açıklayan vektördür, ikinci vektör birincisine dik olmalıdır. | ICA'da amaç, her vektörün verilerinizin bağımsız bir bileşeni olduğu verilerdir, verilerinizi bir sinyal karışımı olarak düşünebilirsiniz. Daha her bağımsız sinyal için bir vektörü olacaktır. |
| Verimizi daha az değişkenle açıklamak istersek kullanırız. | Verimizi bağımsız alt öğeler olarak bir temsilini bulmak istediğinizde kullanırız. |
| Bazı bileşenler diğerlerinden daha önemlidir. (öz değerleri hatırlayın) | Tüm bileşenler eşit derecede önemlidir. |
| Vektörler diktir. | Bağımsız Bileşen Analizi vektörleri dikey değildir. |
| Temel bileşenler ile ilgilenir. | Bağımsız bileşenler ile ilgilenir. |
| Varyansı en üst düzeye çıkarmaya odaklanır. | Veri noktaları arasındaki sapma konusuna odaklanmaz. |

ICA, temel bileşen analizi (PCA) tekniğinin bir uzantısı olarak kabul edilir. Bununla birlikte, PCA ikinci dereceden istatistikleri temsil eden verilerin kovaryans matrisini optimize ederken, ICA basıklık gibi yüksek dereceli istatistikleri optimize eder Bu nedenle, PCA ilişkisiz bileşenler bulurken, ICA bağımsız bileşenler. Sonuç olarak, karışım verilerinin üst düzey korelasyonları küçük veya önemsiz olduğunda PCA bağımsız kaynakları çıkarabilir.

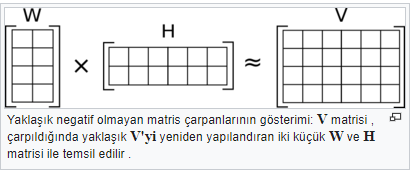
**KULLANIM ALANLARI**

ICA fiziksel olmayan sinyalleri analiz etmek için kullanılabilir. Örneğin, yakın tarihte ICA haber listesi arşivlerinin bir çanta üzerinde tartışma konuları keşfetmeye uygulanmıştır. Örneğin; ses sinyalleri ayrıştırma, nöronların optik görüntülemesi, yüz tanıma, borsa fiyatları tahmin etme, sigortacılık sektörü, cep telefonu iletişimi, EEG verileri vb. alanlarda sıkça kullanılır.

1. **NEGATİF OLMAYAN MATRİS FAKTORİZASYONU**

Negatif Olmayan Matris Faktorizasyonu (NMF), matrisleri negatif olmayan olarak kısıtladığımız bir matris çarpanlara ayırma yöntemidir. NMF, bir dizi negatif olmayan veri vektöründen seyrek ve anlamlı özellikleri otomatik olarak ayıkladığı için yüksek boyutlu verilerin analizi için yaygın olarak kullanılan bir araç haline gelmiştir.

Negatif olmayan matris çarpanlara ayırma (NMF), görüntü ve metin tanıma gibi birçok uygulamada popüler bir araçtır. Negatif olmayan elementlere sahip UxV matrisi A olarak verildiğinde, negatif olmayan, rank-k matrisleri W (UxK) ve H (KxV) bulmak istiyoruz.



Yani A matrisinin W ve H çarpanlarına ayrıştırıldığı yöntemdir. Öyle ki hem A matrisi, hem de çarpanları negatif olmayan değerler içerir. W ve H matrisleri yinelemeli olarak hesaplanır, ana yöntemlerden biri doğrusal regresyondur.

A ≈ WH

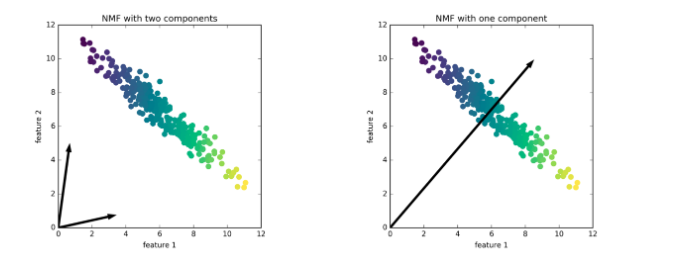
Başka bir deyişle, NMF bir boyut küçültme şeklidir. K ne kadar büyükse, yaklaşımımız o kadar iyi olur.

Çift sayılı yinelemelerde, H'yi bildiğimizi, ancak W bilmediğimizi varsayalım.

A '≈ H' W '

**NEGATİF OLMAYAN MATRİS FAKTORİZASYONU’NUN DEZAVANTAJI**

1. Yalnızca negatif olmayan verilere uygulanabilir.
2. Gelişigüzel yorumlanabilir.
3. Dışbükey olmayan optimizasyon başlatma gerektirir.
4. Dik değildir.



**KULLANIM ALANLARI**

Negatif olmayan matris faktorizasyonu da yüksek boyutlu verilerde tercih edildiği için birçok farklı kullanım alanı vardır. Örneğin; astronomi, metin madenciliği, spektral veri analizi, ölçeklenebilir internet mesafe tahmini (gidiş-dönüş süresi), biyoinformatik, yüz tanıma, nükleer görüntüleme vb. alanlarda kullanılır.

1. **KULLANILAN KÜTÜPHANELER VE FONSİYONLAR**

**library(dplyr);** R'deki veri kümelerini verimli bir şekilde işlemek için bir dizi araç sağlayan bir pakettir. **V**erileri çözmenize yardımcı olan tutarlı bir dizi fiil sağlayan en yaygın veri manipülasyonu dil bilgisidir.

**library(FactoMineR);** Veri kümelerini özetlemek, görselleştirmek ve tanımlamak için açıklayıcı veri analiz yöntemlerini içerir. Ana temel bileşen yöntemleri mevcuttur. (PCA için kullandık.)

**PCA();** Belirtilen veri temel bileşenler analizi gerçekleştirir. Eksik değerler, sütun ortalaması ile değiştirilir.

**prcomp ();** Veri matrisi üzerinde temel bileşenler analizini gerçekleştirir.

**library(factoextra);** PCA(Temel Bileşen Analizi) dahil olmak üzere çok değişkenli veri analizlerinin çıktısını çıkarmak ve görselleştirmek için kullanımı kolay bazı fonksiyonlar sağlar. Ayrıca, bazı kümeleme analizi adımlarını basitleştirmek için işlevler içerir ve 'ggplot2' tabanlı zarif veri görselleştirmesi sağlar.

**library(ggplot2);** Görselleştirme ve grafikler için kullanılan paket.

**library(RColorBrewer);** Renklerin bulundugu paket

**library(fastICA);** Bağımsız Bileşen Analizinde FastICA algoritmasının uygulanması için gerekli paket.

**fastica();** Bağımsız Bileşen Analizi için gerekli algoritmaları yapan fonksiyon.

**library("MASS");** Veri kümeleri içeren kütüphane.

**library(NMF);** Negatif Olmayan Matris Faktorizasyonunu (NMF) gerçekleştirmek için bir çerçeve sağlar.

**nmf();** NMF algoritmalarını NMF paketinde tanımlanan çerçeve içinde çalıştırmak için ana arayüzü tanımlar. NMF algoritmalarının uygulanmasını, geliştirilmesini ve test edilmesini kolaylaştıran birçok yönteme sahiptir.

**Kaynakça**

<http://www.zafercomert.com/IcerikDetay.aspx?zcms=78>

<https://www.sciencedirect.com/science/article/pii/S2210832718301819>

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>

<https://www.r-bloggers.com/reconstructing-principal-component-analysis-matrix/>

<https://uc-r.github.io/pca>

<http://compneurosci.com/wiki/images/4/42/Intro_to_PCA_and_ICA.pdf>

<https://blog.paperspace.com/dimension-reduction-with-independent-components-analysis>[/](https://blog.paperspace.com/dimension-reduction-with-independent-components-analysis/)

[https://medium.com/@datalabtr/r-ile-principal-component-analizi-pca-2b3034caf10a#:~:text=Principal%20component%20analysis%20(PCA)%20%C3%A7ok,buradaki%20toplam%20de%C4%9Fi%C5%9Fkenlik%20ile%20a%C3%A7%C4%B1klanmaktad%C4%B1r.](https://medium.com/@datalabtr/r-ile-principal-component-analizi-pca-2b3034caf10a)

<https://www.r-bloggers.com/quick-intro-to-nmf-the-method-and-the-r-package/>

[file:///C:/Users/ThinkPad/AppData/Local/Packages/microsoft.windowscommunicationsapps\_8wekyb3d8bbwe/LocalState/Files/S0/467/Attachments/ICA-Example\_DATE[1564].html](file:///C:\Users\ThinkPad\AppData\Local\Packages\microsoft.windowscommunicationsapps_8wekyb3d8bbwe\LocalState\Files\S0\467\Attachments\ICA-Example_DATE%5b1564%5d.html)

<https://mlexplained.com/2017/12/28/a-practical-introduction-to-nmf-nonnegative-matrix-factorization/>

<https://www.veribilimiokulu.com/makine-ogrenmesine-cok-degiskenli-istatistiksel-yaklasimlar-temel-bilesenler-analizi/>

<https://www.geeksforgeeks.org/ml-independent-component-analysis/>