T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZDE ROBUST İSTATİSTİKLERİN KULLANIMI

YÜKSEK LİSANS TEZİ

Hasan BULUT

İstatistik Anabilim Dalı

EYLÜL 2014
SAMSUN

T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZDE ROBUST İSTATİSTİKLERİN KULLANIMI

YÜKSEK LİSANS TEZİ

Hasan BULUT
(12211033)

Tezin Savuma Tarihi : 08/09/2014

Tez Danışmanı : Doç. Dr. Yüksel ÖNER

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalında

Hasan BULUT Tarafından Hazırlanan

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZDE ROBUST İSTATİSTİKLERİN KULLANIMI

başlıklı bu çalışma jürimiz tarafından 08/09/2014 tarihinde yapılan sınav ile YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Başkan : Doç. Dr. Mehmet GÜRCAN

Jüri Üyeleri : Doç. Dr. Kamil ALAKUŞ

Doç. Dr. Yüksel ÖNER

.../.../2014

Prof. Dr. Hüseyin DEMİR

Enstitü Müdürü

Aileme, Eşime ve Tüm Öğretmenlerime,

ÖNSÖZ

Çalışmam boyunca değerli yardım ve katkılarıyla beni yönlendiren saygı değer hocam Doç. Dr. Yüksel ÖNER'e, beni her zaman destekleyen sevgili Aileme ve Eşime, ayrıca 2211 Yurt İçi Lisansüstü Burs Programı desteğinden dolayı TÜBİTAK'a teşekkürü bir borç bilirim.

Eylül 2014

Hasan BULUT
(Araştırma Görevlisi)

İÇİNDEKİLER

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

KISALTMALAR

SD	**:** Standart Sapma (Standard Deviation)
Med	**:** Medyan (Median)
MAD	**:** Medyan Etrafında Mutlak Sapmaların Medyanı (The Median Absolute Deviation about Median)
MADN	**:** Normalleştirilmiş MAD (Normalized MAD)
SDE	**:** Stahel-Donoho Tahmin Edicisi (Stahel- Donoho Estimator)
MVT	**:** Çok değişkenli Budama (Multivariate Trimming)
MVE	**:** Minimum Hacim Elipsoidi (Minimum Volume Ellipsoid)
MCD	**:** Minimum Kovaryans Determinantı (Minimum Covariance Determinant)
RMCD	**:** Yeniden Ağırlıklandırılmış MCD (Reweighted MCD)
PP	**:** İzdüşüm İzleme (Projection Pursuit)
ROBPCA	**:** Robust Temel Bileşenler Analizi (Robust Principal Component Analysis)
TBA	**:** Temel Bileşenler Analizi
SEGE	**:** Sosyo-Ekonomik Gelişmişlik Endeksi

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZDE ROBUST İSTATİSTİKLERİN KULLANIMI

ÖZET

Bu çalışmada, bölgeler arasındaki gelişmişlik farklarının azaltılması amacıyla hizmet veren kalkınma ajansları kapsamındaki illerin sosyo-ekonomik gelişmişliklerinin değerlendirilmesi amacıyla klasik ve robust temel bileşenler analizi yöntemleri kullanılmıştır. Robust yöntemlerin kullanılma nedeni, bölgeler arasındaki gelişmişlik farklarının büyük olmasından dolayı ortaya çıkan aykırı değer sorunudur. Veri setinde aykırı değer olup olmadığı da yine klasik ve robust istatistiksel yaklaşımlarla incelenmiştir. Klasik temel bileşenler analizinde değişken sayısı gözlem sayısından daha az olmalıdır. Aksi takdirde kovaryans matrisinin determinantı sıfır olacaktır. Bu durumda klasik yaklaşımda değişken sayısı en fazla gözlem sayısının bir eksiği olabilir. Yeni bir yaklaşım olan Robust Temel Bileşenler Analizi (ROBPCA) yöntemi ile değişken sayısı gözlem sayısından fazla olsa da temel bileşenler analizi uygulanabilmektedir.

Yapılan uygulamada 26 kalkınma ajansı bölgesi öncelikle 19 değişken bakımından klasik ve robust korelasyon matrislerine dayanan temel bileşenler analizi ile, daha sonra da 46 değişken bakımından ROBPCA yöntemi ile değerlendirilmiştir.

Anahtar Kelimeler: Robust Temel Bileşenler Analizi; Robust Çok Değişkenli Tahmin Ediciler, Sosyoekonomik Gelişmişlik Endeksi.

THE USING OF ROBUST STATISTICS IN MULTIVARIATE STATISTICAL ANALYSIS

ABSTRACT

In this study, classic and robust principal component analyses are used to evaluate socioeconomic development of regions of development agencies that give service on the purpose of decreasing development difference among regions. Because development differences among regions are high, outlier problem is discussed. For this reason, robust statistical methods are used. Whether there is any outlier in the data set is also investigated by robust statistical methods. In the classic principal component analysis, the number of observation should be higher than the number of variable. Otherwise, determinant of covariance matrix is zero. In the ROBPCA that is a new approach, even if the number of variable is higher than the number of variable, principal component analysis is applied.

In this thesis, firstly 26 development agencies are evaluated with 19 variables by using principal component analysis based on classical and robust scatter matrices. Secondly, they are evaluated with 46 variables by using ROBPCA method.

Key Words: Robust Principal Component Analysis; Robust Multivariate Estimator; Socioeconomic Development Index.

1. GİRİŞ

1.1 Tezin Amacı

Toplanmış bir veride genel eğilimin dışına çıkan tipik olmayan gözlemler bulunabilir. Bu gözlemlere aykırı değer adı verilir. Bu gözlemlerin diğerlerinden aykırılık gösterme nedeni ölçüm hatası, yanlış veri girişi ya da gözlemlerin geldiği dağılımdaki farklılaşma olabilir. Fakat kaynağı ne olursa olsun, veri setindeki aykırı bir gözlem sonuçlar üzerinde olumsuz etkilere sahip olmaktadır. Aykırı değeri tespit etmek için geliştirilen yöntemler klasik tahminlere dayandıklarında, bu aykırı değerlerden etkilendiklerinden dolayı başarısız olabilirler.

Çok değişkenli veri yapılarında aykırı değer durumu tek değişkenli veri yapılarına göre daha karmaşıktır. Çünkü bir gözlem değişkenler tek tek incelendiğinde aykırılık göstermezken, tüm değişkenler aynı anda ele alındığında aykırı olabilir. Örneğin, 20 yaşında ve 3 kez boşanmış bir birey düşünülsün. 3 kez boşanmak boşanma sayısı değişkeni bakımından aykırı değildir. 20 yaşında olmak da evlilik için aşırı bir yaş değildir. Ancak bu iki değişken birlikte incelendiğinde, 20 yaşında 3 kez boşanmış bir bireyin toplumdaki genel eğilime nazaran oldukça sıra dışı olduğu söylenebilir (Alpar, 2011).

Bu çalışmanın asıl amacı, çok değişkenli veri yapılarında aykırı değer olması durumunda klasik konum ve yayılım parametrelerine önerilen robust alternatiflerini ele almak ve bu robust yayılım tahminlerinden elde edilen robust korelasyon matrislerini kullanarak temel bileşenleri elde etmektir. Ayrıca klasik yada robust korelasyon matrisine dayanan temel bileşenler analizinde mümkün olmayan yüksek boyutsal $(n < p)$ veri setleri için analizi mümkün kılan ROBPCA yöntemini kullanmaktır.

İkinci bölümde robust literatürüyle ilgili genel bilgilere yer verilmiştir. Genel bilgiler verilirken terimlerin daha rahat anlaşılabilmesi amacıyla tek değişkenli veri setleri ile örneklendirmeler yapılmıştır. Bu amaçla bir tahmin edicinin robustlık ölçüsü olarak kullanılan kırılma noktasına, birçok analiz için gerekli olan eş değişkenlik (equivariance) özelliğine değinilmiştir. Ayrıca tek ve çok değişkenli veri

setlerinde aykırı değer olup olmadığının incelemesi için genel olarak kullanılan yöntemler ele alınmıştır.

Üçüncü bölümde çok değişkenli veri setinin konum ve yayılım parametrelerinin tahmin edilmesinde kullanılan klasik ve bazı robust tahmin edicilere yer verilmiştir.

Dördüncü bölümde çok değişkenli veri yapılarında boyut indirgemek amacıyla kullanılan klasik temel bileşenler analizi tanıtılmış ve bu analize sunulan robust alternatifleri ele alınmıştır.

Beşinci bölümde ülkemizde bulunan 26 istatistiki bölgenin her birisinde bir tane olmak üzere kurulmuş olan kalkınma ajansı bölgeleri, bölgeler arasında aykırı değer sorunu tespit edildiğinden dolayı sosyoekonomik gelişmişlik düzeyleri bakımından klasik temel bileşenler analizi ve robust alternatifleri kullanılarak değerlendirilmiştir.

Altıncı bölümde ise sonuç ve önerilere yer verilmiştir.

1.2 Literatür Taraması

Huber PJ (1964) kirlenmiş bir veri setinde tek değişkenli konum parametresinin robust tahminini elde etmek için yeni bir yaklaşım önermiştir. Bu yaklaşımda en küçük kareler yöntemi üzerinde durulmuştur. En küçük kareler yönteminin amaç fonksiyonu olan $min \sum_i (x_i - T)^2$ ifadesini sağlayan T değeri örnek ortalamasıdır. Yeni yaklaşımda amaç fonksiyonu $min \sum_i \rho(x_i - T)$ şeklinde sabit olmayan bir ρ fonksiyonu ile güncellenerek aykırı değerlere karşı robust olan tahminlerin elde edilmesi sağlanmıştır. Böylelikle tek değişkenli konum parametresi için M tahmin edicisi elde edilmiştir.

Granadesikan ve Kettenring (1972) robust çok değişkenli konum ve yayılım parametreleri, çok değişkenli artıkların iki tipi ve çok değişkenli aykırı değerlerin tespiti ile ilgilenmişlerdir.

Devlin, Granadesikan ve Kettenring (1975) örnek korelasyon katsayısı üzerinde etkisi olan gözlemleri tespit etmek için iki grafiksel yöntem önermişlerdir ve robust korelasyon katsayısı geliştirmişlerdir. Daha sonra (1981) korelasyon matrisi ve bu matristen elde edilen temel bileşenlerin tahmin edilmesi için geliştirilen yöntemlerin karşılaştırılması amacıyla Monte Carlo yöntemlerini kullanmışlardır.

Maronna (1976) çok değişkenli bir veri setinin konum ve yayılım parametrelerinin tahmini için M tahmin edicilerini önermiştir.

Donoho (1982) çok değişkenli konum tahmin edicilerinin kırılma özellikleri ile ilgilenmiştir. Bu amaçla budanmış ortalama vektörü, ağırlıklandırılmış ortalama vektörü ve M tahmin edicilerinin kırılma özelliklerini ele almıştır. Çok değişkenli konum parametresi için ilk %50 kırılma noktasına ve afin eş değişkenlik özelliğine sahip tahmin edici Stahel (1981) ve Donoho (1982) tarafından birbirlerinden bağımsız olarak elde edilmiştir. Bu tahmin edici "aykırı değerlik ağırlıklandırılmış ortalama" olarak isimlendirilmiştir.

Rousseeuw (1983) çok değişkenli konum parametresi için geliştirilen ikinci bir tahmin edici geliştirmiştir. Bu tahmin edici minimum hacim elipsoidi (MVE) olarak adlandırılmaktadır. Daha sonra (1984) regresyon analizinde karesi alınmış artıklar toplamının minimizasyon problemine robust bir yaklaşım önererek, toplam yerine keresi alınmış artıkların medyanını kullanmıştır.

Walczak ve Massart (1995) elipsoidsel çok değişkenli budama (MVT) ve en küçük medyan kareler (LMS) yöntemlerine dayanan ve aykırı değerleri belirleme aracı olarak kullanılacak robust temel bileşenler regresyon yöntemini önermişlerdir.
Huber, Rousseeuw ve Branden (2005) robust temel bileşenler analizine yeni bir yaklaşım getirerek ROBPCA yöntemini önermişlerdir. Bu yöntem yüksek boyutsal ($n<p$) veriler içinde temel bileşenler analizini mümkün kılmaktadır.

Yaycılı (2006) temel bileşenler analizine alternatif olarak kullanılan robust temel bileşenler analizi ve bu analiz yardımıyla aykırı değerlerin belirlenmesi ile ilgilenmiştir.

Er ve Sönmez (2006) öğrencilerin başarı notları için robust faktör analizini kullanmıştır. Bu amaçla faktör analizinde kullanılan kovaryans ve korelasyon matrislerinin robust alternatiflerini kullanarak faktörleşme yapıları üzerinde durmuşlardır.

Koç (2007) klasik regresyon analizinde aykırı değerlerin varlığından dolayı varsayımların sağlanmaması durumunda en küçük kareler yöntemine alternatif olarak sunulan robust regresyon yöntemlerini ele almıştır.

Yazar, Yavuz ve Çay (2009) klasik temel bileşen analizi yöntemi yerine çeşitli klasik ve robust tahmin ediciler kullanılarak Yale, ORL ve AR yüz veri tabanları üzerinde yüz tanıma uygulamaları yapmışlardır.

Gümüş (2013) En Küçük Kovaryans Determinantı tahminlerine dayalı sağlam Mahalanobis uzaklıklarının dağılımına ilişkin yaygın olarak kullanılan ki-kare dağılımının aksine, aykırı gözlemlere karşılık gelen uç Mahalanobis uzaklıklar için F dağılımının daha uygun olduğu yönündeki çalışmaların sonuçlarını, benzetim yolu ile irdelemiştir.

2. GENEL BİLGİLER

Çok değişkenli veri yapıları ve parametre tahminlerinden bahsetmeden önce tek değişkenli veri yapıları ile ilgilenmek, temel oluşturmak bakımından önemlidir. Bu amaçla öncelikle Çizelge 2.1'de ki veriyi dikkate alarak tek değişkenli robustlıktan bahsedilecektir.

Çizelge 2.1. Kepekli undaki (bir milyon başına parçalar halinde) bakır miktarları

2,20	2,20	2,40	2,40	2,50	2,70	2,80	2,90
3,03	3,03	3,10	3,37	3,40	3,40	3,40	3,50
3,60	3,70	3,70	3,70	3,70	3,77	5,28	28,95

2.1 Konum ve Yayılım Parametrelerinin Tahmini

Mevcut bir veri setinin öncelikle konumu (merkezi) ve yayılımı (saçılımı, dağılışı) hakkında özetleyici bilgilere ihtiyaç duyulur. Bunların en yaygın olarak kullanılanları örnek ortalaması ve örnek standart sapmasıdır. $x_1, x_2, \dots, x_n$ gözlem değerleri olmak üzere örnek ortalaması ($\overline{x}$) ve örnek standart sapması (SD) şöyle tanımlanmaktadır:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ SD = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} \qquad \textbf{(2.1)}$$

Örnek ortalaması, verinin konumu ya da merkezi için; standart sapma ise verinin yayılımı (saçılımı, dağılışı) için iyi birer tahmindirler. Çizelge 2.1' de ki veri için örnek ortalaması ve standart sapması sırasıyla $\overline{x} = 4{,}28$ ve SD=5,30 olarak hesaplanmıştır. Eğer 28,95 değeri veriden silinirse örnekten hesaplanan bu değerler $\overline{x} = 3{,}21$ ve SD=0,69 olarak değişirler. Dolayısıyla aykırı değerin varlığında bu iki istatistiğin olumsuz etkilendiği söylenebilir.

Şimdi örnek ortalamasının ve örnek standart sapmasının aykırı değerlerden ne kadar etkilendiğini incelemek için, 28,95 değeri veriden çıkarılsın ve $(-\infty, \infty)$ aralığında herhangi bir değer bu gözlemin yerine alınsın. Tanımı gereği örnek ortalamasının mümkün değer aralığı da $(-\infty, \infty)$ olur. Dolayısıyla örnek standart sapmasının değer aralığı $(0, \infty)$ olacaktır. O halde, tek bir aykırı değerin dahi bu iki

istatistik üzerinde sınırsız bir etkiye sahip olduğu söylenebilir (Maronna ve diğ., 2006).

Bu nedenle veride aykırı değer olduğunda bu değerleri belirleyip onları veriden silmeyi düşünmek bir yol olabilir. Bu yöntem hiçbir şey yapmamaktan daha iyi olmasına rağmen bir aykırı değeri silme konusunda bazı sorunlar vardır.

- Silme işlemi ne zaman doğrudur? Bir gözlem, ne zaman silinecek kadar yeterli aykırılığa sahip olur?
- Araştırmacı bir gözlem bir gözlemdir diye düşünebilir ve verisinden herhangi bir eksilmeyi kabul etmeyebilir.
- Genellikle bir gözlemin gerçekten tipik olmayan olup olmadığı konusunda belirsizlikler vardır ve bu durum iyi bir gözlemi silme riskini ortaya çıkartır.
- Araştırmacının öznel kararına dayanan sonuçlardan dolayı, genel bir istatistiksel davranış belirlemek zordur.

Aykırı değer olduğunda kullanılan bir diğer yöntem ise "kırpma" olarak adlandırılır. Tipik gözlemlerin $[a, b]$ aralığında içerildiği varsayılsın. Örneğin $a = \overline{x} - 2\sigma$ ve $b = \overline{x} + 2\sigma$ şeklinde tanımlansın. Dolayısıyla bu aralığın dışındaki gözlemler aykırı değer olarak adlandırılacaktır. Silme yöntemi uygulandığında $[a, b]$ aralığının dışındaki tüm gözlemler veriden atılırken, kırpma işlemi uygulandığında $x_i < a$ olan gözlemlerin yerine a, $x_i > b$ olan gözlemlerin yerine b konulur. Diğer gözlemler üzerinde herhangi bir değişiklik yapılmaz. Yani bu yöntemde tipik olmayan değerler en yakın tipik olanlarla değiştirilir (Maronna ve diğ., 2006).

Verinin merkezini (konumunu) belirlemek için klasik yöntemlerden bir diğeri örnek medyanını kullanmaktır. Herhangi bir t sayısı, iki tarafında eşit sayıda gözlem olacak şekilde alınırsa veri setinin medyanı olarak adlandırılır ve veri seti $x_1, x_2, \dots, x_n$ olmak üzere eğer $s(x_i < t) = s(x_i > t)$ ise

$$t = Med(x) \qquad \textbf{(2.2)}$$

şeklinde ifade edilir. Sıra istatistikleri de medyanın bulunmasında kullanılabilir. $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ olmak üzere;

Eğer n tek sayı ise; $m = (n+1)/2$ olup

$$Med(x) = x_{(m)} \qquad \textbf{(2.3)}$$

dir.

Eğer n çift sayı ise; $m = n/2$ olup örnek medyanı $x_{(m)}$ ile $x_{(m+1)}$ arasında herhangi bir noktadır ve genel olarak

$$Med(x) = \frac{x_{(m)}+x_{(m+1)}}{2} \quad \textbf{(2.4)}$$

şeklinde hesaplanır.

Çizelge 2.1'de ki veri için tüm örneğin örnek medyanı 3,38 iken; aykırı değer atıldığında elde edilen verinin örnek medyanı 3,37'dir. Dolayısıyla örnek medyanının aykırı değere karşı duyarsız olduğu söylenebilir.

Daha önce örnek ortalaması ve dolayısıyla standart sapmanın aykırı değerlerden ne kadar etkilendiği incelenmişti. Aynı düşünce örnek medyanı için uygulanırsa; yani, 28,95 değeri veriden çıkarılır ve yerine $(-\infty, \infty)$ aralığından herhangi bir değer alınırsa bu durumda $x \to -\infty$ iken örnek medyanı 3,38 den 3,23'e düşer ve $x \to \infty$ iken 3,38 değerinde herhangi bir değişiklik olmaz. Dolayısıyla örnek medyanı üzerinde aykırı değerlerin çok az (istatistiksel olarak anlamsız) bir etkiye sahip olduğu söylenebilir (Maronna ve diğ., 2006).

Aykırı değerli ve aykırı değersiz durumlar için örnek ortalaması ve örnek medyanının durumlarını Şekil 2.1'de incelemek faydalı olacaktır. Şekil 2.1'de görüldüğü gibi aykırı değerin varlığı örnek ortalaması üzerinde çok büyük değişikliğe neden olurken, örnek medyanın da anlamsız bir fark oluşur. Aykırı değer varken verinin merkezini tahmin etmede kullanılan örnek ortalamasının 22 gözlem değerinden büyük ve sadece iki değerden küçük olduğu, yani verinin merkezini tahmin etmede başarısız olduğu görülmektedir.

Örnek medyanı aykırı değerlerden etkilenmediği ve verinin merkezi (konumu) için daha uygun bilgiler verdiğinden dolayı; örnek medyanı örnek ortalamasının iyi bir robust alternatifidir (Maronna ve diğ., 2006).

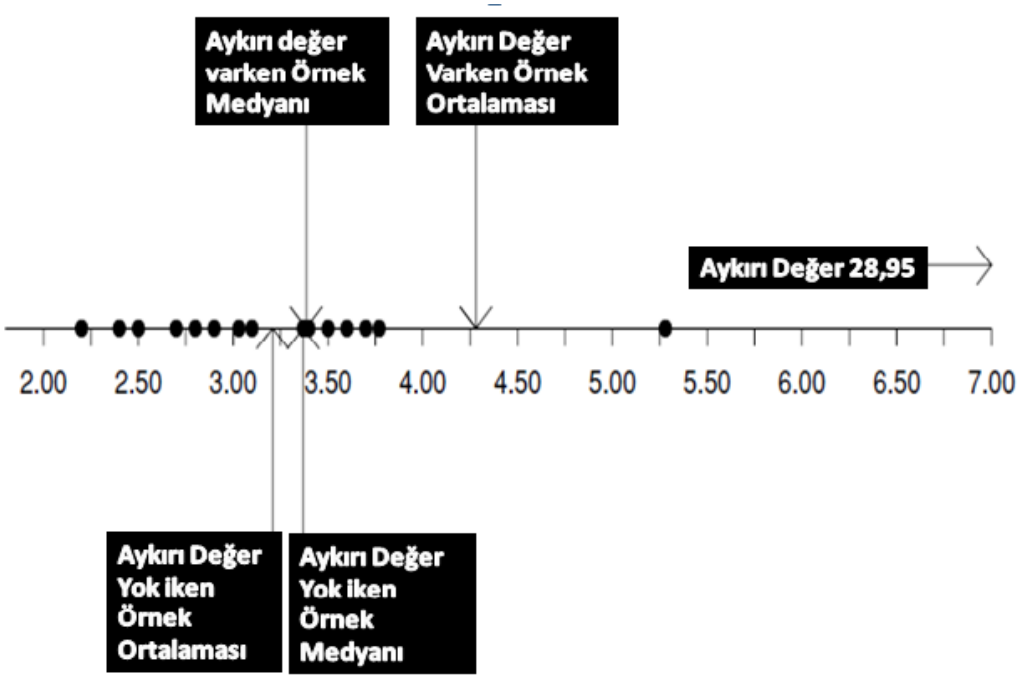
Benzer şekilde örnek standart sapması için iyi bir robust alternatifi medyan etrafında mutlak sapmaların medyanı (MAD) olup şu şekilde tanımlanır:

$$MAD(x) = Med\{|x - Med(x)|\} \quad \textbf{(2.5)}$$

Bu tahmin edici örnek medyanını iki defa kullanmaktadır. MAD'ı standart sapmayla karşılaştırabilmek için normalleştirilmiş MAD $(MADN)$

$$MADN(x) = \frac{MAD(x)}{0{,}6745} \quad \textbf{(2.6)}$$

şeklinde tanımlanır (Maronna ve diğ., 2006).



Şekil 2.1 Kepekli undaki (bir milyon başına parçalar halinde) bakır miktarına ilişkin gözlemlerin örnek ortalaması ve örnek medyanı

Çizelge 2.1'de ki veri için aykırı değerin olduğu durumda $SD = 5{,}30$ iken $MADN(x) = 0{,}53$'tür ve aykırı değer veriden çıkarıldığında $SD = 0{,}69$ iken $MADN(x) = 0{,}50$ olmaktadır.

Böylece $MADN(x)$, örnek standart sapmasının iyi bir robust alternatifidir (Maronna ve diğ., 2006). Tek değişkenli konum ve yayılım parametrelerinin tahmini için geliştirilen diğer tahminlere (Maronna ve diğ., 2006)'den ulaşılabilir.

2.2 Niçin Her Zaman Robust İstatistikleri Kullanılmıyor?

Eğer veri aykırı değer içermiyorsa bu robust tahminlerin istatistiksel performansı, örnek ortalaması ve standart sapmasından daha zayıftır. İdeal çözüm; veri aykırı değer içermediğinde klasik tahminler gibi davranan, ancak aykırı değer olduğu durumda da bu değerlere duyarsız tahminleri bulmaktır. Aynı şekilde klasik tahminler, veri varsayılan dağılıma uyduğunda optimal sonucu verirken; verinin dağılımı varsayılan dağılımdan küçük bir miktar farklılaştığında yetersiz kalırlar. Öte yandan robust tahminleri varsayılan modelden ufak bir farklılaşma olduğunda bile yaklaşık olarak optimal bir performans gösterir.

2.3 Bir Tahmin Edicinin Kırılma Noktası (ε^*)

Bir tahmin edicinin sağlamlığını ölçmenin yollarından biri kırılma noktasına bakmaktır. Bir tahmin ediciyi parametreyi tahmin etmede kullanışsız hale getirecek en küçük örnek miktarının, gözlem sayısına bölünmesi ile elde edilen değer, o tahmin edicinin kırılma noktası olarak adlandırılır (Moller ve diğ., 2006). Eğer verideki kirlenme %50'den fazla olursa, yani verideki gözlemlerin yarısından daha fazlası aykırı olarak ele alınırsa, sezgisel olarak hangi gözlemlerin aykırı olduğu ayırt edilemez ve verideki kirlenmeye neden olan kısmın geri kalan gözlemler olduğu düşünülür. Bu durumda da verideki kirlenmenin alacağı en büyük değer %50 olacaktır. Dolayısıyla bir tahmin edici en fazla %50 oranında kirlenmiş veriyle karşılaşabilir ve bu yüzden bir tahmin edicinin kırılma noktasının alacağı değer en fazla %50'dir. %50 kırılma noktasına sahip tahmin ediciler yüksek kırılma noktalı (high breakdown point estimator) olarak adlandırılırlar.

Örneğin; $x_1, x_2, \dots, x_n$ bağımsız olarak çekilen n çaplı bir örnek olsun. Böyle bir veri setinin konum parametresini tahmin etmek için kullanılacak istatistiklerden birisi örnek ortalaması olup,

$$\overline{x} = \frac{x_1+x_2+\cdots+x_n}{n} \qquad \textbf{(2.7)}$$

şeklinde tanımlanır. Bu tahmin edicinin kırılma noktası sıfırdır. Çünkü örnek ortalamasını verinin konumunu tahmin etmede kullanışsız hale getirmek için tek bir aykırı değer yeterlidir. Dolayısıyla tahmin edicinin verideki en ufak bir kirlenmeye karşı direnci yoktur. Bu nedenle örnek ortalamasının kırılma noktası, yukarıda ki tanım gereği $\varepsilon^* = {}^1\!/_n$ olup, $n \to \infty$ için alacağı değer sıfırdır.

Örnek medyanı ise %50 kırılma noktasına sahip bir tahmin edicidir. Çünkü örnek medyanını kullanışsız hale getirebilmek için verinin yarısının değiştirilmesi gerekir.

Elbette %50 kırılma noktası değeri asimptotik bir değerdir ve tahmin edicinin dayanabileceği maksimum kirlenme oranını ifade eder. Genellikle robustlık ve etkinlik birbirleriyle ters orantılı olarak değerlendirilirler ve yapılan çalışmalarda kırılma noktası ile etkinlik arasında optimal bir denge kurulmaya çalışılır. Bu durumda kırılma noktası düşürülerek tahmin edicinin istatistiksel etkinliği arttırılabilir.

2.4 Bir Tahmin Edicinin Eş Değişim (Equivariance) Özellikleri

Bir X rastgele değişkeni $N(\mu_x, \sigma_x^2)$ dağılımına sahip olmak üzere, bu kitleden çekilen n birimlik örnek $x_1, x_2, \dots, x_n$ olsun. Bu dağılımın parametreleri;

$$\mu_x = E(x) \text{ ve } \sigma_x^2 = Var(x) = E((x-\mu)^2) \quad \textbf{(2.8)}$$

şeklinde tanımlanır. Bu dağılım parametrelerini tahmin etmek için kullanılan klasik istatistikler sırasıyla örnek ortalaması ve örnek varyansı olup Eşitlik (2.1)'den yararlanarak elde edilebilirler. A ve b iki sabit olmak üzere bu rastgele değişken $Y = AX + b$ şeklinde afin bir dönüşüme tabi tutulursa, yeni değişkenin parametreleri;

$$\mu_y = E(y) = E(AX + b) = AE(x) + b = A\mu_x + b \quad \textbf{(2.9)}$$

$$\sigma_y^2 = Var(y) = Var(Ax + b) = A^2 Var(x) = A^2\sigma_x^2 \quad \textbf{(2.10)}$$

şeklinde tanımlanır. Böyle bir dönüşüm söz konusu olduğunda Eşitlik (2.1)'den elde edilen örnek ortalaması ve örnek varyansı da Eşitlik (2.9) ve Eşitlik (2.10)'da ki parametrelerle aynı şekilde;

$$\overline{y} = A\overline{x} + b \text{ ve } s_y^2 = A^2 S_x^2 \quad \textbf{(2.11)}$$

halini alır. Örnek ortalaması ve örnek varyansının bu özelliğine afin eş değişim (affine equivariance) denir. Sonuç olarak afin eş değişim özelliği herhangi doğrusal dönüşüm altında sonuçların değişmeden kalmasının istendiği bazı durumlarda gereklidir (Maronna ve diğ., 2006).

Genel olarak eş değişim (equivariance), verinin sistematik bir dönüşüme tabi tutulduğu durumda tahmin edicinin de ilgili dönüşüme uğraması olarak tanımlanır. Tahmin edicinin sahip olduğu eş değişim özelliğinin hangi tipte olduğu, dönüşümün şekline göre adlandırılır. Veriye ortogonal bir dönüşüm (döndürme ve yansıtma) uygulandığında tahmin edici de uygun şekilde dönüşüyorsa bu tahmin edicinin ortogonal eş değişim özelliğine sahip olduğu söylenir. Örneğin bir regresyon tahmin edicisi için regresyon, ölçek (scale) ve afin eş değişim olmak üzere üç tip eş değişim özelliği vardır (Moller ve diğ., 2006).

Afin eş değişim özelliği doğrusal diskriminant analizi, kanonik korelasyon ve faktör analizi gibi herhangi singüler olmayan doğrusal dönüşümler altında sonuçların değişmeden kalmasının istendiği durumlarda doğaldır. Ancak temel bileşenler analizi

sadece ortogonal dönüşümler altında değişmeyen bir metriğe dayandığından dolayı bu durum söz konusu değildir (Maronna ve diğ., 2006).

2.5 Aykırı Değerin Tespit Edilmesi

2.5.1 Tek değişkenli aykırı değerin tespit edilmesi (3σ kuralı)

Bir x_i gözleminin aykırı değerliğini ölçmenin geleneksel bir yöntemi, örnek ortalamasıyla arasındaki farkı örnek standart sapmasına bölmektir.

$$t_i = \frac{x_i - \mu}{s} \qquad \textbf{(2.12)}$$

$|t_i| > 3$ olan gözlem değerleri şüpheli olarak dikkate alınır. Bu yönteme $3\sigma\ Kuralı$ adı verilir. Çizelge 2.1'de ki veride en büyük gözlem değeri için $t_{24} = 4{,}65$ olarak hesaplanmıştır ve bu gözlemin aykırı değer olduğu sonucuna varılmıştır. Uzun tecrübelere rağmen bu kural bazı dezavantajlara sahiptir;

- İyi bir veride örnek çapı (n) çok geniş olduğunda, bazı gözlemler aykırı değer olarak adlandırılıp değiştirilebilir. Örneğin, geniş bir örnek çapındaki normal dağılımlı veriden çekilen 1000 gözlemin 3'ü için $|t_i| > 3$ olacaktır. Çünkü standart normal dağılımda $P(|t_i| > 3) = 0{,}003$ 'tür.
- Kural küçük örneklerde etkili değildir. Mümkün tüm örnek değerleri için $|t_i| = \frac{n-1}{\sqrt{n}}$ olup eğer $n \leq 10$ ise her zaman $|t_i| < 3$ olacaktır.
- Veride birden fazla aykırı değer olduğunda onların etkileri etkileşime geçer. Bu durumda tüm aykırı değerler tespit edilemeyebilir ki bu etkiye maskeleme adı verilmektedir (Maronna ve diğ., 2006).

Çizelge 2.2. Aykırı değerlerin birbirini maskelemesi

28	26	33	24	34	29	22	24	21	25
-44	27	16	40	-2	30	23	29	31	19

En düşük iki değer -44 ve -2 olmak üzere bu değerler veri için şüpheli konumdadır. Ancak bu iki gözlemin t_i değerleri sırasıyla -3,73 ve -1,35 olarak hesaplanır. Bu durumda -2 gözlemi aykırı değer olarak tespit edilmemiştir. Bunun nedeni bu iki gözlem değerinin örnek ortalamasını küçültmesi ve örnek standart sapmasını büyütmesidir. Bu nedenle -2 gözleminin t_i değeri mutlak değerce küçük çıkmıştır. O halde -44 değerinin -2 değerini maskelediği söylenebilir.

Bu maskeleme probleminin önüne geçmek için Eşitlik (2.12)'de örnek ortalaması yerine robust karşılığı olan örnek medyanını; örnek standart sapması yerine de robust alternatifi olan $MADN(x)$'i koyarak Eşitlik (2.12)'nin robust versiyonu;

$$t'_i = \frac{x_i - Med(x)}{MADN(x)} \quad \textbf{(2.13)}$$

şeklinde elde edilir ve en düşük iki değer için t'_i değerleri sırasıyla -11,73 ve -4,64 olarak hesaplanır. Böylece iki gözlemde aykırı değer olarak tespit edilir. O halde t_i için iyi bir robust alternatifi t'_i olarak bulunmuş olur.

Bu durum göstermiştir ki parametre tahmin etmek yerine, sadece aykırı değerler tespit edilmeye çalışılsa bile, robust tahmin edicilerine dayalı yöntemler daha güvenilir sonuçlar vermektedir (Maronna ve diğ., 2006).

2.5.2 Çok değişkenli aykırı değerin tespit edilmesi

Çok değişkenli bir veri setinde aykırı değerin bulunması için geliştirilen birçok yöntem vardır. Bunlardan en çok kullanılanı tek değişkenli durum için geliştirilen kurala benzer olarak Mahalanobis uzaklıklarına bakmaktır. Bir gözlemin Mahalanobis uzaklığı;

$$d_i = \left((\boldsymbol{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})\right)^{1/2} \quad \textbf{(2.14)}$$

şeklinde tanımlanır. Burada $\boldsymbol{\mu}$ ve $\boldsymbol{\Sigma}$ sırasıyla veri setinin ortalama vektörü ve kovaryans matrisidir. Karesi alınmış Mahalanobis uzaklığı, $\chi^2_{p,0.975}$ kritik değerinden büyük olan gözlemler aykırı değer olarak belirlenmektedirler (Rocke ve Woodruff, 1996).

3. ÇOK DEĞİŞKENLİ KONUM VE YAYILIM TAHMİN EDİCİLERİ

3.1 Klasik Tahmin Ediciler

Çok değişkenli bir veriyi karakterize etmek için verinin konumu ve yayılımı hakkında bilgi verecek istatistiklere ihtiyaç duyulur. Bu amaçla kullanılan klasik konum ve yayılım tahmin edicileri sırasıyla klasik örnek ortalama vektörü ve örnek varyans-kovaryans matrisi olup sırasıyla:

$$\widehat{\boldsymbol{\mu}}_c = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} \tag{3.1}$$

ve

$$\widehat{\boldsymbol{C}}_{\boldsymbol{c}} = \frac{1}{n-1}\left[\boldsymbol{X}\boldsymbol{X}' - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}'\right] \tag{3.2}$$

şeklinde tanımlanırlar. Burada p değişken sayısı, n örnek genişliği ve $\boldsymbol{X}: pxn$ boyutlu veri matrisidir. Eşitlik (3.1)'de tanımlanan ortalama vektörü aynı zamanda en küçük kareler tahmin edicisi olarak da düşünülebilir. Çünkü bu tahmin edici;

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|^2 \tag{3.3}$$

ifadesini minimize etmektedir. Burada $\|...\|$, L_2 normdur. Tıpkı en küçük kareler tahmin edicileri gibi bu tahmin edicilerinde kırılma noktası sıfırdır. Bu durum veride bir tane bile aykırı değer olması durumunda tahmin edicilerin farklılaşacağı ve parametreyi tahmin etmede başarısız olacağı anlamına gelmektedir. Bu nedenle bu parametreler için klasik tahmin ediciler yerine kullanılabilecek robust istatistiklerine ihtiyaç duyulmaktadır.

Burada çok değişkenli konum ve yayılım parametreleri için geliştirilmiş bazı robust tahmin edicilere yer verilmiştir.

3.2 M Tahmin Edicisi

M tahmin edicileri olabilirlik fonksiyonuna dayanarak elde edilmektedir. Öncelikle tek değişkenli M tahmin edicilerinden bahsetmek faydalı olabilir.

$$x_i = \mu + u_i\ ,\ i = 1,2,\dots,n \qquad \textbf{(3.4)}$$

modeline "konum modeli" denilmektedir. Buradaki $u_1, u_2, \dots, u_n$'ler rastgele değişkenlerdir. Eğer gözlemler aynı şartlar altında aynı deneyin bağımsız tekrarlarıysa $u_1, u_2, \dots, u_n$'lerin aynı F_0 dağılım fonksiyonuna sahip ve bağımsız olduğu varsayılır.

Bu bilgiler ışığında $x_1, x_2, \dots, x_n$ 'lerin de aynı $F(x)$ ortak dağılım fonksiyonuna sahip ve bağımsız oldukları söylenebilir. Bu dağılım fonksiyonu;

$$F(x) = F_0\,(x - \mu) \qquad \textbf{(3.5)}$$

şeklinde ifade edilir.

Bir $\hat{\mu}$ tahmini gözlemlerin fonksiyonu olup; $\hat{\mu} = \hat{\mu}(x_1, x_2, \dots, x_n)$ şeklinde gösterilir ve yüksek olasılıkla $\hat{\mu} \approx \mu$ olacak şekilde tahminler aranır.

Konum modelindeki u_i'nin dağılım fonksiyonunun F_0 ve yoğunluk fonksiyonu $f_0 = F'_0$ olduğu varsayılsın. O halde gözlemlerin olabilirlik fonksiyonu ;

$$L(x_1, \dots, x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu) \qquad \textbf{(3.6)}$$

şeklinde yazılabilir.

μ'nün en çok olabilirlik (EÇOB) tahmini $L(x_1, \dots, x_n; \mu)$ değerini maksimize eden ve $x_1, \dots, x_n$ 'lere bağlı $\hat{\mu}$ değeri olup ;

$$\widehat{\mu}\ =\ \widehat{\mu}\,(x_1, \dots, x_n) = argmax\ L(x_1, \dots, x_n; \mu) \qquad \textbf{(3.7)}$$

şeklinde yazılır ve burada argmax maksimize edilmiş değeri ifade etmektedir. Eğer f_0 her yerde pozitifse, logaritma artan bir fonksiyon olduğundan Eşitlik (3.7);

$$\hat{\mu} = argmin\ \textstyle\sum_{i=1}^{n} \rho(x_i - \mu) \qquad \textbf{(3.8)}$$

şeklinde yazılabilir ve burada

$$\rho = -log f_0 \qquad \textbf{(3.9)}$$

dır (Maronna ve diğ., 2006).

Eğer $\rho(x)$ türevlenebilirse μ'ye göre Eşitlik (3.8)'de en küçüklenmeye çalışılan fonksiyonun türevi;

$$\sum_{i=1}^{n} \varphi(x_i - \hat{\mu}) = 0 \quad \textbf{(3.10)}$$

olup, burada $\varphi(x) = \rho'(x)$ 'dir. Sonuç olarak; Bir $\rho(x)$ fonksiyonu verilmişken, konumun bir M tahmin edicisi Eşitlik (3.8)'in çözümüdür (Maronna ve diğ., 2006).

Eğer $F_0 = N(0,1)$ ise olasılık yoğunluk fonksiyonu

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \textbf{(3.11)}$$

şeklindedir ve sabit hariç olmaz üzere $\rho(x) = x^2/2$ olur. Böylece Eşitlik (3.8);

$$\hat{\mu} = argmin \sum_{i=1}^{n} (x_i - \mu)^2 \quad \textbf{(3.12)}$$

olarak yazılabilir. $\rho(x) = x^2/2$ olduğundan $\varphi(x) = x$ olur ve Eşitlik (3.10) $\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0$ halini alır ve bunun çözümü de $\hat{\mu} = \overline{x}$ 'dir. Dolayısıyla dağılım kesin şekilde normal ise, yani veride herhangi bir bozulma ya da kirlenme söz konusu değilse, bu dağılımın konum parametresi için M tahmin edicisi de örnek ortalaması olmaktadır.

Eğer F_0 çift üstel (double exponential) olarak da adlandırılan Laplace Dağılımı ise;

$$f_0(x) = \frac{1}{2} e^{-|x|} \quad \textbf{(3.13)}$$

'dir ve $\rho(x) = |x|$ olur. Böylece Eşitlik (3.8);

$$\hat{\mu} = argmin \sum_{i=1}^{n} |x_i - \mu| \quad \textbf{(3.14)}$$

şeklini alır. Burada $\rho(x) = |x|$ olduğundan $\varphi(x) = sgn(x)$ olup;

$$sgn(x) = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases} \quad \textbf{(3.15)}$$

'dir. Eşitlik (3.14)'te minimize edilen fonksiyon sürekli olduğundan dolayı, türevin işaretinin değiştiği yerde ki μ değerini bulmak yeterlidir. Dolayısıyla;

$$\varphi(x) = sgn(x) = I(x > 0) - I(x < 0) \quad \textbf{(3.16)}$$

olup, burada ki $I(.)$ gösterge fonksiyonu olarak adlandırılır ve şöyle tanımlanır ;

$$I(x > 0) = \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases} \quad \textbf{(3.17)}$$

Eşitlik (3.16), Eşitlik (3.10)'da yerine konursa;

$$\sum_{i=1}^{n} sgn(x_i - \hat{\mu}) = \sum_{i=1}^{n} [I(x_i - \hat{\mu} > 0) - I(x_i - \hat{\mu} < 0)] = 0 \quad \textbf{(3.18)}$$

$$s(x_i > \hat{\mu}) - s(x_i < \hat{\mu}) = 0 \quad \textbf{(3.19)}$$

olur ve böylece

$$s(x_i > \hat{\mu}) = s(x_i < \hat{\mu}) \quad \textbf{(3.20)}$$

elde edilir, ki bu durumda $\hat{\mu}$ herhangi bir örnek medyanıdır. O halde Laplace dağılımının konum parametresinin M tahmin edicisi örnek medyanıdır.

Bu düşünceyi çok değişkenli tahmin edicilere genellersek M tahmin edicisinin amaç fonksiyonu;

$$Minimize \sum_{i=1}^{n} \rho(\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_M\|) \quad \textbf{(3.21)}$$

ile verilmekte olup, burada $\widehat{\boldsymbol{\mu}}_M$ konumun M tahminidir (Moller ve diğ., 2006).

Bu tahmin edicilerin hepsi afin eş değişim özelliğini göstermezler. Bu özelliği göstermeyenlere örnek olarak L_1 konum tahmin edicisi gösterilebilir. "Konumsal (spatial) medyan" ya da "medyan merkez (center)" olarak da adlandırılan bu L_1 tahmin edicisi tek değişkenli durumdaki medyanın çok değişkenli versiyonudur ve tek değişkenlide olduğu gibi %50 kırılma noktasına sahiptir (Huber, 1964). L_1 tahmin edicisi sadece ortogonal eş değişim özelliğini sağlar. Afin eş değişimli bir çok değişkenli medyanı tanımlamanın bir yolu Mahalanobis uzaklıkları toplamını minimize etmektir (Roelant ve Aelst, 2007).

Afin eş değişimli konum ve yayılım parametreleri için M tahminleri Maronna tarafından önerilmiştir (Maronna, 1976). Bu tahmin edicilerin büyük bir dezavantajı kırılma noktalarının en fazla $1/(p+1)$ olmasıdır (Maronna ve diğ., 2006). Yani değişken sayısı çok fazla olmasa bile kırılma noktası düşüktür ve sadece p=1 olduğunda tahmin ediciler %50 kırılma noktasına sahip olabilmektedir.

Çok değişkenli konum ve yayılım parametrelerinin Eşitlik (3.21) fonksiyonunu sağlayan M tahminleri sırasıyla;

$$\widehat{\boldsymbol{\mu}}_M = \frac{\sum_{i=1}^{n} w_1[d(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_M, \boldsymbol{C}_M)]\boldsymbol{x}_i}{\sum_{i=1}^{n} w_1[d(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_M, \boldsymbol{C}_M)]} \quad \textbf{(3.22)}$$

$$\widehat{\boldsymbol{C}}_M = \frac{1}{n}\sum_{i=1}^{n} w_2\left[d^2(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_M, \widehat{\boldsymbol{C}}_M\right](\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_M)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_M)' \quad \textbf{(3.23)}$$

'dır (Croux ve Haesbroeck, 2000). Burada $d(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_M, \boldsymbol{C}_M) = \sqrt{(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_M)' \widehat{\boldsymbol{C}}_M{}^{-1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_M)}$ şeklinde hesaplanmakta olup, $\widehat{\boldsymbol{C}}_M$ pozitif tanımlı bir matris olmak üzere $\boldsymbol{x}_i$ ve $\widehat{\boldsymbol{\mu}}_M$ vektörleri arasındaki istatistiksel

uzaklığı ifade eder. Burada ki w_1 ve w_2 uzaklıkları özel ağırlık fonksiyonlarıdır. Huber'in önerisine göre;

$$w_1(y) = \frac{\psi_H(y,\sqrt{q_r})}{y} \text{ ve } w_2(y) = \frac{\psi_H(y,q_r)}{\beta y} \quad \textbf{(3.24)}$$

olup, buradaki $\psi_H(y,k) = \max\{-k, \min\{y,k\}\}$ 'dır ve bu fonksiyona Huber'in psi fonksiyonu denir. β, yayılım matrisini (tahminini) normal modellerde Fisher tutarlılığına sahip olacak bir tahmin edici yapan bir katsayıdır ve $q_r = \chi^2_{p;0.9}$ 'dur (Croux ve Haesbroeck, 2000).

Burada Fisher tutarlılığından bahsetmek faydalı olacaktır. $x_1, x_2, \dots, x_n$ hepsi bilinmeyen bir θ parametresine bağlı F_θ dağılım fonksiyonuna sahip bir yığından çekilmiş olsun. Bu örneklemin deneysel dağılım fonksiyonu F_n ile gösterilsin. θ parametresinin F_θ dağılım fonksiyonuna bağlı olarak ifade edildiği fonksiyon ile $\hat{\theta}$ tahmin edicisi F_n fonksiyonuna bağlı olarak ifade edilebiliyorsa; yani, $\theta = h(F_\theta)$ iken $\hat{\theta} = h(F_n)$ yazılabiliyorsa, bu tahmin edici Fisher tutarlılığına sahiptir denir (Lauritzen, 2004).

3.3 Stahel-Donoho Tahmin Edicileri (SDE)

Maronna ve Huber'in çalışmaları gösteriyor ki afin eş değişim ve yüksek kırılma ihtiyaçları M tahminleri kullanılarak elde edilemez (Donoho, 1982). Çok değişkenli konum ve yayılım parametreleri için yüksek kırılma noktalı ilk afin eş değişimli tahmin edici Stahel (1981) ve Donoho (1982) tarafından birbirlerinden bağımsız olarak elde edilmiştir. Bu tahmin edicisi Stahel-Donoho tahmin edicisi (SDE) ya da " aykırılık ağırlıklandırılmış medyan (outlyingness-weighted median)" olarak adlandırılır (Rousseeuw, 1985).

Burada iki adımlı bir yöntemle $\boldsymbol{x_i}$'lere karşılık gelen w_i ağırlıkları belirlenir;

Adım 1: $\boldsymbol{x_i}$' nin örneğin merkezine en uzak olduğu izdüşümü bulunur ve "aykırılık" değeri;

$$r_i = \sup_{\|\boldsymbol{v}\| = 1} \frac{|\boldsymbol{v}'\boldsymbol{x_i} - Med(\boldsymbol{v}'\boldsymbol{X})|}{MAD(\boldsymbol{v}'\boldsymbol{X})} \quad \textbf{(3.25)}$$

şeklinde ölçülür. Burada medyan ($Med(.)$) ve mutlak medyan sapma ($MAD(.)$) sırasıyla tek boyutlu (değişkenli) konum ve yayılım parametrelerinin robust tahminleri olarak kullanılır.

Adım 2: Aykırılığına (r_i) göre $\boldsymbol{x}_i$'ye ağırlıklar verilir;

$$w_i = w(r_i) \tag{3.26}$$

olmak üzere, burada $w(r)$, $r = 0$ olduğunda 1 değerini alan ve $r \to \infty$ iken düzgün olarak 0 değerini alacak şekilde hareket eden bir fonksiyondur (Donoho, 1982).

Bu ağırlıklar kullanılarak, Stahel-Donoho tahmin edicileri (SDE) konum ve yayılım parametreleri için sırasıyla ağırlıklı ortalama ve ağırlıklı kovaryans matrisi olarak;

$$\widehat{\boldsymbol{\mu}}_{SDE} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i} \tag{3.27}$$

$$\widehat{\boldsymbol{C}}_{SDE} = \frac{\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{SDE})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{SDE})'}{\sum_{i=1}^{n} w_i} \tag{3.28}$$

şeklinde tanımlanır. Burada ki $w_i = w(r_i)$ olup, r_i değeri afin eş değişimlidir. Yani; herhangi bir singüler olmayan **A** matrisi ve $\boldsymbol{b} \epsilon R^p$ için $r(\boldsymbol{x}_i, \boldsymbol{X}) = r(\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b})$. Bu durum $\left(\widehat{\boldsymbol{\mu}}_{SDE}, \widehat{\boldsymbol{C}}_{SDE}\right)$ tahminlerinin de afin eş değişimli olmasını ifade etmekte olup;

$$\widehat{\boldsymbol{\mu}}_{SDE}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\widehat{\boldsymbol{\mu}}_{SDE}(\boldsymbol{X}) + \boldsymbol{b} \; ve \; \widehat{\boldsymbol{C}}_{SDE}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\widehat{\boldsymbol{C}}_{SDE}\boldsymbol{A}' \tag{3.29}$$

şeklinde yazılabilir (Maronna ve Yohai, 2002).

Konum ve yayılım tahminlerinde hiçbir terimin etkin olmaması için ağırlık fonksiyonu şu şartları sağlamalıdır:

$r \geq 0$ için $w(r)$ ve $r^2 w(r)$ sınırlıdır ve bu şart altında Stahel-Donoho tahmin edicilerinin asimptotik olarak kırılma noktası 0,5'tir (Maronna ve diğ., 2006). SDE'nin kırılma özellikleri ile ilgili detaylı bilgiye (Maronna ve Yohai, 2002)'den ulaşılabilir.

3.4 Çok Değişkenli Budama (MVT)

Çok değişkenli konum ve yayılım parametreleri için robust bir alternatif çok değişkenli budamaya dayanmaktadır. MVT iteratif bir yöntemdir. Yönteme ait algoritma şu şekilde özetlenebilir:

Adım 1: **X** veri matrisinden elde edilen herhangi bir konum ya da yayılım tahmin edicisi sırasıyla $\widehat{\boldsymbol{\mu}}^*$ ve $\widehat{\boldsymbol{C}}^*$ olmak üzere tüm $\boldsymbol{x}_i$ gözlemleri için karesi alınmış Mahalanobis uzaklıkları hesaplanır.

$$d_i^2 = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}^*)' \, \hat{\boldsymbol{C}}^{*^{-1}} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}^*), \text{ i=1,2,...,n.} \quad \textbf{(3.30)}$$

Adım 2: Belli bir α budama oranında en yüksek d_i^2 değeri alan gözlemler göz ardı edilerek, kalan gözlemlerden tekrar ortalama vektörü ve kovaryans matrisi hesaplanarak $\hat{\boldsymbol{\mu}}^*$ ve $\hat{\boldsymbol{C}}^*$ tahminleri güncellenir ve Adım 1'e dönülür. Burada $\hat{\boldsymbol{\mu}}^*$ ve $\hat{\boldsymbol{C}}^*$ tahminleri durağanlaştığında işlemlere son verilir ve bulunan son $\hat{\boldsymbol{\mu}}^*$ ve $\hat{\boldsymbol{C}}^*$ çok değişkenli budama (MVT) tahmin edicileri olarak adlandırılır.

Budamanın etkisiyle büyük d_i^2'li gözlemler işlemlere katılmaz ve dolayısıyla tahmin ediciler üzerinde herhangi bir etkiye sahip değillerdir. Pek çok yaklaşımda ilk adımda $\hat{\boldsymbol{\mu}}^*$ ve $\hat{\boldsymbol{C}}^*$ için klasik ortalama vektörü ve kovaryans matrisi kullanılır. Deneysel olarak MVT, genellikle bir ya da iki adım gibi yüksek bir hızda durağanlaşır (Devlin ve diğ., 2014). Burada önemli hususlardan birisi, budama oranı (α) en fazla $\%50$ değerini alabilir.

Devlin ve arkadaşları (2014**),** MVT'nin kırılma noktasının budama oranına eşit olduğunu, dolayısıyla değişken sayısının artmasıyla azalmayacağını ileri sürmüşlerdir. Fakat Donoho (1982), MVT'nin kırılma noktasının en fazla $1/p$ olduğunu, bu yüzden bu yöntemin düşük kırılma noktasına sahip olmasından dolayı cazip olmadığını iddia etmiştir (Moller ve diğ., 2006).

MVT yöntemi sadece budamadan sonra **X** matrisinde ki gözlem sayısı, değişken sayısından fazlaysa uygulanabilir. Bu kısıt, **X** veri matrisinin temel bileşenler analizinden elde edilen **T** skor matrisine uygulanarak önlenebilir (Walczak ve Massart, 1995). Ancak bu durumda sonuç, TBA modelinden elde edildiği için, uygulanan modelin sağlamlık özelliklerine bağlıdır ve kovaryans matrisinin afin eş değişim özelliği kaybedilecektir (Moller ve diğ., 2006).

3.5 Minimum Hacim Elipsoidi Tahmin Edicisi (MVE)

Çok değişkenli konum ve yayılımın yüksek kırılma noktalı afin eş değişimli bir diğer tahmin edicisi Minimum Hacim Tahmin Edicisi (Minimum Volume Estimator)'dir.

$$\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}' in \; (en \; az) \; h \; gözlemlik \; bir \; kısmını \; kapsayan \quad \textbf{(3.31)}$$

$$minimum \; hacimli \; elipsoidin \; merkezi$$

olup, burada $[\![\ldots]\!]$ içindeki değerin tam kısmını alan bir fonksiyon olmak üzere $h = [\![n/2]\!] + 1$ olabilir (Rousseeuw, 1985).

MVE tahmin edicisi afin eş değişim özelliğine sahiptir. Şöyle ki; herhangi bir E elipsoidi için, singüler olmayan afin bir $f(x) = \boldsymbol{Ax} + \boldsymbol{b}$ dönüşümünden elde edilen $f(E)$ görüntüsü de bir elipsoidtir ve

$$Hacim\big(f(E)\big) = |\boldsymbol{A}|\ Hacim\ (E) \qquad \textbf{(3.32)}$$

'dir. Çünkü $|\boldsymbol{A}|$ bir sabittir ve afin dönüşümler altında elipsoidin nispi genişliği değişmez (Rousseeuw, 1985).

MVE tahmin edicisinin kırılma noktası $\left(\left\lfloor\frac{n}{2}\right\rfloor - p + 1\right)/n$ olup, $n \to \infty$ için kırılma noktasının alacağı değer $\%50$'dir (Rousseeuw, 1985).

MVE tahmin edicisi yavaş yakınsama oranlı $\left(n^{-1/3}\right)$ bir algoritmaya sahiptir. Bu yüzden hesaplama zorluğu ve yüksek maliyetli olmasından dolayı büyük verilerde kullanışlı değildir (Ammann, 1993).

3.6 Minimum Kovaryans Determinantı (MCD)

Minimum kovaryans determinant (Minimum Covariance Determinant) tahmin edicisi MVE tahmin edicisi ile benzerliklere sahiptir. Çok değişkenli bir veri setinin konum parametresi için MCD tahmini;

$$\boldsymbol{T}(\boldsymbol{X}) = Kovaryans\ matrisinin\ determinantı\ minimum\ olacak \qquad \textbf{(3.33)}$$

$$şekilde\ \boldsymbol{X}'in\ h\ gözleminin\ ortalama\ vektörü$$

şeklinde tanımlanır (Rousseeuw, 1985). Yani MCD, n gözlemli tüm veri içerisinde h gözlemli tüm mümkün alt setlerden kovaryans matrisini minimum yapacak alt seti tespit etmeye çalışır. Elde edilen bu optimal alt setin ortalaması konum parametresinin, kovaryans matrisi de yayılım parametresinin tahmini olarak kullanılır.

Bu yöntemde şöyle bir kısıtla karşı karşıya kalınabilir. Alt setin genişliği h olmak üzere, eğer bu değer değişken sayısından düşükse $(h < p)$ bu durumda alt setin kovaryans matrisinin determinantı sıfır olacaktır ve dolayısıyla minimize edilemeyecektir. Bu yüzden, yüksek boyutlu veri setleri öncelikle değişken seçme ya da temel bileşenler analizi kullanılarak indirgenmelidir (Moller ve diğ., 2006).

Bu yöntemde, α budama oranı olmak üzere, $h = (1 - \alpha) * n$ olarak tanımlanır. Bu durumda MCD tahmin edicisinin kırılma noktası da budama oranına (α) eşit

olacaktır. Dolayısıyla kırılma noktasının alabileceği en büyük değer $h = 0{,}5 * n$ olarak alındığında elde edilecek olup, %50'dir. Fakat robustlıkla istatistiksel etkinlik arasında denge sağlanması için genellikle $h = 0,75 * n$ olarak alınmaktadır ve bu durumda da kırılma noktası %25'tir (Rousseeuw ve Driessen, 1999).

Öte yandan aykırı değer kirlenmesi önceden tespit edilebilirse, α $(0 < \alpha \leq 0{,}5)$ budama oranı bakımından h, $h = \lfloor n(1-\alpha) \rfloor + 1$ şeklinde formüle edilebilir ve bu durumda da kırılma noktası yine α olacaktır. Buradan elde edilen h, Eşitlik (2.31) ve Eşitlik (2.33) de kullanılabilir. $\alpha \to 0$ için MVE tüm veriyi kapsayan en küçük hacimli elipsoidin merkezini verirken, MCD aritmetik ortalama vektörüne eşit olacaktır (Rousseeuw, 1985).

Butler ve arkadaşlarına göre (1993) MCD asimptotik olarak normal olduğundan ve MVE'den daha hızlı bir yakınsama oranına sahip olduğundan, MCD MVE'den daha iyi istatistiksel özelliklere sahiptir. MCD için geliştirilmiş hızlı bir algoritma mevcuttur ve literatürde FAST-MCD olarak adlandırılır. FAST-MCD, hesaplama karmaşası olan tüm MVE algoritmalarından daha hızlı yakınsadığından daha kullanışlıdır (Moller ve diğ., 2006).

MCD tahmin edicilerinin yeniden ağırlıklandırılmasıyla istatistiksel etkinlikleri arttırılabilir (Rousseeuw ve Driessen, 1999). Bu ağırlıklandırma işleminde öncelikle çok değişkenli veriye ait konum ve yayılımın MCD tahmin edicileri sırasıyla $\widehat{\boldsymbol{\mu}}_{MCD}$ ve $\widehat{\boldsymbol{C}}_{MCD}$ olmak üzere Mahalanobis uzaklıkları;

$$d\left(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_{MCD}, \widehat{\boldsymbol{C}}_{MCD}\right) = \sqrt{(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{MCD})' \widehat{\boldsymbol{C}}_{MCD}^{-1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{MCD})} \qquad \textbf{(3.34)}$$

şeklinde elde edilir. Burada tüm gözlemler için hesaplanan uzaklıklar kullanılarak ağırlıklar;

$$w_i = \begin{cases} 1, d^2\left(\boldsymbol{x}_i, \widehat{\boldsymbol{\mu}}_{MCD}, \widehat{\boldsymbol{C}}_{MCD}\right) \leq \chi^2_{p,1-\delta} \\ 0, \ diğer\ haller \end{cases} \qquad \textbf{(3.35)}$$

şeklinde hesaplanır ve daha sonra yeniden ağırlıklandırılmış MCD (RMCD) tahmin edicileri ağırlıklandırılmış ortalama vektörü ve ağırlıklandırılmış kovaryans matrisi olarak:

$$\widehat{\boldsymbol{\mu}}_{RMCD} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i} \qquad \textbf{(3.36)}$$

$$\widehat{\boldsymbol{C}}_{RMCD} = c_1 \frac{\sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{RMCD})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{RMCD})'}{\sum_{i=1}^{n} w_i} \quad \textbf{(3.37)}$$

şeklinde hesaplanmaktadır. Burada $c_1 = (1-\delta)/F_{\chi^2_{p+2}}(\chi^2_{p,1-\delta})$'dır. RMCD'nin kırılma noktası da tıpkı MCD'de olduğu gibi α'ya eşittir ve budama parametresi olan $\delta = 0{,}025$ olarak alınır (Croux ve Haesbroeck, 2000).

3.7 S Tahmin Edicileri

Çok değişkenli regresyon S tahmin edicilerini Davies (1987) ve Lopuhaä (1989) çok değişkenli konum ve yayılım parametrelerine genişletmiştir (Moller ve diğ., 2006). Çok değişkenli konum ve yayılım parametreleri için yüksek kırılma noktalı ve afin eş değişimli tahmin ediciler sınıfında bulunan S tahmin edicileri;

$$\frac{1}{n}\sum_{i=1}^{n} \rho(d_i) = b_0 \quad \textbf{(3.38)}$$

şartı altında $\det(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{S}})$'yi minimize eden $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{S}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{S}})$ çifti olup, burada $d_i = \sqrt{(\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{S}})' \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{S}}^{-1} (\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{S}})}$ şeklinde tanımlanan Mahalanobis uzaklığıdır. Pozitif bir kırılma noktası elde etmek için ρ şu şartları sağlamalıdır:

- ρ sıfır etrafında simetrik ve iki kez türevlenebilir bir fonksiyondur.
- ρ, $c > 0$ için $[0, c]$ aralığında artan, $[c, \infty]$ aralığında sabittir ve $\rho(0) = 0$'dır. (Hubert ve diğ., 2013).

Burada b sabiti çok değişkenli normal dağılım varsayımının elde edilmesi için çoğu zaman $\rho(d_i)$'nin beklenen değeri olarak alınır (Campell ve diğ., 1998).

ρ'nun seçimi için sık sık şu fonksiyon kullanılır:

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4}, & |x| \le c \\ \frac{c^2}{6}, & |x| > c \end{cases} \quad \textbf{(3.39)}$$

ve burada c uygun bir ayar sabitidir (Hubert ve diğ., 2013). Burada fonksiyonun yapısına dikkat edilirse, belli bir c değerinden daha büyük Mahalanobis uzaklığına sahip olan gözlemlerin ρ fonksiyon değerleri $c^2/6$ olarak sabitlenmektedir ve uzaklık ne kadar artarsa artsın, yani gözlem ne kadar aykırılaşırsa aykırılaşsın, Eşitlik (3.38)'de ifade edilen toplama sabit bir katkıda bulunacaktır.

Çok değişkenli konum ve yayılım parametrelerinin S tahmin edicilerinin kırılma noktası $^{b}/_{\rho(c)}$ 'dir ve normal model altında b;

$$b = \frac{p}{2}\chi^2_{p+2}(c^2) - \frac{p(p+2)}{2c^2}\chi^2_{p+4}(c^2) + \frac{p(p+2)(p+4)}{6c^4}\chi^2_{p+6}(c^2) + \frac{c^2}{6}(1 - \chi^2_p(c^2)) \quad \textbf{(3.40)}$$

şeklinde hesaplanabilir (Hubert ve diğ., 2013). Kırılma noktası tanımından hareketle %25 kırılma noktası elde etmek için c, $\rho(c) = {}^{b}/_{0,25}$ olacak şekilde seçilir (Croux ve Haesbroeck, 2000).

Çizelge 3.1. Çok değişkenli konum ve yayılım parametreleri için robust istatistikleri

Tahmin edici	Kırılma noktası (ε^*)	Yorumlar
M	$\leq 1/(p+1)$	Tüm M tahminleri afin eş değişimli değildir. (Örneğin L_1) Kısıt: $n > p$
Stahel-Donoho	$n \geq 2p+1$ için %50	Hesaplama maliyeti yüksektir.
MVT	%50	Hızlı durağanlaşır. Kısıt: Budamadan sonra $n > p$
MVE	%50	Yavaş durağanlaşır $(n^{-1/3})$. Kısıt: $n > p$
MCD	%50	Hızlı bir algoritmaya sahiptir. Kısıt: $n > p$
S	%50	Kısıt: $n > p$

Çok değişkenli konum ve yayılım parametrelerinin robust tahmin edicileri ile ilgili bahsedilen bilgiler Çizelge 3.1'de özetlenmiştir. Çizelge 3.1 incelendiğinde M tahmin edicilerinin tümünün afin eş değişimli olmadığı söylenebilir. Ayrıca tahmin edicinin sağlamlık ölçüsü olarak kullanılan kırılma noktası değerlerine bakıldığında da M ve Stahel-Donoho tahmin edicileri dışındaki tahmin edicilerin %50 kırılma noktasına sahip olabildiği görülmektedir. Stahel-Donoho tahmin edicisi ise belli bir kısıt altında %50 kırılma noktasına sahip olabilmektedir. Elbette %50 değeri asimptotik olarak tahmin edicinin kırılma noktasının alabileceği en büyük değerdir ve istatistiksel etkinlik ile sağlamlık arasında bir denge kurmak amacıyla zaman zaman isteğe bağlı olarak düşürülebilir. Bu azalma verideki kirlenme oranından fazla olmadığı takdirde sağlamlıkta herhangi bir azalma yaratmayacaktır. Örneğin MCD tahmin edicisinde $h = 0.75 * n$ alınırsa kırılma noktası %25 olacaktır ve verideki kirlenme %10 ise bu kirlenmenin etkisini ortadan kaldırmak için bu kırılma değeri yeterli olacaktır.

4. TEMEL BİLEŞENLER ANALİZİ

4.1 Klasik Temel Bileşenler Analizi

İstatistiksel analizde çok sayıda değişken olması işlemlerin yapılmasını güçleştirebileceği gibi değişkenler arasındaki bağımlılık veya ilişki olması bir takım sorunların ortaya çıkmasına neden olabilir. Örneğin regresyon analizinde değişkenler arasında ilişki olması durumunda çoklu bağlantı problemi ortaya çıkar. Bu sorunları gidermek amacıyla Temel Bileşenler Analizi uygulanabilmektedir.

Temel bileşenler analizi, başlangıç sisteminde yer alan ve birbiriyle ilişkili olan p tane değişkenden daha az sayıda ve birbirleriyle ilişkisiz, başlangıçtaki değişkenlerin doğrusal fonksiyonları olacak şekilde yeni değişkenler türetmeyi amaçlar. Bu yeni değişkenlere temel bileşenler adı verilir.

Genel olarak temel bileşenler analizinin amaçları;

- Boyut indirgemek (değişken sayısını azaltmak)
- Değişkenler arasındaki ilişki yapısını ortadan kaldırmak
- Başka istatistiksel analizler için veri hazırlamak

şeklinde ifade edilebilir.

4.1.1 Uygun verinin seçimi

Temel bileşenler analizinde orijinal değişkenlerle işlem yapılabildiği gibi standartlaştırılmış değişkenlerle de işlem yapılabilir. $\boldsymbol{X}: pxn$ ham veri matrisi ve $\boldsymbol{Z}: pxn$ standartlaştırılmış veri matrisi olarak tanımlansın. Analizde ham veri matrisi kullanılması durumunda varyans-kovaryans matrisinden, standartlaştırılmış veri matrisinin kullanılması durumunda ise korelasyon matrisinden yararlanılmaktadır. Oldukça farklı sonuçlar verebilen bu iki yoldan hangisinin seçileceği konusunda en önemli belirleyici değişkenlerin ölçü birimleridir. Eğer değişkenlerin ölçü birimleri ve varyansları birbirine yakınsa kovaryans matrisinden, aksi takdirde korelasyon matrisinden yararlanılması gerekir (Tatlıdil, 1996).

4.1.2 Temel bileşenlerin elde edilmesi

4.1.2.1 Temel bileşenlerin başlangıç verisinden elde edilmesi

$\boldsymbol{X}: px1$ rastgele değişken vektörü için kovaryans matrisi $cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$ olup, bu matrisin özdeğerleri;

$$\left|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}_p\right| = 0 \qquad \textbf{(4.1)}$$

denkleminin kökleri olan $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ şeklindeki λ_j'lerdir. Temel bileşenlerin orijinal ya da standart değişkenlerin doğrusal fonksiyonu şeklinde yazıldığı belirtilmişti. O halde temel bileşenler;

$$\left.\begin{aligned} Y_1 &= \boldsymbol{t}_1'\boldsymbol{X} = t_{11}X_1 + t_{21}X_2 + \cdots + t_{p1}X_p \\ Y_2 &= \boldsymbol{t}_2'\boldsymbol{X} = t_{12}X_1 + t_{22}X_2 + \cdots + t_{p2}X_p \\ &\vdots \\ Y_p &= \boldsymbol{t}_p'\boldsymbol{X} = t_{1p}X_1 + t_{2p}X_2 + \cdots + t_{pp}X_p \end{aligned}\right\} \qquad \textbf{(4.2)}$$

şeklinde tanımlanır. Her bir temel bileşen için varyanslar ve kovaryanslar hesaplanacak olursa Y_j temel bileşeni için ($j \neq k = 1,2,\ldots,p$);

$$Var(Y_j) = Var(\boldsymbol{t}_j'\boldsymbol{X}) = \boldsymbol{t}_j' Var(\boldsymbol{X})\boldsymbol{t}_j = \boldsymbol{t}_j'\boldsymbol{\Sigma}\boldsymbol{t}_j \qquad \textbf{(4.3)}$$

$$Cov(Y_j, Y_k) = Cov(\boldsymbol{t}_j'\boldsymbol{X}, \boldsymbol{t}_k'\boldsymbol{X}) = \boldsymbol{t}_j' Cov(\boldsymbol{X})\boldsymbol{t}_k = \boldsymbol{t}_j'\boldsymbol{\Sigma}\boldsymbol{t}_k \qquad \textbf{(4.4)}$$

olarak yazılır. Temel bileşenler analizinin amaçlarından birisi ilişkisiz değişkenler elde etmek olduğundan dolayı Eşitlik (4.4)'ün sonucunun sıfır olması istenir.

Temel bileşenleri oluştururken en büyük varyansa sahip olan temel bileşen 1.temel bileşen olarak adlandırılır ve varyansların büyüklük sırasına göre temel bileşenler oluşturulur. O halde 1. Temel bileşen $Var(Y_1) = \boldsymbol{t}_1'\boldsymbol{\Sigma}\boldsymbol{t}_1$ değerini $\boldsymbol{t}_1'\boldsymbol{t}_1 = 1$ yan şartı altında maksimize eden, $\boldsymbol{Y}_1 = \boldsymbol{t}_1'\boldsymbol{X}$ doğrusal bağıntısıdır. Aynı şekilde 2. Temel bileşen $Var(Y_2) = \boldsymbol{t}_2'\boldsymbol{\Sigma}\boldsymbol{t}_2$ değerini $\boldsymbol{t}_2'\boldsymbol{t}_2 = 1$ ve $\boldsymbol{t}_1'\boldsymbol{t}_2 = 0$ yan şartları altında maksimize eden $\boldsymbol{Y}_2 = \boldsymbol{t}_2'\boldsymbol{X}$ doğrusal bağıntısıdır. Bu şekilde devam edilirse j_nci temel bileşen $Var(Y_j) = \boldsymbol{t}_j'\boldsymbol{\Sigma}\boldsymbol{t}_j$ değerini $\boldsymbol{t}_j'\boldsymbol{t}_j = 1$ ve tüm $j \neq k$ için $\boldsymbol{t}_j'\boldsymbol{t}_k = 0$ yan şartları altında maksimize eden $\boldsymbol{Y}_j = \boldsymbol{t}_j'\boldsymbol{X}$ doğrusal bağıntısıdır.

Başlangıç sisteminin toplam varyansı;

$$\sigma_{top}^2 = iz(\Sigma) = \sigma_{11} + \cdots + \sigma_{pp} = \lambda_1 + \cdots + \lambda_p \qquad \textbf{(4.5)}$$

olup, $j = 1,2,\ldots,p$ için temel bileşenlerin varyansları toplamı da;

$$\sum_{j=1}^{p} Var(Y_j) = \lambda_1 + \cdots + \lambda_p \qquad \textbf{(4.6)}$$

olacaktır.

$\boldsymbol{\Sigma}$ matrisinin özdeğer-özvektör çiftleri $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \dots, (\lambda_p, \boldsymbol{e}_p)$ olsun. Bu özdeğer ve özvektörler için $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ ve $\boldsymbol{e}_j'\boldsymbol{e}_k = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$ özellikleri yazılabilir. Böylece Eşitlik (4.2)'de verilen denklemlere göre j_nci temel bileşen;

$$Y_j = \boldsymbol{e}_j'\boldsymbol{X} = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{pj}X_p \quad \textbf{(4.7)}$$

şeklinde tanımlanır. Bu durumda bu temel bileşenin varyansı

$$Var(Y_j) = Var(\boldsymbol{e}_j'\boldsymbol{X}) = \boldsymbol{e}_j' Var(\boldsymbol{X})\boldsymbol{e}_j \quad \textbf{(4.8)}$$

olarak yazılır.

$\boldsymbol{\Sigma}$ matrisinin özdeğer matrisi $\boldsymbol{\Lambda}$ ve özvektörler matrisi $\boldsymbol{P}$ olmak üzere, $\boldsymbol{\Sigma}$ matrisinin spektral ayrışımı $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ şeklinde yazılmakta olup, buradan $\boldsymbol{P}'\boldsymbol{\Sigma}\mathbf{P} = \boldsymbol{\Lambda}$ elde edilebilir. Böylece $\boldsymbol{e}_j'\boldsymbol{\Sigma}\boldsymbol{e}_j = \lambda_j$ olup, Eşitlik (4.8)'in çözümünü vermektedir. O halde j_nci temel bileşenin varyansı j_nci özdeğerdir.

Buna göre 1. Temel bileşen, varyansı en büyük olan temel bileşen olarak 1. Özdeğere karşılık gelen özvektörün kullanılmasıyla elde edilir. Temel bileşenler, ilgili özvektörün rastgele değişken vektörüyle çarpılması sonucunda doğrusal bir fonksiyon olarak ifade edilir. Kovaryans matrisinin özdeğer-özvektör çiftlerinin kullanılmasıyla yukarıdaki kısıtlar altında varyanslar maksimize edilir.

Temel bileşenler elde edildikten sonra bunların varyans açıklama oranları ve temel bileşen skorları elde edilmek istenir. j_nci temel bileşenin varyans açıklama oranı;

$$VAO = \frac{Var(Y_j)}{\sigma^2_{Top}} = \frac{\lambda_j}{\sigma^2_{Top}} \quad \textbf{(4.9)}$$

şeklinde ifade edilmektedir. i_nci birimin $(i = 1,2, \dots, n)$ temel bileşen skorları ise o birime ait değişken değerlerinin Eşitlik (4.7)'de yerine yazılmasıyla hesaplanır.

4.1.2.2 Temel bileşenlerin standartlaştırılmış veriden elde edilmesi

Standart verilerden temel bileşenlerin elde edilmesinde orijinal değişkenlerin korelasyon matrisinden yararlanılması gerektiğinden bahsedilmişti. Öncelikle veri;

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} = \left(\boldsymbol{\Sigma}^{1/2}\right)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad \textbf{(4.10)}$$

şeklinde standartlaştırılır. Daha sonra bu standart rastgele değişken vektörünün kovaryans matrisi;

$$Cov(\boldsymbol{Z}) = Cov\left[\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right] = \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1} Cov(\boldsymbol{x} - \boldsymbol{\mu}) \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}$$

$$= \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1} \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1} = Kor(\boldsymbol{X}) = \boldsymbol{R} \quad \textbf{(4.11)}$$

olarak bulunur. O halde temel bileşenler bu korelasyon matrisine göre oluşturulur. Eğer korelasyon matrisinin özdeğer-özvektör çiftleri $(\lambda_j, \boldsymbol{e}_j)$ ise j_nci temel bileşen;

$$Y_j = \boldsymbol{e}_j'\boldsymbol{Z} = e_{1j}Z_1 + e_{2j}Z_2 + \cdots + e_{pj}Z_p \, , j = 1,2, \ldots, p \quad \textbf{(4.12)}$$

şeklinde ifade edilir ve $Var(Y_j) = \lambda_j$, $Cov(Y_j, Y_k) = 0$'dır. Ayrıca sistemin toplam varyansı

$$\sigma^2_{Top} = iz(\boldsymbol{R}) = \lambda_1 + \lambda_2 + \cdots \lambda_p = p \quad \textbf{(4.13)}$$

'dir.

Orijinal değişkenlerde olduğu gibi temel bileşenlerin standart veriden elde edildiği durumlarda da varyans açıklama oranları ve temel bileşen skorları ile ilgilenilir. j_nci temel bileşenin varyans açıklama oranı;

$$VAO = \frac{Var(Y_j)}{\sigma^2_{Top}} = \frac{\lambda_j}{p} \quad \textbf{(4.14)}$$

olarak ifade edilmektedir. i_nci birimin $(i = 1,2, \ldots, n)$ temel bileşen skorları ise o birime ait standart değişken değerlerinin Eşitlik (4.12)'de yerine konulmasıyla hesaplanır.

4.1.3 Önemli temel bileşen sayısının belirlenmesi

4.1.3.1 Kaiser kriteri

Temel bileşenler standart veriler üzerinden elde edildiğinde bu kriter yaygın bir şekilde kullanılır. Buna göre **R** matrisinin birden büyük özdeğer $(\lambda \geq 1)$ sayısı önemli temel bileşen sayısı olarak alınır (Özdamar, 2013).

4.1.3.2 Scree plot grafiği (Catell scree test)

Bileşen sayısı 1,2,…,p biçiminde x- ekseninde ve özdeğerler y- ekseninde olmak üzere özdeğerlerin büyüklük sırasına göre bir xy koordinat sisteminde çizgi eğim grafiği çizilir. Bu grafiğe bileşen sayısı arttıkça özdeğerlerin azalışını gösteren yamaç eğim grafiği adı verilir. Çizgi grafiğinde eğimin kaybolmaya başladığı noktanın işaret ettiği bileşen sayısı, önemli temel bileşen sayısı olarak alınır (Özdamar, 2013).

4.1.3.3 Açıklanan varyans kriteri

Özdeğerlerin açıkladıkları toplam varyansın en az %80 olacak biçimde özdeğer sayısı kadar temel bileşenin önemli kabul edilmesi esasına dayanır. Bazı kaynaklar bu oranın en az 2/3 (%67) olması gerektiğini belirtmektedir (Özdamar, 2013).

4.1.3.4 Joliffe kriteri

Korelasyon matrisinin 0.7 ve daha büyük değerli özdeğer ($\lambda \geq 0.7$) sayısını önemli temel bileşen sayısı olarak kabul eden bir yöntemdir. Bu yöntemde tespit edilen önemli temel bileşen sayısı Kaiser kriterinde tespit edilenin yaklaşık iki katıdır (Özdamar, 2013).

4.2 Temel Bileşenler Analizine Robust Yaklaşımlar

4.2.1 Robust kovaryans matrislerine dayanarak temel bileşenler analizi

Robust temel bileşenler analizinin amacı aykırı değerlerden etkilenmeyen temel bileşenleri elde etmektir. Bu amaçla uygulanan en yaygın yöntem, klasik kovaryans matrisi yerine Bölüm 3'de söz edilen robust alternatiflerini kullanarak temel bileşenleri elde etmektir.

Bu amaçla $\widehat{\boldsymbol{\mu}}_r, \widehat{\boldsymbol{\Sigma}}_r$ herhangi bir robust konum ve yayılım tahmini olsun. Eğer verideki değişkenler ölçü birimleri açısından farklılık göstermiyorsa bu $\widehat{\boldsymbol{\Sigma}}_r$ matrisinin özvektörleri, farklılık gösteriyorsa

$$\boldsymbol{R}_r = \left(\boldsymbol{D}_r^{\frac{1}{2}}\right)^{-1} \widehat{\boldsymbol{\Sigma}}_r \left(\boldsymbol{D}_r^{\frac{1}{2}}\right)^{-1} \qquad \textbf{(4.15)}$$

ile elde edilen robust korelasyon matrisinin özvektörleri kullanılarak temel bileşenler elde edilir. Burada $\boldsymbol{D}_r^{\frac{1}{2}}$ matrisi köşegen elemanları $\widehat{\boldsymbol{\Sigma}}_r$ matrisinin köşegen elemanlarının kareköklerinden (standart sapmalardan) oluşan köşegen bir matristir. Bu yöntemle bulunan temel bileşenler aykırı değerlerden etkilenmezler. Ancak klasik durumda olduğu gibi bu yöntemde de gözlem sayısı değişken sayısından büyük olmalıdır $(n > p)$. Aksi takdirde kovaryans matrisinin determinantı sıfır olacaktır. Ayrıca burada başlangıç örnek çapı n, değişken sayısından fazla olsa bile yukarıdaki kısıt sağlanmayabilir. Örneğin MCD tahmin edicisinde n tane gözlem kovaryans matrisinin determinantını minimum yapacak $h = 0{,}5 * n$ ya da $h = 0{,}75 * n$ tane gözleme indirgenmektedir. Bu durumda kovaryans matrisinin hesaplanacağı örnek çapı azalır ve $h < p$ olabilir. Eğer değişken sayısı gözlem sayısından büyük olursa

kovaryans matrisinin determinantı sıfır olacaktır ve minimize edilemeyecektir. Dolayısıyla bu yaklaşım sadece düşük boyutsal $(n > p)$ verilerle sınırlıdır.

4.2.2 Projeksiyon izleme (Projection pursuit-PP) yöntemi

Projeksiyon izleme (PP) yöntemi yüksek boyutsal $(p > n)$ bir veride, bu veri noktalarını projeksiyon indeksi olarak kullanılan bir yayılım ölçüsünü maksimize eden daha düşük boyutsal bir uzaya izdüşümleyerek verinin yapısını inceler. Burada projeksiyon indeksi olarak kullanılan yayılım ölçüsü MAD gibi bir robust yayılım ölçüsü olabilir. PP yöntemi maksimizasyon problemini çözerek robust özvektör tahminlerini elde eder. Öncelikle ilk $\boldsymbol{v}$ yön (özvektör) bulunur ve daha sonraki adımda her yeni yön önceki tüm yönlere ortogonal olacak şekilde oluşturulur. Yöntem robust temel bileşenler ve robust kovaryans matrisi elde ederek sonuçlanır. Temel bileşenler ikinci aşamada hesaplandığından dolayı PP yaklaşımı yüksek boyutsal $(p > n)$ verilerde de kullanılabilir. Klasik temel bileşenler analizi de PP algoritmasının özel bir hali olup projeksiyon indeksi olarak varyansı kullanır (Moller ve diğ., 2006). PP yöntemiyle ilgili kapsamlı bilgiye (Huber, 1985)'den ulaşılabilir.

4.2.3 ROBPCA yöntemi

ROBPCA yöntemi hem robust kovaryans tahmini hem de projeksiyon izleme (PP) düşüncelerini birlikte kullanır. PP yöntemi ilk boyut indirgeme aşamasında kullanılır. Daha sonra elde edilen düşük boyutsal veri uzayından, MCD tahmin edicisine dayanarak temel bileşenler elde edilir. Birleştirilmiş yaklaşım ham PP algoritmalarından daha hızlı sonuçlar verir (Hubert ve diğ., 2005).

Gözlem sayısı n ve orijinal değişken sayısı p olmak üzere, $\boldsymbol{X}: nxp$ orijinal veri matrisi olsun. ROBPCA üç aşamada uygulanır. Öncelikle; veri en fazla $(n-1)$ boyutlu bir alt uzaya dönüştürülür. Sonra; başlangıç kovaryans matrisi $\boldsymbol{\Sigma}_0$ elde edilir ve k boyutlu alt uzay veriye iyi uyan bir alt uzay olacak şekilde bileşen sayısı k belirlenir. Son aşamada; konum vektörünün ve k tane sıfır olmayan $\lambda_1, \lambda_2, \ldots, \lambda_k$ özdeğerin hesaplandığı yayılım matrisinin robust bir şekilde tahmin edildiği bu alt uzay üzerine veri noktalarının iz düşümleri alınır. Bu özdeğerlere karşı gelen özvektörler ise k tane robust temel bileşeni verecektir (Hubert ve diğ., 2005).

Orijinal p boyutlu uzayda, bu k bileşen k boyutlu bir uzayı kapsar. Özvektörler sütun şeklinde yan yana getirilerek $\boldsymbol{P}: pxk$ matrisi elde edilir. Konum tahmini, p

değişkenli bir sütun vektörü olmak üzere $\widehat{\boldsymbol{\mu}}$ ile ifade edilir ve robust konumu olarak adlandırılır. Skor matrisi ise;

$$\boldsymbol{T}_{n,k} = \left(\boldsymbol{X}_{n,p} - \mathbf{1}_n \widehat{\boldsymbol{\mu}}'\right)\boldsymbol{P}_{p,k} \tag{4.16}$$

şeklinde ifade edilmekte olup, burada $\mathbf{1}_n$ 1'lerden oluşan $nx1$ boyutlu bir sütun vektördür. Robust yayılım matrisi;

$$\boldsymbol{\Sigma}_{p,p} = \boldsymbol{P}_{p,k}\boldsymbol{L}_{k,k}\boldsymbol{P}'_{k,p} \tag{4.17}$$

olup, burada $\boldsymbol{L}_{k,k}$ köşegen elemanları $\lambda_1, \lambda_2, \ldots, \lambda_k$ olan köşegen matristir (Hubert ve diğ., 2005). Klasik temel bileşenler analizinde olduğu gibi ROBPCA ortogonal eş değişim özelliğine sahiptir. Yani veriye ortogonal bir dönüşüm (döndürme) uygulandığında yükler ve böylece temel bileşen skorları değişmez (Hubert ve diğ., 2005).

5. BULGULAR VE TARTIŞMA

Bölgeler arası gelişmişlik farklarının azaltılmasına yönelik olarak bölgelerin sosyoekonomik analizlerinin yapılması ve Avrupa Birliği (AB) ile karşılaştırılabilir veriler üretilmesi amacıyla, AB bölgesel sınıflandırması olan NUTS kriterlerine göre Türkiye'de İstatistiki Bölge Birimleri Sınıflaması (İBBS) tanımlanmış ve 2002 yılında uygulamaya konulmuştur. İBBS üç düzeyden oluşmaktadır. İlk aşamada idari yapıya uygun olarak 81 il, 3. düzeyde bölge birimleri olarak tanımlanmıştır. Ekonomik, sosyal, kültürel ve coğrafi yönlerden benzer illerin belirli bir nüfus büyüklüğü de dikkate alınarak gruplanması ile 26 bölge, 2. düzeyde bölge birimleri olarak tanımlanmıştır. Yine aynı kritere göre 2. düzey bölge birimlerinin gruplanması sonucu 12 birim 1. düzeyde bölge birimleri tanımlanmıştır (TÜİK, 2005).

2006 yılında AB'ye uyum yasaları çerçevesinde Devlet Planlama Teşkilatı (DPT) bünyesinde kurulmaya başlanan kalkınma ajansları, günümüzde bölgesel kalkınmayı hızlandırmak amacıyla 26 istatistiki bölgenin her birinde bir tane olmak üzere yaygınlaştırılmıştır.

Uygulamanın asıl amacı, Türkiye'de illerin sosyoekonomik gelişmişlik düzeyini belirlemek için dönem dönem yapılan çalışmaları bu kalkınma ajanslarına ya da istatistiki bölgelere göre güncellemektir. İllerin bireysel olarak incelenmesi yerine bu bölgelerin tek bir idari bölge olarak düşünülüp ele alınmasının bölgeler arası gelişmişlik seviyelerini daha iyi açıklayacağı düşünülmektedir. Böylece aynı bölgedeki bir ilin gelişmişlik seviyesi artsa bile diğer illerde böyle bir gelişme söz konusu değilse, bölgenin gelişmekte olduğu ve kalkınma ajansının doğru politikalar izlediği yönündeki iddiaların doğruluğu tartışılabilir olacaktır.

Ülkemizde bölgeler arasında gelişmişlik farklarının olduğu aşikârdır. Ancak bazı bölgeler diğerlerinden olumlu ya da olumsuz anlamda çok farklıdır. Bu durum sosyoekonomik gelişmişlik endeksi (SEGE) çalışmalarında aykırı değerlerin söz konusu olmasına neden olur. Bu nedenle bu çalışmada öncelikle veri setinde aykırı değer olup olmadığı klasik ve MCD tahminlerine dayanan Mahalanobis uzaklıklarına

göre incelenmiş, daha sonra klasik ve robust yöntemlere dayanan temel bileşenler analizi kullanılarak sosyoekonomik gelişmişlik değerlendirmesi yapılmıştır.

Bölgeler arası gelişmişlik farklarını değerlendirmek amacıyla kalkınma ajansları kapsamında yer alan illere ait bazı sosyoekonomik göstergelerden yararlanarak her bir kalkınma ajansı bölgesi için söz konusu göstergelere ait değerler bulunmuştur. Daha sonra, elde edilen çok değişkenli veri yapısına klasik temel bileşenler, robust (MCD ve S) yayılım matrislerine dayanan temel bileşenler ve robust temel bileşenler analizi (ROBPCA) uygulanarak kalkınma ajansı bölgeleri değerlendirilmiştir. Çalışmada ilk aşamada 19 değişken kullanılmış olup söz konusu değişkenler Çizelge 5.1'de verilmiştir. 26 kalkınma ajansı için 19 değişken kullanıldığında elde edilen veri düşük boyutsal $(n > p)$ yapıya sahip olduğundan klasik ve robust yayılım matrislerine dayanan temel bileşenler analizi için uygundur.

Çizelge 5.1. İlk aşamada kullanılan 19 sosyoekonomik gösterge

Değişkenler	Açıklama
X_1	Bölgedeki banka kredilerinin Türkiye içindeki payı
X_2	Yeni kurulan şirketlerin toplam sermayesinin Türkiye içindeki payı
X_3	İmalat sanayi işyerlerinin Türkiye içindeki payı
X_4	Bölge ihracatının Türkiye içindeki payı
X_5	Nüfus yoğunluğu
X_6	On bin kişiye düşen yabancı sermayeli şirket sayısı
X_7	Bölge vergi gelirinin Türkiye içindeki payı
X_8	Yüz bin kişiye düşen marka başvuru sayısı
X_9	Kişi başına düşen ihracat tutarları
X_{10}	Genç bağımlı nüfus oranı (0-14 yaş arası nüfus)
X_{11}	15-49 yaş arası 1000 kadın başına düşen canlı doğum sayısı
X_{12}	Genel ortaöğretim net okullaşma oranı
X_{13}	SGK'da aktif çalışanların toplam nüfusa oranı
X_{14}	Çalışma çağındaki (15-64) nüfusun toplam nüfusa oranı
X_{15}	Okuryazar kadın nüfusun toplam kadın nüfusa oranı
X_{16}	İstihdam oranı
X_{17}	Okuryazar nüfur oranı
X_{18}	İmalat sanayi istihdamının sigortalı istihdam içindeki oranı
X_{19}	Kişi başı imalat sanayi elektrik tüketimi

Daha sonraki aşamada 26 kalkınma ajansı bölgesi 46 değişken bakımından değerlendirmeye alınmıştır. Bu durumda elde edilen çok değişkenli veri seti yüksek boyutsal $(p > n)$ olup klasik ya da robust yayılım matrislerine dayanan yöntemler uygulanamaz. Bu nedenle elde edilen bu yüksek boyutsal veriye sadece ROBPCA uygulanmıştır. Bu aşamada kullanılan 46 değişken Çizelge 5.2'de verilmiştir. İlk aşamada kullanılan 19 değişken de bu göstergeler arasındadır.

Çizelge 5.2. İkinci aşamada kullanılan 46 sosyoekonomik gösterge

Değişkenler	Açıklama
X_1	Nüfus yoğunluğu
X_2	15-49 yaş arası 1000 kadın başına düşen canlı doğum sayısı
X_3	Genç bağımlı nüfus oranı (0-14 yaş nüfus)
X_4	Net göç hızı
X_5	Şehirleşme oranı
X_6	Okuryazar nüfus oranı
X_7	Okuryazar kadın nüfus oranının toplam kadın nüfusuna oranı
X_8	Genel ortaöğretim net okullaşma oranı
X_9	Yüksekokul veya fakülte mezunu nüfusun 22+ yaş nüfusa oranı
X_{10}	İşsizlik oranı
X_{11}	İşgücüne katılma oranı
X_{12}	Çalışma çağındaki nüfusun (15-64) toplam nüfus içindeki oranı
X_{13}	İmalat sanayi istihdamının sigortalı istihdam içindeki oranı
X_{14}	SGK'da aktif çalışanların toplam nüfusa oranı
X_{15}	İstihdam oranı
X_{16}	Bölge ihracatının Türkiye içindeki payı
X_{17}	Kişi başına düşen ihracat tutarları
X_{18}	İmalat sanayi işyerlerinin Türkiye içindeki payı
X_{19}	Yeni kurulan şirketlerin toplam sermayesinin Türkiye içindeki payı
X_{20}	Turizm yatırım-işletme ve belediye belgeli yatak sayısının Türkiye içindeki payı
X_{21}	Teşvik belgeli yatırım tutarının Türkiye içerisindeki payı
X_{22}	Kişi başı imalat sanayi elektrik tüketimi
X_{23}	On bin kişiye düşen yabancı sermayeli şirket sayısı
X_{24}	Yüz bin kişiye düşen marka başvuru sayısı
X_{25}	Yüz bin kişiye düşen patent başvuru sayısı
X_{26}	Yüksek lisans ve doktora sahibi nüfusun 30 yaş üstü nüfusa oranı
X_{27}	Yüz bin kişiye düşen hastane yatak sayısı
X_{28}	On bin kişiye düşen hekim sayısı
X_{29}	On bin kişiye düşen diş hekimi sayısı
X_{30}	On bin kişiye düşen eczane sayısı
X_{31}	Yeşil kart sahibi nüfusun il nüfusuna oranı
X_{32}	Bölgedeki banka kredilerinin Türkiye içindeki payı
X_{33}	Bölgedeki tasarruf mevduatının Türkiye içindeki payı
X_{34}	Kişi başına düşen banka mevduatı tutarı
X_{35}	Bin kişiye düşen internet bankacılığında aktif bireysel müşteri sayısı
X_{36}	Bin kişiye düşen internet bankacılığında aktif kurumsal müşteri sayısı
X_{37}	Kişi başına düşen bütçe gelirleri
X_{38}	Bölge vergi gelirinin Türkiye içindeki payı
X_{39}	Kişi başına düşen GSM abone sayısı
X_{40}	Toplam demiryolunun yüzölçümüne oranı
X_{41}	Bin kişiye düşen AVM brüt kiralanabilir alan oranı
X_{42}	Kişi başına mesken elektrik tüketimi
X_{43}	Bin kişiye düşen özel otomobil sayısı
X_{44}	Sosyal güvenlik kapsamı dışında kalan nüfusun toplam nüfusa oranı

Çizelge 5.2.'nin devamı

X_{45}	Yüz bin kişiye düşen ceza infaz kurumuna giren hükümlü sayısı
X_{46}	Yüz bin kişiye düşen intihar vakası sayısı

5.1 Aykırı Değerlerin İncelenmesi

5.1.1 Klasik tahminlere dayanan Mahalanobis uzaklıklarının kullanımı

Çok değişkenli bir veri setinde aykırı değer incelemesi için Mahalanobis uzaklıklarının nasıl kullanılabileceğine Bölüm 2.5.2'de yer verilmişti. Öncelikle Eşitlik (2.14)'ten yararlanarak 26 kalkınma ajansı bölgesi için Mahalanobis uzaklıkları hesaplanmış ve Çizelge 5.3'te verilmiştir. Çalışmada 19 değişken olduğu için bu uzaklıklar $\chi^2_{19,0.975} = 32{,}8523$ kritik değeri ile karşılaştırılmıştır. Tüm bölgeler için ilgili Mahalanobis uzaklıkları bu değerden küçük olduğu için klasik tahminlere dayanan Mahalanobis uzaklıklarına göre veri setinde aykırı değer yoktur.

5.1.2 MCD tahminlerine dayanan Mahalanobis uzaklıklarının kullanımı

Burada klasik yaklaşımdan farklı olarak Eşitlik (3.14)'te konum ve yayılım parametresinin robust tahminlerinin kullanılmasıyla, robust tahminlerine dayanan Mahalanobis uzaklıkları elde edilir. Bu uzaklıklar $\chi^2_{19,0.975} = 32{,}8523$ kritik değeri ile karşılaştırılarak, Mahalanobis uzaklığı kritik değerden büyük olan gözlemler aykırı değer olarak kabul edilir. Bu düşünceden hareketle 26 kalkınma ajansı bölgesi için söz konusu uzaklıklar MCD tahminleri kullanılarak bulunmuş ve Çizelge 5.3'te yer verilmiştir. Klasik tahminlere dayanan uzaklıkların aksine, MCD tahminlerine dayanan uzaklıklardan TR10 (İstanbul) ve TR51 (Ankara) bölgelerine ait olanlar kritik değeri aştığından dolayı aykırı değer olarak tespit edilmişlerdir. Veri setinde aykırı değer olduğundan robust yaklaşımların kullanılması uygun olacaktır.

Burada klasik tahminlere dayanan Mahalanobis uzaklıklarının bu aykırı değerleri tespit edememe nedeni, klasik tahminlerin verideki bozulmaya/kirlenmeye uyma eğiliminde olmasıdır. Verideki aykırılık ortalama değerlerini yükselttiğinden Mahalanobis uzaklığı bağıntısında ifade edilen farklar ve dolayısıyla uzaklıklar da küçülecektir. Ancak MCD tahminlerinde aykırı değerlerin etkisi arındırıldığı için ortalamalar durağandır ve ilgili gözlem değerleri ile aralarındaki farklar büyük olacaktır. Gözlem aykırılık eğilimindeyse Mahalanobis uzaklıkları büyüyecek ve aykırı değerler belirlenebilecektir.

Çizelge 5.3. Klasik ve robust tahminlerine dayanan Mahalanobis uzaklıkları

BÖLGEDEKİ İLLER	MAH_{klasik}	MAH_{robust}
İstanbul	24,0245	**44239,60**
Tekirdağ, Edirne, Kırklareli	20,3834	18,8606
Balıkesir, Çanakkale	20,5394	19,5989
İzmir	21,4768	20,8742
Aydın, Denizli, Muğla	15,9978	16,2937
Manisa, Afyonkarahisar, Kütahya, Uşak	17,5093	20,9132
Bursa, Eskişehir, Bilecik	20,2121	21,6676
Kocaeli, Sakarya, Düzce, Bolu, Yalova	22,5142	21,1610
Ankara	23,5942	**1368,7959**
Konya, Karaman	16,7065	19,1341
Antalya, Isparta, Burdur	23,1457	21,2671
Adana, Mersin	21,6639	21,7789
Hatay, Kahramanmaraş, Osmaniye	21,6593	20,9132
Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	18,5234	19,8158
Kayseri, Sivas, Yozgat	8,1037	21,0783
Zonguldak, Karabük, Bartın	19,7686	18,4648
Kastamonu, Çankırı, Sinop	20,3300	20,8796
Samsun, Tokat, Çorum, Amasya	13,9103	12,7590
Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	11,6429	11,1892
Erzurum, Erzincan, Bayburt	13,6748	12,5990
Ağrı, Kars, Iğdır, Ardahan	19,1553	17,6662
Malatya, Elazığ, Bingöl, Tunceli	9,5493	10,7736
Van, Muş, Bitlis, Hakkari	18,9512	19,5681
Gaziantep, Adıyaman, Kilis	21,2518	19,5752
Şanlıurfa, Diyarbakır	17,3308	17,5569
Mardin, Batman, Şırnak, Siirt	13,3808	12,6118

5.2 Klasik Temel Bileşenler Analizi ile SEGE Değerlendirmesi

Daha önce de belirtildiği gibi Çizelge 5.1'de verilen 19 değişken kullanılarak 26 kalkınma ajansı bölgesi sosyoekonomik gelişmişlik bakımından değerlendirilmiştir. Değişkenler ölçü birimleri bakımından farklı olduklarından dolayı korelasyon matrisi üzerinden temel bileşenler elde edilmelidir. Bu nedenle ilk olarak korelasyon matrisinin özdeğerleri hesaplanmış ve varyans açıklama oranlarına bakılarak önemli temel bileşen sayısına karar verilmiştir. Sonuçlar Çizelge 5.4'te gösterilmiştir.

Çizelge 5.4. Klasik Korelasyon matrisinin özdeğerleri ve varyans açıklama oranları

Özdeğerler	Varyans Açıklama Oranı	Toplam Açıklanan Varyans Oranı
9,5703	0,5037	0,5037
6,2283	0,3278	0,8315
1,4828	0,078	0,9095
0,8058	0,0424	0,952

Çizelge 5.4. ‘ün devamı

Özdeğerler	Varyans Açıklama Oranı	Toplam Açıklanan Varyans Oranı
0,3206	0,0169	0,9688
0,2260	0,0119	0,9807
0,1594	0,0084	0,9891
0,0772	0,0041	0,9932
0,0403	0,0021	0,9953
0,0227	0,0012	0,9965
0,0218	0,0011	0,9976
0,0146	0,0008	0,9984
0,0095	0,0005	0,9989
0,0071	0,0004	0,9993
0,0052	0,0003	0,9996
0,0040	0,0002	0,9998
0,0028	0,0001	0,9999
0,0011	0,0001	1
0,0006	0	1

Çizelge 4.4 incelendiğinde ilk üç temel bileşenin açıkladığı toplam varyansın %90,95 olduğu görülmektedir. Dolayısıyla önemli temel bileşen sayısı 3 olarak alınabilir. Bu üç özdeğere karşılık gelen özvektörler Çizelge 5.5’te verilmiştir.

Çizelge 5.5. Önemli temel bileşen katsayıları (özvektörler)

1.TB	2.TB	3.TB
0,2565	0,2318	-0,0580
-0,2076	0,2828	0,1520
-0,1981	0,2918	0,1565
0,1949	-0,2650	0,0510
0,2146	-0,2502	0,0272
0,2247	-0,2761	-0,0293
0,2461	-0,2412	-0,0640
0,1665	-0,1218	0,5805
0,1517	-0,3024	-0,2120
0,1210	-0,2589	-0,3148
0,2689	0,2159	0,0180
0,2506	0,1111	0,3372
0,2805	0,1936	-0,0431
0,2594	0,2253	-0,1056
0,1063	-0,2125	0,5509
0,2584	0,2227	-0,1399
0,2885	0,1284	0,0316
0,2715	0,2089	-0,0925
0,2745	0,1965	0,0048

Temel bileşenler analizi korelasyon matrisi üzerinden yapıldığından temel bileşen skorlarını elde etmek için öncelikle standart veri matrisi hesaplanmalıdır. Standart veri matrisi;

$$\boldsymbol{Z}_c = \left(\boldsymbol{\Sigma}_c{}^{1/2}\right)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) \quad \textbf{(5.1)}$$

şeklinde hesaplanmıştır.

Temel bileşenler analizi değişkenler arasındaki korelasyonların önemli temel bileşenlerden kaynaklandığını varsayarken, değişkenler arasındaki ilişkilerin büyük bir bölümü ise tek bir temel faktörün etkisi sonucu ortaya çıkmaktadır. Söz konusu temel bileşene genel nedensel bileşen denilmektedir (Albayrak, 2005). Araştırmada bölgelerin sosyoekonomik gelişmişlik düzeylerinin kullanılan tüm göstergelere etki eden ve birlikte değişimlerini sağlayan genel nedensel bir bileşen olduğu kabul edilmiştir. Özetlemek gerekirse, genel nedensel bileşen bölgelerin sosyoekonomik gelişmişlik düzeyleridir.

Bu düşünceden hareketle en büyük özdeğere ve dolayısıyla varyans açıklama oranına sahip olan 1. Temel bileşen genel nedensel bileşen olarak kabul edilmiş ve bu temel bileşene göre hesaplanan skorlar, bölgelerin sosyoekonomik gelişmişlik endeksi olarak alınarak bölgeler arası sıralama yapılmış ve sonuçlar Çizelge 5.6'da gösterilmiştir. 1. Temel bileşen toplam varyansın % 50,37'sini tek başına açıklamaktadır.

Çizelge 5.6. Klasik TBA'ne göre bölgelerin SEGE sıralaması

SIRA	BÖLGEDEKİ İLLER	SEGE DEĞERİ
1	İstanbul	11,4638
2	İzmir	2,6597
3	Kocaeli, Sakarya, Düzce, Bolu, Yalova	2,3142
4	Bursa, Eskişehir, Bilecik	2,0643
5	Ankara	1,8664
6	Tekirdağ, Edirne, Kırklareli	1,3715
7	Aydın, Denizli, Muğla	1,1337
8	Manisa, Afyonkarahisar, Kütahya, Uşak	0,7947
9	Antalya, Isparta, Burdur	0,7557
10	Balıkesir, Çanakkale	0,5915
11	Zonguldak, Karabük, Bartın	0,2734
12	Konya, Karaman	0,0196
13	Adana, Mersin	-0,0136
14	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	-0,3322
15	Kayseri, Sivas, Yozgat	-0,4482
16	Samsun, Tokat, Çorum, Amasya	-0,4525
17	Kastamonu, Çankırı, Sinop	-0,7396
18	Hatay, Kahramanmaraş, Osmaniye	-0,8671
19	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	-0,8969
20	Gaziantep, Adıyaman, Kilis	-1,0341
21	Malatya, Elazığ, Bingöl, Tunceli	-1,3492
22	Erzurum, Erzincan, Bayburt	-2,2181
23	Ağrı, Kars, Iğdır, Ardahan	-3,7326

Çizelge 5.6. 'in devamı

SIRA	BÖLGEDEKİ İLLER	SEGE DEĞERİ
24	Van, Muş, Bitlis, Hakkari	-4,0643
25	Mardin, Batman, Şırnak, Siirt	-4,4507
26	Şanlıurfa, Diyarbakır	-4,7096

Analiz sonucunda en gelişmiş üç bölge TR10 (İstanbul), TR31 (İzmir), TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) bölgeleridir. En gelişmemiş bölgeler ise TRB2 (Van, Muş, Bitlis, Hakkari), TRC3 (Mardin, Batman, Şırnak, Siirt) ve TRC2 (Şanlıurfa, Diyarbakır) bölgeleridir.

Bölgelerin sıralanmasında kullanılan 19 değişken ve metodoloji 81 il için uygulandığında bazı çarpıcı sonuçlara ulaşılmaktadır. 81 il için 19 sosyoekonomik gösterge bakımından SEGE sıralaması EK-A'da verilmiştir.

İl bazında değerlendirme yapıldığında örneğin Kayseri ili 18. sırada yer alarak ilk %25'lik dilime girmiştir. Ancak Kayseri, Sivas ve Yozgat'ın oluşturduğu TR72 bölgesi tek bir idari yapı olarak ele alındığında 26 bölge içinde 15. olarak %60'lık dilime girebilmiştir. Dolayısıyla il bazında gelişmişlik incelendiğinde bölgesel kalkınmayı hedefleyen kurumlar için yanıltıcı sonuçlar elde edilebilir.

5.3 MCD Yayılım Matrisine Dayanan TBA ile SEGE Değerlendirimesi

MCD yayılım matrisine dayanan temel bileşenleri elde etmek için öncelikle MCD kovaryans matrisi ve MCD korelasyon matrisi elde edilmelidir. MCD kovaryans matrisinin elde edilmesinde MATLAB R2009b programında LIBRA toolboxı kullanılmıştır (Verboven, 2005). Daha sonra elde edilen kovaryans matrisi kullanılarak MCD korelasyon matrisi Eşitlik (3.15) ile elde edilmiştir. Elde edilen korelasyon matrisinin özdeğer ve özvektörleri hesaplanmıştır. İlgili özdeğerler ve varyans açıklama oranları Çizelge 5.7'de verilmiştir.

Çizelge 5.7. MCD korelasyon matrisinin özdeğerleri ve varyans açıklama oranları

Özdeğerler	Varyans Açıklama Oranı	Açıklanan Toplam Varyans
10,1927	0,5365	0,5365
3,9191	0,2063	0,7427
1,7486	0,0920	0,8348
0,8359	0,0440	0,8788
0,6377	0,0336	0,9123
0,5380	0,0283	0,9406
0,3827	0,0201	0,9608
0,2509	0,0132	0,9740

Çizelge 5.7'nin devamı

Özdeğerler	Varyans Açıklama Oranı	Açıklanan Toplam Varyans
0,1846	0,0097	0,9837
0,1198	0,0063	0,9900
0,0645	0,0034	0,9934
0,0605	0,0032	0,9966
0,0283	0,0015	0,9981
0,0152	0,0008	0,9989
0,0008	0,0000	0,9989
0,0022	0,0001	0,9990
0,0041	0,0002	0,9993
0,0082	0,0004	0,9997
0,0060	0,0003	1,0000

MCD korelasyon matrisinin 1'den büyük özdeğer sayısı üç olduğundan Kaiser kriterine göre önemli temel bileşenlerin sayısı üçtür. Bu üç temel bileşen toplam varyansın %83,47'sini açıklamaktadır.

Temel bileşenler analizi korelasyon matrisi kullanılarak yapıldığından standartlaştırılmış veri matrisine ihtiyaç vardır. Burada önemli olan husus standartlaştırma işleminde MCD konum ve yayılım tahminlerinin kullanılmasıdır. Buna göre MCD konum ve yayılım matrislerini kullanarak elde edilen standartlaştırılmış veri matrisi;

$$\boldsymbol{Z}_{MCD} = \left(\boldsymbol{\Sigma}_{MCD}{}^{1/2}\right)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{MCD}) \qquad \textbf{(5.2)}$$

şeklinde hesaplanmıştır.

Klasik temel bileşenler analizinde olduğu gibi genel nedensel faktör olarak en büyük özdeğere sahip 1. temel bileşen skorları, bölgelerin sosyoekonomik gelişmişlik göstergesi olarak alınmış ve bölgelerin bu değerlere göre sıralaması Çizelge 5.8'de verilmiştir. 1. Temel bileşen toplam varyansın % 53,65'ini tek başına açıklamaktadır.

Çizelge 5.8. MCD korelasyon matrisine dayanan TBA ile SEGE sıralaması

SIRA	İLLER	SEGE DEĞERİ
1	İstanbul	39,0306
2	İzmir	6,3427
3	Ankara	6,0479
4	Kocaeli, Sakarya, Düzce, Bolu, Yalova	5,0281
5	Bursa, Eskişehir, Bilecik	4,3118
6	Aydın, Denizli, Muğla	2,5010
7	Antalya, Isparta, Burdur	2,4953
8	Tekirdağ, Edirne, Kırklareli	2,3562
9	Manisa, Afyonkarahisar, Kütahya, Uşak	1,6345
10	Adana, Mersin	1,5915

Çizelge 5.8. 'in devamı

SIRA	İLLER	SEGE DEĞERİ
11	Balıkesir, Çanakkale	1,1423
12	Konya, Karaman	0,5920
13	Zonguldak, Karabük, Bartın	0,5818
14	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	-0,0175
15	Samsun, Tokat, Çorum, Amasya	-0,1427
16	Kayseri, Sivas, Yozgat	-0,1768
17	Hatay, Kahramanmaraş, Osmaniye	-0,2340
18	Gaziantep, Adıyaman, Kilis	-0,3042
19	Kastamonu, Çankırı, Sinop	-1,0273
20	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	-1,0327
21	Malatya, Elazığ, Bingöl, Tunceli	-1,6544
22	Erzurum, Erzincan, Bayburt	-2,9013
23	Ağrı, Kars, Iğdır, Ardahan	-4,7911
24	Van, Muş, Bitlis, Hakkari	-5,1200
25	Mardin, Batman, Şırnak, Siirt	-5,5143
26	Şanlıurfa, Diyarbakır	-5,6608

Analiz sonucunda en gelişmiş üç bölge TR10 (İstanbul), TR31 (İzmir), TR51 (Ankara) bölgeleridir. En gelişmemiş bölgeler ise TRB2 (Van, Muş, Bitlis, Hakkari), TRC3 (Mardin, Batman, Şırnak, Siirt) ve TRC2 (Şanlıurfa, Diyarbakır) bölgeleridir.

5.4 S Yayılım Matrisine Dayanaran TBA ile SEGE Değerlendirimesi

S yayılım matrisine dayanan temel bileşenleri elde etmek için bir önceki kısımda olduğu gibi öncelikle S kovaryans matrisi ve S korelasyon matrisi elde edilmiştir. S kovaryans matrisinin elde edilmesinde MATLAB R2009b programında FSDA toolboxı kullanılmıştır (Riani ve diğ., 2012). Daha sonra elde edilen kovaryans matrisi kullanılarak S korelasyon matrisi Eşitlik (3.15) ile elde edilmiştir. Elde edilen korelasyon matrisinin özdeğer ve özvektörleri hesaplanmıştır. İlgili özdeğerler ve varyans açıklama oranları Çizelge 5.9'da verilmiştir.

Çizelge 5.9. S korelasyon matrisinin özdeğerleri ve varyans açıklama oranları

Özdeğerler	Varyans Açıklama Oranı (VAO)	Kümülatif VAO
10,8799	0,5726	0,5726
3,4566	0,1819	0,7545
1,6655	0,0877	0,8422
0,8388	0,0441	0,8864
0,6214	0,0327	0,9191
0,4834	0,0254	0,9445
0,3595	0,0189	0,9634
0,2546	0,0134	0,9768
0,1687	0,0089	0,9857
0,1215	0,0064	0,9921
0,0589	0,0031	0,9952

Çizelge 5.9.'un devamı

Özdeğerler	Varyans Açıklama Oranı (VAO)	Kümülatif VAO
0,0536	0,0028	0,9980
0,0150	0,0008	0,9988
0,0116	0,0006	0,9994
0,0048	0,0003	0,9997
0,0038	0,0002	0,9999
0,0022	0,0001	1,0000
0,0000	0,0000	1,0000
0,0002	0,0000	1,0000

S korelasyon matrisinin 1'den büyük özdeğer sayısı üç olduğundan Kaiser kriterine göre önemli temel bileşenlerin sayısı üçtür. Bu üç temel bileşen toplam varyansın %84,22'sini açıklamaktadır.

Temel bileşenler analizi korelasyon matrisi kullanılarak yapıldığından standartlaştırılmış veri matrisine ihtiyaç vardır. MCD yönteminde olduğu gibi burada da standartlaştırma işleminde S konum ve yayılım tahminlerinin kullanılması gerekmektedir. S konum ve yayılım matrislerini kullanarak elde edilen standartlaştırılmış veri matrisi,

$$\boldsymbol{Z}_S = \left(\boldsymbol{\Sigma}_S^{1/2}\right)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_S) \qquad \textbf{(5.3)}$$

şeklinde hesaplanmıştır.

S yayılım matrisine dayanan temel bileşen analizinde de en büyük özdeğere sahip 1. temel bileşen skorları, bölgelerin sosyoekonomik gelişmişlik göstergesi olarak alınmış ve bölgelerin bu değerlere göre sıralaması Çizelge 5.10'da verilmiştir. Bu temel bileşen tek başına toplam varyansın % 57,26'sını açıklamaktadır.

Çizelge 5.10. S korelasyon matrisine dayanan TBA ile SEGE sıralaması

SIRA	İLLER	SEGE DEĞERİ
1	İstanbul	6,0426
2	Ankara	3,7147
3	Bursa, Eskişehir, Bilecik	3,4427
4	Kocaeli, Sakarya, Düzce, Bolu, Yalova	3,1905
5	İzmir	2,7402
6	Antalya, Isparta, Burdur	2,5481
7	Aydın, Denizli, Muğla	2,4965
8	Tekirdağ, Edirne, Kırklareli	2,3982
9	Manisa, Afyonkarahisar, Kütahya, Uşak	2,0330
10	Balıkesir, Çanakkale	1,5619
11	Konya, Karaman	1,4796
12	Adana, Mersin	1,4047
13	Kayseri, Sivas, Yozgat	0,9378

Çizelge 5.10.'un devamı

SIRA	İLLER	SEGE DEĞERİ
14	Zonguldak, Karabük, Bartın	0,8556
15	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	0,6218
16	Samsun, Tokat, Çorum, Amasya	0,5928
17	Kastamonu, Çankırı, Sinop	0,3572
18	Hatay, Kahramanmaraş, Osmaniye	0,1991
19	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	0,1686
20	Gaziantep, Adıyaman, Kilis	0,0236
21	Malatya, Elazığ, Bingöl, Tunceli	-0,2361
22	Erzurum, Erzincan, Bayburt	-0,9661
23	Ağrı, Kars, Iğdır, Ardahan	-2,2090
24	Van, Muş, Bitlis, Hakkari	-2,5654
25	Mardin, Batman, Şırnak, Siirt	-3,2074
26	Şanlıurfa, Diyarbakır	-3,3995

Analiz sonucunda en gelişmiş üç bölge TR10 (İstanbul), TR51 (Ankara) ve TR41 (Bursa, Eskişehir, Bilecik) bölgeleridir. En gelişmemiş bölgeler ise TRB2 (Van, Muş, Bitlis, Hakkari), TRC3 (Mardin, Batman, Şırnak, Siirt) ve TRC2 (Şanlıurfa, Diyarbakır) bölgeleridir.

5.5 ROBPCA ile SEGE Değerlendirimesi

Daha önce yapılan analizlerde değişken sayısı gözlem sayısından daha düşük olduğundan bu veriler düşük boyutsal veri olarak nitelenmiş ve klasik ya da robust korelasyon matrislerine dayanarak temel bileşenler analizi yapılmış, böylece kalkınma ajanslarına ait 26 bölge 19 sosyoekonomik göstergeden hesaplanan gelişmişlik endeksleri ile sıralanmıştır.

SEGE çalışmaları genellikle 81 il için uygulandığından daha fazla sosyoekonomik gösterge kullanılmaktadır. Ancak gözlem sayısı 81'den 26'ya düşürüldüğünde klasik yöntemde değişken sayısı en fazla 25 olmaktadır ki bu durum bölgelerin çok fazla gösterge kullanılmaksızın değerlendirmeye alınmasına neden olmaktadır. Örneğin, Çizelge 5.1 incelendiğinde bölgelerin sağlık hizmetleriyle ilgili herhangi bir değişkene rastlanmamaktadır. Bu sorunları ortadan kaldırmak amacıyla klasik temel bileşenler analizine sunulan alternatif bir robust yaklaşım ROBPCA yöntemidir. Bu yöntem değişken sayısının gözlem sayısından daha yüksek olduğu yüksek boyutsal ($p > n$) verilerde de kullanılabilmektedir.

ROBPCA'nın sağladığı bu yarar göz önüne alınarak uygulamanın bu aşamasında 26 kalkınma ajansı bölgesi Çizelge 5.2'de verilen 46 sosyoekonomik

gösterge bakımından değerlendirilmiştir. ROBPCA sonuçları MATLAB R2009b programında FSDA toolbox'ı kullanılarak elde edilmiştir (Riani ve diğ., 2012).

Analiz sonucunda elde edilen 1. Temel bileşen skorları bölgelerin sosyoekonomik gelişmişlik endeksi olarak kabul edilmiş ve bölgeler aldıkları SEGE değerlerine göre Çizelge 5.11'de sıralanmışlardır. 1. Temel bileşen toplam varyansın %75,148'ini açıklamaktadır.

Çizelge 5.11. ROBPCA yöntemi ile SEGE sıralaması

SIRA	İLLER	SEGE DEĞERİ
1	İstanbul	19324,9049
2	Ankara	19022,3395
3	İzmir	8046,9567
4	Kocaeli, Sakarya, Düzce, Bolu, Yalova	5087,7513
5	Antalya, Isparta, Burdur	3281,1568
6	Aydın, Denizli, Muğla	3078,9385
7	Bursa, Eskişehir, Bilecik	3053,3108
8	Tekirdağ, Edirne, Kırklareli	2631,3506
9	Zonguldak, Karabük, Bartın	2499,6713
10	Balıkesir, Çanakkale	1771,7221
11	Adana, Mersin	1436,2605
12	Kayseri, Sivas, Yozgat	482,1030
13	Hatay, Kahramanmaraş, Osmaniye	170,0197
14	Manisa, Afyonkarahisar, Kütahya, Uşak	149,4580
15	Konya, Karaman	-15,9938
16	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	-141,1040
17	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	-339,4714
18	Kastamonu, Çankırı, Sinop	-527,0589
19	Samsun, Tokat, Çorum, Amasya	-674,4044
20	Gaziantep, Adıyaman, Kilis	-936,5357
21	Malatya, Elazığ, Bingöl, Tunceli	-1160,1362
22	Erzurum, Erzincan, Bayburt	-1604,2760
23	Ağrı, Kars, Iğdır, Ardahan	-2640,2809
24	Mardin, Batman, Şırnak, Siirt	-2734,8023
25	Şanlıurfa, Diyarbakır	-2757,9177
26	Van, Muş, Bitlis, Hakkari	-2905,2349

Analiz sonucunda en gelişmiş üç bölge TR10 (İstanbul), TR51 (Ankara) ve TR31 (İzmir) bölgeleridir. En gelişmemiş bölgeler ise TRC3 (Mardin, Batman, Şırnak, Siirt), TRC2 (Şanlıurfa, Diyarbakır) ve TRB2 (Van, Muş, Bitlis, Hakkari) bölgeleridir.

Nihai olarak tüm yöntemlere göre bölgelerin SEGE sıralaması Çizelge 5.12'de özetlenmiştir.

Çizelge 5.12. TBA yöntemlerine göre SEGE sıralamalarının karşılaştırılması

BÖLGE KODU	İLLER	Klasik TBA	MCD TBA	S TBA	ROBPCA
TR10	İstanbul	1	1	1	1
TR21	Tekirdağ, Edirne, Kırklareli	6	8	8	8
TR22	Balıkesir, Çanakkale	10	11	10	10
TR31	İzmir	2	2	5	3
TR32	Aydın, Denizli, Muğla	7	6	7	6
TR33	Manisa, Afyonkarahisar, Kütahya, Uşak	8	9	9	14
TR41	Bursa, Eskişehir, Bilecik	4	5	3	7
TR42	Kocaeli, Sakarya, Düzce, Bolu, Yalova	3	4	4	4
TR51	Ankara	5	3	2	2
TR52	Konya, Karaman	12	12	11	15
TR61	Antalya, Isparta, Burdur	9	7	6	5
TR62	Adana, Mersin	13	10	12	11
TR63	Hatay, Kahramanmaraş, Osmaniye	18	17	18	13
TR71	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir	19	20	19	16
TR72	Kayseri, Sivas, Yozgat	15	16	13	12
TR81	Zonguldak, Karabük, Bartın	11	13	14	9
TR82	Kastamonu, Çankırı, Sinop	17	19	17	18
TR83	Samsun, Tokat, Çorum, Amasya	16	15	16	19
TR90	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane	14	14	15	17
TRA1	Erzurum, Erzincan, Bayburt	22	22	22	22
TRA2	Ağrı, Kars, Iğdır, Ardahan	23	23	23	23
TRB1	Malatya, Elazığ, Bingöl, Tunceli	21	21	21	21
TRB2	Van, Muş, Bitlis, Hakkari	24	24	24	26
TRC1	Gaziantep, Adıyaman, Kilis	20	18	20	20
TRC2	Şanlıurfa, Diyarbakır	26	26	26	25
TRC3	Mardin, Batman, Şırnak, Siirt	25	25	25	24

6. SONUÇ VE ÖNERİLER

İnsan hakları evrensel beyannamesinin 21/2 maddesinde "Herkesin ülkesinin kamu hizmetlerinden eşit olarak yararlanma hakkı vardır" denilmektedir. Bu beyannamenin onaylandığını ifade eden Bakanlar Kurulu Kararı resmi gazetede 1949 yılında yayımlanmış ve beyanname yürürlüğe girmiştir. Dolayısıyla ülkenin gelişmesi belirli bölgelerle sınırlı kalmamalı ve tüm yurttaşların faydalanacağı şekilde hizmetler sunulmalıdır.

Bölgeler arasındaki farklılıkların azaltılması ve herkesin kamu hizmetlerinden eşit şekilde yararlanması düşüncesi ile ülkemizde kalkınma ajansları kurulmuştur. Ajansların amacı çalışma bölgelerini kalkındırarak bölgeler arasındaki eşitsizliği ortadan kaldırmaktır. Dolayısıyla geri kalmış bölgelerdeki kalkınma ajanslarına daha fazla devlet desteği verilmesi gerekmektedir.

Fakat bölgelerin sosyoekonomik gelişmişliği literatürde çok fazla ilgilenilen bir konu olmamıştır. Bunun yerine 81 ilin sosyoekonomik gelişmişliği incelenmiştir. Ancak bu durum daha önce de örneklendirildiği gibi bölgeler arasındaki gelişmişliği açıklamada yeterli olamayabilir. Bu nedenle çalışmada 26 kalkınma ajansına ait bölgelerin sosyoekonomik gelişmişlik endeksleri hesaplanarak bölgeler değerlendirilmiştir.

Bölgeler arasında gelişmişlik farklarının çok yüksek olması veri setinde aykırı değer incelemesi yapılmasını gerekli kılmıştır. Bu amaçla öncelikle klasik daha sonra da robust tahminler kullanılarak inceleme yapılmıştır. Klasik yönteme göre aykırı değer tespit edilemezken, robust yöntemde iki gözlem aykırı değer olarak belirlenmiştir. Dolayısıyla tek ya da çok değişkenli bir veri setinde aykırı değerin varlığından şüphe ediliyorsa ve inceleme yapılmak isteniyorsa, klasik tahmin edicilerin bu değerlerden etkileneceği ve gerçeği tam olarak yansıtmayacağı bilinmelidir. Bu nedenle aykırı değer incelemelerinde robust tahmin edicilerine dayanan yöntemler kullanılmalıdır.

Daha sonra bölgelerin sosyoekonomik gelişmişliğinin incelenmesinde temel bileşenler analizi kullanılmıştır. Veride aykırı değer varlığından dolayı klasik

yöntemin yanı sıra robust yöntemler kullanılmıştır. Analiz sonucunda benzer sonuçlara ulaşılmış olmasına rağmen sosyoekonomik gelişmişlik endeks değeri olarak alınan 1. Temel bileşenler bakımından açıklanan varyans miktarları klasik yönteme göre robust alternatiflerinde artış göstermiştir. Bu varyans açıklama oranları Çizelge 6.1'de özetlenmiştir.

Klasik yönteme göre robust yöntemlerde birinci temel bileşenin açıkladığı varyansın artma nedeni, robust istatistiklerin varyansı arttıran aykırı değerlerin etkisi arındırıldıktan sonra elde edilmesidir.

Çizelge 6.1. Farklı yöntemlere göre 1. Temel bileşenin varyans açıklama oranları

YÖNTEM	Varyans Açıklama Oranı
Klasik TBA	0,5037
MCD TBA	0,5365
S TBA	0,5726
ROBPCA	0,7515

Klasik temel bileşenler analizi ve robust kovaryans matrislerine dayanan temel bileşenler analizinde bölge birimi sayısı düşük olduğundan az sayıda değişken ile çalışılabilmiştir. Ancak bu durumda çalışmaya sağlık hizmetleri gibi önemli göstergeler dâhil edilememiş, değerlendirmenin kapsamı genişletilememiştir. Klasik temel bileşenler analizine alternatif olarak sunulan ve yeni bir robust yaklaşım olan ROBPCA yöntemi ile klasik yöntemin aksine bölge sayısından daha fazla değişken ile inceleme yapılabilmiştir. Böylelikle daha fazla gösterge ve bilgi kullanılarak daha güvenilir sonuçlara ulaşılmıştır.

Nihai olarak bundan sonraki SEGE çalışmalarında illerin yanı sıra bölgelerinde sosyo-ekonomik gelişmişliğinin incelenmesi, bölgeler arasındaki gelişmişlik farkları düşürülüp elde edilecek veri yapılarında aykırı değer sorunu ortadan kaldırılıncaya kadar söz konusu sosyoekonomik gelişmişlik endeksi çalışmalarında robust yöntemlerin kullanılması gerektiği ve böylece daha az bilgi kaybıyla daha güçlü sonuçların elde edileceği düşünülmektedir.

KAYNAKLAR

Albayrak A. S., 2005. Türkiye'de İllerin Sosyoekonomik Gelişmişlik Düzeylerinin Çok Değişkenli İstatistik Yöntemlerle İncelenmesi, *ZKÜ Sosyal Bilimler Dergisi*, 1, 153-176.

Alpar C. R., *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, 2011. Detay Yayıncılık, Ankara.

Ammann L. P., 1993. Robust singular value decompositions: A new approach to projection pursuit, *Journal of the American Statistical Association*, 88, 505-514.

Butler R. W., Davies P. L., Jhun M., 1993. Asymptotics for the minimum covariance determinant estimator, *The Annals of Statistics*, 21, 1385-1400.

Campell N. A., Lopuhaa H. P., Rousseeuw P. J., 1998. On the calculation of a robust S-estimator of a covariance matrix, *Statistics in Medicine*, 17, 2685-2695.

Croux C., Haesbroeck G., 2000. Principal Component Analysis based on Robust Estimator of the Covariance or Correlation Matrix, *Biometrica*,87, 603-618.

Davies P. L., 1987. Asymptotic behaviour of S-estimators of multivariate location and dispersion matrices, *The Annals of Statistics*, 15, 1269-1292.

Davies P. L., 1992. The asymptotics of Rousseeuw's minimum colume elipsoid estimator, *The Annals of Statistics*, 20, 1828-1843.

Devlin S. J., Granadesikan R., Kettenring J. R., 1975. Robust estimation and outlier detection with correlation coefficients, *Biometrika Trust*, 62, 531-545.

Devlin S. J., Granadesikan R., Kettenring J. R., 1981. Robust Estimation of Dispersion Matrices and Principal Components, *American Statistical Association*, 76, 354-362.

Devlin S. J., Granadesikan R., Kettenring J. R., 2014. Robust Estimation of dispersion matrices and principal components, *Journal of the American Statistical Association*, 76, 354-362.

Donoho D. L., 1982. Breakdown properties of multivariate location estimators, Doktora tezi, Harvard University.

Er F., Sönmez H., 2006. Öğrenci Başarı Notları İçin Robust Faktör Analizi Uygulaması, *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 7, 149-155.

Granadesikan R., Kettenring J. R., 1972. Robust Estimates, Residuals and Outlier Detection with Multiresponse Data, *International Biometric Society*, 28, 81-124.

Gümüş E., 2013. Çok Değişkenli Veride Aykırı Gözlemlerin Tespiti İçin En Küçük Kovaryans Determinantına Dayalı Test İstatistiğinin 1. Tip Hata Bakımından Sağlamlığının İncelenmesi, Yüksek Lisans Tezi, Gazi Üniversite, Fen Bilimleri Enstitüsü, ANKARA, 342830.

Huber P. J., 1964. Robust Estimation of a Location Parameter, *The Annals of Mathematical Statistics*,35, 73-101.

Huber P. J., 1985. Projection Pursuit, *The Annals of Statistics*, 13, 435-475.

Hubert M., Rousseeuw P. J., Branden K. V., 2005. ROBPCA: A New Approach to Robust Principal Component Analysis, *Technometrics*, 47, 64-79.

Hubert M., Rousseeuw P. J., Vanpaemel D., Verdonck T., 2013. A deterministic algorithm for S-estimators and MM-estimators of multivariate location and scatter, *Department of Mathematics and Leuven Statistics Research Center (LStat)*, Celestijnenlaan 200B, BE-3001 Heverlee, Belgium.

Lauritzen S., Properties of Estimators, University of Oxford, http://www.stats.ox.ac.uk/~steffen/teaching/bs2siMT04/si2c.pdf (Ziyaret tarihi:15.04.2014).

Lopuhaa H. P., 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance, *The Annals of Statistics*, 17, 1662-1683.

Maronna R. A., 1976. Robust M-estimators of multivariate location and scatter, *The Annals of Statistics*, 4, 51-67.

Maronna R. A., Martin R. D., Yohai V. J., *Robust Statistics Theroy and Methods*, John Wiley & Sons, England, 2006.

Maronna R. A., Yohai V. J., 2002. The behaviour of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, 90, 330-341.

Moller S. F., Frese J. V., Bro R., 2006. Robust methods for multivariate data analysis, *Journal of Cheomemetrics*,19, 549-563.

Özdamar K., *Paket Programlar ile İstatistiksel Veri Analizi,* 9. Baskı, Nisan Kitabevi, Ankara, 2013.

Riani M., Perrotta C., Torti F., 2012. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration, *Chemometrics and Intelligent Laboratory Systems*, 116, 17-32.

Rocke D. M., Woodruff D. L., 1996. Identification of Outlier in Multivariate Data, *Journal of the American Statistical Association*, 91, 1047-1061.

Roelant E., Aelst S.V., 2007. An L1-type estimator of multivariate location and shape, *Statistical Methods and Applications*, 15, 381-393.

Rousseeuw P. J., 1985. Multivariate Estimation With High Breakdown Point, *Mathematical Statistics and Applications*, 283-297.

Rousseeuw P. J., van Driessen K., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, 41, 212-223.

Tatlıdil H., *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* Akademi Matbaası, Ankara, 1996.

Verboven S., Hubert M., 2005. LIBRA: a MATLAB Library for Robust Analysis, *Chemometrics and Intelligent Laboratory Systems*, 75, 127-136.

Walczak B., Massart D. L., 1995. Robust Principal components regression as a detection tool for outliers, *Chemometrics and Intelligent Laboratory Systems*, 27, 41-54.

Yaycılı A. Ö., 2006. Temel Bileşenler Analizi için Robust Algoritmaları, Yüksek Lisans Tezi, Gazi Üniversite, Fen Bilimleri Enstitüsü, ANKARA, 180176.

Yazar I., Yavuz H. S., Çay M. A., 2009. Temel Bileşen Analizi Yönteminin Ve Bazı Klasik Ve Robust Uyarlamalarının Yüz Tanıma Uygulamaları, *Eskişehir Osmangazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 22, 49-63.

URL – 1: **http://tuikapp.tuik.gov.tr/DIESS/SiniflamaSurumDetayAction.do?surumId=164** (Ziyaret tarihi: 18.03.2014)

EKLER

EK A: Klasik TBA'ne göre illerin SEGE sıralaması

EK A: Klasik TBA'ne göre illerin SEGE sıralaması

SIRA	İL	SEGE	SIRA	İL	SEGE
1	İstanbul	22,07031	42	Kırıkkale	-0,40965
2	Kocaeli	4,464798	43	Artvin	-0,46004
3	Ankara	4,348531	44	Malatya	-0,50833
4	İzmir	4,102506	45	Sinop	-0,52834
5	Bursa	3,281792	46	Çorum	-0,55353
6	Antalya	2,137454	47	Bartın	-0,60112
7	Denizli	1,91897	48	Elazığ	-0,64681
8	Tekirdağ	1,76586	49	Kahramanmaraş	-0,6726
9	Eskişehir	1,281891	50	Osmaniye	-0,69655
10	Gaziantep	1,15356	51	Çankırı	-0,72209
11	Manisa	1,059835	52	Erzincan	-0,74365
12	Çanakkale	0,984741	53	Niğde	-0,77446
13	Sakarya	0,982404	54	Aksaray	-0,79407
14	Muğla	0,956722	55	Giresun	-0,82334
15	Yalova	0,832796	56	Kastamonu	-0,90736
16	Bilecik	0,829838	57	Tunceli	-0,95483
17	Kırklareli	0,722294	58	Tokat	-0,96288
18	Kayseri	0,63516	59	Sivas	-0,98012
19	Konya	0,617222	60	Gümüşhane	-1,03564
20	Adana	0,578735	61	Ordu	-1,07304
21	Mersin	0,531314	62	Bayburt	-1,15926
22	Uşak	0,528791	63	Yozgat	-1,18607
23	Karaman	0,525339	64	Kilis	-1,35222
24	Balıkesir	0,488469	65	Erzurum	-1,54113
25	Rize	0,409336	66	Ardahan	-1,71121
26	Düzce	0,396832	67	Adıyaman	-1,74036
27	Aydın	0,331558	68	Iğdır	-1,75531
28	Hatay	0,308263	69	Kars	-1,97029
29	Bolu	0,278377	70	Batman	-1,97929
30	Trabzon	0,256673	71	Hakkari	-2,09227
31	Edirne	0,206517	72	Diyarbakır	-2,15562
32	Kütahya	0,095958	73	Bingöl	-2,23625
33	Karabük	0,053568	74	Mardin	-2,26115
34	Burdur	0,00948	75	Bitlis	-2,5829
35	Samsun	-0,02013	76	Şırnak	-2,71588
36	Isparta	-0,05901	77	Şanlıurfa	-2,79282
37	Zonguldak	-0,13013	78	Siirt	-2,86922
38	Nevşehir	-0,21902	79	Van	-2,91287
39	Kırşehir	-0,26297	80	Ağrı	-2,991
40	Amasya	-0,28047	81	Muş	-3,01926
41	Afyonkarahisar	-0,30133			

ÖZGEÇMİŞ

Adı Soyadı :	Hasan BULUT
Doğum Yeri veTarihi :	SAMSUN / 15.04.1989
Adres :	Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü ATAKUM/ SAMSUN
E-Posta :	hasan.bulut@omu.edu.tr
Lisans :	Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü 2007-2011
Mesleki Deneyim:	Araştırma Görevlisi, Ondokuz Mayıs Üniversitesi İstatistik Bölümü, 2012-…