



**DAYANIKLI BULANIK TEMEL
BİLEŞENLER ANALİZİ İÇİN YENİ BİR
YAKLAŞIM**

SEVGİ GANIK

YÜKSEK LİSANS TEZİ

MATEMATİK ANABİLİM DALI

T.C.
SİNOP ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**DAYANIKLI BULANIK TEMEL BİLEŞENLER ANALİZİ İÇİN YENİ BİR
YAKLAŞIM**

SEVGİ GANIK

YÜKSEK LİSANS TEZİ
MATEMATİK ANABİLİM DALI

DANIŞMAN
DOÇ. DR. B. BARIŞ ALKAN

SİNOP-2018

TEZ KABUL

SEVGİ GANIK tarafından hazırlanan “DAYANIKLI BULANIK TEMEL BİLEŞENLER ANALİZİ İÇİN YENİ BİR YAKLAŞIM ” başlıklı bu çalışma, 13.07.2018 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak, jürimiz tarafından **YÜKSEK LİSANS tezi** olarak kabul edilmiştir.

Başkan Prof. Dr. Kamil DEMİRCİ
Sinop Üniversitesi/ Matematik Bölümü



Üye Doç. Dr. B.Barış ALKAN
Sinop Üniversitesi/ İstatistik Bölümü



Üye Doç. Dr. Pelin KASAP
Ondokuz Mayıs Üniversitesi/ İstatistik Bölümü



ETİK BEYANI

Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu, bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.



Sevgi GANIK

TEŞEKKÜR

Yüksek lisans eğitimim boyunca çalışmalarına yönlendirme yapan, araştırmalarımın bütün aşamasında bilgilerini ve yardımlarını esirgemeyerek akademik ortamda gelişmeye katkıda bulunan, akademik anlamdaki katkılarının yanı sıra; sevgi ve merhamet dolu kişiliği ile tez hazırlama sürecim boyunca gösterdiği her türlü destek ve yardımdan dolayı saygıdeğer danışman hocam Doç. Dr. B.Barış ALKAN' a en içten dileklerle teşekkür ederim.

Hayatım boyunca ve çalışmamın her aşamasında birçok fedakarlık göstererek her zaman beni destekleyen, bana güç veren başta annem ve babam olmak üzere tüm aileme ve arkadaşlarıma sevgi, saygı ve sonsuz teşekkür ederim.

ÖZET

DAYANIKLI BULANIK TEMEL BİLEŞENLER ANALİZİ İÇİN YENİ BİR YAKLAŞIM

Veri kümesinde aykırı gözlemlerin varlığından, klasik istatistik yöntemlerin birçoğu gibi temel bileşenler analizi yöntemi de ciddi şekilde etkilenmektedir. Bu nedenle veri kümelerinde aykırı gözlemlerin olması durumunda araştırmacılar alternatif yöntemlerin kullanımına yönelmektedir. Bu yöntemlerin başında dayanıklı ve bulanık yöntemlerin kullanımı gelmektedir. Bu çalışmada, aynı anda hem dayanıklı hem de bulanık yöntemlerin gücünü birleştiren ve bu iki yaklaşımı Temel Bileşenler Analizi (TBA) çatısı altında toplayan Dayanıklı Bulanık Temel Bileşenler Analizi için yeni bir yaklaşım önerilmektedir. Önerilen bu yeni yaklaşım, bulanık kodlanmış veriye dayanan dayanıklı temel bileşenler analizi olarak adlandırılır. Bu yaklaşımın performansının aykırı gözlem oranındaki değişimler ve örneklem genişliğindeki artışlardan nasıl etkilendiği, üç farklı senaryo düşünülerek oluşturulan yapay veri kümeleri ve bir gerçek veri kümesi üzerinden incelenmiştir. Veri kümesinde aykırı gözlemlerin varlığında elde edilen bulgulardan hareketle önerilen yaklaşımın klasik ve dayanıklı temel bileşen analizine göre daha iyi sonuç verdiği görülmektedir.

Çalışmanın birinci bölümünde, klasik temel bileşenler analizi ve dayanıklı temel bileşenler analizi ile ilgili ön bilgiler ve literatür özetine yer verilmiştir. İkinci bölümde ise çalışmada yer alan bazı temel kavramlar ve yöntemler kısaca açıklanmıştır. Üçüncü bölümde, dayanıklı bulanık temel bileşenler analizi için önerilen yeni yaklaşım algoritması ve çalışmada kullanılan veri kümeleri tanıtılmıştır. Sonraki bölümde ise, önceki bölümlerde ele alınan yöntemlerin gerçek ve yapay veri kümelerine uygulanması ile elde edilen bulgular değerlendirilmiştir. Son bölümde, çalışmadan elde edilen sonuçlar ve öneriler tartışılmıştır.

Anahtar Kelimeler: Dayanıklı Temel Bileşen Analizi, Bulanık Kodlanmış Veri, Aykırı Gözlemler

ABSTRACT

A NEW APPROACH FOR ROBUST FUZZY PRINCIPAL COMPONENT ANALYSIS

The principal component analysis is negatively affected in the presence of outliers in the data. Therefore, we need to select a robust method, if the data include outliers. In this study, a new approach to robust fuzzy principal component analysis is proposed. This new approach bring together the strong sides of both robust and fuzzy methods. The performance of this approach is checked over a set of artificial data sets and an actual data set. The findings obtained in this study indicate that this new approach provide better results than the classical and robust principal component analysis.

In the first section of the study, preliminary information on the analysis of classical principal component analysis and analysis of robust principal components and the literature summary are included. In the second section, some basic concepts and methods in the study are briefly explained. In the third section, a new approach algorithm for analyzing robust fuzzy principal components analysis and the data sets used in the study are introduced. In the next section, the findings obtained by applying the methods discussed in the previous sections are evaluated over the real and artificial data sets. In the last section, the results and suggestions obtained from the study will be discussed.

Key Words: Robust Principal Component, Fuzzy Code Data, Outliers

İÇİNDEKİLER

Sayfa No

TEŞEKKÜR.....	ii
ÖZET	iii
ABSTRACT.....	iv
SEMBOLLER LİSTESİ	vi
KISALTMALAR LİSTESİ.....	vii
ÇİZELGELER LİSTESİ	viii
1. GİRİŞ	1
2. GENEL BİLGİLER	3
2.1. Aykırı Gözlemler	3
2.2. Bozulma Noktası.....	3
2.3. Dayanıklı İstatistik	3
2.4. Klasik Temel Bileşenler Analizi	4
2.5. Bulanık Kodlanmış Veri	5
2.6. Dayanıklı Temel Bileşenler Analizi.....	5
3. MATERYAL VE YÖNTEM.....	7
3.1. Bulanık Kodlanmış Verinin Dayanıklı Temel Bileşenler Analizi.....	7
3.1.1. Önerilen Algoritma	7
3.2. Yapay ve Gerçek Veri Uygulamaları.....	7
4. BULGULAR.....	9
4.1. Yapay Veri 1, Düşük Boyutlu Durum ($p = 5$, $n = 30$ ve $n=100$).....	9
4.2. Yapay Veri 2, Düşük Boyutlu Durum ($p=5$, $n=30$ ve $n=100$).....	10
4.3. Yapay Veri 3, Yüksek Boyutlu Durum ($p=30$, $n=1000$).....	11
4.4. Gerçek Veri, Daudin'in Süt Kompozisyon Veri Kümesi ($p=8$, $n=86$), %20 Aykırı Gözlem.....	11
5. TARTIŞMA ve SONUÇ	13
6. KAYNAKLAR	14
ÖZGEÇMİŞ	18

SEMBOLLER LİSTESİ

k	: Temel bileşen sayısı
n	: Gözlem sayısı
p	: Değişken sayısı
P	: Yükler matrisi tahmini
S	: Örneklem kovaryans matrisi
T	: Skor matrisi
X	: $n \times p$ boyutlu veri matrisi
$\mathbb{R}^{n \times p}$	: $n \times p$ boyutlu reel sayılar kümesi
$\hat{\mu}$	: Ortalama vektörü tahmini
1_n	: Elemanları 1'lerden oluşan $n \times 1$ boyutlu sütun vektörü
$\hat{\mu}_k$	: Konum parametresi tahmini
$\hat{\Sigma}_k$	: Ölçek parametresi tahmini

KISALTMALAR LİSTESİ

BKVDTBA	: Bulanık Kodlanmış Verinin Dayanıklı Temel Bileşen Analizi
DTBA	: Dayanıklı Temel Bileşenler Analizi
EKKD	: En Küçük Kovaryans Determinant
KTBA	: Klasik Temel Bileşenler Analizi
TB	: Temel Bileşenler
TBA	: Temel Bileşenler Analizi
TVAO	: Toplam Varyans Açıklama Oranı
VAO	: Varyans Açıklama Oranı

ÇİZELGELER LİSTESİ

Sayfa No

Çizelge 4.1.1. Yapay Veri 1, Düşük Boyutlu Durum Tablosu ($p=5$, $n=30$ ve $n=100$)	9
Çizelge 4.2.1. Yapay Veri 2, Düşük Boyutlu Durum Tablosu ($p=5$, $n=30$ ve $n=100$)	10
Çizelge 4.3.1. Yapay Veri 3, Yüksek Boyutlu Durum Tablosu ($p=30$, $n=1000$)	11
Çizelge 4.4.1. Gerçek Veri ($p=8$, $n=86$), %20 Aykırı Gözlem	12



1. GİRİŞ

Yüksek boyutlu veri kümelerinde bir boyut indirgeme yöntemi, diğer çok değişkenli istatistiksel yöntemlerin başlangıç analizi veya çoklu bağlantı gibi problemlerin bir çözüm adımı olarak araştırmacıların yaygın olarak kullandığı Klasik Temel Bileşenler Analizi (KTBA)'nin bir çok araştırma alanı için hayati öneme sahip olduğu düşünülmektedir.

Literatürde oldukça popüler bir kullanım alanına sahip olan KTBA, veri kümesinde aykırı gözlemlerin varlığından negatif etkilendiği için alternatiflerinin geliştirilmesi üzerine bilim insanlarının ilgisini çekmektedir. Bu nedenle, KTBA'nin alternatif yaklaşımlarını içeren birçok bilimsel çalışma yapılmıştır. Croux ve Haesbroeck (2000), KTBA'nın, dayanıklı bir tahmin edicisinin özdeğerler ve özvektörler yardımıyla hesaplanabileceğini göstermişlerdir. Croux ve Ruiz-Gazen (2005) tarafından izdüşüm takibine dayanan dayanıklı bir alternatif yaklaşım önerilmiştir. Bu yaklaşımın, gözlem sayısının değişken sayısından küçük olduğu veri yapılarında uygun olduğunu ifade etmişlerdir. Çünkü böyle durumlarda dayanıklı kovaryans matrisi tahminini bulmak imkansızdır. Dayanıklı TBA üzerine diğer çalışmalar incelendiğinde, Locantore ve ark. (1999)'nın önerdiği küresel temel bileşenler analizi ve Maronna (2005)'in önerdiği dik temel bileşenler analizi yöntemleri çalışmaları görülmektedir (Farcomeni & Greco, 2015). Alkan ve ark. (2015) ise kayıp değer atama metotlarıyla dayanıklı TBA'ya bir alternatif bulunup bulunamayacağını araştırmışlardır. Daha sonra, Alkan (2016)'da, jackknife yeniden örnekleme yaklaşımına dayanan en küçük kovaryans determinantı yardımıyla elde edilecek dayanıklı TBA tahminlerin daha iyi bir performans gösterdiğini vurgulamıştır.

DTBA elde edilmesi için en basit yol konum ve ölçek parametrelerinin klasik tahminleri ile dayanıklı tahminlerinin yer değiştirilmesidir. Devlin ve ark. (1981) ve Campbell (1980) konum ve ölçek parametrelerinin M tahmin edicilerini bu doğrultuda kullanmışlardır. Fakat, M tahmin edicileri yüksek boyutta düşük bozulma noktasına sahip oldukları için iyi performans göstermediklerinden dolayı popülaritesini yitirmişlerdir. Rousseeuw (1984) de verilen en küçük kovaryans determinant (EKKD) metodu çok değişkenli konum ve ölçek parametrelerinin oldukça dayanıklı bir tahmin edicisidir ve %50 lik bir aykırı gözlem oranına kadar sağlam tahminler verir. TBA 'nin

dayanıklı versiyonu, μ ve Σ parametrelerinin, $\hat{\mu}$ ve $\hat{\Sigma}$ dayanıklı tahminleriyle yer değiştirilmesi ile bulunabilir (Todorov ve Filzmoser, 2009).

Veri kümesinde aykırı gözlemlerin varlığında önerilen diğer yaklaşımlar bulanık mantık temeline dayanmaktadır. Bulanık modelleme ve bulanık istatistik yaklaşımları belirsizlik, kayıp veri ve aykırı gözlemlerin varlığı gibi problemlerle karşılaşılması durumunda başvurulabilecek alternatif yaklaşımlardandır. Son yıllarda bulanık regresyon, bulanık kümeleme, bulanık TBA gibi birçok bulanık istatistiksel yaklaşım önerilmiştir. Temel bileşenler analizi genelde kesin veri kümelerine uygulanır. Fakat, Lauro ve Palumbo (2000), Zimmermann ve Ravi (2001), Taheri (2003), Douzal-Chouakria ve ark. (2011), Viertl (2011), Calcagni ve ark. (2016) çalışmalarında bulanık, aralıklı ve sembolik veri kümelerinin analizi için KTBA yöntemini bir takım uyarlamalarla genişletmişlerdir. Bu çalışmalara ek olarak bulanık kümelemeye dayanan bulanık TBA üzerine literatürde birçok çalışmaya rastlamak mümkündür. Bu konuyu Bezdek ve ark. (1984), Dumitrescu ve ark. (1994), Pop ve ark. (1996), Sarbu ve Pop (2000), Yang ve Wang (1999), Sarbu ve Pop (2001), Sarbu ve Pop (2004), Sarbu ve Pop (2005) çalışmalarında ele almışlardır.

Bu çalışmada, aynı anda hem dayanıklı hem de bulanık yöntemlerin gücünü birleştiren ve bu iki yaklaşımı Temel Bileşen Analizi çatısı altında toplayan Dayanıklı Bulanık Temel Bileşen Analizine yeni bir yaklaşım önerilmektedir.

Bu çalışmada, önerilen yaklaşımın aykırı gözlem oranındaki değişimlerden ve örneklem genişliğindeki artışlardan nasıl etkilendiğini farklı aykırı gözlem oranları ve örneklem genişliklerine sahip üç yapay, bir gerçek veri kümesinin analiz edilmesiyle yorumlanacaktır.

Bu tez çalışmasının İkinci bölümde ise çalışmada yer alan bazı temel kavramlar ve yöntemler kısaca açıklanmıştır. Üçüncü bölümde, dayanıklı bulanık temel bileşenler analizi için önerilen yeni yaklaşım algoritması ve çalışmada kullanılan veri kümeleri tanıtılmıştır. Sonraki bölümde ise, önceki bölümlerde ele alınan yöntemlerin gerçek ve yapay veri kümelerine uygulanması ile elde edilen bulgular değerlendirilmiştir. Son bölümde, çalışmadan elde edilen sonuçlar ve öneriler tartışılacaktır.

2. GENEL BİLGİLER

2.1. Aykırı Gözlemler

Aykırı gözlem, bir kitleden rasgele bir örneklemden farklı değerlerden anormal bir uzaklıkta olan bir gözlemdir. Aykırı gözlemler için, örneklemin çoğunun sahip olduğu dağılımdan farklı bir dağılıma veya aynı dağılıma fakat farklı parametre değerlerine sahip oldukları düşünülür.

Veri kümesinde aykırı gözlemlerin varlığında, dayanıklı olmayan tahmin edicilere bağlı olarak elde edilecek tahminler, hipotez testleri ve güven aralıklarının tutarlı ve güvenilir sonuçlar vermesi beklenemez.

2.2. Bozulma Noktası

Bozulma noktası, aykırı gözlem sayısının toplam gözlem sayısına oranı olarak bilinmektedir. Bu oran bir tahmin edicinin dayanıklılığının önemli bir göstergesidir.

İlk olarak Hampel (1971) tarafından bozulma noktasının genel bir tanımı verilmiştir. Daha sonra Donoho ve Huber (1983) sonlu örneklem için daha basit bir tanım vermiştir (Leroy ve Rousseeuw, 1987).

Bu tanıma göre, iyi bir dayanıklı tahmin edicinin yüksek bir bozulma noktasına sahip olması gerektiği söylenmektedir. Bozulma noktası için elde edilebilecek en yüksek değer, % 50'dir. Çünkü, % 50'yi aşan bozulma noktası ile normal gözlemlerle aykırı gözlemler arasında ayırım yapılamamaktadır.

2.3. Dayanıklı İstatistik

Dayanıklı istatistik üzerine çalışmalar son 10 yılda önem kazanmıştır. Birçok araştırmacı dayanıklı istatistiğin teorisini geliştirmek ve klasik istatistiksel metotların dayanıklı versiyonlarını elde etmek üzerine çalışmaktadır.

Normallik, bağımsızlık ve doğrusallık gibi klasik varsayımlar genellikle sağlanamamaktadır. Bu nedenle, bu varsayımları temel alan istatistiksel tahmin ediciler ve testler sürecin duyarlılığına ve sapmanın büyüklüğüne bağlı olarak yanlış sonuçlar verecektir.

Güvenilir sonuçlar elde etmek için parametrik modellerden sapmanın türünü açıklayan istatistiksel bir teori gereklidir. Dayanıklı istatistik parametrik modellere adeta bir komşu gibi çalışır. Parametrik modellerin avantajlarını kullanır. Bu yüzden dayanıklı istatistik, parametrik modellere bir yaklaşım teorisi olarak görülebilir.

2.4. Klasik Temel Bileşenler Analizi

Klasik Temel Bileşenler Analizi (KTBA), p orjinal değişkenli bir yapıyı k boyutlu lineer birleşimlerini bularak boyut indirgemeyi hedefleyen çok değişkenli istatistiksel bir tekniktir.

$\mathbf{X} = \mathbf{X}_{n,p} \in \mathbb{R}^{n \times p}$ olsun. $\mathbf{X}$ 'deki p boyutlu gözlemler $x_1, \dots, x_n$ ile gösterilsin. Verilen tahmin edilmiş yükler matrisi $\mathbf{P}$ ve ortalama vektörü $\hat{\boldsymbol{\mu}}$, yeni vektörler üzerinde merkezleştirilmiş $\mathbf{X}$ 'in izdüşümü $\mathbf{T} = (\mathbf{X} - \mathbf{1}_n \hat{\boldsymbol{\mu}}') \mathbf{P}$ skor matrisini verir. Burada $\mathbf{1}_n$, elemanları 1'lerden oluşan $n \times 1$ boyutlu sütun vektördür. KTBA, skorların maksimum varyansa sahip ve ilişkisiz olmasını sağlayan bir $\hat{\boldsymbol{\mu}}$ ve $\mathbf{P}$ araştırması olarak tanımlanabilir. KTBA vektörleri, $\mathbf{X}$ 'in klasik örneklem kovaryans matrisi $\mathbf{S}$ 'nin özvektörlerine karşılık gelirken, özvektör üzerine izdüşürülmüş verinin varyansı $\mathbf{S}$ 'nin özdeğerlerine karşılık gelmektedir. Orijinal değişkenlerin varyansı büyük farklılıklar gösterdiği zaman, öncesinde veri matrisinin standartlaştırılması gerekmektedir (Hubert ve ark., 2016).

Genellikle KTBA'da veri kümesindeki bilgiyi açıklamak için $k < p$ boyut gereklidir. Burada temel bileşen sayısı k 'nin seçimi oldukça önemli olup, bu seçim için literatürde kullanılan çeşitli yaklaşımlar vardır. En basit ve en popüler yaklaşımlardan birisi scree plot grafiğini kullanmaktır. Diğerleri ise toplam varyans açıklama oranına göre k 'yı belirlemektir. Daha sonra $\mathbf{P}$ 'nin ilk k sütunu alınır ve $P_{p,k} = [P_1, \dots, P_k]$ olarak gösterilir (Hubert ve ark., 2016).

2.5. Bulanık Kodlanmış Veri

Bulanık kodlama, Guitonneau ve Roux (1977) tarafından ortaya atılmıştır. Guitonneau ve Roux (1977), çalışmalarını uygunluk analizi üzerine yapmışlardır. Bulanık kodlama fikri sonraki yıllarda çok değişkenli istatistiksel yöntemlerin uygulama alanlarında da kullanılmaktadır (Aşan ve Greenacre, 2011). Bu çalışmada, bulanık kodlanmış veriyi elde etmek için üç parametrelili üçgen üyelik fonksiyonu kullanılmıştır. Bu fonksiyon,

$$f(x; a, b, c) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ \frac{c-x}{c-b}, & b < x < c \\ 0, & c \leq x \end{cases}$$

ile tanımlanmaktadır. Burada, a, b ve c parametreleri sırasıyla X 'in minimum, ortalama ve maksimum değerlerine karşılık gelmektedir. Bu fonksiyon yardımıyla veri kümesinin 0-1 aralığında değişen bulanık kodlara dönüştürülür .

Literatürde üçgen üyelik fonksiyonuna alternatif olabilecek Trapezoidal, Gaussian ve Cauchy gibi üyelik fonksiyonları da vardır (Jang ve ark., 1997; Smithson, 2006; Aşan ve Greenacre, 2011).

2.6. Dayanıklı Temel Bileşenler Analizi

Temel bileşenler analizi, örneklem kovaryans matrisindeki her türlü değişimden etkilenebileceğinden dolayı, veri yapısında yer alan aykırı gözlemlerden oldukça olumsuz bir yönde etkilendiği için yanlış ve tutarsız bulguların bulunmasına yol açabilir. Veri kümesinde yer alabilecek sadece bir tane aykırı gözlem bile bütün süreci negative yönde etkileyebilir. Bu durum gerçekte olduğundan daha yüksek bir varyasyona sebep olur. Örneğin, uç noktalarda özdeğerler ve bunlara bağlı olarak da gerçekte var olamayacak kadar fazla toplam varyans açıklama oranlarına (VAO) neden olabilir. Temel bileşenler analizi için dayanıklı metotların seçilmesi ile bu tür sıkıntılar yüksek oranda giderilebilir (Farcomeni ve Greco, 2015).

Gözlem sayısı (n), değişken sayısı (p)'ndan büyük olduğu durumlarda kovaryans matrisinin dayanıklı tahminini bulurken EKKD metodundan faydalanılmaktadır (Rousseeuw, 1984). Bu metot, aykırı gözlemlere karşı oldukça sağlam ve hesaplama kolaylığından dolayı son yıllarda önerilen en çok tercih edilen metotlardan biri olmuştur (Rousseeuw ve Van Driessen, 1999).

EKKD tahmin edicisini ifade etmek için n gözlemden oluşan tüm veri kümesi ve bu veri kümesinin h boyutlu altkümeleri ele alınır. Yani kombinasyon($n;h$) sayısı kadar alt küme ile çalışılır. Burada h değeri, tahmin edicinin sağlamlığını ifade eder ve en azından bir alt sınır olarak $[(n + p + 1)/2]$ seçilmelidir. EKKD tahmin edicisi bu alt kümeler içerisinde kovaryans determinantı minimum olan optimal h -alt kümesini elde etmeye çalışır. $\hat{\mu}_{EKKD}$, konum parametre tahmini ve $\hat{\Sigma}_{EKKD}$, ölçek parametre tahmini olarak tanımlanır. Bu tahmin edici $(n - h)$ tane aykırı gözlemin varlığında dayanıklılığını muhafaza edebilir. Bazı çalışmalarda h 'ın değerinin yaklaşık $[0.75n]$ olarak seçildiği görülmektedir (Todorov ve Filzmoser, 2009).

3. MATERYAL VE YÖNTEM

3.1. Bulanık Kodlanmış Verinin Dayanıklı Temel Bileşenler Analizi

Bulanık kodlanmış verinin dayanıklı temel bileşenler analizi (BKVDTBA) için, üçgen üyelik fonksiyonu yardımıyla veri kümesi yeniden kodlanarak kodlanmış veri kümesi elde edilecektir. Daha sonra orijinal veri kümesi, elde edilen bu bulanık kodlar ile ağırlıklandırılacaktır. Elde edilen yeni veri kümesine EKKD'ye dayanan DTBA yöntemi uygulanacaktır. Bu bağlamda önerilen algoritma 3.1.1 de verilmektedir.

3.1.1. Önerilen Algoritma

- Adım1. Üçgen üyelik fonksiyonu yardımıyla veri kümesini yeniden kodla
Adım2. Orijinal veri kümesini yeniden kodlanarak elde edilen bulanık kodlarla ağırlıklandır.
Adım 3. Ağırlıklandırılmış yeni veri kümesi için ;
Adım3.1. n'in h'lı kombinasyonu değeri elde edilir.
Adım3.2. h örneğe sahip her alt küme için ;
Adım 3.2.1. Örnek kovaryans matrisi hesaplanır.
Adım 3.2.2. Örnek kovaryans matrisinin determinanı hesaplanır,
Adım 4. EKKD'ya sahip olan alt küme seçilir ve alt kümenin örneklem ortalama vektörü ve örneklem kovaryans matrisini bulunur.

Buradan, $\hat{\mu}_{BKVDTBA}$ ve $\hat{\Sigma}_{BKVDTBA}$ çok değişkenli konum ve ölçek tahminleri elde edilmektedir. Bu tahminlere dayanan bu yaklaşım BKVDTBA olarak adlandırılır.

3.2. Yapay ve Gerçek Veri Uygulamaları

Bu çalışmada, veri kümesinde aykırı gözlemlerin varlığında KTBA, EKKD'ye dayanan DTBA ve önerdiğimiz BKVDTBA'nın performanslarının karşılaştırması için, Daudin'in süt kompozisyon veri kümesi gerçek veri kümesi olarak kullanılmıştır. Ayrıca farklı aykırı gözlem oranlarına göre oluşturulan senaryolar doğrultusunda elde edilen iki yapay veri kümesi üzerinden de yöntemlerin değerlendirilmesi sağlanmıştır. Yapay veri olarak 3 farklı senaryo düşünülerek elde edilen veriler kullanılmıştır. Bu

senaryolardan ilk ikisi düşük boyutlu durum, ikinci senaryo ise yüksek boyutlu durumu ifade etmektedir. Yöntemler kullanılarak veri analizi için R programında bulunan rrcov , robustbase, ve rrcovHD kütüphanelerinden yararlanılmıştır (R Development Core Team, 2011; Todorov ve Filzmoser, 2009). Önerilen Dayanıklı Bulanık Temel Bileşenler Analizi ise R 'da fonksiyon oluşturularak yapılmıştır.



4. BULGULAR

4.1. Yapay Veri 1, Düşük Boyutlu Durum ($p = 5$, $n = 30$ ve $n=100$)

Yapay veri 1 kullanılarak, yöntemlerin düşük boyutlu ($p = 5$, $n = 30$ ve $n=100$) durumdaki performanları incelenecektir. Bu ilk veri setinde $p=5$ değişkeni çok değişkenli normal dağılım $N_5([5 \ 10 \ 7 \ 2 \ 1], \text{diag}[100 \ 20 \ 8 \ 4 \ 2])$ ve farklı oranlardaki aykırı gözlemler (% 10, % 20, % 30), 5 serbestlik dereceli çok değişkenli t dağılımından üretilmiştir.

Çizelge 4.1.1. Yapay Veri 1, Düşük Boyutlu Durum Tablosu ($p=5$, $n=30$ ve $n=100$)

		%10 Aykırı Gözlem			%20 Aykırı Gözlem			%30 Aykırı Gözlem		
		TB1	TB2	VAO *	TB1	TB2	VAO *	TB1	TB2	VAO *
n=30	KTBA	0.37064	0.2934	0.6640	0.4330	0.24227	0.6753	0.4577	0.1974	0.6552
	DTBA	0.5656	0.2919	0.8575	0.7085	0.2133	0.9217	0.5792	0.2801	0.8592
	BKVDTBA	0.4346	0.4077	0.8423	0.8708	0.07149	0.9422	0.6916	0.2212	0.9128
n=100	KTBA	0.33378	0.2203	0.5541	0.3616	0.21496	0.5766	0.4587	0.1861	0.6449
	DTBA	0.764	0.1495	0.9135	0.6612	0.2327	0.8938	0.7996	0.1041	0.9036
	BKVDTBA	0.597	0.3490	0.9460	0.7729	0.1797	0.9527	0.7452	0.1857	0.9309

* VAO (Varyans Açıklama Oranı)

*TB (Temel Bileşen)

Yapay veri 1 KTBA, DTBA ve BKVDTBA yöntemleri ile analiz edildiğinde Çizelge 4.1.1 'de verilen bulgular elde edilmiştir. Çizelge 4.1.1 incelendiğinde; %10 aykırı gözlem varlığında, $n=30$ durumunda, ilk iki boyut kümülatif toplam varyans açıklama oranlarına (TVAO) göre BKVDTBA, DTBA'ya göre bir iyileşme gösterememiştir. Fakat aykırı gözlem oranı %20 ve %30'a çıktığında önerdiğimiz yaklaşım olan BKVDTBA'nın daha iyi bir performans gösterdiği söylenir. Bu durumda küçük örneklemelerde %10 gibi bir aykırı gözlem olması durumunda DTBA, %20 civarında bir aykırı gözlem olması durumunda ise BKVDTBA'nın tercih edilmesi önerilmektedir. $n= 100$ durumunda, %10, %20 ve %30 aykırı gözlem oranlarında, BKVDTBA iyi bir performans göstermiştir. Buradan gözlem sayısının 30'dan fazla

oduğu ve veri kümesinde aykırı gözlemlerin varlığında BKVDTBA'nın kullanımı önerilmektedir.

4.2. Yapay Veri 2, Düşük Boyutlu Durum ($p=5$, $n=30$ ve $n=100$)

Yapay veri 2 kullanılarak yöntemlerin düşük boyutlu ($p = 5$, $n = 30$ ve $n=100$) durumlarını inceleyeceğiz. $p=5$ değişkenli çok değişkenli standart normal dağılım $N_5(\mathbf{0}, \mathbf{I})$, ve farklı oranlardaki aykırı gözlemler (% 10,% 20,% 30), 5 serbestlik dereceli çok değişkenli t dağılımından (T_5) üretilmiştir.

Çizelge 4.2.1. Yapay Veri 2, Düşük Boyutlu Durum Tablosu ($p=5$, $n=30$ ve $n=100$)

		%10 Aykırı Gözlem			%20 Aykırı Gözlem			%30 Aykırı Gözlem		
		TB1	TB2	VAO *	TB1	TB2	VAO *	TB1	TB2	VAO *
n=30	KTBA	0.3368	0.2837	0.6206	0.2735	0.2453	0.5188	0.3359	0.2233	0.5593
	DTBA	0.4415	0.3318	0.7733	0.3226	0.2409	0.5635	0.4143	0.2311	0.6455
	BKVDTB	0.4133	0.2793	0.6927	0.4327	0.2495	0.6822	0.3537	0.2990	0.6527
n=100	KTBA	0.2548	0.2358	0.4907	0.2600	0.2165	0.4766	0.2828	0.2292	0.5120
	DTBA	0.2891	0.2224	0.5115	0.3121	0.2412	0.5533	0.3268	0.2458	0.5726
	BKVDTB	0.3459	0.2564	0.6022	0.3492	0.2905	0.6397	0.3611	0.2404	0.6014

* VAO (Varyans Açıklama Oranı)

*TB (Temel Bileşen)

Yapay veri 2 KTBA, DTBA ve BKVDTBA yöntemleri ile analiz edildiğinde Çizelge 4.2.1 'de verilen bulgular elde edilmiştir. Çizelge 4.2.1 incelendiğinde elde edilen sonuçların, yapay veri 1'den elde edilen Çizelge 4.2.1 'de sunulan sonuçları güçlü bir şekilde desteklediği görülmektedir. %10 aykırı gözlem varlığında, $n=30$ durumunda, ilk iki boyut kümülatif toplam varyans açıklama oranlarına göre BKVDTBA, DTBA'ya göre bir iyileşme gösterememiştir. Fakat aykırı gözlem oranı %20 ve %30'a çıktığında önerdiğimiz yaklaşım olan BKVDTBA daha iyi bir performans gösterdiği söylenir. Bu durumda küçük örneklerde %10 gibi bir aykırı gözlem olması durumunda DTBA, %20 ve %30 civarında bir aykırı gözlem olması durumunda ise BKVDTBA'nın tercih edilmesi önerilmektedir. $n= 100$ durumunda, %10, %20 ve %30 aykırı gözlem varlığı durumlarının hepsinde BKVDTBA oldukça iyi performans sergilemiştir.

4.3. Yapay Veri 3, Yüksek Boyutlu Durum (p=30, n=1000)

. Yapay veri 3'te, yöntemlerin yüksek boyut ($p = 30$, $n = 1000$) durumdaki performansını inceleyeceğiz. $p=30$ değişkenli çok değişkenli standart normal dağılım $N_{30}(\mathbf{0}, \mathbf{I})$, ve % 30 aykırı gözlem oranı, 5 serbestlik dereceli çok değişkenli t dağılımından (T_5) üretilmiştir.

Yapay veri 3 KTBA, DTBA ve BKVDTBA yöntemleri ile analiz edildiğinde Çizelge 4.3.1'de verilen bulgular elde edilmiştir. Çizelge 4.3.1'de sunulan bulgular incelendiğinde, yüksek boyutlu veri kümelerinin analizinde de önerilen BKVDTBA yaklaşımının iyi çalıştığı görülmektedir.

Çizelge 4.3.1. Yapay Veri 3, Yüksek Boyutlu Durum Tablosu (p=30, n=1000)

	TB1	TB2	TB3	TB4	TB5	TB6	TB7	TB8	TB9	TB10	VAO*
KTBA	0.0509	0.0442	0.0440	0.0426	0.0415	0.0408	0.0394	0.0388	0.0374	0.0366	0.4164
DTBA	0.0471	0.0454	0.0442	0.0426	0.0418	0.0403	0.0396	0.0380	0.0377	0.0366	0.4138
BKVDTBA	0.0504	0.0477	0.0449	0.0448	0.0422	0.0402	0.0399	0.0393	0.0379	0.0367	0.4243

* VAO (Varyans Açıklama Oranı)

* TB (Temel Bileşen)

4.4. Gerçek Veri, Daudin'in Süt Kompozisyon Veri Kümesi (p=8, n=86), %20 Aykırı Gözlem

Çalışmada, önerilen Bulanık Kodlanmış Verinin Dayanıklı TBA yönteminin performansının değerlendirilmesi için, gerçek veri olarak Daudin ve ark. (1988) tarafından verilen 8 değişken ve 86 gözlem içeren süt kompozisyon verisi kullanılmıştır. Todorov ve ark.(1994), Atkinson (1994), Rocke ve Woodruff (1996) gibi bir çok araştırmacı aykırı gözlemlerin belirlenmesi ve dayanıklı istatistiksel çıkarımlar üzerine önerdikleri yeni metotların performanslarını göstermek için klasik metotlarla mukayeselerinde Daudin'in süt kompozisyon verisini örnek veri olarak seçmişlerdir. Bu yüzden çalışmada bu veri kümesi uygulama için seçilmiştir. Bu veri kümesinde yer alan çok değişkenli aykırı gözlemlerin belirlenmesi için düzeltilmiş kartil yöntemi kullanılmaktadır. Düzeltilmiş kartil yöntemi, ki-kare dağılımının dağılım fonksiyonu ve

karesel dayanıklı uzaklığın empirik dağılımı arasındaki farkı karşılaştırma yoluyla süreci tamamlar (Filzmoser, Reimann & Garrett, 2003).

Veri kümesi sırasıyla KTBA, DTBA ve Bulanık Kodlanmış Verinin Dayanıklı TBA yöntemleri ile analiz edilmiş olup elde edilen sonuçlar Çizelge 4.4.1’de sunulmuştur. Çizelge 4.4.1, klasik temel bileşenler analizi’nin ilk üç temel bileşenle toplam varyansın %77.12’si açıklandığı görülmüştür. Ancak, veri kümesinde 18 aykırı gözlem (%20) olması ve aykırı gözlemlerin varlığında klasik temel bileşenler analizi’nin VAO’nda gerçekte varolmayan büyümeler olabileceği ve aykırı gözlemlerin özellikle de birinci temel bileşen yönünü değiştirebileceğinden KTBA ile elde edilen açıklama oranını bir ölçüt olarak kullanmak uygun olmaz. Dayanıklı TBA yöntemi ile en büyük ilk iki temel bileşen ile TVAO %87.13 olarak elde edilir. Dayanıklı TBA veri kümesinde aykırı gözlemlerin varlığında kullanılacak sağlam bir metottur. KTBA ile karşılaştırma yapıldığında bize daha dayanıklı bir yaklaşım göstermektedir. Çizelge 4.4.1’de verilen diğer bir sonuç ise önerilen BKVDTBA yaklaşımı ile bulunan TVAO, % 92.45 ile en büyük TVAO’na sahip olmasıdır. Bu durumda Çizelge 4.4.1’deki sonuçlar arasında karşılaştırma yapıldığında veri kümesinde aykırı gözlemlerin varlığında Klasik TBA ve Dayanıklı TBA yöntemlerinin yerine Bulanık Kodlanmış Verinin Dayanıklı Temel Bileşenler Analizi yaklaşımının kullanılması ile daha dayanıklı sonuçlar vereceği görülmektedir. Elde edilen bu sonucun üç farklı senaryo ile ele alınan farklı yapay veri analiz sonuçlarını da desteklediği görülmektedir.

Çizelge 4.4.1. Gerçek Veri (p=8, n=86), % 20 Aykırı Gözlem

	TB1	TB2	VAO*
KTBA	0.6306	0.1406	0.7712
DTBA	0.5929	0.2785	0.8713
BKVDTB	0.8265	0.0980	0.9245

*VAO (Varyans Açıklama Oranı)

*TB (Temel Bileşen)

5. TARTIŞMA ve SONUÇ

Veri kümesinde aykırı gözlemlerin varlığından Temel Bileşen Analizinin etkilendiği bilinmektedir. Bu nedenle bu çalışmada aykırı gözlemlerin varlığından etkilenmesini minimuma indirerek daha dayanıklı sonuçlar veren yeni bir yaklaşım önerilmektedir. Çalışmada, çok değişkenli veri kümesinde aykırı gözlemlerin varlığında kullanılan dayanıklı temel bileşenler analizinin sonuçlarını iyileştiren BKVDTBA yaklaşımı önerilmiştir. Klasik ve EKKD'ye dayanan dayanıklı TBA ile karşılaştırıldığında, önerilen BKVDTBA yaklaşımının aykırı gözlemlerin varlığında, hem düşük boyutlu hem de yüksek boyutlu veri kümelerinin analizinde daha dayanıklı sonuçlar verdiği tespit edilmiştir.

Önerilen yaklaşımın etkinliği bir gerçek ve üç farklı yapay veri kümesi üzerinden incelenmiştir. Bulanık kodlama sürecinde, önerilen yaklaşım üzerine farklı üyelik fonksiyonlarının kullanımının negatif veya pozitif etkilerinin incelenmesinin gelecek bir çalışma olarak planlanmasının bu alana önemli katkılar sağlayacağı düşünülmektedir.

6. KAYNAKLAR

- Alkan, B. B. 2016. Robust Principal Component Analysis Based On Modified Minimum Covariance Determinant In The Presence Of Outliers (in Turkish). *Alphanumeric Journal*, 4(2).
- Alkan, B. B., Atakan, C., Alkan, N. 2015. A comparison of different procedures for principal component analysis in the presence of outliers. *Journal of Applied Statistics*, 42(8): 1716-1722.
- Aşan, Z., Greenacre, M. 2011. Biplots of fuzzy coded data. *Fuzzy sets and Systems*, 183(1): 57-71.
- Atkinson, A.C. 1994. Fast Very Robust Methods for the Detection of Multiple Outliers. *J. Amer. Statist. Assoc.* 89, 1329–1339.
- Bezdek, J. C., Ehrlich, R., Full, W. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers., Geosciences*, 10(2-3): 191-203.
- Calcagni, A., Lombardi, L., Pascali, E. 2016. A dimension reduction technique for two-mode non-convex fuzzy data. *Soft Computing*, 20(2): 749-762.
- Campbell, N. A., 1980. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, 231-237.
- Croux, C., Haesbroeck G., 2000. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87: 603–618.
- Croux, C., Ruiz-Gazen, A. 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95: 206–226.
- Daudin, J.J., Duby, C. and Trecourt, P. (1988). Stability of Principal Component Analysis Studied by the Bootstrap Method; *Statistics* **19**, 241–258
- Devlin, S. J., Gnanadesikan, R., Kettenring, J. R. 1981. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374): 354-362.

- Donoho, D.L. ve Huber, P.J. 1983. The Notion of Breakdown Point, in A Festschrift for Erich Lehmann (P.J. Bickel, K. A. Doksum ve J. L. Hodges, eds.) 157-184. Wadsworth, Belmont, CA.
- Douzal-Chouakria A., Billard, L, Diday, E. 2011. Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining*, 4(2): 229–246.
- Dumitrescu, D., S [acaron] rbu, C., Pop, H. 1994. A fuzzy divisive hierarchical clustering algorithm for the optimal choice of sets of solvent systems. *Analytical letters*, 27(5): 1031-1054.
- Farcomeni, A. , Greco, L. 2015. Robust methods for data reduction, CRC press.
- Filzmoser, P., Reimann, C., Garrett, R.G. 2003. Multivariate outlier detection in exploration geochemistry. Technical Report TS 03–5, Department of Statistics. Vienna University of Technology, Austria.
- Guitonneau, G. G., Roux, M. 1977. Sur la taxinomie du genre *Erodium*. *Les cahiers de l'analyse des données*, 2(1): 97-113.
- Hampel, F. R. 1971, A general Qualitative Definition of Robustness, *Annals of Mathematical Statistics*, vol. 42, 1887–1896
- Hubert, M., Engelen, S., Branden, K. V. 2016. A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics*, 34(2): 117-126.
- Jang, J. S. R., Sun, C. T., Mizutani, E. 1997. Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence.
- Lauro CN, Palumbo, F. 2000. Principal component analysis of interval data: a symbolic data analysis approach. *Comput Stat* 15(1): 73–87
- Leroy, A. M., Rousseeuw, P. J. 1987. Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics, New York
- Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K. 1999. Robust principal components for functional data. *Test* 8: 1–28.

- Maronna, R., 2005. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47(3): 264-273.
- Pop, H. F., Sârbu, C., Horowitz, O., Dumitrescu, D. 1996. A Fuzzy Classification of the Chemical Elements. *Journal of chemical information and computer sciences*, 36(3): 465-482.
- R Development Core Team, (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rocke, D. M., Woodruff, D. L. 1996. Identification of Outliers in Multivariate Data. *J. Amer. Statist. Assoc.* 91 (435): 1047–1061.
- Rousseeuw, P. J., 1984. “Least median of squares regression”, *Journal of the American statistical association*, 79(388): 871-880.
- Rousseeuw, P. J., Driessen, K. V., 1999. A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41(3): 212-223
- Sârbu, C., Pop, H. F. 2000. Fuzzy clustering analysis of the first 10 MEIC chemicals. *Chemosphere*, 40(5): 513-520.
- Sârbu, C., Pop, H. F. 2001. Fuzzy robust estimation of central location. *Talanta*, 54(1): 125-130.
- Sarbu, C., Pop, H. F. 2004. Fuzzy soft-computing methods and their applications in chemistry. *Reviews in Computational Chemistry*, 20, 249.
- Sarbu, C., Pop, H. F. 2005. Principal component analysis versus fuzzy principal component analysis: a case study: the quality of Danube water (1985–1996). *Talanta*, 65(5): 1215-1220.
- Smithson, M., Verkuilen, J. 2006. Fuzzy set theory: Applications in the social sciences (No. 147). Sage.
- Taheri, S.M. 2003. Trends in fuzzy statistics. *Austrian journal of statistics*, 32(3): 239–257

- Todorov, V., Neyko, N., Neytchev, P. 1994. Stability of High Breakdown Point Robust PCA, in Short Communications, COMPSTAT'94; Physica Verlag, Heidelberg.
- Todorov, V., Filzmoser, P. 2009. An object-oriented framework for robust multivariate analysis. J. Statist. Softw., 32(3): 1–47.
- Viertl, R. 2011. Statistical methods for fuzzy data. Wiley
- Yang, T. N., Wang, S. D. 1999. Robust algorithms for principal component analysis. Pattern Recognition Letters, 20(9), 927-933.
- Zimmermann, H. J., Ravi, V. 2001. A neural network and fuzzy rule base hybrid for pattern classification. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 5(2): 152-159.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	: Sevgi GANIK
Doğum Tarihi ve Yeri	: 25.10.1992 / SAMSUN
Yabancı Dili	: İngilizce
E-posta	: sevgignk@outlook.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Lisans	İstatistik	Sinop Üniversitesi	2015

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2017	BANTAŞ	Sayım Memuru

YAYINLARI

Makale

1. ALKAN B. Barış, **GANIK Sevgi** (2017). Robust Principal Component Analysis Based on Fuzzy Coded Data, *Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering*, 18(3),754-762.

Proje

1. **GANIK, Sevgi (Araştırmacı)**, BAP Projesi (No:FEF-1901-16-06). Aykırı gözlemlerin varlığında uyarlanmış en küçük kovaryans determinant tahminine dayalı dayanıklı lineer diskriminant analizi, *Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi*, Tamamlandı, 2017.