



**DAYANIKLI SEYREK TEMEL
BİLEŞENLER ANALİZİ**

İŞİL ÜNALDI

YÜKSEK LİSANS TEZİ

MATEMATİK ANABİLİM DALI

T.C.
SİNOP ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DAYANIKLI SEYREK TEMEL BİLEŞENLER ANALİZİ

İŞİL ÜNALDI

YÜKSEK LİSANS TEZİ
MATEMATİK ANABİLİM DALI

DANIŞMAN
YRD. DOÇ. DR. B.BARIŞ ALKAN

SİNOP-2017

T.C.
SİNOP ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bu çalışma, jürimiz tarafından 17/07/2017 tarihinde yapılan sınav ile Matematik Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Başkan: Prof. Dr. Cemal ATAKAN



Üye: Prof. Dr. Kamil DEMİRCİ



Üye: Yrd. Doç. Dr. B.Barış ALKAN



ONAY:

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

04.10.2017



Doç. Dr. Turgay KÖRKÜT
Fen Bilimleri Enstitüsü Müdürü

DAYANIKLI SEYREK TEMEL BİLEŞENLER ANALİZİ

ÖZET

Temel Bileşenler Analizi (TBA), çok boyutlu veri kümesi olduğunda genellikle ilk başvurulmuş boyut indirgeme ve veri işleme yöntemidir. Bu nedenle TBA, hemen hemen bütün bilimsel alanlarda kullanımı oldukça yaygın olan bir yöntemdir. Ancak TBA’da her bir temel bileşenin tüm orijinal değişkenlerin doğrusal birleşimi olmasından dolayı analiz sonuçlarının yorumlama aşamasında genellikle birtakım güçlüklerle karşılaşılır. Bu sıkıntıların çözümü için önerilen yaklaşımlardan biri Seyrek Temel Bileşenler Analizi (STBA) olarak adlandırılmaktadır. Fakat seyrek yaklaşımlar da TBA’da olduğu gibi veri kümesinde aykırı gözlemlerin olması durumuna karşı dayanıklı değildir. Bu çalışmada, STBA ve Dayanıklı Temel Bileşenler Analizinin (DTBA) avantajlı özelliklerini bir araya getiren, Croux ve ark. (2013) tarafından önerilen yaklaşımın performansı iki gerçek ve üç yapay veri kümesi üzerinden incelenecektir. Çalışmanın 1. bölümünde TBA, DTBA ve STBA hakkında ön bilgilerin yer aldığı giriş bölümü verilecektir. 2. bölümünde çalışmada kullanılan bazı temel kavramların tanımları ve açıklamaları ile birlikte literatür özeti yer alacaktır. 3. bölümünde incelenen veri kümeleri ve yöntemlere, 4. bölümde ise önceki bölümlerde ayrıntılı olarak incelenen yöntemlerin gerçek ve yapay veri kümeleri üzerinde uygulamalarına yer verilecektir. Son bölümde ise çalışmadan elde edilen sonuçlar tartışılacaktır.

Anahtar Kelimeler: Temel Bileşenler, Aykırı Gözlemler, Seyreklik, Dayanıklılık.

ROBUST SPARSE PRINCIPAL COMPONENT ANALYSIS

ABSTRACT

The Principal Component Analysis (PCA) is the first method of size reduction and data processing when the dataset is of a high-dimensional. Therefore PCA is a widely used method in almost all scientific fields. However, since all the original variables of each principal component in the PCA are linear combination, the interpretation process of the analysis results is often encountered with some difficulties. The approaches proposed for solving these problems are referred to as Sparse Principal Component Analysis (SPCA). However, sparse approaches are not resistant to the fact that there are outliers in the data set, such as in the PCA. In this study, Croux et al. (2013), which combines the advantageous properties of SPCA and Robust Principal Component Analysis (RPCA), the performance of the proposed approach will be examined through two real and three artificial datasets. Preliminary information on PCA, RPCA and SPCA will be given in Section 1 of the study. In the Section 2, a summary of the literature will be included with definitions and explanations of some basic concepts used in the study. The datasets and methods examined in Section 3 will be included, and in Section 4, the methods examined in detail in the previous sections will be applied to real and artificial datasets. In the last section, the results obtained in the study will be discussed.

Key Words: Principal Component, Outliers, Sparsity, Robustness.

TEŞEKKÜR

Bu tez konusunu bana vererek, çalışmalarım boyunca yardımlarını esirgemeyen danışman hocam Sayın Yrd. Doç. Dr. B. Barış ALKAN'a saygı ve teşekkürlerimi sunarım. Ayrıca çalışmalarım boyunca desteğini esirgemeyen aileme ve bu çalışmada emeği geçen herkese teşekkür ederim.



İÇİNDEKİLER

	Sayfa No
ÖZET	i
ABSTRACT.....	ii
TEŞEKKÜR.....	iii
İÇİNDEKİLER	iv
SEMBOLLER VE KISALTMALAR LİSTESİ	vi
ÇİZELGELER	viii
1.GİRİŞ.....	1
2. GENEL BİLGİLER VE LİTERATÜR ÖZETİ	3
2.1. Klasik Temel Bileşenler Analizi.....	3
2.2. Dayanıklı Skor Uzaklığı	3
2.3. Dik Uzaklık.....	4
2.4. En Küçük Kovaryans Determinant	4
2.5. İzdüşüm Takip Yaklaşımı.....	4
2.6. Literatür Özeti.....	5
3. MATERYAL VE YÖNTEM.....	7
3.1. Materyal.....	7
3.1.1. Yapay Veri Kümesi 1	7
3.1.2. Yapay Veri Kümesi 2	7
3.1.3. Yapay Veri Kümesi 3 ($p > n$ durumu)	7
3.1.4. Gerçek Veri 1-Borsa Verisi	8
3.1.5. Gerçek Veri 2-Kibler Verisi	8
3.2. Yöntem.....	8
3.2.1. Seyrek Temel Bileşenler Analizi	8
3.2.2. İzdüşüm Takibine Dayanan Dayanıklı Temel Bileşenler Analizi	9
3.2.3. Dayanıklı Seyrek Temel Bileşenler Analizi	10
3.2.4. Genişletilmiş Grid Algoritması.....	11
3.2.5. Seyreklik Parametresi λ 'nın Seçimi.....	13
4. BULGULAR.....	15

4.1. Yapay ve Gerçek Veri Kümesi Uygulamaları	15
4.1.1. Yapay Veri Kümesi 1 Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması	15
4.1.2. Yapay Veri Kümesi 2 Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması	19
4.1.3. Yapay Veri Kümesi 3 ($p > n$ durumu) Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması	20
4.1.4. Gerçek Veri 1-Borsa Verisi Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması	22
4.1.5. Gerçek Veri 2-Kibler Verisi Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması	23
5. TARTIŞMA VE SONUÇ	25
KAYNAKLAR	27
EKLER.....	29
ÖZGEÇMİŞ	41

SEMBOLLER VE KISALTMALAR LİSTESİ

SEMBOLLER

- h : En küçük kovaryans determinant yöntemindeki alt küme boyutu
- k : Temel bileşen sayısı
- l_j : L köşegen matrisin j . elemanı
- n : Gözlem sayısı
- p : Değişken sayısı
- $\mathbf{p}$: Yük vektörü
- $\mathbf{P}_k^s$: İlk k temel bileşeni içeren seyrek temel bileşenlerin yükler matrisi
- $\mathbf{P}_k^c$: İlk k temel bileşeni içeren seyrek olmayan temel bileşenlerin yükler matrisi
- $\mathbf{R}^s$: Seyrek artık matrisi
- $\mathbf{R}^c$: Seyrek olmayan artık matrisi
- r_j^s : Seyrek artık matrisinin j . sütunu
- r_j^c : Seyrek olmayan artık yükler matrisinin j . sütunu
- $\mathbb{R}^{n \times p}$: $n \times p$ boyutlu reel sayılar kümesi
- $\mathbf{X}$: $n \times p$ boyutlu veri matrisi
- $\hat{\mu}$: Ortalama vektörü tahmini
- λ_j : SCoTLASS için seyreklik parametresi
- λ_j^{TPO} : TPO için λ_j değerleri
- λ^{BIC} : BIC yaklaşımına göre seyreklik parametresi değeri
- λ_{max} : k bileşenli seyrek temel bileşenler analizi sonuçlarının tam seyrekliğindeki sonuçlar

KISALTMALAR

DTBA	: Dayanıklı temel bileşenler analizi
EKKD	: En küçük kovaryans determinant
İT	: İzdüşüm takip
İT-DTBA	: İzdüşüm takip dayanıklı temel bileşenler analizi
İT-DSTBA	: İzdüşüm takip dayanıklı seyrek temel bileşenler analizi
STBA	: Seyrek temel bileşenler analizi
STB	: Seyrek temel bileşenler
TB	: Temel bileşenler
TBA	: Temel bileşenler analizi
VAO	: Varyans açıklama oranı

ÇİZELGELER

Sayfa No

Çizelge 4.1. %10 aykırı gözlem varlığında, ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 7.8$) temel bileşen yükleri	16
Çizelge 4.2. %20 aykırı gözlem varlığında, ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 12.14$) temel bileşen yükleri	17
Çizelge 4.3. %30 aykırı gözlem varlığında, ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 3.87$) temel bileşen yükleri	18
Çizelge 4.4. Yapay veri 2’de değişkenlerin ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 11.03$) temel bileşen yükleri	19
Çizelge 4.5. Yapay veri 3’te değişkenlerin ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 50.69$) temel bileşen yükleri	21
Çizelge 4.6. Gerçek veri 1-Borsa verisinde değişkenlerin ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 3.31$) temel bileşen yükleri	22
Çizelge 4.7. Gerçek veri 2-Kibler verisinde değişkenlerin ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 5.5$) temel bileşen yükleri	24

DAYANIKLI SEYREK TEMEL BİLEŞENLER ANALİZİ

1.GİRİŞ

Temel bileşenler analizi (TBA), çok boyutlu veri kümesi olduğunda genellikle ilk başvurulanan boyut indirgeme yöntemidir. TBA, veri kümesini daha iyi özetlemeyi ve yorumlamayı sağlayan, p orijinal değişkenin indirgenmiş boyutlu bir uzayda k tane lineer birleşimlerini bularak, $k \leq p$, boyut indirgemeyi amaçlayan çok değişkenli bir yöntemdir. Temel bileşenler (TB), bu k tane lineer birleşim üzerine izdüştürülen verinin varyansını maksimum yapan yönler doğrultusundaki vektörlere karşılık gelir (Croux ve ark., 2013). İlk k temel bileşen orijinal verinin toplam varyansının yeterli bir bölümünü açıklamaktadır. TBA, örneklem kovaryans (veya korelasyon) matrisi ve ortalama vektörünü temel almaktadır.

TBA, hemen hemen bütün bilimsel alanlarda kullanımı oldukça yaygın bir çok değişkenli istatistik analiz yöntemi olmasına rağmen eksik yönleri de azımsanamayacak kadar çok olan bir yöntemdir. TBA’da temel bileşenleri tanımlayan ve temel bileşenleri yorumlamak için kullanılan bir dönüşüm matrisi var olup, bu matris *yükler matrisi* olarak adlandırılır. Yüklerin birçoğu mutlak değerce ne çok küçük ne de çok büyük elde edildiğinden, TB’lerin yorumlanması zordur. TBA sonuçlarının yorumlanabilirlik probleminin çözümüne yönelik literatürde çeşitli öneriler ortaya atılmıştır. Bu önerilerden birisi, yükler matrisinin basit yapısını elde etmek için döndürme yöntemlerini kullanmaktır (Jolliffe, 1995). Araştırmacılar, özellikle bir temel bileşene zayıf veya belirsiz katkı sağlayan orta yük değerlerine sahip bir değişkenin katkısından daha ziyade, değerce küçük (mutlak) veya değerce büyük yüklere sahip değişkenlerin katkısıyla ilgilenmek isterler. Aksi takdirde elde edilen temel bileşenleri yorumlamak çok zor olabilmektedir. TBA sonuçlarının yorumlanabilirliğini artırmak için çoğu sıfır yüklere sahip temel bileşenleri tahmin etmek için seyrek temel bileşenler analizi (STBA) olarak adlandırılan birçok yaklaşım sunulmuştur. Fakat tüm seyrek ve klasik yaklaşımlar veri kümesinde aykırı gözlemlerin olması durumuna karşı dayanıklı değildir. Bu bağlamda dayanıklı temel bileşenler analizi (DTBA) adı verilen birçok yöntem geliştirilmiştir. DTBA elde etmek için kullanılan yaklaşımlardan birisi de Croux

ve Ruiz-Gazen (2005) tarafından ortaya atılan izdüşüm takip (İT) yaklaşımıdır. Seyrek ve dayanıklı TBA'nın avantajlı özelliklerini bir araya getirmek için ise ilk çalışma Croux ve ark.(2013) tarafından yapılmıştır. Croux ve ark. (2013) çalışmalarında, izdüşüm takip TBA algoritmasının içine seyreklik temellerini entegre ederek, dayanıklı izdüşüm takip TBA'nın seyrek versiyonunu geliştirmişlerdir.

Bu çalışma, Croux ve ark. (2013) tarafından önerilen izdüşüm takip yaklaşımına dayanan dayanıklı seyrek temel bileşenler analizi yönteminin, farklı aykırı gözlem oranındaki değişimlerden nasıl etkilendiğini incelemeyi amaçlamaktadır.

Çalışmanın 2. bölümünde çalışmada kullanılan bazı temel kavramların tanımları ve açıklamaları ile birlikte literatür özeti yer alacaktır. 3. bölümünde incelenen veri kümeleri ve yöntemlere, 4. bölümde önceki bölümlerde ayrıntılı olarak verilen yöntemlerin gerçek ve yapay veri kümeleri üzerinde uygulamalarına yer verilecektir. Son bölümde ise çalışmadan elde edilen sonuçlar tartışılacaktır.

2. GENEL BİLGİLER VE LİTERATÜR ÖZETİ

2.1. Klasik Temel Bileşenler Analizi

Bir veri matrisi için TBA'yı tanımlayalım. $\mathbf{X} = \mathbf{X}_{n,p} \in \mathbb{R}^{n \times p}$ olsun. $\mathbf{X}$ 'deki p boyutlu gözlemler $\mathbf{x}_1, \dots, \mathbf{x}_n$ ile gösterilsin. Temel bileşenlerin yükleri, dik yükler matrisi $\mathbf{P}$ 'nin sütunlarında yer alır. Verilen tahmin edilmiş yükler matrisi $\mathbf{P}$ ve ortalama vektörü $\hat{\mu}$, yeni vektörler üzerinde merkezleştirilmiş $\mathbf{X}$ 'in izdüşümü $\mathbf{T} = (\mathbf{X} - \mathbf{1}_n \hat{\mu}')\mathbf{P}$ skor matrisini verir. Burada $\mathbf{1}_n$, elemanları 1'lerden oluşan $n \times 1$ boyutlu sütun vektörüdür. TBA, skorların maksimum varyansa sahip ve ilişkisiz olmasını sağlayan bir $\hat{\mu}$ ve $\mathbf{P}$ araştırması olarak tanımlanabilir. TBA 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 standartlaştırılmalıdır.

Genellikle veri kümesindeki bilgiyi açıklamak için $k \leq p$ boyut gereklidir. Temel bileşen sayısı k 'nın seçimi için çeşitli yaklaşımlar vardır. En basit ve en popüler yaklaşımlardan birisi scree plot grafiğini kullanmaktır. Bu grafik azalan özdeğerlere karşılık onların indekslerini verir. Seçilecek temel bileşen sayısı grafikteki kırılma noktalarına karşılık gelir ve önemli temel bileşen sayısı buna göre belirlenir. Temel bileşen sayısının seçiminden sonra, sadece $\mathbf{P}$ 'nin ilk k sütunu kullanılır ve $\mathbf{P}_{p,k} = [\mathbf{p}_1, \dots, \mathbf{p}_k]$ olarak ifade edilir (Hubert ve ark. 2016).

2.2. Dayanıklı Skor Uzaklığı

Dayanıklı skor uzaklığı (SU), bir temel bileşen skorunun, skorların merkezine uzaklığını ölçer. $\mathbf{x}_i$ gözlemi için dayanıklı skor uzaklığı ($i = 1, 2, \dots, n$ için),

$$SU_i = \sqrt{\sum_{j=1}^k \frac{(t_i)_j^2}{l_j}} = \sqrt{\mathbf{t}_i' \mathbf{L}^{-1} \mathbf{t}_i} \quad (2.2.1)$$

ile verilir. Burada k ; temel bileşenlerin sayısı, $(t_i)_j$; i . skor t_i 'nin j . değişkeni ve $\mathbf{L}$, dayanıklı temel bileşenlere karşılık gelen dayanıklı özdeğerleri içeren köşegen matristir.

k serbestlik derecesine sahip bir ki-kare dağılımının %97.5'lik çeyreğinin karekökü olan

$c_{SU} = \sqrt{\chi_{k,0.975}^2}$ değerinde, yüksek SU değerlerine sahip gözlemler için bir kesme

noktası belirlenmiştir. Skorlar yaklaşık olarak normal dağıldığında bu bilgi doğrulanmaktadır.

2.3. Dik Uzaklık

Bir $\mathbf{x}_i$ gözleminin TBA alt uzayına dik uzaklığı,

$$DU_i = \|\mathbf{x}_i - \hat{\mu} - \mathbf{P}_{p,k} \mathbf{t}_i\| \quad (2.3.1)$$

eşitliği ile bulunur. Burada $\hat{\mu} + \mathbf{P}_{p,k} \mathbf{t}_i$; $\mathbf{P}_{p,k}$ ve $\hat{\mu}$ ile tanımlanan TBA alt uzayı üzerinde $\mathbf{x}_i$ 'nin izdüşümüdür. Dik uzaklıklar için bir kesme noktası Hubert ve ark. (2005) tarafından önerilen yol izlenerek bulunur.

2.4. En Küçük Kovaryans Determinant

En küçük kovaryans determinant (EKKD) tahmin edicisini tanımlamak için tüm veri kümesinin (n gözlemden oluşan) h boyutlu alt kümeleri düşünülür. Yani n 'in h 'li kombinasyonu kadar alt kümeyle ilgilenilir. h değeri, tahmin edicinin dayanıklılığını belirler ve bir alt sınır olarak en azından $[(n + p + 1)/2]$ alınmalıdır. EKKD tahmin edicisi bu alt kümeler içerisinde kovaryans determinantı minimum olan optimal h alt kümesini bulmaya çalışır. EKKD konum parametresi tahmin edicisi $\hat{\mu}_{EKKD}$, optimal h alt kümenin ortalaması ve EKKD ölçek parametresi tahmin edicisi $\hat{\Sigma}_{EKKD}$ ise onun kovaryans matrisi olarak verilir. EKKD tahmin edicisi $(n - h)$ aykırı gözleme dayanabilir. Daha genel olarak, EKKD tahmin edicisi $(n - h + 1)/n$ bozulma noktası değerine sahiptir. h değerinin varsayılan değeri yaklaşık olarak $[0.75]$ olarak alınmaktadır (Hubert ve Engelen, 2004).

2.5. İzdüşüm Takip Yaklaşımı

İzdüşüm takip (İT) yaklaşımı dayanıklı tahminin çok daha kolay olduğu daha düşük boyutlu bir uzay üzerine çok değişkenli verinin izdüşümü fikrini kullanır. İT yaklaşımı ilk olarak Friedman ve Tukey (1974) tarafından önerilmiştir. Bu yaklaşımın amacı çok değişkenli veri yapısını gösteren vektörleri takip etmektir. Bu bir İT indeksini maksimize ederek yapılır. Bu indeksin maksimumunda elde edilen vektörler ilginç veri yapılarını ortaya çıkarmak için düşünülür.

Huber (1985) çalışmasında temel bileşenler analizinin, İT yaklaşımının özel bir durumu olduğunu açıklamıştır. Burada İT indeksi izdüşürülen verinin varyansdır ve diklik kısıtları maksimizasyon sürecine dahil edilmelidir. Li ve Chen (1985), dayanıklı

bir ölçek tahmin edicisi olarak TBA'nın dayanıklılığı için bu yaklaşımı kullanmıştır. Croux ve Ruiz-Gazen (2005) bu dayanıklı TBA yaklaşımının dayanıklılık özelliklerini araştırmışlardır. Daha sonra hızlı hesaplama için bir algoritma önermişlerdir.

İT yaklaşımının avantajları aşağıda verilmiştir.

- (a) Daha önce de bahsedildiği gibi dayanıklı tahminin düşük boyutta hesaplanması kolay ve hızlıdır.
- (b) Dayanıklı kovaryans tahmini, gözlem sayısının (n) değişken sayısından (p) oldukça fazla olduğu veri kümeleri ile sınırlıdır. İT yaklaşımı $p > n$ olması durumunda da kullanılması önerilen bir yaklaşımdır.
- (c) İzdüşüm vektörlerini araştırma ardışık bir süreçtir. Bu yüzden araştırmacı ilgilendiği vektörlerin sayısını belirleyebilir. İT yaklaşımının kullanımı çok boyutlu verilerde hesaplama zamanını ciddi şekilde azaltmaktadır.

2.6. Literatür Özeti

STBA, veri kümesi çok boyutlu olduğu zaman, değişkenlerin sadece bir alt kümesinin analiz edilmesi gerekebileceğinden kullanışlıdır. STBA, temel bileşenlerin daha kolay bir şekilde yorumlanması avantajını araştırmacılara sunmaktadır. Bunu yapmak için en kolay yol, belli bir eşik değerinden mutlak değerce daha küçük değerlere sahip olan yükleri sıfır olarak ayarlamaktır. Bu metot basit eşikleme olarak adlandırılır. Cadima ve Jolliffe (1995), bu yöntemin yanıltıcı olabileceğini fark etmişlerdir. Bu bağlamda, belli bir temel bileşene bir değişkenin katkısının belirlenmesi için değişkenlerin standart sapmalarının incelenmesi gerektiğini vurgulamışlardır.

TBA kapsamında seyreklik kavramı ilk olarak SCoTLASS yöntemini ortaya atan Jolliffe ve ark. (2003) tarafından ele alınmıştır. Daha sonra, Zou ve ark. (2006) SCoTLASS'ın verdiği sonuçlara göre daha iyi sonuçlar veren elastic-net regresyona dayanan STBA algoritmasını önermişlerdir. Metodun uyarlamaları ve iyileştirmeleri Wang ve ark. (2009) tarafından yapılmıştır. D'Aspremont ve ark. (2007) STBA için direkt bir formülasyon önermişlerdir. Sigg ve ark. (2008) STBA'yı çözmek için olasılıksal TBA'ya dayanan EM-TBA'yı tasarlamışlardır. Witten ve ark. (2009) ise, ceza matrisi ayrıştırması için genel bir prosedür geliştirmişler ve STBA için bu prosedürün nasıl uygulandığını göstermişlerdir. Journée ve ark.(2010) STBA'yı formüle eden GPower algoritmasını üretmişlerdir. Guo ve ark. (2010) değişkenlerin içinde blok

yapıları yakalayan bir kaynaşma cezası ortaya atmışlardır. STBA'nın elde edilmesi için diğer öneriler Jenatton ve ark. (2010) ve Bien ve ark. (2010) tarafından ele alınmıştır.

Veri kümesindeki aykırı gözlemlerin varlığında, TBA için birçok dayanıklı alternatif çözüm önerilmiştir. Bunlardan en önemlileri sırasıyla Li ve Chen (1985), Hubert ve ark. (2002), Croux ve Ruiz-Gazen (2005) tarafından ele alınan izdüşüm takip TBA yaklaşımı, Locantore ve ark. (1999) tarafından ele alınan küresel TBA yaklaşımı ve Hubert ve ark. (2005) tarafından sunulan dayanıklı TBA yaklaşımı olarak verilebilir. Croux ve Haesbroeck (2000), dayanıklı TBA (DTBA)'nın, korelasyon veya kovaryans matrisinin dayanıklı bir tahmin edicisinin özdeğerlerinin ve özvektörlerinin hesaplanmasıyla kolayca yapılabileceğini göstermişlerdir. Bu yaklaşım, çok değişkenli konum ve ölçek dayanıklı tahmini mümkün (değişken sayısı yeterince küçük olduğunda) olduğu sürece iyi çalışır. Croux ve Ruiz-Gazen (2005) tarafından ortaya atılan, izdüşüm takibini temel alan dayanıklı TBA yaklaşımı değişken sayısının gözlem sayısından fazla olduğu durumlarda ve yüksek boyutlu veri kümelerinin analizinde uygundur. İzdüşüm takip yönteminin avantajı TBA için kovaryans matrisinin tahmin edilmesinin gerekli olmamasıdır. Art arda elde edilen özdeğerler ve özvektörlerin dayanıklı tahminleri elde edilir.

Literatürde, temel bileşenler analizinin dayanıklı versiyonları ve onların dayanıklılık özellikleri hakkında birçok çalışmaya rastlamak mümkündür. Ayrıca, seyrek temel bileşenler analizi de son yıllarda özellikle internet teknolojileri, gen açıklama ve makine öğrenme gibi alanlarda, artan hızlı teknolojik gelişmeler sonucunda ortaya çıkan çok boyutlu (büyük veri) veri kümelerinin analizinde önem kazanmıştır.

Croux ve ark. (2013)'nın seyrek ve dayanıklı TBA'nın avantajlı özelliklerini bir araya getiren çalışmalarından sonra He ve ark. (2014), STBA'nın dayanıklılığını iyileştirmek için bir algoritma önermişlerdir. Önerdikleri algoritmanın dayanıklılığını ve etkinliğini yapay ve gerçek veri kümeleri üzerinden incelemişlerdir. STBA üzerine son çalışma Hubert ve ark. (2016) tarafından yapılmıştır. Çalışmalarında probleme farklı açıdan yaklaşmışlardır. Bu son çalışmanın Croux ve ark. (2013)'nın çalışmasından ana farkı, seyreklik adımı ile aykırı gözlem belirleme adımlarının birbirlerinden ayrılmasıdır.

3. MATERİYAL VE YÖNTEM

3.1. Materyal

3.1.1. Yapay Veri Kümesi 1

%10 aykırı gözlem durumu: 100 gözlem ($n=100$) ve 5 değişkenden ($p=5$) oluşan, 100 gözlemin 10 tanesi serbestlik derecesi 5 olan çok değişkenli t dağılımından (T_5), 90 tanesi $N_5(0, I)$ 'dan bir veri seti üretilmiştir.

%20 aykırı gözlem durumu: 100 gözlem ($n=100$) ve 5 değişkenden ($p=5$) oluşan, 100 gözlemin 20 tanesi serbestlik derecesi 5 olan çok değişkenli t dağılımından (T_5), 80 tanesi $N_5(0, I)$ 'dan bir veri seti üretilmiştir.

%30 aykırı gözlem durumu: 100 gözlem ($n=100$) ve 5 değişkenden ($p=5$) oluşan, 100 gözlemin 30 tanesi serbestlik derecesi 5 olan çok değişkenli t dağılımından (T_5), 70 tanesi $N_5(0, I)$ 'dan bir veri seti üretilmiştir.

3.1.2. Yapay Veri Kümesi 2

215 gözlem ($n=215$), 6 değişkenden ($p=6$) oluşan 215×6 boyutlu bir veri seti üretilmesi planlanmıştır. Bu amaçla 215 gözlemin 200'ü çok değişkenli normal dağılımdan, 15'i de yine çok değişkenli normal dağılımdan aykırı gözlem olarak üretilmiştir. 200 gözlemin ortalama vektörü (μ) ve kovaryans matrisi (Σ) şu şekildedir:

$$\mu = [0 \ 0 \ 0 \ 0 \ 0 \ 0], \Sigma = k\ddot{o}\ddot{s} \begin{bmatrix} 5 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

15 gözlemin ortalama vektörü (μ) ve kovaryans matrisi (Σ) şu şekildedir:

$$\mu = [0 \ 20 \ 20 \ 20 \ 20 \ 20], \Sigma = k\ddot{o}\ddot{s} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

3.1.3. Yapay Veri Kümesi 3 ($p > n$ durumu)

18 gözlem 20 değişkenden oluşan 18×20 boyutlu bir veri seti üretilmiştir. 18 gözlemin 15'i çok değişkenli normal dağılımdan, 3'ü de yine çok değişkenli normal dağılımdan aykırı gözlem olarak üretilmiştir. Verideki aykırı gözlem oranı %16'dır. 15 gözlemin ortalama vektörü (μ) ve kovaryans matrisi (Σ) şu şekildedir:

$$\mu = [0 \ 0 \ 0 \ \dots \ 0], \Sigma = k\ddot{o}\ddot{s} \begin{bmatrix} 5 & 1 & 1 & \dots & 1 \end{bmatrix}$$

3 gözlemin ortalama vektörü (μ) ve kovaryans matrisi (Σ) şu şekildedir:

$$\mu = [0 \ 20 \ 20 \ \dots \ 20], \Sigma = k\ddot{o}\mathring{s} [1 \ 1 \ 1 \ \dots \ 1]$$

3.1.4. Gerçek Veri 1-Borsa Verisi

İstanbul Menkul Kıymetler Borsası'nda Ocak 2005 ve Mart 2013 yılları arasında işlem gören sekiz hisse senedinin üç aylık getiri oranlarını içeren veri kümesi kullanılmıştır. Analiz edilen veri kümesi 33×8 boyutludur.

3.1.5. Gerçek Veri 2-Kibler Verisi

Kibler veri seti (Kibler ve ark., 1989), 205 farklı araba modeli için teknik ve sigorta ile ilgili verileri içeren 26 değişkenden oluşmaktadır. Yalnızca sürekli değişkenler ve eksik değerler içermeyen gözlemler burada kullanılmıştır ve bu da 195×14 boyutlu bir veri kümesi üzerinden analiz yapılmasını sağlamıştır. Değişkenlerin ölçeğini karşılaştırılabilir hale getirmek için, veri matrisinin her bir sütunu standart sapma ($V = var$ ise) veya dayanıklı bir ölçek ölçüsü ($V = Q_n$ ise) ile bölünmüştür.

3.2. Yöntem

3.2.1. Seyrek Temel Bileşenler Analizi

X veri matrisinin satırlarında yer alan, $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$, n tane çok değişkenli gözlem için ilk TBA vektörü,

$$\operatorname{argmax}_{\|\mathbf{p}\|=1} V(\mathbf{p}'\mathbf{x}_1, \dots, \mathbf{p}'\mathbf{x}_n) \quad (3.2.1.1)$$

ile verilir. Burada V varyansın bir ölçüsüdür. Standart dayanıklı olmayan durumda V deneysel varyanstır (var) ve optimal vektör $\mathbf{p}_1$, örneklem kovaryans matrisinin ilk özvektörüne karşılık gelir. (3.2.1.1) eşitliği ilk TB'yi bulmak için izdüşüm takip formülasyonudur.

Dayanıklı TBA vektörleri, V için dayanıklı bir varyans ölçüsü alınarak kolayca elde edilebilir. Bu ölçü karesel medyan mutlak sapma ya da daha etkin olan karesel Q_n tahmin edicisi olarak seçilebilir.

Q_n ölçek tahmin edicisi, tek değişkenli bir veri kümesi $y_1, \dots, y_n$ için, $|y_i - y_j|$, $1 \leq i < j \leq n$, tüm ikili uzaklıkların ilk çeyreği olarak tanımlanır. Bu ilk çeyrek daha sonra normal bir dağılım ölçeği için tutarlı bir tahmin edici elde etmek için 2.219 sabiti ile çarpılmalıdır. Bu yüzden Q_n^2 varyansın tutarlı ve dayanıklı tahmin edicisidir. Croux

ve Ruiz-Gazen (2005), temel bileşenler için dayanıklı ve etkin tahminler sağlayan izdüşüm takip indeksi olarak Q_n^2 tahmin edicisini kullanmayı önermişlerdir. Bu çalışmada Q_n^2 , dayanıklı varyans tahmin edicisi olarak alınmıştır.

İlk $j - 1$ TBA vektörlerinin bulunduğu varsayalım ($j > 1$). Bu durumda j . vektör ($j \leq p$),

$$\operatorname{argmax}_{\|p\|=1, p \perp p_1, \dots, p \perp p_{j-1}} V(p'x_1, \dots, p'x_n) \quad (3.2.1.2)$$

olarak tanımlanır.

TBA kapsamında seyreklik ilk olarak Jolliffe ve ark. (2003) tarafından SCoTLASS metodu ile ele alınmıştır. SCoTLASS, seyrek yükler sağlamak için bir L_1 kısıtını TBA ile birleştirmiştir. Sonuçta elde edilen amaç fonksiyonu,

$$\max_{\|p\|=1, p \perp p_1, \dots, p \perp p_{j-1}} p'Sp, \quad \text{kısıt } \|p\|_1 \leq t, \quad (3.2.1.3)$$

şeklinde yazılır ve bu amaç fonksiyonu, açıklanan varyansı maksimize eden p_j dik yüklerini bulmaya çalışır. Burada $\|p_j\|_1$, p_j 'nin L_1 normudur, $\|p\|_1 = \sum_{j=1}^p |p_j|$ dir.

Yukarıda verilen formun dual formülasyonu,

$$\max_{\|p\|=1, p \perp p_1, \dots, p \perp p_{j-1}} (p'Sp - \lambda_1 \|p\|_1) \quad (3.2.1.4)$$

ile verilir. Burada p_j , j . TBA vektörüdür. λ_1 , SCoTLASS için seyreklik parametresi olup, t yerine kullanılmıştır. λ_j 'nin büyük değerleri daha büyük seyrekliğe, sıfır değeri ise seyreklik yok anlamına karşılık gelmektedir.

3.2.2. İzdüşüm Takibine Dayanan Dayanıklı Temel Bileşenler Analizi

Klasik TBA ve seyrek TBA'da tahmin edilen yükler matrisi aykırı gözlemlere karşı çok duyarlıdır. Literatürde dayanıklı TBA üzerine yapılan çalışmalar bu konuyu ele almaktadır. Dayanıklı temel bileşenler elde etmek için en kolay yol konum ve ölçek parametrelerinin klasik tahminleriyle, onların dayanıklı tahminlerinin yer değiştirilmesidir. Bu yaklaşım dayanıklı çok değişkenli veri analizinde oldukça yaygın kullanılan bir yaklaşımdır (Hubert ve ark., 2008). Fakat yüksek boyutlarda özellikle de değişken sayısının gözlem sayısından büyük olduğu durumlarda dayanıklı kovaryans matrislerini hesaplamak zordur. Hatta böyle durumlarda tahmin edici var olmayabilir. Yaygın olarak kullanılan dayanıklı TBA metotlarından en bilinen iki tanesi dayanıklı

izdüşüm takip TBA (İT-TBA) ve ROBPCA olarak verilebilir. İT-TBA verinin izdüşürüldüğü ardışık vektörleri bulmak için yayılımın dayanıklı bir ölçüsünü maksimize eder (Croux ve Ruiz-Gazen, 2005). Hubert ve ark. (2005) tarafından önerilen ROBPCA yöntemi izdüşüm takip ve dayanıklı kovaryans tahmininden elde ettikleri fikirleri birleştirmiştir. TBA’da aykırı gözlemleri belirlemek için dayanıklı skor uzaklıkları ve dik uzaklıklar kullanılır.

Değişken sayısı (p), gözlem sayısından (n) küçük olduğu durumlarda kovaryans matrisinin dayanıklı tahminini bulurken en küçük kovaryans determinant (EKKD) yöntemi kullanılmaktadır (Rousseeuw, 1984, 1985; Hubert ve Engelen, 2004). Bu yöntem, konum ve ölçek tahmin edicisi aykırı gözlemlere karşı yüksek derecede dayanıklı ve hesaplanması açısından son zamanlarda geliştirilen en hızlı algoritmaya sahip olmasından dolayı oldukça popülerdir (Rousseeuw ve Van Driessen, 1999).

3.2.3. Dayanıklı Seyrek Temel Bileşenler Analizi

Croux ve ark. (2013) İT yaklaşımını ve seyrek TBA’yı birleştiren dayanıklı seyrek TBA yöntemini önermişlerdir. Yaklaşımları İT denklemleri içine L_1 cezasının eklenmesiyle oluşmaktadır. Bu yöntem, vektör yüklerinin çok büyük olmaması kısıtı altında bu vektörler üzerine izdüşürülen verinin ölçeğini maksimize eden vektörleri arar. Temel bileşenler, bir kovaryans matrisi tahmini olmaksızın doğrudan İT yaklaşımı kullanılarak elde edilir.

İlk seyrek TBA vektörünü bulmak için L_1 kısıtı eklenir ve

$$\tilde{\mathbf{p}}_1 = \operatorname{argmax}_{\|\mathbf{p}\|=1} (V(\mathbf{p}'x_1, \dots, \mathbf{p}'x_n) - \lambda_1 \|\mathbf{p}\|_1) \quad (3.2.3.1)$$

eşitliği elde edilir. Burada $\tilde{\mathbf{p}}_1$ vektörü ilk seyrek TBA vektörü olup, onun seyrekliği λ_1 parametresiyle kontrol edilmektedir. $\lambda_1 = 0$ olarak düzenlenirse ilk TBA vektörü $\mathbf{p}_1$ ’e kısıt eklenmemiş olur. Benzer şekilde, j . seyrek TBA vektörü ($1 < j \leq p$),

$$\tilde{\mathbf{p}}_j = \operatorname{argmax}_{\|\mathbf{p}\|=1, \mathbf{p} \perp \tilde{\mathbf{p}}_1, \dots, \mathbf{p} \perp \tilde{\mathbf{p}}_{j-1}} (V(\mathbf{p}'x_1, \dots, \mathbf{p}'x_n) - \lambda_j \|\mathbf{p}\|_1) \quad (3.2.3.2)$$

olarak tanımlanır. Burada V , ölçeğin bir ölçüsü olan varyans tahmin edicisidir. Klasik temel bileşenler için bu ampirik (deneysel) varyanstır. Dayanıklı seyrek TBA için bu varyans tahmin edicisi karelendirilmiş Q_n gibi bir dayanıklı varyans tahmin edicisidir. Eğer $V = \text{var}$ ise eşitlik (3.2.1.4) ve eşitlik (3.2.3.1) aynı olur. Croux ve ark. (2013) tarafından önerilen dayanıklı seyrek TBA’da ölçeğin dayanıklı tahmin edicisi için

Rousseeuw ve Croux (1993) tarafından önerilen dayanıklı Q_n tahmin edicisi kullanılmıştır. Bu Q_n tahmin edicisi, bir vektörün elemanları arasındaki ikili uzaklıkların ilk çeyreğidir.

(3.2.3.1) ve (3.2.3.2) eşitliklerinde verilen optimizasyon problemlerini çözmek kolay değildir. Croux ve ark. (2013) seyrek vektörleri bulmak için, değişkenlerin ikili çiftleri ile gerilen uzaylarda iteratif grid aramalarına dayanan bir algoritma önermişlerdir. Croux ve ark. (2007) tarafından önerilen grid algoritması İT yaklaşımında TBA vektörlerini elde etmek için kullanılan iyi çalışan bir algoritmadır. Bu algoritmayı Croux ve ark. (2013) seyrek TBA için genişletmişlerdir.

3.2.4. Genişletilmiş Grid Algoritması

Eşitlik (3.2.1.1) ve (3.2.1.2)'deki optimal vektörleri bulmak için p boyutlu bir uzay üzerinden maksimizasyon gereklidir. Bu konveks olmayan bir optimizasyon problemidir. Hatta $V = var$ için bile optimal vektörlerin analitik çözümlerini bulmak mümkün değildir. Üstelik, V argümanlarına göre diferansiyellenebilir olmayabileceği için, gradient temelli metotlar her zaman mümkün değildir. Temel bileşenlere dayanan İT'nin iyi yaklaşımlarını bulmak için Hubert ve ark. (2002), Croux ve Ruiz-Gazen (2005), Croux ve ark. (2007), Croux ve ark. (2013) çalışmalar yapmıştır. Croux ve ark. (2013) çalışmalarında, Croux ve ark. (2007) tarafından önerilen grid algoritmasını seyrek çözümler elde etmek için genişletmişlerdir. Bu algoritma eşitlik (3.2.3.1) ve eşitlik (3.2.3.2)'yi çözmek için oldukça hızlı hesaplama yapan bir algoritmadır. Bu algoritmanın tam bir uygulaması R (version 3.3.2) paketi pcaPP kütüphanesinde mevcuttur (Filzmoser ve ark., 2010). Aşağıda bu algoritmanın bir taslağı verilmiştir.

Hesaplama gerektiren temel bileşenlerin sayısı k olsun. İlk $j - 1$ seyrek TBA vektörleri $\tilde{\mathbf{p}}_{j-1}$ 'in elde edildiğini ve $\tilde{\mathbf{P}}_{j-1}$ yükler matrisinin ilk $j - 1$ sütununda toplandığını varsayalım, $1 \leq j \leq k - 1$. Şimdi $\tilde{\mathbf{p}}_j$ 'yi hesaplamak istiyoruz. Notasyon tutarlılığı için, $\tilde{\mathbf{P}}_0^\perp$ 'yi birim matris olarak gösterelim. $j > 1$ için, $\tilde{\mathbf{P}}_{j-1}^\perp$, ilk $j - 1$ seyrek TBA vektörleri ile gerilen uzaya dik alt uzay için ortonormal bir temeli sütunlarında içeren bir matris olsun ve $\mathbf{x}_i^{(j-1)} = (\tilde{\mathbf{P}}_{j-1}^\perp)' \mathbf{x}_i$, $i = 1, \dots, n$ için, eşitlik (3.2.3.2)'nin çözümü o zaman,

$$f(\mathbf{p}) = V(\mathbf{p}'\mathbf{x}_1^{(j-1)}, \dots, \mathbf{p}'\mathbf{x}_n^{(j-1)}) - \lambda_j \|\tilde{\mathbf{P}}_{j-1}^\perp \mathbf{p}\|_1 \quad (3.2.4.1)$$

ile verilen amaç fonksiyonunun maksimize edilmesi ile eşdeğerdir. Eşitlik (3.2.4.1), $\|\mathbf{p}\| = 1$ ve $\mathbf{p}$ 'nin $\mathbb{R}^{p-j+1}$ boyutlu alt bir uzaya sahip olması kısıtı altında yazılabilir. Seyreklik başlangıç değişkenlerinin uzayında vektör bileşenleriyle ilişkili olduğu ve daha az boyutlu bir uzayda $\mathbf{p}$ için böyle bir durum söz konusu olmadığı için, eşitlik (3.2.4.1)'de L_1 normunu almadan önce bu $\mathbf{p}$ vektörünün orijinal uzaya geri dönüşümüne ihtiyacımız vardır. Eşitlik (3.2.4.1)'i optimize etmek için grid algoritması kullanılır. Bu algoritmanın temel fikri, bu problemi birim norm kısıtı altında optimizasyonların bir dizisine indirgemektir. Yani, $[-\pi, \pi]$ üzerinden bir grid araması ile çözülebilen basitçe tek değişkenli bir maksimizasyon problemi olan birim çember üzerinde bir fonksiyonun maksimumlarının bir dizisi özüne indirgemektir. Eşitlik (3.2.4.1)'in optimizasyonu, $1 \leq j \leq k$ 'nın verilen bir değeri için düşünüldüğünde, aşağıdaki adımları uygulayalım:

Adım 1- $\mathbf{X}^{(j)}$, satırlarında $\mathbf{x}_i^{(j-1)}$ vektörlerine sahip olan bir matris olsun. Bu matrisin sütun sayısını $p_j = p - j + 1$ olarak gösterelim. Matrisin her bir s sütunu için $V_s = V(\mathbf{x}_{1s}^{(j)}, \dots, \mathbf{x}_{ns}^{(j)})$ izdüşüm indeks değeri hesaplanır. İzdüşüm indeksleri s 'de azalan sıradadır, $V_1 \geq V_2 \geq \dots \geq V_{p_j}$. Bu durumda, algoritmanın başlangıç değeri p_j bileşene sahip $\mathbf{p} = (1, 0, \dots, 0)$ vektörüdür.

Adım 2- Bu adım, $l = 1, \dots, \text{maxiter}$ (dış döngü) için mevcut en iyi çözüm $\mathbf{p} = (p^1, \dots, p^{p_j})$ aşağıdaki alt adımlar takip edilerek güncellenir.

$1 \leq s \leq p_j$ (iç döngü)

Adım 2.1- Tek değişkenli grid arama fonksiyonu üzerinden maksimizasyon

$$\gamma \rightarrow f(p^1 b(\gamma), \dots, p^{s-1} b(\gamma), \cos \gamma, p^{s+1} b(\gamma), \dots, p^{p_j} b(\gamma)),$$

burada, $b(\gamma) = \sin(\gamma)/\sqrt{1 - (p^s)^2}$, f fonksiyonunun argümanı üzerinde birim norm koşulu (kısıtı) tutulmaya devam ediyor. γ değerleri $\left[\arccos(p^s) - \frac{\pi}{2^{l-1}}, \arccos(p^s) + \pi/(2^{l-1}) \right]$ aralığında değişir. Bu yüzden $l = 1$ için, grid arama tüm birim çember üzerinden yapılır.

İterasyon adımı 1 arttıkça, daha kısıtlanmış bir arama yaparız. Bu fonksiyon tek sayı olan bir Ngrid değerlendirme noktalarını kullanan bir grid arama ile maksimize edilir. En iyi çözüme karşılık gelen her zaman $\gamma = \arccos(p^s)$ gibi bir grid noktasıdır.

Adım 2.2- Mevcut en iyi çözüm,

$\mathbf{p} \leftarrow f(p^1 b(\gamma^*), \dots, p^{s-1} b(\gamma^*), \cos \gamma^*, p^{s+1} b(\gamma^*), \dots, p^p b(\gamma^*))'$ olarak tanımlanır. Burada γ^* grid aramanın maksimumuna karşılık gelir.

Adım 3- Sonuç olarak grid algoritması ile j . temel bileşen için elde edilen $\mathbf{p}$ optimal seyrek vektörünün $\tilde{\mathbf{p}}_j = \tilde{\mathbf{P}}_{j-1}^\perp \mathbf{p}$ 'yi sağlayan orijinal uzaya geri dönüştürülmesi gerekmektedir.

Grid algoritması koordinat kökenli yöntemlerin bir türüdür. Genişletilmiş grid algoritması uygun bir hesaplama zamanı ile her zaman dayanıklı ve seyrek çözümler verir (Croux ve ark., 2013).

3.2.5. Seyreklik Parametresi λ 'nın Seçimi

Seyreklik parametresi λ 'nın seçimi n 'nin bir kümesi için hesaplanan amaç fonksiyonu optimize edilerek yapılır. λ , sıfır ile alabileceği maksimum değer arasında değişen farklı değerler alabilir. Eğer λ 'nın alabileceği maksimum değer belirlenemezse, o zaman tam seyreklik sağlayan minimum λ değeri tercih edilir. Tam seyreklik durumu, yükler matrisinde genelde sadece sıfır ve birlerden oluşan yüklerin minimum mutlak toplamını ifade etmektedir. Araştırmacılar, λ seçimi için iki optimizasyon yaklaşımından bir tanesini kullanarak bu süreci tamamlayabilirler. Bu yaklaşımlar sırasıyla TPO (Tradeoff Çarpım Optimizasyonu) ve BIC (Bayes Bilgi Kriteri) olarak verilir. BIC yaklaşımı, tüm TB'ler için aynı λ değerini seçer ve düşünülen TB'ler sayısı k 'nın özel bir seçimini ifade eder. Ancak TPO yaklaşımı her bir TB için ayrı olarak optimize edilir ve böylece k üzerinden bir karara bağlı olmayan bir model içinde farklı λ değerleri türetilir. TPO yaklaşımında, STBA bağlamında yapılan *tradeoff*, açıklanan varyans kaybına karşı kazanılan seyrekliği ifade etmektedir. TPO, j .inci seyrek temel bileşenin sıfır yüklerinin sayısı ile açıklanan varyansının çarpımını maksimize eder. Buradan TPO yaklaşımına göre k tane TB için λ_j değerleri,

$$\lambda_j^{TPO} = V((STB)_j) \alpha_j, \quad j = 1, 2, \dots, k \quad (3.2.5.1)$$

eşitliği ile elde edilir. Burada α_j , j . STB için sıfır yüklerin sayısıdır. Croux ve ark. (2013), seyreklik parametresi λ 'yı seçmek için BIC tipi bir kriter kullanmışlardır. $\mathbf{P}_k^s$ ve $\mathbf{P}_k^c$ sırasıyla ilk k TB'yi içeren seyrek ve seyrek olmayan temel bileşenlerin yükler matrisleri olsun. Buradan, $\mathbf{R}^s = \mathbf{X} - \mathbf{X}\mathbf{P}_k^s(\mathbf{P}_k^s)'$ ve $\mathbf{R}^c = \mathbf{X} - \mathbf{X}\mathbf{P}_k^c(\mathbf{P}_k^c)'$ artık matrisleri

tanımlanır. Artık matrislerin j .inci sütunları sırasıyla $r_j^s = (r_{1j}^s, r_{2j}^s, \dots, r_{nj}^s)$ ve $r_j^c = (r_{1j}^c, r_{2j}^c, \dots, r_{nj}^c)$ olarak tanımlanır. BIC kriteri,

$$\lambda^{BIC} = \frac{\sum_{j=1}^k V(r_j^s)}{\sum_{j=1}^k V(r_j^c)} + \beta \frac{\log n(n)}{n} \quad (3.2.5.2)$$

eşitliği ile tanımlanır. Burada, β sıfır olmayan yüklerin sayısıdır.

Pratikte λ 'nın seçimi bir grid $[0, \lambda_{max}]$ aralığında λ^{BIC} minimize edilerek yapılır. Burada, λ_{max} , k bileşenli seyrek TBA sonuçlarının tam seyrekliğindeki sonuçlardır. λ 'ya ek olarak, ayrıca k 'nın bileşenlerinin sayısını seçmek gerekir. k 'nın uygun seçimi TBA'da eski ve yaygın bir problemdir. Bu problemin giderilmesi doğrultusunda birçok öneri sunulmuştur. Bu çalışmada k sayısına, dayanıklı varyans açıklama oranı (VAO),

$$VAO_k = \frac{V(T_k)}{V(X)} \quad (3.2.5.3)$$

ile karar verilir. Burada T_k , temel bileşen skorlarını içerir. $V = var$ için, VAO_k , k en büyük özdeğerin toplamının, örneklem kovaryans matrisinin tüm özdeğerlerinin toplamına oranıdır. Bu seyrek metot çalışırken, temel bileşenlerin sayısı korunmuştur. k 'nın bu değeri için, seçilmiş bir λ , daha seyrek bir yükler matrisine ve daha küçük bir dayanıklı varyansa sahip olmalıdır.

4. BULGULAR

Bu bölümde izdüşüm takibine dayanan dayanıklı TBA ve dayanıklı seyrek TBA yöntemlerinin performansları iki gerçek ve üç yapay veri kümesi üzerinden karşılaştırılacaktır. İzdüşüm takip dayanıklı temel bileşenler analizi (İT-DTBA) ve izdüşüm takip dayanıklı seyrek temel bileşenler analizi (İT-DSTBA) yöntemlerini gerçek ve yapay veri kümelerine uygulamak için R programında (version 3.3.2) pcaPP, pls, lars, lasso2, spls, lattice, elasticnet, stats, stats4, robust, robustbase, rrcov, rrcovHD kütüphaneleri yardımıyla yazılan kodlar (Ek-1, Ek-2, Ek-3, Ek-4, Ek-5) kullanılmıştır.

4.1. Yapay ve Gerçek Veri Kümesi Uygulamaları

4.1.1. Yapay Veri Kümesi 1 Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması

Materyal bölümünde bahsedilen Yapay Veri Kümesi 1, yazılan R (version 3.3.2) kodu ile analiz edilmiş olup elde edilen sonuçlar %10, %20, %30'luk aykırı gözlem oranlarına göre sırasıyla Çizelge 4.1, Çizelge 4.2, Çizelge 4.3'te sunulmuştur.

%10 aykırı gözlem oranı durumu için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri TPO kriterine göre $\lambda = 7.8$ olarak seçilmiştir.

Çizelge 4.1 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %88'ini açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk dört temel bileşen ile toplam varyansın %84'ünün açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %4'lük varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.1'de İT-DTBA'da sıfır olmayan yüklerin sayısı 20 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 4'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

Çizelge 4.1. Değişkenlerin %10 aykırı gözlem varlığında, ilk dört seyrek olmayan

dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 7.8$) temel bileşen yükleri

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.59	0.30	-0.59	-0.07	1.00	0.00	0.00	0.00
X2	0.18	-0.09	0.65	0.24	0.00	0.00	0.00	1.00
X3	0.74	0.03	0.30	0.22	0.00	0.00	0.00	0.00
X4	-0.14	0.92	0.31	-0.21	0.00	0.00	1.00	0.00
X5	-0.21	0.25	-0.22	0.92	0.00	1.00	0.00	0.00
VAO %	0.25	0.24	0.23	0.16	0.25	0.21	0.19	0.18
Birikimli VAO %	0.25	0.48	0.72	0.88	0.25	0.46	0.66	0.84

Çizelge 4.2. Değişkenlerin %20 aykırı gözlem varlığında, ilk dört seyrek olmayan

dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 12.14$) temel bileşen yükleri

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.04	0.73	-0.67	0.01	0.00	1.00	0.00	0.00
X2	-0.03	0.57	0.71	0.00	0.00	0.00	1.00	0.00
X3	0.68	-0.27	-0.12	-0.29	0.00	0.00	0.00	1.00
X4	-0.28	-0.24	-0.14	0.74	1.00	0.00	0.00	0.00
X5	0.67	0.15	0.12	0.61	0.00	0.00	0.00	0.00
VAO %	0.31	0.23	0.19	0.15	0.23	0.23	0.21	0.19
Birikimli VAO %	0.31	0.54	0.73	0.88	0.23	0.46	0.67	0.86

%20 aykırı gözlem oranı durumu için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri TPO kriterine göre $\lambda = 12.14$ olarak seçilmiştir.

Çizelge 4.2 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %88'ini açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk dört temel bileşen ile toplam varyansın %86'ünün açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %2'lük varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.2'de İT-DTBA'da sıfır olmayan yüklerin sayısı 19 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 4'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

Çizelge 4.3. Değişkenlerin %30 aykırı gözlem varlığında, ilk dört seyrek olmayan

dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 3.87$) temel bileşen yükler

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	-0.50	0.20	0.14	-0.48	0.00	0.00	0.00	1.00
X2	0.77	-0.07	-0.17	-0.57	0.00	-0.01	0.00	0.00
X3	0.23	0.31	-0.36	0.61	1.00	0.00	0.01	0.00
X4	0.32	0.27	0.89	0.17	-0.01	0.00	1.00	0.00
X5	0.00	0.89	-0.19	-0.20	0.00	1.00	0.00	0.00
VAO %	0.31	0.23	0.22	0.15	0.21	0.20	0.20	0.20
Birikimli VAO %	0.31	0.54	0.76	0.91	0.21	0.42	0.62	0.81

%30 aykırı gözlem oranı durumu için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri TPO kriterine göre $\lambda = 3.87$ olarak seçilmiştir.

Çizelge 4.3 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %91'ini açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk dört temel bileşen ile toplam varyansın %81'inin açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %10'luk varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.3'te İT-DTBA'da sıfır olmayan yüklerin sayısı 19 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 7'ye düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

4.1.2. Yapay Veri Kümesi 2 Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması

Materyal bölümünde bahsedilen Yapay Veri Kümesi 2 yazılan R (version 3.3.2) kodu ile analiz edilmiş olup elde edilen sonuçlar Çizelge 4.4'te sunulmuştur.

Çizelge 4.4. Yapay Veri Kümesi 2'de değişkenlerin ilk dört seyrek olmayan dayanaklı ($\lambda = 0$) ve seyrek dayanaklı ($\lambda = 11.03$) temel bileşen yükleri

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.99	-0.06	0.03	0.11	1.00	0.00	0.00	0.00
X2	0.04	0.26	0.86	-0.18	0.00	0.00	0.00	1.00
X3	0.08	0.79	-0.23	-0.47	0.00	0.00	0.00	0.00
X4	0.07	0.00	0.08	-0.18	0.00	0.00	1.00	0.00
X5	0.07	0.40	-0.36	0.39	0.00	0.00	0.00	0.00
X6	-0.11	0.37	0.27	0.74	0.00	1.00	0.00	0.00
VAO %	0.23	0.18	0.17	0.16	0.18	0.18	0.18	0.16
Birikimli VAO %	0.23	0.41	0.59	0.74	0.18	0.36	0.54	0.71

Yapay Veri Kümesi 2 için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri BIC kriterine göre $\lambda = 11.03$ olarak seçilmiştir.

Çizelge 4.4 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %74'ünü açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk

dört temel bileşen ile toplam varyansın %71'inin açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %3'lük varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.4'te İT-DTBA'da sıfır olmayan yüklerin sayısı 23 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 4'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

4.1.3. Yapay Veri Kümesi 3 ($p > n$ durumu) Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması

Materyal bölümünde bahsedilen Yapay Veri Kümesi 3 yazılan R (version 3.3.2) kodu ile analiz edilmiş olup elde edilen sonuçlar Çizelge 4.5'te sunulmuştur.

Yapay Veri Kümesi 3 için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri BIC kriterine göre $\lambda = 50.69$ olarak seçilmiştir.

Çizelge 4.5 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %87'sini açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk dört temel bileşen ile toplam varyansın %84'ünün açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %3'lük varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.5'te İT-DTBA'da sıfır olmayan yüklerin sayısı 71 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 4'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

Çizelge 4.5. Yapay Veri Kümesi 3’te değişkenlerin ilk dört seyrek olmayan dayanaklı

($\lambda = 0$) ve seyrek dayanaklı ($\lambda = 50.69$) temel bileşen yükleri

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.62	-0.18	0.36	-0.25	0.00	0.00	0.00	1.00
X2	0.35	0.03	-0.10	0.13	0.00	0.00	0.00	0.00
X3	-0.22	0.25	-0.22	-0.10	0.00	0.00	0.00	0.00
X4	-0.04	-0.15	0.00	-0.02	0.00	0.00	0.00	0.00
X5	-0.24	-0.14	-0.13	0.33	1.00	0.00	0.00	0.00
X6	0.00	0.00	0.04	0.04	0.00	0.00	0.00	0.00
X7	0.00	-0.14	-0.07	0.00	0.00	0.00	0.00	0.00
X8	0.13	0.03	-0.22	-0.07	0.00	0.00	0.00	0.00
X9	0.02	0.50	0.23	0.13	0.00	0.00	0.00	0.00
X10	0.00	-0.13	-0.41	-0.50	0.00	1.00	0.00	0.00
X11	-0.20	0.16	-0.11	-0.08	0.00	0.00	0.00	0.00
X12	-0.08	0.13	0.05	-0.03	0.00	0.00	0.00	0.00
X13	0.11	-0.24	-0.12	0.05	0.00	0.00	0.00	0.00
X14	0.00	0.50	0.07	0.00	0.00	0.00	0.00	0.00
X15	-0.03	-0.21	-0.28	0.55	0.00	0.00	1.00	0.00
X16	0.01	-0.08	-0.07	0.01	0.00	0.00	0.00	0.00
X17	0.00	-0.22	0.14	0.03	0.00	0.00	0.00	0.00
X18	-0.08	0.06	-0.34	-0.41	0.00	0.00	0.00	0.00
X19	0.17	0.25	-0.10	0.07	0.00	0.00	0.00	0.00
X20	-0.53	-0.23	0.51	-0.22	0.00	0.00	0.00	0.00
VAO %	0.28	0.24	0.21	0.14	0.27	0.23	0.19	0.16
Birikimli VAO %	0.28	0.53	0.73	0.87	0.27	0.50	0.68	0.84

4.1.4. Gerçek Veri 1-Borsa Verisi Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması

Materyal bölümünde bahsedilen Gerçek Veri 1-Borsa Verisi yazılan R (version 3.3.2) kodu ile analiz edilmiş olup elde edilen sonuçlar Çizelge 4.6'da sunulmuştur.

Çizelge 4.6. Gerçek Veri 1-Borsa Verisinde değişkenlerin ilk dört seyrek olmayan dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 3.31$) temel bileşen yükleri

Değişkenler	İT-DTBA				İT-DSTBA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.36	-0.12	-0.48	0.19	0.00	0.00	0.00	1.00
X2	-0.02	0.56	-0.22	-0.26	0.79	-0.01	0.00	0.00
X3	0.29	0.39	0.19	-0.45	0.00	0.00	0.17	0.01
X4	0.22	0.00	-0.34	0.18	0.00	0.00	0.00	0.00
X5	0.15	0.54	0.11	0.02	0.00	0.00	0.99	0.00
X6	0.59	0.00	-0.35	0.01	0.01	0.10	0.00	0.00
X7	0.46	0.09	0.61	0.55	0.00	1.00	0.00	0.00
X8	-0.39	0.47	-0.26	0.61	0.62	0.01	0.00	0.00
VAO %	0.34	0.24	0.15	0.15	0.28	0.19	0.19	0.10
Birikimli VAO %	0.34	0.58	0.73	0.87	0.28	0.47	0.66	0.76

Gerçek Veri 1-Borsa Verisi için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri TPO kriterine göre $\lambda = 3.31$ olarak seçilmiştir.

Çizelge 4.6 incelendiğinde İT-DTBA'nın ilk dört temel bileşenle toplam varyansın %87'sini açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk dört temel bileşen ile toplam varyansın %76'sının açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %11'lik varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.6'da İT-DTBA'da sıfır olmayan yüklerin sayısı 30 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 11'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

4.1.5. Gerçek Veri 2-Kibler Verisi Üzerinden İT-DTBA ve İT-DSTBA'nın Karşılaştırılması

Materyal bölümünde bahsedilen Gerçek Veri 2-Kibler Verisi yazılan R (version 3.3.2) kodu ile analiz edilmiş olup elde edilen sonuçlar Çizelge 4.7'de sunulmuştur.

Gerçek Veri 2-Kibler Verisi için TPO ve BIC kriterleri kullanılarak λ 'lar belirlenmiş, en optimal λ değeri TPO kriterine göre $\lambda = 5.5$ olarak seçilmiştir.

Çizelge 4.7 incelendiğinde İT-DTBA'nın ilk üç temel bileşenle toplam varyansın %90'ını açıkladığı görülmektedir. İT-DSTBA sonuçları incelendiğinde ilk üç temel bileşen ile toplam varyansın %81'inin açıklandığı görülmektedir. Yani bu durumda seyreklik özelliğini kullanmanın maliyeti %9'luk varyans açıklama oranı kaybına eşdeğerdir. Fakat Çizelge 4.7'de İT-DTBA'da sıfır olmayan yüklerin sayısı 38 iken İT-DSTBA'da sıfır olmayan yüklerin sayısının 8'e düştüğü görülmektedir. Yükler matrisine eklenen seyreklik özelliğinin temel bileşenlerin yorumlanmasını son derece iyi bir performansla ve düşük bir hata oranıyla kolaylaştırdığı görülmektedir.

Çizelge 4.7. Gerçek Veri 2-Kibler Verisinde değişkenlerin ilk üç seyrek olmayan

dayanıklı ($\lambda = 0$) ve seyrek dayanıklı ($\lambda = 5.5$) temel bileşen yükleri

Değişkenler	İT-DTBA			İT-DSTBA		
	PC1	PC2	PC3	PC1	PC2	PC3
X1	-0.12	0.05	0.05	0.00	0.00	0.00
X2	0.27	0.09	-0.12	0.00	0.00	0.00
X3	0.22	0.04	-0.12	0.00	0.00	0.00
X4	0.36	-0.11	0.10	0.04	0.00	0.00
X5	0.00	0.64	-0.35	0.00	0.00	0.00
X6	0.30	-0.18	-0.12	0.00	-0.02	0.00
X7	0.45	0.33	0.23	1.00	0.00	0.00
X8	0.00	-0.45	-0.20	0.01	0.00	0.00
X9	0.00	0.18	0.76	0.00	0.00	0.02
X10	0.33	0.04	0.17	0.00	0.00	0.00
X11	0.00	-0.31	0.30	0.00	0.00	0.00
X12	-0.32	-0.04	0.10	0.00	0.00	1.00
X13	-0.33	-0.01	0.14	0.00	0.99	0.00
X14	0.35	-0.30	-0.04	0.00	-0.13	0.00
VAO %	0.65	0.13	0.12	0.31	0.29	0.22
Birikimli VAO %	0.65	0.77	0.90	0.31	0.59	0.81

5. TARTIŞMA VE SONUÇ

Seyrek temel bileşenler analizi, açıklanan varyansı maksimum yapan yönler doğrultusundaki vektörleri bulmak ve temel bileşenlerin yorumlanabilirliğini artırmak anlamında çok kullanışlıdır. Dayanıklı seyrek temel bileşenler analizi de (DSTBA) yine yorumlanabilirliği kolaylaştıran ve aykırı gözlemlerden etkilenmeyen temel bileşenler analizi vektörlerinin belirlenmesini sağlar. Ele alınan izdüşüm takip yaklaşımı temel bileşenler analizi vektörlerini bulmak için dayanıklı bir varyansı maksimum yapmaktadır. İzdüşüm takip yaklaşımının avantajlarından birisi temel bileşenlerin sırayla hesaplanmasıdır ve istenilen temel bileşen sayısına ulaşıldığında algoritma durur. Bu özellik yüksek boyutlu veri kümelerinde değişken sayısının gözlem sayısından daha büyük ($p > n$) olduğu durumlarda oldukça kullanışlıdır.

Hem yorumlanabilirlik hem de varyans açıklama oranı bakımından optimal sonucu elde etmek için kullanılan seyreklik parametresi λ 'nın optimal değeri BIC ve TPO yaklaşımlarıyla bulunabilir.

Bu çalışmada, çok değişkenli bir veri kümesinde aykırı gözlemlerin varlığında kullanılan dayanıklı TBA ve TBA sonuçlarını daha net ve anlaşılır bir formata dönüştüren seyrek TBA yöntemlerinin güçlerini birleştiren Croux ve ark. (2013)'nın önerdiği dayanıklı seyrek TBA yönteminin performansı farklı aykırı gözlem oranlarına sahip iki gerçek ve üç yapay veri kümesi üzerinden incelenmiştir. Yapay Veri Kümesi 3, değişken sayısı (p) > gözlem sayısı (n) durumunu içermektedir. Bu nedenle böyle bir veri kümesinin İT-DSTBA ile çözümünde elde edilecek bulgular oldukça önemlidir. İzdüşüm takip yöntemi de bu $p > n$ durumunda kullanımı oldukça yaygın bir dayanıklı yöntemdir.

Bu çalışmada, λ 'nın belirlenmesinde kullanılan BIC ve TPO kriterlerinin birbirlerine göre fazla bir üstünlük göstermediği görülmüştür. Fakat her bir uygulamada, λ değerini veren λ belirleme kriterlerinin (BIC ve TPO) en optimali seçilmiştir.

Gelecekteki çalışmada, bu çalışmada ele alınan izdüşüm takibine dayanan DSTBA, farklı dayanıklı varyans tahmin edicilerine dayanan DSTBA sonuçları ile karşılaştırılabilir. Küçük örneklem ve büyük örneklem durumlarında farklı aykırı gözlem oranları için yöntemlerin performansı değerlendirilebilir. İzdüşüm takip

yaklaşımı yerine EKKD gibi farklı yaklaşımların kullanımlarının da DSTBA üzerine etkileri incelenebilir.



KAYNAKLAR

- Bien, J., Xu, Y., Mahoney, M. W. 2010. CUR from a sparse optimization viewpoint. In *Advances in Neural Information Processing Systems* (pp. 217-225).
- Cadima, J., Jolliffe, I. T. 1995. Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2): 203-214.
- Croux, C., Haesbroeck, G. 2000. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3): 603-618.
- Croux, C., Ruiz-Gazen, A. 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1): 206-226.
- Croux, C., Filzmoser, P., Oliveira, M. R. 2007. Algorithms for Projection–Pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2): 218-225.
- Croux, C., Filzmoser, P., Fritz, H. 2013. Robust sparse principal component analysis. *Technometrics*, 55(2): 202-214.
- D'Aspremont, A., El Ghaoui, L., Jordan, M. I., Lanckriet, G. R. 2007. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3): 434-448.
- Filzmoser, P., Fritz, H., Kalcher, K. 2010. pcaPP: Robust PCA by Projection Pursuit. R package version 1.9-0.
- Friedman, J. H., Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9): 881-890.
- Guo, J., James, G., Levina, E., Michailidis, G., Zhu, J. 2010. Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4): 930-946.
- He, Y., Zhou, W., Qian, G., Chen, B. 2014. Methane storage in metal–organic frameworks. *Chemical Society Reviews*, 43(16): 5657-5678.
- Huber, P. J. 1985. Projection pursuit. *The annals of Statistics*, 435-475.
- Hubert, M., Rousseeuw, P. J., Verboven, S. 2002. A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1): 101-111.
- Hubert, M., Engelen, S. 2004. Robust PCA and classification in biosciences. *Bioinformatics*, 20(11): 1728-1736.
- Hubert, M., Rousseeuw, P. J., Vanden Branden, K. 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1): 64-79.
- Hubert, M., Rousseeuw, P. J., Van Aelst, S. 2008. High-breakdown robust multivariate methods. *Statistical Science*, 92-119.

Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T. 2016. Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58(4): 424-434.

Jenatton, R., Obozinski, G., Bach, F. 2010, March. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 366-373).

Jolliffe, I. T. 1995. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1): 29-35.

Jolliffe, I. T., Trendafilov, N. T., Uddin, M. 2003. A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics*, 12(3): 531-547.

Journée M., Nesterov Y., Richtárik P., Sepulchre R. 2010 Generalized power method for sparse principal component analysis. *J Mach Learn Res* 11: 517–553.

Kibler, D., Aha, D. W., Albert, M. K. 1989. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5(2): 51-57.

Li, G., Chen, Z. 1985. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80(391): 759-766.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., ..., Fan, J. 1999. Robust principal component analysis for functional data. *Test*, 8(1): 1-73.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American statistical association*, 79(388): 871-880.

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8: 283-297.

Rousseeuw, P. J., Croux, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424): 1273-1283.

Rousseeuw, P. J., Driessen, K. V. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3): 212-223.

Sigg, C. D., Buhmann, J. M. 2008, July. Expectation-maximization for sparse and non-negative PCA. In *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 960-967.

Wang, H., Li, B., Leng, C. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 671-683.

Witten, D. M., Tibshirani, R., Hastie, T. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, kxp008.

Zou, H., Hastie, T., Tibshirani, R. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2): 265-286.

EKLER

Ek-1

Yapay Veri Kümesi 1 için R programında (version 3.3.2) yazılan kodlar:

%10 aykırı gözlem durumu:

```
x<-read.table("C:/%10.txt")

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, qn)

k<-3

maxiter <- 25

method <- "qn"

oTPO <- opt.TPO (xqn, k = k, method = method, maxiter = maxiter)

par(mfrow=c(2,2))

for (i in 1:k) plot (oTPO, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=5)

srpca <- SPcaGrid(xqn, lambda=7.8, k=5)

tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)
```

```

cvars1 <- cumsum(vars1)

cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)

%20 aykırı gözlem durumu:

x<-read.table("C:/%20.txt")

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, qn)

k<-3

maxiter <- 25

method <- "Qn"

oTPO <- opt.TPO (xqn, k = k, method = method, maxiter = maxiter)

for (i in 1:k) plot (oTPO, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=5)

srpca <- SPcaGrid(xqn, lambda=12.14, k=5)

tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

```

```

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)

cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)

%30 aykırı gözlem durumu:

x<-read.table("C:/%30.txt")

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, qn)

k<-3

maxiter <- 25

method <- "Qn"

oTPO <- opt.TPO (xqn, k = k, method = method, maxiter = maxiter)

par(mfrow=c(2,2))

for (i in 1:k) plot (oTPO, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=5)

srpca <- SPcaGrid(xqn, lambda=3.87, k=5)

```

```
tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)

cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)
```

Ek-2

Yapay Veri Kümesi 2 için R programında yazılan kodlar:

```
library(mvtnorm)

x <- rbind(rmvnorm(200, rep(0, 6), diag(c(5, rep(1,5))))),
rmvnorm( 15, c(0, rep(20, 5)), diag(rep(1, 6))))

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, qn)

k<-3

maxiter <- 25

method <- "qn"

oBIC <- opt.BIC (xqn, k = k, method = method, maxiter = maxiter)

for (i in 1:k) plot (oBIC, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=6)

srpca <- SPcaGrid(xqn, lambda=11.03, k=6)

tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)
```

```
cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)
```



Ek-3

Yapay Veri Kümesi 3 için R programında yazılan kodlar:

```
library(mvtnorm)

x <- rbind(rmvnorm(15, rep(0, 20), diag(c(5, rep(1,19)))),
           rmvnorm( 3, c(0, rep(20, 19)), diag(rep(1, 20))))

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, Qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, Qn)

k<-4

maxiter <- 25

method <- "qn"

oBIC <- opt.BIC (x, k = k, method = method, maxiter = maxiter)

for (i in 1:k) plot (oBIC, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=5)

srpca <- SPcaGrid(xqn, lambda=50.69, k=5)

tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)
```

```
cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)
```



Ek-4

Gerçek Veri 1-Borsa Verisi için R programında yazılan kodlar:

```
x=read.table("C:/BORSAVERİ.txt")

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, Qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, Qn)

k<-4

maxiter <- 25

method <- "qn"

oTPO <- opt.TPO (x, k = k, method = method, maxiter = maxiter)

usr<-par(mfrow=c(1,2))

for (i in 1:k) plot (oTPO, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=8)

srpca <- SPcaGrid(xqn, lambda=3.31, k=8)

tab <- cbind(round(getLoadings(rpca)[,1:5], 2), round(getLoadings(srpca)[,1:5], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)

cvars2 <- cumsum(vars2)
```

```
ev <- round(c(vars1[1:4], vars2[1:4]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)
```



Ek-5

Gerçek Veri 2-Kibler Verisi için R programında yazılan kodlar:

```
data(kibler)

x<-kibler

x.sd <- apply(x,2,sd)

xsd <- sweep(x, 2, x.sd, "/", check.margin = FALSE)

apply(xsd, 2, sd)

x.qn <- apply(x, 2, Qn)

xqn <- sweep(x, 2, x.qn, "/", check.margin = FALSE)

apply(xqn, 2, Qn)

k <- 4

maxiter <- 25

method <- "qn"

oTPO <- opt.TPO (x, k = k, method = method, maxiter = maxiter)

par (mfrow = c (2, 2))

for (i in 1:k) plot (oTPO, k = i, f.x = "lambda")

rpca <- SPcaGrid(xqn, lambda=0, k=4)

srpca <- SPcaGrid(xqn, lambda=5.5, k=4)

tab <- cbind(round(getLoadings(rpca)[,1:3], 2), round(getLoadings(srpca)[,1:3], 2))

vars1 <- getEigenvalues(rpca); vars1 <- vars1/sum(vars1)

vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)

cvars1 <- cumsum(vars1)
```

```
cvars2 <- cumsum(vars2)

ev <- round(c(vars1[1:3], vars2[1:3]),2)

cev <- round(c(cvars1[1:4], cvars2[1:4]),2)

rbind(tab, ev, cev)
```



ÖZGEÇMİŞ

Işıl ÜNALDI, 1987 yılında Sinop'ta doğdu. İlk, orta ve lise öğrenimini Sinop'ta tamamladı. 2013 yılında Hacettepe Üniversitesi İstatistik Bölümünden mezun oldu. 2015 yılında Sinop Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalında yüksek lisans öğrenimine başladı ve halen yüksek lisans öğrenimine devam etmektedir.

