Available online at www.alphanumericjournal.com

alphanumeric journal
The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 4, Issue 2, 2016

Received : May 16, 2016
Accepted : August 25, 2016
Published Online : September 29, 2016

AJ ID: 2016.04.02.STAT.02
DOI : 10.17093/aj.2016.4.2.5000189525

Robust Principal Component Analysis Based On Modified Minimum Covariance Determinant In The Presence Of Outliers

B. Barış Alkan | Department of Statistics, Sinop University, Turkey, bbalkan@sinop.edu.tr

ABSTRACT

Principal component analysis (PCA) is not resistant to outliers existing in multivariate data sets. The results which are obtained by using classical PCA are far from real values in the presence of outliers. Therefore, using robust versions of PCA is favorable. The easiest way to obtain robust principal components is to replace classical estimates of the location and scale parameters with their robust versions. Robust estimations of location and scale parameters can be found with minimum covariance determinant (MCD) providing high breakdown point. In this study, algorithm of MCD is modified using Jackknife resampling approach and results of this modification are examined. Proposed robust principal component analysis (RPCA) based on modified MCD (MMCD) method that is modified using Jaccknife resampling are evaluated over two real data with different outlier ratios. In the light of obtained results, it can be said that RPCA based on MMCD is better than RPCA based on MCD in the presence of outliers.

Keywords: Minimum covariance determinant, Robust principal component analysis, Outliers

Aykırı Gözlemlerin Varlığında Uyarlanmış En Küçük Kovaryans Determinant Tahminine Dayalı Dayanıklı Temel Bileşenler Analizi

ÖZET

Klasik temel bileşenler analizi (KTBA), çok değişkenli veri kümelerinde yer alabilen aykırı gözlemlere karşı dayanıklı değildir. Aykırı gözlemlerin varlığında KTBA kullanılarak elde edilen sonuçlar gerçekte olması gerekenden oldukça farklı çıkabilir. Bu yüzden, aykırı gözlemlerin varlığında PCA'nın dayanıklı versiyonlarının kullanımı tercih edilmelidir. Dayanıklı temel bileşenleri elde etmek için en kolay yol konum ve ölçek parametrelerinin klasik tahminleriyle, onların dayanıklı tahminlerinin yer değiştirilmesidir. Çok değişkenli veri kümesi için konum ve ölçek parametrelerinin dayanıklı tahmini, yüksek bozulma noktası sağlayan en küçük kovaryans determinant (EKKD) yöntemi ile yapılabilir. Bu çalışmada, EKKD yöntemi, jackknife yeniden örnekleme yaklaşımı kullanılarak uyarlanıp, bu uyarlamadan kaynaklanan değişimlerin dayanıklı temel bileşenler analizi (DTBA) üzerindeki etkilerin incelenmesi amaçlanmaktadır. Jackknife yeniden örnekleme yöntemine dayanan EKKD'nin aykırı gözlem oranındaki değişmelerden nasıl etkilendiği iki gerçek veri kümesi için değerlendirilmektedir. Elde edilen bulgular ışığında, önerilen uyarlanmış en küçük kovaryans determinant (UEKKD) tahminine dayalı DTBA, klasik EKKD'ye dayanan DTBA'ya göre veri kümesinde aykırı gözlemlerin varlığında daha iyi sonuçlar verdiği görülmektedir

Anahtar Kelimeler: En küçük kovaryans determinant, Dayanıklı temel bileşenler analizi, Aykırı gözlemler

1. Giriş

Temel bileşenler analizi (TBA), veri kümesi yüksek boyutlu olduğunda genellikle ilk başvurulan boyut indirgeme yöntemidir. Fakat, klasik TBA (KTBA) yöntemi de diğer klasik istatistik yöntemler gibi veri kümesinde aykırı gözlemlerin varlığından, hatta bazen tek bir aykırı gözlemden bile, negatif olarak etkilenmektedir. Aykırı gözlemlerin çoğu %97.5 lik tolerans elipsoid'i içinde olduğu için klasik kovaryans tahmini üzerinde etkilidir (Filzmoser &Todorov, 2011). KTBA'da örneklem kovaryans (veya korelasyon) matrisi ve ortalama vektörü temel aldığından, veri kümesinde aykırı gözlemlerin varlığında güvenli ve tutarlı sonuç vermemektedir. Bu nedenle, aykırı gözlemlerin varlığında temel bileşenler analizinin dayanıklı versiyonlarının kullanılmasının gereği literatürde birçok araştırmacı tarafından vurgulanmaktadır. 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 ile çok değişkenli konum ve ölçek dayanıklı parametre tahmini mümkün (değişken sayısı yeterince küçük olduğunda) olduğu sürece iyi çalışır. DTBA elde etmek için farklı bir yaklaşım ise Croux ve Ruiz-Gazen (2005) tarafından ortaya atılmıştır. Bu yaklaşım izdüşüm takibini (projection pursuit) temel alan dayanıklı TBA olarak ifade edilmektedir. Değişken sayısının gözlem sayısından fazla olduğu durumlarda ve yüksek boyutlu veri kümelerinin analizinde dayanıklı kovaryans matrisi tahminini bulmak imkansız olduğundan, Croux ve Ruiz-Gazen (2005) tarafından önerilen yaklaşım uygundur. Dayanıklı TBA için diğer önerilere bakıldığında, Locantore v.d. (1999) tarafında geliştirilen küresel TBA ve Maronna (2005) tarafından geliştirilen dik TBA yaklaşımları ile karşılaşılmaktadır (Farcomeni & Greco, 2015). Ayrıca Alkan v.d. (2015) çalışmalarında, veri kümesinde yer alan aykırı gözlemlerin hangi değişkeninde aldığı değerden dolayı aykırı gözlem olarak belirlendiğini tespit edip, o değeri kayıp sayarak kayıp değer atama yöntemleri ile onun yerine atama yapılmasının dayanıklı TBA analizine bir alternatif olup olamayacağını incelemişlerdir.

Dayanıklı temel bileşenleri elde etmek için en kolay yol konum ve ölçek parametrelerinin klasik tahminleriyle, onların dayanıklı tahminlerinin yer değiştirilmesidir. Devlin v.d. (1981) ve Campbell (1980) çalışmalarını ilk başta konum ve ölçek parametrelerinin M tahmin edicilerini bu yönde kullanmışlardır. Ancak, M tahmin edicilerinin yüksek boyutlarda düşük bozulma noktasına sahip olmaları, bu tahmin edicilerinin kullanımını gündemden düşürmüştür. Daha sonra, yüksek boyutlu veri kümelerinde yüksek bozulma noktası avantajı sunan en küçük kovaryans determinant (EKKD) tahmin edicisi Todorov v.d. (1994) tarafından kullanılmıştır. Croux ve Haesbroeck (2000), EKKD tahmin edicisinin etki fonksiyonu ve etkinlik özelliklerini incelemişlerdir. Aykırı gözlemlere karşı yüksek bozulma noktası değeri ile dayanıklılık gösteren EKKD yöntemi bu özelliğinin yanı sıra, veri kümesinde yer alan aykırı gözlemlerin belirlenmesi için de oldukça kullanışlıdır. Uygulamada tıp, mühendislik, finans, kemometri gibi birçok alanda yaygın bir şekilde kullanılmaktadır (Hubert & Debruyne, 2010). Rousseeuw (1984) tarafından ortaya atılan en küçük kovaryans determinant (EKKD) yöntemi çok değişkenli konum ve ölçek parametrelerinin oldukça sağlam bir tahmin edicisidir ve %50 lik bir aykırı gözlem oranına kadar dayanıklı sonuçlar verebilmektedir. Çok değişkenli veri kümesi için konum ve ölçek parametrelerinin tahmini, yüksek bozulma noktası sağlayan EKKD

yöntemi ile yapılabilir. TBA'nin dayanıklı versiyonu, μ ve Σ parametrelerinin, (μ) ̂ve Σ̂ dayanıklı tahminleriyle yer değiştirilmesi ile elde edilebilir (Todorov & Filzmoser, 2009).

Bu çalışmada EKKD yöntemi, jackknife yeniden örnekleme yaklaşımı kullanılarak uyarlanıp, bu uyarlamadan kaynaklanan değişimlerin incelenmesi amaçlanmaktadır. Jackknife yeniden örnekleme yöntemine dayanan EKKD'nin aykırı gözlem oranındaki değişmelerden nasıl etkilendiği farklı aykırı gözlem oranına sahip iki gerçek veri kümesi üzerinden değerlendirilecektir.

Çalışmanın ikinci bölümünde, KTBA hakkında temel teorik kavramlardan bahsedilmiştir. Daha sonraki bölümlerde sırasıyla DTBA ve uyarlanmış en küçük kovaryans determinant (UEKKD)'a dayanan dayanıklı TBA için temel kavramlar ve gerekli matematiksel teori verilmiştir. Çalışmanın gerçek veri uygulamaları bölümünde ise ekonomi ve gıda alanından veri kümeleri üzerinden önceki bölümlerde ayrıntılı olarak incelenen yöntemlerin uygulamalarına yer verilmektedir. Son bölümde ise, çalışmadan elde edilen sonuçlar tartışılmaktadır.

2. Klasik Temel Bileşenler Analizi

Temel bileşenler analizi (TBA), veri kümesini daha iyi özetlemeyi ve yorumlamayı sağlayan, p-tane orjinal değişkenin indirgenmiş boyutlu bir uzayda k-tane lineer birleşimlerini bularak boyut indirgemeyi amaçlayan çok değişkenli bir yöntemdir. Temel bileşenler, bu k-tane lineer birleşim üzerine izdüşürülen verinin varyansını maksimum yapan yönler doğrultusundaki vektörlere karşılık gelir (Croux, Filzmoser & Fritz, 2013).

$\mathbf{X}$, $n \times p$ boyutlu bir veri matrisi, $\bar{\mathbf{x}}$, veri kümesinin ortalama vektörü ve $\mathbf{1}$, elemanları 1 olan $n \times 1$ tipinde bir sütun vektörünü göstersin. $\mathbf{t}_j$, ortalamaya göre merkezileştirilmiş verinin bir $\mathbf{u_j}$ vektörü doğrultusu üzerine izdüşümünden oluşan lineer birleşimleri olmak üzere, $\mathbf{t}_j$ ve $\mathbf{u_j}$,

$$\mathbf{t}_j = \left(\mathbf{X} - 1\bar{\mathbf{x}}^T\right)\mathbf{u}_j \tag{1}$$

$$\mathbf{u}_j = \arg\max_{\mathbf{u}} \ \mathrm{Var}\left(\mathbf{Xu}\right) \tag{2}$$

eşitlikleri yardımıyla elde edilir. Burada kısıtlar, $\|\mathbf{u}_j\| = 1$ ve $\mathrm{Cov}(\mathbf{Xu}_j, \mathbf{Xu}_i) = 0$, i < j ve $j = 1, \ldots, k,\ k \leq \min(n, p)$ olarak verilir. Bu maksimazyon probleminin çözümü için Lagrange problem çözüm yöntemi kullanılır. Sonuç olarak, $\mathbf{X}$'in kovaryans matrisinin özdeğerleri $\mathbf{Xu}_j$'nin varyanslarına eşittir. Yani, $\lambda_j = \mathrm{Var}(\mathbf{Xu}_j)$ olarak ifade edilir. Burada, $\mathbf{t}_j$ vektörleri, $n \times k$ boyutlu $\mathbf{T}$ *skor matrisinin* sütunlarını oluştururken, $\mathbf{u_j}$ vektörleri ise $\mathbf{U}$ *yükler matrisinin* sütunlarını oluşturur (Filzmoser & Todorov, 2011).

k'nın uygun değerini belirlemek için bir çok kriter vardır. Bu kriterlerden en yaygın olarak kullanılan, ilk k boyut tarafından açıklanan toplam varyansa dayalı olan, $\left(\sum_{j=1}^{k} \lambda_j\right)/\left(\sum_{j=1}^{p} \lambda_j\right) \geq \%80$ olması kriteridir (Hubert & Engelen 2004; Johnson & Wischern 1998). Orjinal $\mathbf{X}$ matrisi, verinin temel yapısını koruyarak orjinal koordinat sistemindeki (k temel bileşeni kullanarak) $\mathbf{T}$ skorlarından yeniden oluşturulabilir:

$$\mathbf{X} = 1\bar{\mathbf{x}}^T + \mathbf{T}\mathbf{U}^T + \mathbf{E} \tag{3}$$

3. Dayanıklı Kovaryans Matrisine Dayanan Temel Bileşenler Analizi

TBA, örneklem kovaryans matrisine dayandığı için, veri kümesine yer alan ve verinin genel yapısından oldukça farlı hareket eden gözlemler olduğunda tamamen yanlı ve güvenilmez sonuçların elde edilmesine sebep olabilir. Hatta tek bir aykırı gözlem bile tüm bu süreci bozabilir. Bozulma durumunda, en büyük varyans açıklama oranına sahip birinci temel bileşen aykırı gözlemlere doğru yön değiştirir. Bu durum gerçekte varolduğundan daha şişmiş bir değişkenliğe neden olabilir. Yani, aşırı iyimser özdeğerler ve bunlara bağlı olarak da gerçekte varolmayacak kadar yüksek toplam varyans açıklama oranlarının elde edilmesine yol açar. TBA için dayanıklı yöntemlerin kullanılmasıyla bu problemlerin büyük ölçüde üstesinden gelinebilir (Farcomeni & Greco, 2015).

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

EKKD tahmin edicisini tanımlamak için tüm veri kümesinin (n gözlemden oluşan) h boyutlu altkümeleri düşünülür. Yani n'in h'lı kombinasyonu kadar alt kümeyle ilgileniriz. 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-altkümesini bulmaya çalışır. EKKD konum parametresi tahmini $\widehat{\boldsymbol{\mu}}_{\boldsymbol{EKKD}}$, optimal h-altkümenin ortalaması ve EKKD ölçek parametresi tahmini $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{EKKD}}$, ise onun kovaryans matrisi ile verilir. EKKD tahmin edicisi $(n-h)$ tane aykırı gözleme dayanabilir. Daha genel olarak, EKKD tahmin edcisi $(n-h+1)/n$ bozulma noktası değerine sahiptir. h değerinin varsayılan değeri yaklaşık olarak $[0.75n]$ olarak alınmaktadır (Hubert & Engelen 2004).

4. Uyarlanmış En Küçük Kovaryans Determinant (UEKKD)'A Dayanan Dayanıklı TBA

Jackknife yeniden örnekleme yöntemi, aynı anda her seferinde örneklemden bir gözlemi sırayla atarak her biri $(n-1)$ büyüklüğünde olan n tane örneklem üretmektedir. Jackknife yönteminin aykırı gözlemlerin belirlenmesinde de kullanışlı olduğunu Riu ve Bro (2003) çalışmalarında göstermişlerdir. EKKD yöntemi Jackknife yeniden örnekleme yaklaşımına göre uyarlanarak elde edilen UEKKD algoritması aşağıda verilmiştir.

UEKKD Algoritması

- *Adım1. Veri kümesinde yer alan i. gözlemi dışarda bırak.*
- *Adım2. (n-1) gözlem için* $h = [0.75\,(n-1)]$ *değerini bul.*
- *Adım3. Kombinasyon (n-1,h) değerini bul.*
- *Adım3.1. Her bir h örneğe sahip alt kümeler için,*
- *Adım 3.1.1. Örneklem kovaryans matrisini hesapla*
- *Adım 3.1.2. Örneklem kovaryans matrisi determinantını hesapla*

- *Adım 4. En küçük determinanta sahip olan alt kümeyi seç. Bu alt kümenin Örneklem ortalama vektörü ve örneklem kovaryans matrisini bul.*
- *Adım 5. Adım 1-4, i=1,2…,n için tekrarlanır ve buradan elde edilen Örneklem ortalama vektörlerinin ortalamasından ve örneklem kovaryans matrislerinin ortalamasından sırasıyla,* $\hat{\boldsymbol{\mu}}_{\boldsymbol{UEKKD}}$ *ve* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{UEKKD}}$ *çok değişkenli konum ve ölçek tahminleri elde edilir.*

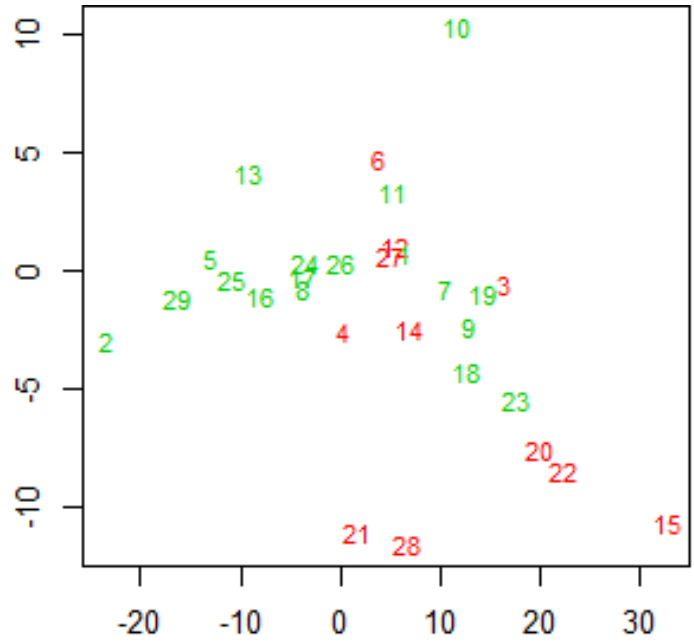
5. Gerçek veri uygulamaları

Bu çalışmada, KTBA, EKKD'ye dayanan DTBA (DTBA_EKKD) ve UEKKD'ye dayanan TBA (DTBA_UEKKD)'nın karşılaştırması için, Dünya Bankası kalkınma göstergeleri veri tabanından derlenen Avrupa ve merkez Asya ülkelerinin temel makroekonomik göstergeleri 2008 veri kümesi ve Daudin'in süt kompozisyon veri kümesi kullanılmıştır. Klasik TBA ve EKKD'ye dayanan DTBA analizleri için R istatistik yazılımında yer alan robustbase ve rrcov kütüphaneleri kullanılmıştır (R Development Core Team, 2011; Rousseeuw, Croux, Todorov, Ruckstuhl, Salibian-Barrera, Verbeke & Maechler, 2009; Todorov, 2009). Önerdiğimiz UEKKD'ye dayanan TBA için ise R'da yazdığımız fonksiyon kullanılmıştır.

5.1. Avrupa ve merkez Asya ülkelerinin temel makroekonomik göstergeleri 2008 veri kümesi üzerinden Klasik TBA, EKKD'ye dayanan DTBA ve UEKKD'ye dayanan TBA'nın karşılaştırması

Bu uygulamada, veri kümesi elde edilebilir olan ülkelerden derlenen 29 Avrupa ve Merkez Asya (tüm gelir grupları için) ülkeleri için kişi başına düşen gayri safi yurt içi hasıla (GSYİH), doğurganlık hızı (DH), tüketici fiyatları enflasyonu (TÜFE), kentsel nüfus (KN), ölüm oranı (ÖLO), toplam işsizlik oranı (TİO), hane halkı nihai tüketim harcamaları (HHNTH) temel göstergeleri 2008 yılı verileri alınmıştır. 2008 yılının alınmasının sebebi 2008 yılının bir kriz dönemi olması ve dolayısıyla veri kümesinin aykırı gözlem içermesinin muhtemel olmasıdır.

Veri kümesinde yer alan çok değişkenli aykırı gözlemlerin belirlenmesi için düzeltilmiş kartil yöntemi kullanılmıştır. Düzeltilmiş kartil yöntemi, ki-kare dağılımının dağılım fonksiyonu ve karesel dayanıklı uzaklığın empirik dağılımı arasındaki farkı karşılaştırır (Filzmoser, Reimann & Garrett, 2003). Bu yöntemin kullanılmasıyla 11 gözlem (%37) aykırı olarak tespit edilmiştir. Aykırı gözlemler Şekil 1'de kırmızı noktalar olarak görünmektedir



Şekil 1. Düzeltilmiş kartil yöntemine göre belirlenen aykırı gözlemler (kırmızı noktalar), 11 aykırı gözlem (%37)

Avrupa ve merkez Asya ülkelerinin temel makroekonomik göstergeleri 2008 veri kümesine sırasıyla KTBA, DTBA_EKKD ve DTBA_UEKKD yöntemleri ile analiz edilmiş ve elde edilen sonuçlar Tablo 1'de sunulmuştur. Tablo 1 inceleğinde, klasik TBA'nın en önemli ilk üç temel bileşenle toplam varyansın %95.24'ünü açıkladığı görülmektedir. Fakat, veri kümesin 11 aykırı gözlem (toplam gözlem sayısının %37'si) olması ve aykırı gözlemlerin varlığında klasik TBA'nın varyans açıklama oranlarında şişmeler olabileceği ve aykırı gözlemlerin özellikle birinci temel bileşenin yönünü değiştirebileceğinden önceki bölümlerde bahsedilmişti. Bu nedenle klasik TBA ile elde edilen yüksek açıklama oranı bir iyilik ölçütü olarak düşünülemez. DTBA_EKKD yöntemi ile en önemli ilk üç temel bileşenle toplam varyansın %80.34'ünü açıkladığı görülmektedir. Bu yöntem, veri kümesinde aykırı gözlem olması durumunda kullanılacak dayanıklı bir yöntemdir. Klasik ile karşılaştırdığımızda bize daha sağlam bir yaklaşım sunacağı önceki bölümlerde verilen bilgiler doğrultusunda açıktır. Tablo 1'de verilen diğer bir sonuç ise önerdiğimiz DTBA_UEKKD yönteminden elde edilen toplam varyans açıklama oranının %82.21 ile en fazla toplam varyans açıklama oranına sahip olmasıdır. Bu durumda Tablo 1'deki sonuçlar karşılaştırıldığında veri kümesinde aykırı gözlemlerin varlığında KTBA ve DTBA_EKKD yöntemleri yerine DTBA_UEKKD kullanılmasının daha uygun olacağı görülmektedir.

	Kümülatif toplam varyans açıklama oranı		
	TB 1	TB 2	İlk üç TB ile
Klasik TBA	0.7269	0.8634	0.9524
EKKD dayanan DTBA	0.3769	0.6390	0.8034
UEKKD dayanan DTBA	0.4260	0.6616	0.8221

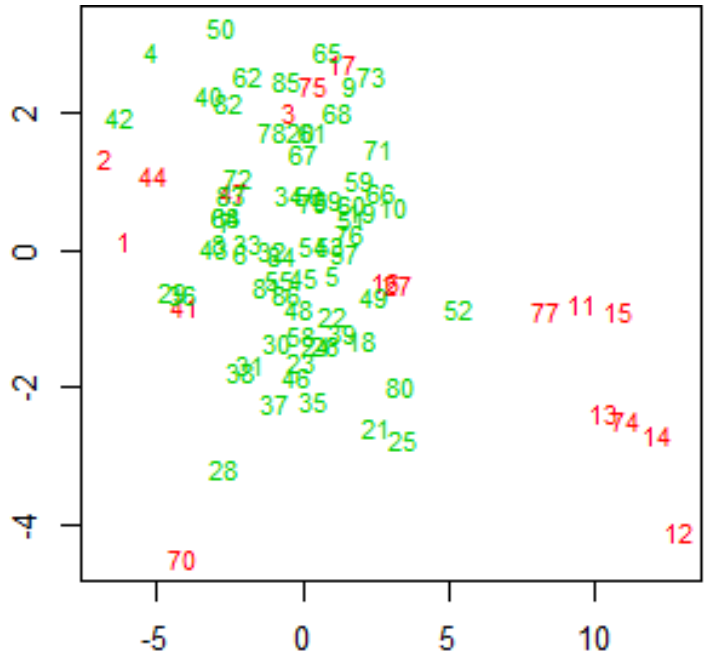
Tablo 1. Avrupa ve merkez Asya ülkelerinin temel makroekonomik göstergeleri 2008 veri kümesi için KTBA, DTBA_EKKD ve DTBA_UEKKD sonuçları

5.2. Daudin'in süt kompozisyon veri kümesi üzerinden Klasik TBA, EKKD'ye dayanan DTBA ve UEKKD'ye dayanan TBA'nın karşılaştırması

Çalışmada, önerilen UEKKD yönteminin işlevselliğinin gösterilmesinde, ikinci gerçek veri uygulaması için Daudin v.d. (1988) tarafından verilen 86 gözlem ve 8 değişken içeren süt kompozisyon veri kümesi seçilmiştir. Todorov v.d. (1994), Atkinson (1994), Rock & Woodruff (1996) gibi bir çok araştırmacı aykırı gözlemlerin belirlenmesi ve dayanıklı istatistiksel çıkarımlar üzerine önerdikleri yaklaşımların geçerliliklerini gösterebilmek için klasik yöntemlerle karşılaştırmalarında örnek veri kümesi olarak Daudin'in süt kompozisyon veri kümesini kullanmışlardır. Bu nedenle çalışmamızda bu veri kümesini uygulama için seçilmiştir.

Veri kümesinde yer alan çok değişkenli aykırı gözlemlerin belirlenmesi için düzeltilmiş kartil yöntemi uygulanmıştır. Bu yöntemin kullanılmasıyla 18 gözlem (toplam gözlem sayısının %20'si) aykırı olarak tespit edilmiştir. Aykırı gözlemler Şekil 2'de kırmızı noktalar olarak görülmektedir.

Daudin'in süt kompozisyon veri kümesi sırasıyla KTBA, DTBA_EKKD ve DTBA_UEKKD yöntemleri ile analiz edilmiş ve elde edilen sonuçlar Tablo 2'de sunulmuştur. Tablo 2 inceleğinde, klasik TBA'nın en önemli ilk üç temel bileşenle toplam varyansın %94.42'sini açıkladığı görülmektedir. Fakat, veri kümesin 18 aykırı gözlem (%20) olması ve aykırı gözlemlerin varlığında klasik TBA'nın varyans açıklama oranlarında şişmeler olabileceği ve aykırı gözlemlerin özellikle birinci temel bileşenin yönünü değiştirebileceğinden klasik TBA ile elde edilen yüksek açıklama oranını bir ölçüt olarak kullanmak mantıklı olmaz. DTBA_EKKD yöntemi ile en önemli ilk üç temel bileşenle toplam varyansın %82.67'sini açıkladığı görülmektedir. Bu yöntem, veri kümesinde aykırı gözlemler olması durumunda kullanılacak dayanıklı bir yöntemdir. Klasik ile karşılaştırdığımızda bize daha sağlam bir yaklaşım sunmaktadır. Tablo 2'de verilen diğer bir sonuç ise önerdiğimiz DTBA_UEKKD yönteminden elde edilen toplam varyans açıklama oranının %84.19 ile en fazla toplam varyans açıklama oranına sahip olmasıdır. Bu durumda Tablo 2'deki sonuçlar karşılaştırıldığında veri kümesinde aykırı gözlemlerin varlığında klasik TBA ve DTBA_EKKD yöntemleri yerine DTBA_UEKKD kullanılmasının daha uygun olacağı görülmektedir. Elde edilen bu sonucun birinci uygulamadaki sonucu da desteklediği görümektedir.



Şekil 2. Düzeltilmiş kartil yöntemine göre belirlenen aykırı gözlemler (kırmızı noktalar), 18 aykırı gözlem (%20)

	Kümülatif toplam varyans açıklama oranı		
	TB 1	TB 2	İlk üç TB ile
Klasik TBA	0.7549	0.8862	0.9442
EKKD dayanan DTBA	0,5306	0.6920	0.8267
UEKKD dayanan DTBA	0.5307	0.7046	0.8419

Tablo 2. Daudin'in süt kompozisyon veri kümesi için KTBA, DTBA_EKKD ve DTBA_UEKKD sonuçları

6. Sonuçlar

Bu çalışmada, gözlem sayısının (n) değişken sayısından (p) büyük olduğu durumlarda kullanılan çok değişkenli istatistik analiz yöntemlerinden klasik TBA ve veri kümesinde aykırı gözlemlerin varlığında EKKD'a dayanan dayanıklı TBA yöntemleri gözden geçirilmiş ve Jackknife yeniden örnekleme yaklaşımı ile EKKD algoritmasının uyarlanmış versiyonu UEKKD yöntemi önerilmiştir.

Önerilen UEKKD'nin aykırı gözlem oranındaki değişmelerden nasıl etkilendiğinin belirlenmesi amacıyla, sırasıyla %20 ve %37'lik aykırı gözlem içeren iki gerçek veri kümesi üzerinden değerlendirme yapılmıştır. Bu veri kümeleri için analizlerin sonuçları ışığında, veri kümesinde aykırı gözlemlerin varlığında literatürde kullanılan EKKD'ye dayanan dayanıklı TBA yerine UEKKD'ya dayanan dayanıklı TBA'nin kullanılmasıyla daha sağlam bulguların elde edilebileceği görülmüştür.

References

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.

Atkinson, A.C., (1994). "Fast Very Robust Methods for the Detection of Multiple Outliers", J. Amer. Statist. Assoc. 89, 1329–1339.

Campbell, N. A., (1980). "Robust procedures in multivariate analysis I: Robust covariance estimation", Applied statistics, 231-237.

Croux, C., Filzmoser, P., & Fritz, H. (2013). Robust sparse principal component analysis. Technometrics, 55(2), 202-214.

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., 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.

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 ReportTS 03–5, Department of Statistics, Vienna University of Technology, Austria.

Filzmoser, P., Todorov, V., (2011). "Review of robust multivariate statistical methods in high dimension", Analytica chimica acta, 705(1), 2-14.

Hubert, M., Debruyne, M., (2010). "Minimum covariance determinant", Wiley interdisciplinary reviews: Computational statistics, 2(1), 36-43.

Hubert, M., Engelen, S., (2004). "Robust PCA and classification in biosciences", Bioinformatics, 20(11), 1728-1736.

Johnson, R., Wichern, D. (1992). "Applied multivariate statistical methods", 3rd Edition., Prentice Hall, Englewood Cliffs, NJ.

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.

R Development Core Team, (2011). "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna.

Riu, J., Bro, R., (2003). "Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models", Chemometrics and Intelligent Laboratory Systems, 65(1), 35-49.

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., (1985). "Multivariate estimation with high breakdown point", Mathematical statistics and applications, 8, 283-297.

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

Rousseeuw, P.J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke T., Maechler, M., (2009). "Robustbase: basic robust statistics", R package version 0.4–5. Available at http://CRAN. R-project. org/package = robustbase.

Todorov ,V. and Filzmoser, P., (2009). "An object-oriented framework for robust multivariate analysis", J. Statist. Softw. 32(3) (2009), 1–47.

Todorov, V., (2009). "rrcov: Scalable Robust Estimators with High Breakdown Point", R package version 0.5–03, Availableat http://CRAN. R-project. org/package = rrcov.

Todorov, V., Neyko, N., Neytchev, P., (1994). "Stability of High Breakdown Point Robust PCA", in Short Communications, COMPSTAT'94; Physica Verlag, Heidelberg.

This page intentionally left blank