İstatistik Araştırma Dergisi Cilt 03 No 03 Sayfa 55-66 Aralık 2004

Çok Değişkenli Veri Kümeleri Üzerinde Tanımlı Aykırı Değer Belirleme Tekniklerinin Simülasyon Çalışması ile Karşılaştırılması

Gülsen KIRAL* Nedret BİLLOR**

ÖZET

Temel bileşenler analizi ilişkisiz değişkenler kümesinin kurulması ve/ya boyut indirgenmesi amacı ile kullanılan istatistiksel bir tekniktir. Bazen tek başına bir analiz olarak kullanıldığı gibi bazen de başka analizler için veri hazırlama tekniği olarak kullanılmaktadır. Veri boyutunda indirgeme yaptığından özellikle yüksek boyutlu verilerin analiz edilmesinde tercih edilen bir tekniktir. Fakat klasik varyans ve kovaryans matrisine dayalı olarak hesaplandığından aykırı değerlerin varlığı durumunda sağlıklı sonuç vermemektedir. Bu nedenle aykırı değer olması olasılığına karşın dayanıklı temel bileşenler analiz teknikleri kullanımı önerilmektedir.

Bu çalışmada temel bileşenler analizi çerçevesinde tanımlanan dayanıklı tekniklerden, Dayanıklı Temel Bileşenler Analizi (DTBA) (Hubert ve ark.,2002) ve BACON Dayanıklı Temel Bileşenler Analizi (BDTBA) (Kıral ve Billor, 2001) ve Klasik Temel Bileşenler Analizinin (TBA) performansları simülasyon çalışması yapılarak karşılaştırılmıştır.

Anahtar Kelimeler: *Klasik Temel Bileşenler Analizi, Dayanıklı Temel Bileşenler Analizi, BACON Algoritması, Aykırı Değer*

1. GİRİŞ

Bir veri kümesinde gözlemlerin çoğu tarafından önerilen modele uymayan gözlemlere aykırı değer denir. Aykırı değerlerin belirlenmesi kullanılan istatistiksel analizin doğru yorumlanması açısından önemlidir. Tek boyutlu veri kümesindeki aykırı değerler tek değişkenli analiz yöntemleriyle kolayca belirlenebilir, ancak çok boyutlu veri kümesinde işlemler zorlaşır. Çünkü bu tip veriler hesaplama zorluğu getirmektedir ve pek çok istatistiksel analiz ilişkili değişkenlerin varlığında sağlıklı sonuç vermemektedir. Aykırı değer belirlemeye ilişkin üç farklı yaklaşım vardır: Klasik aykırı değer belirleme yöntemleri, dayanıklı yöntemler ve birleştirilmiş *(combined)* yöntemler.

Klasik aykırı değer belirleme yöntemleri aykırı değerlere hassas klasik kestiricilere dayalı olup etkin değildirler. Masking (aykırı değer olan bir gözlemin normal gözlem gibi görülmesi) ve bulandırma *(swamping)* (normal olan bir gözlemin aykırı değer gibi görülmesi) hataları yanlış yorumlara sebeb olabilirler.

Dayanıklı yöntemler dayanıklı kestiricilere dayalı olup maskeleme ve bulandırma problemlerinden etkilenmezler, fakat hesaplanması ve uygulaması zor yöntemlerdir.

*Çukurova Ünv. İ.İ.B.F. Ekonometri Böl. Balcalı/ADANA
**Auburn Unv. Discrete and Statistical Science, Auburn, USA

Birleştirilmiş yöntemler ise dayanıklı ve klasik aykırı değer belirleme tekniklerinin birleşimi ile tanımlanmışlardır. Dayanıklı ve/ya klasik kestiricilere dayalıdırlar. Dayanıklı yöntemlere göre uygulamada daha pratik olup bir çoğu maskeleme ve bulandırma problemlerinden etkilenmemektedirler.

Veri kümeleri içerisinde karşılaşabileceğimiz en önemli problemlerden biri büyük veri kümeleri ile çalışıldığında ortaya çıkmaktadır. Çok sayıda değişkenle çalışmak değişkenlerin bağımsızlık varsayımını bozup, işlem yükünü arttıracağı gibi yapılacak analiz sonuçlarının yorumlanmasını da güçleştirebilmektedir. TBA birbirleri ile ilişkisiz olan yeni bir veri kümesi oluşturarak veri boyutunu indirgediğinden yüksek boyutlu verilerin incelenmesinde tercih edilen bir tekniktir. Yöntem örneklem kovaryans matrisinin özvektörlerinin belirlenmesi ile yürütülür. Ancak örneklem kovaryans matrisi örneklem ortalamasına dayalı olduğundan yöntem aykırı değerlerden büyük ölçüde etkilenmektedir. Bu nedenle aykırı değerlerin varlığı olasılığına karşın dayanıklı temel bileşenler analizinin kullanılması tercih edilmektedir.

Dayanıklı temel bileşenler analizlerinde amaç aykırı değerlerden etkilenmeyen temel bileşenlerin belirlenmesidir. Bu amaçla işlemler iki farklı mantıkla yürütülebilir. Birincisinde klasik kovaryans matrisi dayanıklı kovaryans matrisi ile yer değiştirerek analiz gerçekleştirilebilir (örnek: Campbell, 1980 ve Croux ve Haesbroeck, 2000). Bu grupta sonuçlar dayanıklı olarak elde edilir ama ne yazık ki yüksek boyutlu verilerde hesaplama problemleri nedeniyle kullanışlı değildirler. İkincisinde ise analizi izdüşüm takibi *(projection pursuit)* yöntemini kullanarak gerçekleştirmektedirler (Örnek: Lie ve Chen, 1985, Croux ve Ruiz Gazen, 2000, Hubert ve ark., 2002). Bu yöntemler yansıtılmış veri üzerine dayanıklı yayılım ölçüsünü en büyük yapacak şekilde ardışık adımlarla yeni doğrultuları bulmaya çalışır.

Dayanıklı temel bileşenler ile ilgili ilk çalışma Campbell (1980) tarafından yapılmıştır. Campbell çalışmasında dayanıklı M-kestiricisi kullanarak aykırı değerlerden etkilenmeyen temel bileşenleri belirlemiştir. Yöntem temel bileşenler analizi içerisinde dayanıklı M-kestiricisine ait varyans-kovaryans matrisinin kullanımı ile tanımlanmıştır. Yöntemde amaç; aykırı değerlerin etkisini ortadan kaldıracak gerçek ağırlıkları bularak tüm veri kümesini temsil eden gerçek varyans-kovaryans matrisini elde etmektir. Ardından Lie ve Chen (1985); izdüşüm takibi yöntemine dayalı bir çözüm önerdiler. Lie ve Chen'nin amacı; en büyük dayanıklı ölçeklemeye sahip izüşümü alınmış gözlemlerin doğrultusunu belirlemektir. Birbirini izleyen adımlarda her yeni doğrultu önceki tüm doğrultulara dik olacak şekilde belirlenmektedir. Yüksek boyutlu veri kümelerinde hatta ve hatta parametre sayısı gözlem sayısından büyükken de dahil olmak üzere iyi sonuç veren bir algoritmadır. Fakat hesaplama problemleri içermektedir. İzdüşüm takibi yöntemine dayalı yöntemlerde karşılık gelen etki fonksiyonunun sınırlandırılmamış olması yerel dayanıklılıkta eksikliğe sebep olmaktadır. Bunun yanında izdüşüm takibi yöntemine dayalı kestiricilerin nasıl hesaplanacağı açık değildir. Bu problemleri ortaya çıkaran Croux ve Ruiz-Gazen kısıtlamalar altında bir maksimizasyon probleminin çözümünü önerdiler. C-R algoritması adını verdikleri yöntem küçük boyutlu veri kümelerinde iyi çalışmasına rağmen büyük boyutlu veri kümeleri için hesaplama problemleri içermektedir. Daha sonra Hubert ve ark. (2002) C-R algoritmasını biraz daha geliştirerek, daha hızlı iki adımlı algoritma *(a faster two-step algorithm, DTBA)* adını verdikleri yeni bir yöntem sunmuşlardır. Yöntem sayısal olarak C-R algoritmasından daha kararlıdır ve veride gözlemden çok değişken olması durumunda da sağlıklı sonuç vermektedir. Diğer taraftan Caroni (2000); Campbell'in (1980) yöntemi üzerine bir

simülasyon çalışması yapmıştır. Bu çalışmada $x_i \sim N_p(\mu,\Sigma)$ i = 1,...,n sıfır hipotezi altında DTBA içinde gözlemlerin ağırlıklarına bağlı olarak kritik değerler belirlenmektedir. Bu kritik değerlerin belirlenmesi ile düşük ağırlığa sahip olan gözlemler yöntem tarafından doğru bir şekilde belirlenir. Bu nedenle yöntem aykırı değerler için bir formal test olarak düşünülebilir (Caroni, 2000).

Bu çalışmanın ikinci bölümde klasik temel bileşenler analizi ve BACON Dayanıklı Temel Bileşenler Analizi (BDTBA) anlatıldıktan sonra üçüncü bölümde BDTBA yönteminin etkinliği bir veri kümesi üzerinde gösterilecektir. Dördüncü bölümde Dayanıklı Temel Bileşenler Analizi (Hubert ve ark., 2002) ve BACON Dayanıklı Temel Bileşenler Analizi (Kıral ve Billor, 2001) ve Klasik Temel Bileşenler Analizi yöntemlerinin performansları simülasyon çalışmaları yapılarak karşılaştırılacaktır. Nihai tartışma ve sonuç ise beşinci bölümde verilecektir.

2. YÖNTEMLER

2.1. Temel Bileşenler Analizi

Temel bileşenler analizi; değişkenler arası bağımlılık yapısının yok edilmesi ve (veya) boyut indirgenmesi ya da başka analizler için veri hazırlanması amaçları ile kullanılır. Analizde, veriyi temsil eden X_{nxp} matrisine uygun bir dönüşüm yapılarak, X uzayındaki problemler düzeltilmeye çalışılır. Dönüşüm sonucu birbirleri ile ilişkisiz sütunlardan oluşan bir veri matrisi elde edilmiş olur.

İncelemede X_{nxp} matrisinin sütunlarının birimleri arasında uyuşma söz konusu değilse, bu matris yerine onun standartlaştırılmış formu kullanılır.

Aslında temel bileşenler p tane X_1, X_2,..., X_p rasgele değişkenin özel doğrusal kombinasyonudur. Geometrik olarak, bu doğrusal kombinasyonlar koordinat eksenleri X_1, X_2, ..., X_p ler olan orijinal sistemin döndürülmesiyle elde edilen yeni bir koordinat sistemini temsil eder.

X_{nxp} çok değişkenli veri kümesinin varyans-kovaryans matrisi V nin öz değerleri $(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0)$ ve bu öz değerlere karşılık öz vektörleri u_i olmak üzere $(\lambda_1,u_1),(\lambda_2,u_2),....(\lambda_p,u_p)$ çiftleri için i. temel bileşen

$$y_i = u_i'X \quad i=1,2,...,p$$

eşitliği yardımıyla hesaplanır. Örneğin 1.temel bileşen, $Var(y_1)=u_1'Vu_1=\lambda_1$, maksimum varyanslı doğrusal kombinasyondur. Bu şekilde temel bileşenler belirlenerek X_{nxp} matrisini temsil eden yeni temel bileşenler matrisi Y_{nxp} belirlenmiş olur.

Klasik temel bileşenler analizi; verilerde homojenliğin bozulması durumunda (*yani aykırı değerlerin varlığında*) sağlıklı sonuçlar vermemektedir. Bu durumda aykırı değerlere karşı dayanıklı olan kestiriciler kullanılarak analizin yapılması gerekmektedir.

2.2. BACON Algoritması

Bu yöntemde esas amaç; aykırı değerlerden arındırılmış olacak şekilde gözlemlerin hemen hemen yarısını içeren temel alt kümeyi bulmak, hemen ardından da temel alt küme ile uyumlu gözlemleri bu kümeye dahil etmektir. İşlem sonunda temel alt küme dışında kalan

gözlemler aykırı değer olarak belirlenirler. Temel alt küme dışında hiç gözlem kalmamışsa "*veri kümesi aykırı değer içermemektedir*" denir.

BACON yönteminde (Billor ve ark. 2000); gözlemlerin çok değişkenli eliptik dağılımdan geldiği varsayılarak Mahalanobis uzaklığından yararlanılmakta, kritik değer olarak da düzeltilmiş ki-kare değeri kullanılmaktadır.

BACON yöntemi gözlemlerin bloklanması nedeniyle hesaplamalar için etkin bir yöntemdir. Diğer yöntemlere göre bu yöntemdeki tekrarlama sayısı daha azdır. Tekrarlamaların her biri kovaryans matrisinin hesaplanması ve tersinin alınmasını gerektirir. Fakat tekrarlama sayısı n örneklem büyüklüğünün artması ile büyümez ve hesaplanan n uzaklığın sıralanmasını gerektirmez.

2.3. BACON Dayanıklı Temel Bileşenler Analizi (*BDTBA*)

Dayanıklı kestiricilerle yapılan işlemler çoğu zaman için sağlıklı sonuç verirler ama bilindiği gibi yapılması gereken işlemler problemli ve zaman alıcıdır. Gözlem ve parametre sayılarının artması durumunda hesaplamalar iyice artmaktadır. Bunun yanında kullanılan veri kümesine ve istatistiğe bağlı olarak etkinliklerinde değişikliklerin olabilmesi ve sadece belli tipteki aykırı değerleri ortaya çıkarıyor olmaları da karşılaşılabilecek problemlerdendir. O halde bu problemlerden etkilenmeyen daha hızlı işleyip sağlıklı sonuç veren bir yönteme gereksinim duyulmaktadır. Bu amaçla; bu çalışmada Billor ve ark. (2000) tarafından tanımlanan BACON algoritması kullanılarak dayanıklı temel bileşenlerin belirlenmesini sağlayan bir algoritma tanımlanmıştır (BDTBA) (Kıral ve Billor, 2001).

BDTBA yöntemi; DTBA (Hubert ve ark., 2002) ya da Campbell (1980) tarafından tanımlanan yöntem içinde kullanılan dayanıklı M-kestiricisi yerine BACON algoritmasından elde edilen klasik ortalama ve kovaryans matrisinin kullanılmasına dayalı olarak yürütülmektedir.

BDTBA yönteminde ana düşünce büyük veri kümelerinde etkinliği ispatlanmış BACON algoritması (Billor ve ark., 2000) kullanmak ve hemen ardından temel bileşenler analizini uygulamaktır. Böylece analizci aykırı değerlerden arındırılmış X veri matrisini tanımlayan en önemli bileşenleri belirleyebilir.

BDTBA Algoritması

Adım 1: Temel altküme; BACON algoritmasında tanımlı yaklaşımlardan biri kullanılarak $m=cp$, ($c=4$ veya 5) elemanlı olacak şekilde belirlenir.
Adım 2: Temel alt kümedeki gözlemlerin ortalama ve varyans-kovaryans matrisleri sırasıyla, $\overline{X}_b$ ve S_b olmak üzere

$$d_i(\overline{X}_b, S_b) = \sqrt{(x_i - \overline{X}_b)' S_b^{-1} (x_i - \overline{X}_b)} \qquad i=1,2,\ldots,n$$

uzaklıkları hesaplanır.
Adım 3: $d_i(\overline{X}_b, S_b) < C_{npr} \cdot \chi_{p,\alpha/n}$ olan gözlemlerle yeni temel alt küme belirlenir. $\chi^2_{p,\alpha}$; p serbestlik dereceli, $1-\alpha$ yüzdelikli ki-kare değeri, $C_{npr} = C_{np} + C_{hr}$ olan bir düzeltme faktörü, r; şu an ki

temel alt kümede bulunan eleman sayısı, $C_{hr} = \max\{0,(h-r)/(h+r)\}$ ve $C_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p}$ olarak tanımlıdır $(h = \lfloor (n+p+1)/2 \rfloor)$.

Adım 4: *2.* ve 3. adımlar temel alt kümede değişme olmayana kadar tekrarlanır.

Adım 5: Son adımda elde edilen temel alt küme dışında kalan gözlemler aykırı değer olarak tanımlanır.

Adım 6: Aykırı değer olarak belirlenen gözlemler veri kümesinden atılarak indirgenmiş veri kümesi elde edilir ($X_{(I)}$).

Adım 7: $X_{(I)}$ matrisinin öz değer ve öz vektör çiftleri (λ_i, u_i); $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ olacak şekilde hesaplanır.

Adım 8: İndirgenmiş matrisin varyans-kovaryans matrisine karşılık gelen özvektörler $U=(u_1,u_2,...,u_p)$ olarak tanımlanmak üzere yeni temel bileşenler

$$y = X_{(I)}.U = (y_1^*, y_2^*, ..., y_p^*)$$

elde edilir. y_i^*, y nin *i.* sütunudur.

Grafiksel Yöntem:

Kıral ve Billor (2001) aykırı değerlerin görsel olarak belirlenmesine yardımcı olmak amacı ile birkaç grafiksel yöntem önermiştir. Grafiklerin oluşturulmasında takip edilecek adımlar:

Adım 1: BACON Dayanıklı Temel Bileşenler Uzaklığı (BDTBU) aşağıdaki adımları takip edilerek hesaplanır.

- BTBA'den elde edilen temel bileşenlerin (y) ortalama ($\bar{y}$) ve kovaryans matrisi (s_y) hesaplanır.
- Orijinal veri matrisine (X) ait temel bileşenler x_i^* hesaplanır.
- Aşağıdaki eşitliği kullanarak dayanıklı uzaklıklar hesaplanır.

$$BDTBU_i = d_i = (x_i^* - \bar{y}).(s_y)^{-1}.(x_i^* - \bar{y})' \quad i = 1,2,...,n$$

Adım 2: Birinci adımda elde edilen bilgiler kullanılarak aşağıdaki grafikleri çizilir

- BDTBU nın küp köküne ait Kantil-Kantil *(Quantile-Quantile)* (*Q-Q)* grafiği ya da
- Klasik Mahalanobis uzaklığı (MD) karşın *BDTBU'nın* serpilme grafiği ya da
- *BDTBU* na ait indeks grafiği

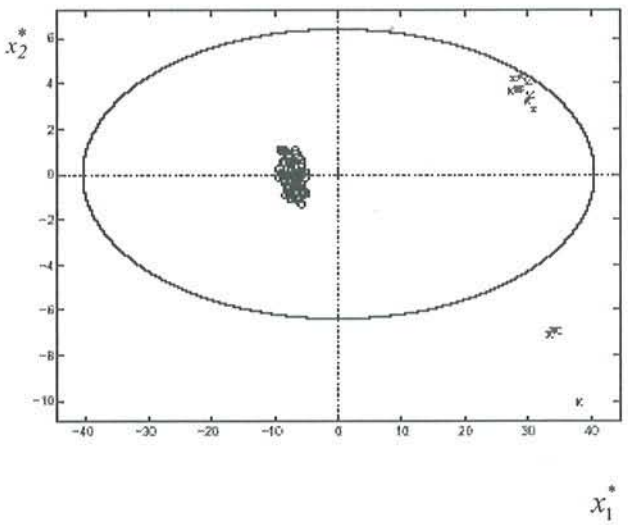
Birinci grafik verinin normal dağılımdan gelip gelmediği hakkında bilgi vermekte ve aykırı gözlemleri belirlemekte.

İkinci grafik BDTBA'ne dayalı ortalama ve kovaryans matrisini klasik ortalama ve kovaryans matrisi ile karşılaştırmaktadır. Bunları karşılaştırmada araştırmacı grafik içerisine $x = C_{npr} \cdot \chi_{p,\alpha/n}$ ve $y = C_{npr} \cdot \chi_{p,\alpha/n}$ doğrularını çizmelidir. Bu doğrular bir dikdörtgen belirlemektedir. Bu dikdörtgen içerisinde bulunan gözlemler temiz dışarısındakiler ise aykırı değerlerdir. Bu grafik yüksek boyutlu veride çok kullanışlı olmaktadır. Çünkü yüksek boyutta verinin ve veriyi temsil eden elipsoidin gözlemlenmesi problemlidir. Veride aykırı değerin olmaması durumunda tüm gözlemler dikdörtgen içerisinde birlikte bulunur.

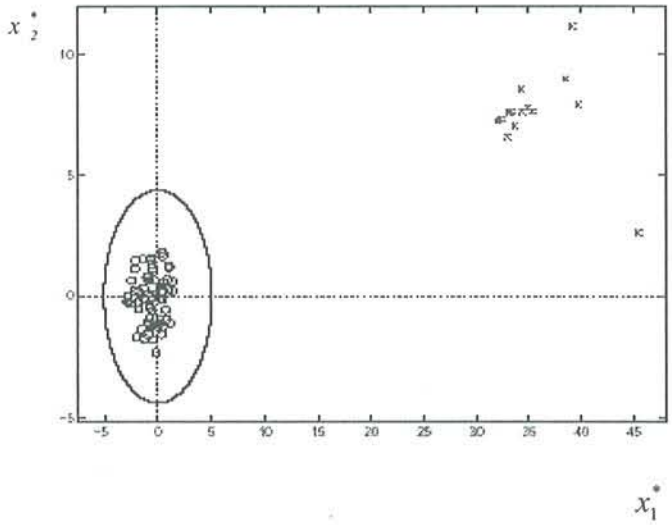
Üçüncü grafik aykırı değer belirlemek amacı ile kullanılmaktadır. Aykırı değerlerin belirlenebilmesi için $y = C_{npr} \cdot \chi_{p,\alpha/n}$ doğrusunun çizilmesi gerekmektedir. Bu doğru üzerinde bulunan gözlemler aykırı değer olarak bilinirler. Araştırmacı aykırı değerlerin atılmasından sonra veride başka aykırı değer olup olmadığını araştırmak isterse y matrisinin bileşenlerine ait Q-Q grafiklerini de inceleyebilir.

3. UYGULAMA

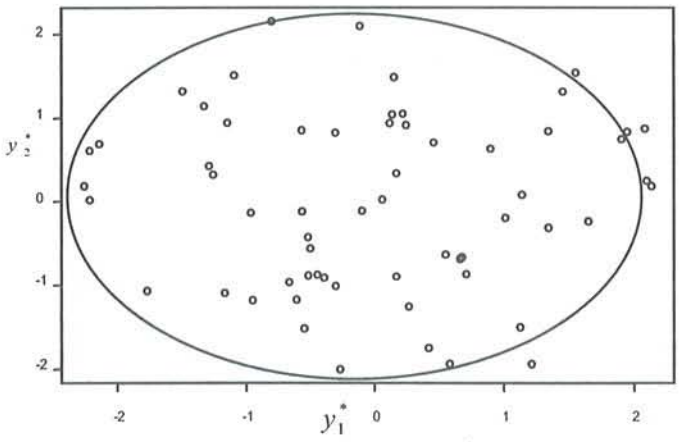
BDTBA nin etkinliği Hawkins, Bradu ve Kass (*HBK*) veri kümesi üzerinde gösterilecektir. *HBK* verisi (Hawkins ve ark., 1984) maskeleme etkisinin incelenmesi amacı ile oluşturulan üç değişken ve toplam 75 gözlem üzerine kurulu tipik bir veri kümesidir. İlk on gözlem xy-uzayında ardından gelen dört gözlem ise x-uzayında aykırı değer olacak şekilde üretilmiştir.



Şekil 1. TBA den elde edilen HBK verisinin ilk iki bileşenine ait serpilme diyagramı



Şekil 2.DTBA ya dayalı HBK nın skorları



Şekil 3. HBK data için BDTBA den elde edilen ilk iki bileşene ait serpilme grafiği

Şekil 1.; tüm değişkenler (x_1, x_2, x_3 ve y) üzerinden elde edilen temel bileşenlerin ilk ikisine ait skorları göstermektedir. Bu grafik içerisine temel bileşenlerin %97.5 lik güven elipsoidini çizersek, klasik temel bileşenler yönteminin yalnızca 4 gözlemi aykırı değer olarak belirlediğini görebiliriz.

TBA yöntemi veri noktalarını ortalamamaktadır çünkü verinin ortalaması temiz gözlemler bulutunun dışarısında bulunmaktadır (Bkz. Şekil 1.). Diğer taraftan DTBA ve BDTBA yöntemleri veri noktalarını ortalamakta ve tüm aykırı değerleri doğru bir şekilde temiz gözlem kümesinden ayırmaktadır (Bkz. Şekil 2. ve 3.).

BDTBA yöntemi aykırı değer olan gözlemleri veri kümesinden attıktan sonra temel bileşenleri hesapladığından elipsoidin dışında gözlem bulunmamaktadır (Bkz. Şekil 3.). BDTBA nin indeks ve Q-Q grafiklerinden de aykırı değerlerin temiz veriden ayrıldıklarını görmekteyiz (Bkz. Şekil 4. ve 5.). Buna karşın TBA sonrası elde edilen Mahalanobis uzaklığına ait indeks ve Q-Q grafikleri aykırı değerleri tam olarak belirleyememektedir (Bkz. Şekil 6. ve 7.).

TBA klasik kestiricilere dayalı olduğundan aykırı değerlerin varlığında sağlıklı sonuçlar vermeyecektir. Klasik Mahalanobis uzaklığının küp küküne ait Q-Q grafiği ve indeks grafikleri bu problemi göstermektedir.

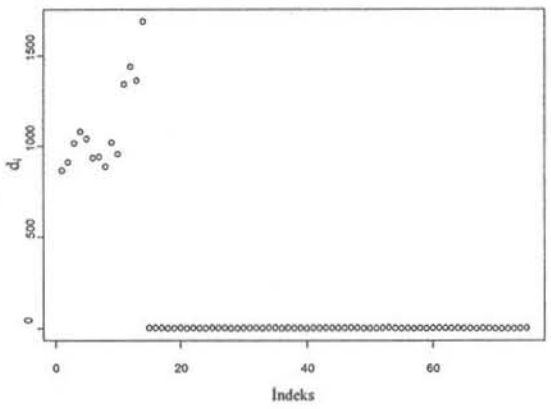
TBA klasik kestiricilere dayalı olduğundan aykırı değerlerin varlığında sağlıklı sonuçlar vermeyecektir. Klasik Mahalanobis uzaklığının küp küküne ait Q-Q grafiği ve indeks grafikleri bu problemi göstermektedir.

TBA, DTBA ve BDTBA yöntemlerin performanslarını karşılaştırmada incelenmek istenen yöntemler uygulandıktan sonra belirlenen verinin varyans-kovaryans matrisinin özdeğerleri hesaplandı. Sonuçlar Tablo 1 de görülmektedir. Bu özdeğerler (λ_i) aynı zamanda i. temel bileşenin varyansına da karşılık gelmektedir. Tüm değerler bire eşitse orijinal X matrisinin sütunları ortogonaldir denir. Bir λ_i tam olarak sıfıra eşitse bu orijinal X matris sütunları arasında mükemmel bir doğrusal ilişki vardır. Bir ya da birden fazla sıfıra yakın değer ise çoklu iç ilişkinin varlığını vurgulamaktadır.

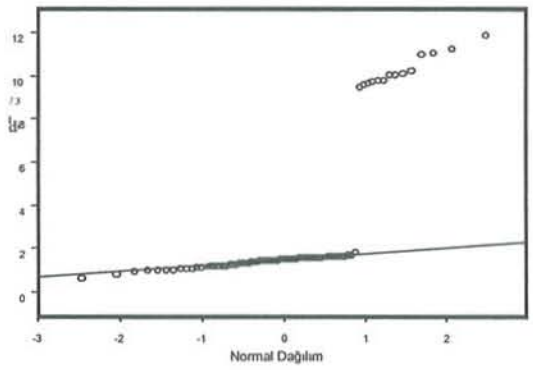
Aykırı değerler verinin yoğunlaştığı kısmın uzağında bulunduğundan TBA den elde edilen özdeğerler oldukça büyük elde edilmiştir (Bkz. Tablo 1). Diğer taraftan DTBA ve BDTBA'ne ait algoritmalar küçük özdeğerlere sahiptir. Fakat BDTBA'nın özdeğerleri DTBA'nınkilerden daha küçüktür. Bu sonuçlar gösteriyor ki BDTBA yöntemi diğerlerinden daha iyi sonuç vermektedir. BDTBA yöntemi aykırı değerleri eledikten sonra temel bileşenleri hesapladığından veri matrisinin boyutunun indirgenmesi daha güvenilir şekilde yapılmaktadır.

Tablo1. TBA, BDTBA, ve DTBA den elde edilen HBK verisinin özdeğerleri

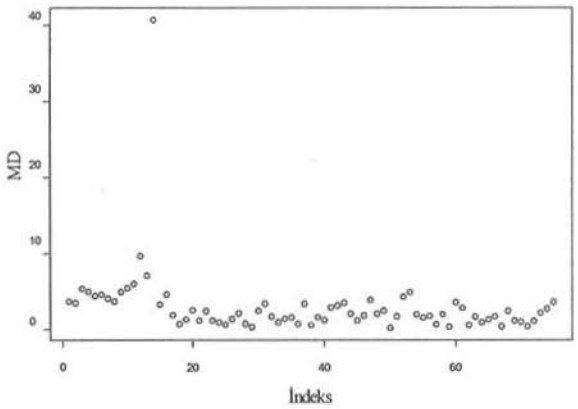
	TBA	**DTBA**	**BDTBA**
λ_1	223.12	3.47	1.326
λ_2	5.538	2.63	1.093
λ_3	1.688	2.47	0.956
λ_4	0.914	0.67	0.2957



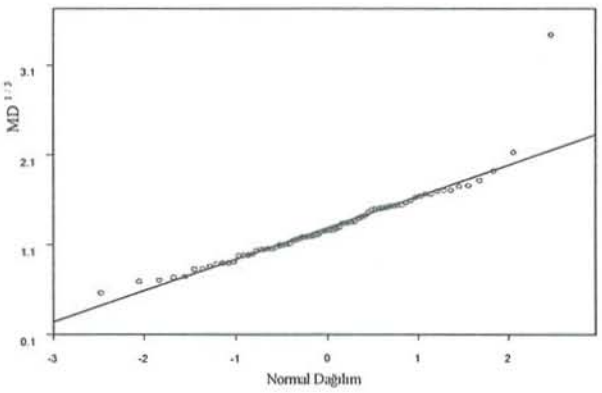
Şekil 4. HBK verisi için BDTBU'na ait indeks grafiği



Şekil 5. HBK verisinin BDTBU nın küpköküne ait Q-Q grafiği



Şekil 6. HBK data için elde edilen Mahalanobis uzaklığının indeks grafiği



Şekil 7. HBK verisi için elde edilen Mahalanobis uzaklığının küpköküne ait Q-Q grafiği

Tablo 2. Simülasyon çalışmasında kullanılacak faktörler

Gözlem Sayısı	100
Parametre sayısı	4 veya 20
Temiz gözlemin varyans kovaryans matrisi (Σ)	diag(4,3,2,1) (p=4 için) diag(5,4,3,2,1,0.15,0.14,...,0.02,0.01) (p=20 için)
Kirletmemiktarı (ε)	%0 veya %10
Kirletilen verinin ortalama vektörü (μ*)	$(0,0,0,0)^T$ veya $(0,0,0,10)^T$ (p=4 için) $(0,0,0,0,0,1)^T$ veya $(0,0,0,0,1,0)^T$ (p=20 için)
Kirletilen verinin varyans kovaryans matrisi (Σ*)	Σ veya 9Σ (p=4 için) Σ veya (1/20)Σ (p=20 için)

4. SİMÜLASYON ÇALIŞMASI

Bu çalışmada TBA, *BDTBA (Kıral ve Billor,2001)* ve *DTBA* (Hubert, ve ark., 2002) yöntemleri altı farklı veri grubu üzerinde karşılaştırılacaktır.

Simülasyon çalışmasında kullanılan temiz ve kirletilmiş veri sırasıyla $N(0,\Sigma)$ ve $N(\mu^*,\Sigma^*)$ dağılımlarından üretilmektedir. BDTBA yönteminin performansını değerlendirme de kullanılan veri kümesine ait gözlem sayısı 100 ve parametre sayısı (düşük boyutlu veri için) 4 veya (büyük boyutlu veri için) 20 olarak alınmıştır. Simülasyon çalışması sonuçları 100 farklı veri kümesi üzerinden elde edilen sonuçların değerlendirilmesi ile elde edilmiştir. Kirletme seviyesi 1. ve 4. veri grupları için %0, 2., 3., 5. ve 6. veri grupları içinse %10 olarak belirlenmiştir. 2., 3., 5. ve 6. veri kümeleri içerisindeki kirletilen veriler sırasıyla $N(0,9\Sigma)$, $N((0,0,0,10)',\Sigma)$, $N(10e_6,\Sigma)$ ve $N(10e_5,(1/20)\Sigma)$ 'dan üretilmiştir, burada e_i i. elemanı 1 diğerleri 0 olan birim vektördür.

Karşılaştırmada kullanılan performans ölçüleri:

- $MSE(\hat{\lambda}_i)=\frac{1}{100}\sum_{i=1}^{100}\left(\lambda_i-\hat{\lambda}_i^{(j)}\right)^2$ değerleri *(MSE: Ortalama kareler hatası)*

(λ_i :Veri kümesinin varyans-kovaryans matrisinin i. özdeğeri ve $\hat{\lambda}_i^{(j)}$ j. tekrarlama sonucu elde edilen tahmini i. özdeğer),

- Veri kümesinin varyans-kovaryans matrisine ait özdeğerlerin ortalamasıdır.

Araştırmada kullanılan faktörler Tablo 2 de özetlenmiştir.

Tablo 3 ve 4 de sırasıyla $p=4$ ve 20 için simülasyon sonuçlarını göstermektedir. $p=4$ olması durumu sonuçlarını veren Tablo 3 de tüm özdeğer ve özvektörler için elde edilen sonuçlar verilmekte, fakat $p=20$ için ki sonuçları gösteren Tablo 4 de sadece en büyük ilk beş özdeğer için elde edilen sonuçlar verildi.

Tablo 3: p=4 için simülasyon sonuçları

	Veri grubu 1			Veri grubu 2			Veri grubu 3		
	TBA	**BDTBA**	**DTBA**	**TBA**	**BDTBA**	**DTBA**	**TBA**	**BDTBA**	**DTBA**
Ort $\hat{\lambda}_1$	4.190	4.187	4.23	8.197	4.565	5.43	10.278	4.162	5.01
Ort $\hat{\lambda}_2$	2.933	2.917	3.18	5.104	3.116	3.94	4.115	2.916	4.06
Ort $\hat{\lambda}_3$	1.915	1.909	2.11	3.225	2.029	2.67	2.879	1.867	3.25
Ort $\hat{\lambda}_4$	0.974	0.967	1.04	1.594	1.037	1.35	1.874	0.944	2.37
MSE $\hat{\lambda}_1$	0.352	0.356	0.39	20.38	0.811	2.61	39.905	0.327	1.38
MSE $\hat{\lambda}_2$	0.120	0.128	0.18	5.190	0.19836	1.20	1.507	0.154	1.36
MSE $\hat{\lambda}_3$	0.072	0.071	0.13	1.838	0.101	0.62	0.895	0.111	1.73
MSE $\hat{\lambda}_4$	0.017	0.017	0.03	0.477	0.040	0.19	0.821	0.038	2.00

Tablo 4 : p=20 için simülasyon sonuçları

	Veri grubu 4			Veri grubu 5			Veri grubu 6		
	TBA	**BDTBA**	**DTBA**	**TBA**	**BDTBA**	**DTBA**	**TBA**	**BDTBA**	**DTBA**
Ort $\hat{\lambda}_1$	5.361	5.361	5.12	9.519	5.532	5.30	10.05	10.05	5.55
Ort $\hat{\lambda}_2$	3.963	3.962	3.83	5.312	4.045	4.07	4.882	4.881	4.65
Ort $\hat{\lambda}_3$	2.887	2.887	2.87	3.940	2.867	3.25	3.627	3.623	3.81
Ort $\hat{\lambda}_4$	1.933	1.933	1.92	2.826	1.910	2.51	2.545	2.542	3.12
Ort $\hat{\lambda}_5$	0.953	0.952	0.96	1.875	0.952	1.78	1.695	1.695	2.24
MSE $\hat{\lambda}_1$	0.602	0.603	0.53	20.54	0.982	0.55	25.59	25.60	0.70
MSE $\hat{\lambda}_2$	0.179	0.179	0.25	2.074	0.293	0.21	1.076	1.070	0.64
MSE $\hat{\lambda}_3$	0.134	0.134	0.15	1.100	0.200	0.24	0.599	0.598	0.81
MSE $\hat{\lambda}_4$	0.063	0.063	0.12	0.793	0.104	0.41	0.401	0.397	1.43
MSE $\hat{\lambda}_5$	0.023	0.023	0.03	0.846	0.041	0.74	0.548	0.548	1.71

5. TARTIŞMA VE SONUÇ

Tablo 3 ve 4 de veri grubu 1-6 için elde edilen Ort $\hat{\lambda}_i$ ve MSE $\hat{\lambda}_i$ değerleri görülmektedir. İdeal olarak küçük Ort $\hat{\lambda}_i$ ve MSE $\hat{\lambda}_i$ değerleri bulmak arzulanır.

Kirletme oranının %0 olduğu, Veri grubu 1 de TBA ve BDTBA nin performans ölçü değerleri küçük ve hemen hepsinde benzer sonuçlar vermektedir. Ama DTBA yöntemi için elde edilen değerler TBA ve BDTBA için elde edilenlere göre daha yüksektir. Bu nedenle veride aykırı değer yokken DTBA yönteminin iyi işlemediğini söyleyebiliriz. Buna rağmen TBA bazı durumlar için iyi performansa sahip fakat kirletme varlığında (Veri grubu 2 ve 3) güvenilmeyen tahminler vermektedir. BDTBA yöntemi parametre sayısının 4 olması durumunda tüm kirletme seviyelerinde iyi sonuç vermektedir (Bkz. Tablo 3). Veri grubu 2 ve 3 de BDTBA ve DTBA yöntemlerinin performans ölçüleri oldukça küçük fakat BDTBA yönteminin değerleri DTBA yöntem değerlerinden biraz daha küçüktür. Bu nedenle veri grubu 1, 2 ve 3 için BDTBA diğer yöntemlerden daha üstün denilebilir.

Diğer taraftan Veri grubu 4 de TBA ve BDTBA yöntemleri hemen hemen aynı sonucu vermektedir. Yani kirletme yok ve parametre sayısı yüksek ise bu iki yöntem aynı performansı vermektedir denilebilir. DTBA değerleri Ort $\hat{\lambda}_i$ için diğerleri ile hemen hemen aynı fakat DTBA'ne ait MSE $\hat{\lambda}_i$ değerleri diğerlerine göre biraz daha yüksektir. Bu nedenle bu konfügürasyon içerisinde DTBA yöntemi diğer yöntemler kadar iyi sonuç vermemektedir.

Veri grubu 5 içinde pek çok bileşen içerisinde BDTBA; DTBA'dan daha iyi performansa sahiptir. Bu nedenle bu veri grubu içinde BDTBA yönteminin çoğunlukla DTBA'den daha iyi olduğu söylenebilir.

Veri grubu 6 da BDTBA'ne ait performans ölçüleri gösterir ki yöntem ilk önemli bileşen içinde aykırı değerlere karşı daha hassastır. Fakat diğer bileşenler için BDTBA yöntemi DTBA'den daha iyi sonuç vermektedir. Bu nedenle BDTBA yönteminin parametre sayısı ve kirletme

seviyesinin artması durumlarında TBA ve DTBA yöntemlerinden daha iyi sonuç verdiği güvenle söylenebilir.

Özet olarak Bu çalışmada dayanıklı temel bileşenleri elde etmede BDTBA yönteminin kullanımının yararları bir simülasyon çalışması yapılarak gösterilmiştir. Çok değişkenli veri kümeleri içerisinde çoklu aykırı değerleri bulmayı amaçlayan bu algoritma; büyük veri kümelerine (*1 milyon gözlem için bile*) uygulanabilmekte, model üzerinde çok küçük etkisi olabilecek gözlemleri belirleyebilmekte, hesaplama zorlukları içermemektedir. Dahası maskeleme ve bulandırma problemlerinden etkilenmemektedir. Bu nedenlerle şimdiye kadar yapılmış dayanıklı temel bileşenlere dayalı yöntemlere alternatif olarak önerilmektedir.

KAYNAKLAR

BILLOR, N. , HADI, A. S. AND VELLEMAN, P. F. (2000). "BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators", Computational Statistics and Data Analysis, 34, 279-298.

CAMPBELL, N. A. (1980). "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation", Applied Statistics, 29 , 231-237.

CARONI,C. (2000). "Outlier Detection by Robust Principal Components Analysis. "Commun.Statist.-Simula.,29(1),139-151.

CROUX, C. RUIZ-GAZEN A. (2000). "High Breakdown Estimators for Principle Components: The Projection-Pursuit Approach" revisited, under revision, http://homepages.ulb.ac.be/~ccroux.

CROUX, C., AND HAESBROECK, G. (2000), ``Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, 87,603—618

HAWKINS, D. M. , BRADU, D. , KASS, G. V. (1984). "Location of Several Outliers in Multiple Regression Data Using Elemental Sets". Technometrics, 26,197-208.

HUBERT, M., ROUSSEEUW, P.J., AND VERBOVEN, S. (2002). "A Fast Robust Method for Principal Components with Applications to Chemometrics". Chemometrics and Intelligent Laboratory Systems, 60, 101-111.

HUBERT, M., ROUSSEEUW, P.J. (2002), "ROBPCA: A New Approach to Robust Principal Component Analysis".

KIRAL, G. ve BILLOR, N. (2001)."BACON Temel Bileşenler Analizi" 5.Ulusal Ekonometri ve İstatistik Sempozyumu".19-22 Eylül, ADANA

KIRAL, G. (2003). "A Comparision of The Recent Algorithms for The Identification Of Outliers in Data". PhD Thesis.Department of Mathematics, Institute of Natural and Applied Sciences, Cukurova University.

LIE, G. AND CHEN, Z. (1985). "Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components:Primary Theory and Monte Carlo", J. Amer. Statistics. Assosc., 80, 759-766.

A Comparison of the Multiple Outlier Detection Method for Multivariate Data by Simulation Study

ABSTRACT

Principle component analysis is a statistical technique used for reducing data dimension and/or constructing a set of uncorrelated variables. In some cases it is solely used as a technique of analysis itself while in other situations used as a data preparation technique for further analysis. In particular, it is preferred in testing high dimensional data since it provides dimensional reduction in data. But it is based on classical variance and covariance matrix therefore it is sensitive to outliers in data. So using robust principle component analysis is preferred to obtain reliable results in the existence of outliers.

In this study classical principal component analysis technique and two robust principle component techniques [Robust Principle Component Analysis (Hubert et al., 2002) and BACON Robust Principle Component Analysis (Kıral and Billor, 2001)] are compared using simulation.

Key Words : *Classic Principle Component Analysis, Robust Principle Component Analysis, BACON Algorithm, Outlier*