BACON YÖNTEMİNE DAYALI TEMEL BİLEŞENLER ANALİZİ

Gülsen KIRAL*

Nedret BİLLOR**

ÖZET

Çok değişkenli veri kümelerinde sapan değerlerin belirlenmesi için pek çok yöntem önerilmiştir. Bu konuda son yıllarda önerilen yöntemlerden BACON algoritması büyük veri kümeleri için hesapsal olarak etkin ve maskeleme ve swamping problemlerine karşı dayanıklı olan bir yöntemdir (Billor, Hadi ve Velleman, 2000). Bu çalışmada şimdiye kadar önerilmiş olan robust temel bileşenler analizi ile ilgili yöntemlere alternatif olan bir yöntem geliştirilmiştir. Bu yöntem BACON algoritmasını kullanarak sapan değerlerden etkilenmeyen temel bileşenlerin belirlenmesine dayalı olan bir yöntemdir. Bu yöntemin uygulanabilirliği iki veri kümesi kullanılarak gösterilmiştir.

ABSTRACT

BACON PRINCIPLE COMPONENTS ANALYSIS

Numerous methods have been suggested for outlier detection in multivariate data. In this paper we propose a new method for robust principle component analysis .The main idea is to obtain robust principal components based on robust covariance estimate of the clean data obtained from BACON algorithm (Billor et al., 2000). This method yields significant components that are free of outliers and that contain most of the information in a data matrix.

The applicability of this method is shown by using two different data sets.

* Dr., Çukurova Üniversitesi İİBF Ekonometri Bölümü

* Prof. Dr. Çukurova Üniversitesi FEF Matematik Bölümü

1.GİRİŞ

Bir veri kümesinde gözlemlerin çoğu tarafından önerilen modele uymayan gözlemlere **sapan değer** denir. Tek bir bileşen içerisinde büyük ölçüde etkili olan gözlem(ler); her bir değişkene tek değişkenli tekniklerin uygulanması ile kolayca belirlenmekle beraber çok değişkenli veri içerisinde bu gözlemlerin belirlenmesi o kadar kolay değildir. Bu gözlemler ancak her bir değişken içerisindeki gözlemin diğer değişkenlerle olan ilişkisi birlikte düşünüldüğü zaman belirlenebilir.

Çok değişkenli veri kümesinde bir veya iki sapan değer her bir gözlem için Mahalanobis uzaklığının hesaplanması ile saptanabilir. Ancak veri kümesinde pek çok sapan değer varsa Mahalanobis uzaklığı sapan değerlerin bulunmasında kullanışlı olmayabilir. Ayrıca çok değişkenli veri kümelerinde sapan değer olan gözlemlerin sapan değer olarak elde edilememesi (maskeleme) veya sapan değer olmayan gözlemlerin sapan değer olarak bulunması (swamping) problemleri ile karşılaşıldığında Mahalanobis uzaklığının kullanışsız bir yöntem olduğu çok iyi bilinmektedir.

Çok değişkenli veri kümelerinde sapan değerlerin belirlenmesi için bir *geleneksel* yaklaşım çok değişkenli normal dağılımdan geldiği varsayılan veri için bilinen *hipotez testi* tekniğidir. Ayrıca Wilks testinin (Wilks ,1963) uygulamaları olan çeşitli yöntemler (Bacon, Shone ve Fung ,1987 ve Caroni ve Prescott ,1992), Simonoff'un (1991) kümelemeye dayalı olarak tanımladığı yöntem, Atkinson ve Mulira (1993) 'nın sarkıt (stalactite) grafiği kullanılan yaklaşımlardandır. Çok değişkenli normal dağılımdan farklı dağılımlar için çok az sayıda çalışma yapılmıştır (Barnett ve Lewis 1994).

Maskeleme ve swamping problemlerinin varlığında sapan değerlerden etkilenmeden çok değişkenli veri kümesi ile ilgili analizi yapabilmek için çoğu zaman robust yöntemlerden yararlanılmaktadır. Bunun yanında son yıllarda birleştirilmiş

yöntemlerin (klasik ve robust yöntemlerin birleşimi ile tanımlı yöntemler) kullanımı da yaygınlaşmıştır.

Rousseeuw'nun (1984) minimum kovaryans determinant yöntemi (MCD), Rousseeuw'nun (1985) minimum hacimli elipsoid yöntemi (MVE), Rousseeuw ve van Zomeren'un (1990) MVE yi kullanarak tanımladığı sapan değer bulma yöntemi ve Rousseeuw ve van Driessen'in (1999) FAST-MCD yöntemi robust olarak bilinen yöntemlerdendir. Birleştirilmiş yöntemlere (hem robust hem de klasik yöntemlerin bir arada kullanılmasıyla elde edilen yöntemler) ise Hadi'nin (1992,1994) adımsal (stepwise) yöntemleri, Billor, Hadi ve Velleman'ın (2000) BACON yaklaşımı örnek olarak verilebilir.

Bu yöntemlerde sapan değerler; dağılımın merkezinden uzakta bulunan gözlemlerin belirlenmesi ile saptanırlar. Genel olarak kullanılan uzaklık

$$MD_i^2 = \left(x_i - T(X)\right)' \left(C(X)\right)^{-1} \left(x_i - T(X)\right) \qquad (1)$$

ile tanımlı Mahalanobis uzaklığıdır. *T(X); X* veri kümesinin ortalama vektörü ve *C(X);* örneklem varyans-kovaryans matrisidir. *T* ve *C* nin farklı tanımlamaları kullanılarak farklı robust ölçüler elde edilebilir.

Çok değişkenli veri kümesi içerisindeki çoklu sapan değerlerin belirlenmesi, incelenmek istenilen veri kümesindeki değişken sayısının fazla olması durumunda problemlidir. Veri matrisi X_{nxp} den elde edilecek bilgilerin daha küçük boyutlu ($k<p$) veri matrisinden elde edilmesi ve çok değişkenli veri kümesinin yorumlanması ve anlaşılmasına yardımcı olması için temel bileşenler analizinden (TBA) yararlanılır. Bu analiz sonucunda elde edilen temel bileşenlerin sapan değerlerin varlığı durumunda etkilendiği bilinmektedir. Temel bileşenlerin sapan değerlerden etkilenmeyecek şekilde elde edilmesi analizin doğruluğu için önemlidir. Ancak klasik kestiricilere dayalı olarak kullanılan yöntemler maskeleme ve swamping problemlerinin varlığında sağlıklı sonuç vermemektedir. Bu gibi durumlarda robust yöntemlerden yararlanılır. Bu konu ile ilgili ilk çalışma Campbell (1980) tarafından yapılmıştır. Campbell; robust M-kestiricisi yardımıyla temel bileşenleri belirleyip, değerlendirmede bilgi verecek olan ağırlıkları hesaplamıştır. Benzer bir çalışma Croux ve Haesbroeck (2000) tarafından yapılmıştır. Ancak bu yöntemler yüksek boyutlu veride etkin olarak kullanılamamaktadır. Li ve Chen (1985); projection

pursuit yöntemini önermiştir. Bu yöntem büyük veri kümelerine uygulanabilir, fakat hesaplaması çok zaman alıcı ve uygulaması pratik olmayan bir yöntemdir. Daha sonra Croux ve Ruiz-Gazen (1996 ve 2000) projection pursuit yöntemine göre daha az zaman alıcı olan robust temel bileşenler analiz yöntemini önerdiler. Ancak bu yöntem büyük boyutlu veri kümelerinde sayısal hesaplama problemi içermektedir. Ayrıca Caroni'de (2000); sapan değerleri belirlemede değerlendirmenin yapılacağı kritik değerlerin hesaplanması üzerinde bir çalışma yaparak Campbell'in yaklaşımının sapan değerlerin formal testi olarak kullanılabileceğini önermiştir.

Bir istatistiksel yöntemin; varsayılan model üzerindeki etkinliği, farklı tipteki sapan değerlere dayanıklılığı ve hesaplama ve uygulamadaki esneklikleri gibi istenen istatistiksel özellikleri bir arada bulundurması gerektirdiğinden, robust yöntemlerin geliştirilmesinde zorluklarla karşılaşılır. Bu nedenle uygulamada çok yaygın olarak kullanılmamaktadırlar.

Bu nedenle son yıllarda robust yöntemlere alternatif olabilecek hesapsal problemi olmayan, maskeleme ve swamping problemlerinden etkilenmeyen, büyük veri kümelerine rahatlıkla uygulanabilen algoritmalar tanımlanmıştır. Billor, Hadi ve Velleman (2000) tarafından tanımlanan BACON algoritması da bu konuda tanımlanmış en son algoritmalardan biridir.

Bu çalışmanın ikinci bölümünde, çok değişkenli veri kümelerinde sapan değerlerin belirlenmesi yöntemlerinden en yaygın olarak kullanılan yöntemlerden bazıları verildi. Üçüncü bölümde ise şimdiye kadar robust temel bileşenler analizi ile ilgili olarak tanımlanmış yöntemler özetlendi. Dördüncü bölümde robust temel bileşenler çerçevesinde BACON algoritması (Billor, Hadi ve Velleman, 2000) kullanılarak tanımladığımız BACON temel bileşenler analizi yöntemi verilmiştir (Kıral ve Billor, 2001). Son bölümde ise bilinen veri kümeleri üzerinde yöntemin uygulanabilirliği gösterilmiştir.

2. ÇOK DEĞİŞKENLİ VERİ KÜMELERİNDE BAZI SAPAN DEĞER BELİRLEME YÖNTEMLERİ

Çok değişkenli veri kümelerinde çoklu sapan değerlerin belirlenmesi problemi, veri matrisinin boyutun büyümesi ile maskeleme ve swamping problemlerini ortaya

çıkarmış ve 1980 lerden itibaren bilgisayar teknolojisindeki hızlı ilerleme ile bu problemlerin üstesinden gelebilecek yöntemler geliştirilmiştir. Halen günümüzde daha hızlı ve çok büyük veri kümelerinde çoklu sapan değerleri ortaya çıkarabilecek algoritmaların geliştirilmesi üzerine çalışmalar devam etmektedir.

Çok değişkenli veri içerisinde sapan değer belirleme ile ilgili çok sayıda yöntem bulunmaktadır (Atkinson, 1994, Barnett ve Lewis, 1994, Gnanadesikan ve Kettenring, 1972, Hadi 1992, Hawkins,1980, Rocke ve Woodruff 1996, Rousseeuw ve van Zomeron, 1990). Bu yöntemlerin çoğu sapan değerleri belirlemede gözlemlerin veri merkezine olan uzaklıkları hakkında bilgi veren Mahalanobis Uzaklığından yararlanmaktadır. Yüksek Mahalanobis uzaklığına sahip gözlemler sapan değer olarak belirlenmektedir. Bu uzaklığın hesabında sağlıklı sonuç elde edebilmek için sapan değerlere karşı dayanıklı kestiricilerin kullanımı tercih edilmektedir. Bu tip kestiricileri hesaplamamıza yardımcı olacak çok sayıda robust yöntem tanımlanmıştır (bkz: Hawkins (1980), Huber (1981), Chatterjee ve Hadi (1988), Barnet ve Lewis (1994), Atkinson ve Raini (2000)).

En yaygın olarak kullanılanları Minimum hacimli elipsoidi (**M**inimum **V**olume **E**llipsoid: **MVE**) (Rousseeuw, 1985), Minimum kovaryans determinantı (MCD) (Rousseeuw, 1984) yöntemleri yardımıyla elde edilen kestiricilerdir.

MVE yöntemi gözlemlerin en azından yarısını içine alan minimum hacimli elipsoidi bulmaya çalışır. Minimum hacmi veren gözlemlerin alt kümesinin ortalaması ve varyans-kovaryans matrisi robust ortalama ve varyans-kovaryans tahminleridir.

MVE hesaplanması için; veri kümesinden (X_{nxp}) rasgele olacak şekilde seçilen $p+1$ gözleme ait alt küme için ortalama ve varyans-kovaryans matrisi yardımıyla karşılık gelen Mahalanobis uzaklıkları hesaplanır. Alt kümedeki gözlem sayısı s ise elde edilen Mahalanobis uzaklıklarından minimum $s+1$ tanesini alarak yeni alt küme belirlenir. Alt kümede $n-h$ gözlem olana kadar yukarıdaki işlemler tekrarlanır $\left(h=\left[\left[n+p+1/2\right]\right]\right)$. Son adımda elde edilen alt kümeye ait Mahalanobis uzaklıkları yardımıyla bu alt kümeye karşılık gelen elipsoidin hacmi hesaplanır. Bu işlem $\binom{n}{p+1}$ kadar seçilen tüm alt kümeler için tekrarlanır. İçlerinden

minimum hacmi veren alt küme belirlenir. Bu alt kümedeki gözlemler temiz, dışında kalanları ise sapan değer olarak bildirilir.

Yöntemin bir avantajı %50 lik kırılma noktasına sahip olmasıdır (Lopuhaa ve Rousseeuw (1991)). Ancak gözlem sayısının yüksek olması durumunda hesaplanması zaman alıcı ve problemlidir.

MVE ye alternatif olarak minimum kovaryans determinant (**M**inimum **C**ovariance **D**eterminant: **MCD**) yöntemi önerilmiştir. Bu yöntemde de amaç n gözlem üzerinden klasik kovaryans matrisinin determinantı en küçük olan h gözlemin bulunmasıdır. Bu durumda yerel ve yayılım parametrelerinin MCD tahminleri sırasıyla bu h gözlemin ortalama ve kovaryans matrisleri olacaktır. Yöntemin kırılma noktası MVE yöntemi ile aynıdır. Ancak MCD nin asimtotik olarak normal olması (Butler, Davies ve Jhun, 1993) nedeniyle MVE ile karşılaştırıldığında avantajlara sahiptir. Yöntem MVE e göre istatistiksel olarak daha etkindir. MCD ye dayalı robust uzaklıklar MVE ye dayalı olarak elde edilenlere göre daha kesindir. Bu nedenle de çok değişkenli veri kümeleri içerisinde problemli gözlemleri belirlemeye daha uygun bir yöntemdir (Rousseeuw ve van Zomeren ,1990).

Tüm bu avantajlarının yanında bu tahmin edicilerin hesaplanması problemli ve zaman alıcıdır. Bu nedenle yaklaşık sonuçlar veren iteratif algoritmalar tanımlanmıştır. Bunlardan biri Rousseeuw ve van Driessen (1999) tarafından geliştirilen FAST-MCD (Fast Minimum Covariance Determinant) yöntemidir. Yöntemde rasgele seçilmiş (p+1) elemanlı alt kümeye dayalı olarak hesaplanan klasik tahmin ediciler yardımı ile Mahalanobis uzaklıkları hesaplanır. Minimum uzaklıklara dayalı olarak klasik tahmin ediciler yakınsama oluncaya kadar tekrar tekrar hesaplanır. K farklı başlangıç ile elde edilen K farklı tahmin edici içerisinden kovaryans matris determinantı en küçük olan kestirici seçilerek tahmin işlemi yapılır.

Yöntem küçük veri kümelerinde MCD ile aynı sonuçları verirken büyük veri kümelerinde benzer olarak tanımlı algoritmalar ile karşılaştırıldığında daha kesin sonuçlar vermektedir. Bu yöntemin bir avantajı da kesin tahmin değerlerini hesaplamasıdır.

Son yıllarda bu yöntemlere alternatif olarak Billor ve ark. (2000) BACON (Blocked Adaptive Computationally Efficient Outlier Nominators) adı verilen bir yöntem tanımlanmıştır. Bu yöntem Hadi (1992, 1994) ve Hadi ve

Simonoff (1993) yöntemlerine dayalı olup hesaplamaları örneklem büyüklüğüne bakmaksızın yapmaktadır. Yöntem ilk olarak sapan değerlerden arındırılmış olduğu varsayılan başlangıç temiz alt küme belirler. Başlangıç alt küme iki farklı şekilde belirlenmektedir. **Birinci yaklaşım** Mahalanobis uzaklığına, **ikinci yaklaşım** ise medyana bağlı olarak tanımlı uzaklığa dayalı olarak hesaplama yapmaktadır. Bu yöntemde esas amaç; sapan 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 sapan değer olarak belirlenirler. Temel alt küme dışında hiç gözlem kalmamışsa "*veri kümesi sapan değer içermemektedir*" denir.

Birinci yaklaşım robust değil fakat affine-equivarianttır (Herhangi bir **b** vektörü ve tekil olmayan bir **A** matrisi için $T(XA+\mathbf{b})=T(X)\mathbf{A}+\mathbf{b}$ eşitliği sağlanıyorsa T kestiricisine affine equivariant denir). Bunun yanında düşük kırılma noktasına sahiptir. Diğer yaklaşım ise affine-equvariant değildir ama medyan kullanılarak yönteme başladığımızdan dolayı robust bir yaklaşımdır. Sapan değerlerin varlığında daha sağlıklı sonuçlar vermektedir. Ayrıca kırılma noktası daha yüksektir (%40 civarında) (Billor ve ark. (2000)). Çalışmada daha robust olması bakımından ikinci yaklaşım kullanılmıştır.

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

Gözlemlerin bloklanması nedeniyle hesapsal açıdan etkin bir yöntemdir. Diğer yöntemlere göre bu yöntemdeki iterasyon sayısı daha azdır. İterasyonların her biri kovaryans matrisinin hesaplanması ve tersinin alınmasını gerektirir. Fakat iterasyon sayısı n örneklem büyüklüğünün artması ile büyümez ve hesaplanan n uzaklığın sıralanmasını gerektirmez.

3. ROBUST TEMEL BİLEŞENLER ANALİZİ

Tek bir bileşen içerisinde sapan değer olan gözlemler tek değişkenli tekniklerin uygulanması ile rahatlıkla bulunabilirken yüksek boyutlu veri de problemler ile karşı karşıya kalınmaktadır. Bunlardan en önemlileri hesaplama ve

çoklu iç ilişki problemi olarak bilinmektedir. Bu problemlerin üstesinden gelmek için temel bileşenler analizinden (TBA) yararlanılmaktadır. Böylece yeni ilişkisiz değişkenlerin bir kümesi oluşturularak boyut indirgenmesi yapılır. Bu işlem örneklem kovaryans matrisinin özvektörlerine dayalı olarak yapılmaktadır. Kovaryans matrisi örneklem ortalamasına dayalı olduğundan sapan değerlerin varlığında etkinliğini kaybetmektedir. Bu nedenle son yıllarda robust kovaryans kestiricisi kullanım yardımı ile tanımlanan robust temel bileşenler analizinin kullanımı yaygınlaşmıştır.

Robust temel bileşenler analizi robust kovaryans ya da korelasyon matrisinin özdeğer ve özvektörlerinin hesaplanması ile elde edilmektedir. Bu konu ile ilgili ilk çalışma Campbell (1980) tarafından yapılmıştır. RPCA (Robust Principle Component Analysis) adını verdikleri yöntem temel bileşenler analizi içerisinde varyans-kovaryans matrisinin robust M-kestiricisinin kullanımı ile tanımlanmıştır. Bu yöntemde amaç; sapan değerlerin etkisini ortadan kaldıracak gerçek ağırlıkları bularak tüm veri kümesini temsil eden gerçek varyans-kovaryans matrisini elde etmektir. Yöntem her gözlemin her bir bileşen üzerindeki robust tahminine katkısını göstermek amacı ile ağırlıklar hesaplar. Düşük ağırlıklar sapan değerleri vurgular. Yöntem veri yapısını incelenme ve sapan değer testini aynı anda yapılmasına izin verdiğinden kullanışlıdır. Fakat M-tahmin edicisine dayalı olduğundan sapan değerlerin küçük bir oranı için koruma sağlamaktadır. (Rousseeuw ve van Zomeron (1990)). Campbell (1980) ayrıca çok değişkenli veride sapan değerlerin kullanılması ile ilgili olarak grafiksel bir yöntem önermiştir. Bu grafik robust M-kestircisi kullanılarak hesaplanan Mahalanobis uzaklığını karesine ait olasılık grafiğidir.

Ardından Li ve Chen (1985); tüm veri kümesi hakkında daha detaylı bilgi edinmemize yarayacak düşük boyutlu veriyi Projection Pursuit (PP)'ye dayalı olarak hesaplayan bir çözüm önerdiler. PP yöntemi; çok değişkenli verinin bir doğru ya da bir düzlem üzerindeki lineer izdüşümleri yardımı ile orijinal verinin yapısını ortaya çıkarmaya çalışır. Burada tüm veri kümesi hakkında en fazla bilgiyi açığa çıkaran küçük boyutlu izdüşümü bulma amacı ile veri kullanılır. Li ve Chen'nin amacı; en büyük robust ölçeklemeye sahip izdüşümü alınmış gözlemlerin doğrultusunu belirlemektir. Birbirini izleyen adımlarda her yeni doğrultu önceki tüm doğrultulara dik olacak şekilde belirlenmektedir. Yüksek boyutlu veri kümelerinde hatta ve hatta

p>n iken de dahil olmak üzere iyi sonuç veren bir algoritmadır. Ama hesapsal problemler içermektedir.

PP ye dayalı yöntemlerde karşılık gelen etki fonksiyonunun sınırlandırılmamış olması yerel robustlıkta eksikliğe sebep olmaktadır (Croux ve Ruiz-Gazen (2000), Croux ve Filzmoser (2001) Pires ve Branco (2001)). Bunun yanında PP ye dayalı kestiricilerin nasıl hesaplanacağı açık değildir. Croux ve Ruiz-Gazen (1996,2000) PP algoritmasından çok daha etkin olan C-R algoritmasını önerdiler. Bu yöntem kısıtlamalar altında bir maksimizasyon probleminin çözümü olup küçük boyutlu veri kümelerinde iyi çalışmasına rağmen büyük boyutlu veri kümeleri için hesapsal problemler içermektedir. Croux ve Ruiz-Gazen (2000) C-R algoritmasını ile PP algoritması'nın etkinliklerini ve birkaç robust ölçek ile robustlıklarını karşılaştırmışlardır. Hubert ve ark.(2001) a faster two-step algorithm adını verdikleri (RAPCA) temel bileşenler analizinde yansımaya dayalı olarak tanımlanan boyut indirgemesine yardımcı olacak C-R algoritmasını geliştirerek sunmuşlardır. Yöntem C-R algoritmasından daha etkin ve veride gözlem sayısından çok değişken olması durumunda da sağlıklı sonuçlar verebilmektedir. Diğer taraftan Caroni (2000); Campbell'in (1980) yapmış olduğu çalışma ile ilgili bir simülasyon çalışması yapmıştır. Bu çalışmada $x_i \sim N_p(\mu,\sigma)$ _i=1,2,...,n sıfır hipotezi altında RPCA içinde gözlemlerin ağırlıkları için kritik değer belirleme amaçlanmıştır. Kritik değerlerin hesaplanması ile düşük ağırlıklara sahip olan gözlemler dolayısıyla sapan değerler belirlenmektedir.

Robust 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 sapan 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 robust temel bileşenlerin belirlenmesini sağlayan bir algoritma tanımlanmıştır.

4. BACON TEMEL BİLEŞENLER ANALİZİ (BTBA)

Algoritma

Adım 1: Temel altküme ; BACON algoritmasında tanımlı yaklaşımlardan biri (*Y1 veya Y2*) 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

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

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 ($h = \left[\left|\frac{n+p+1}{2}\right|\right]$).

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 sapan değer olarak tanımlanır.

Adım 6: Sapan 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); $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ olacak şekilde hesaplanır.

Adım 8: $U=(u_1,u_2,\ldots,u_p)$ olarak tanımlanmak üzere yeni temel bileşenler

$$Y = U^T X$$

elde edilir.

Adım 9:Elde edilen yeni temel bileşenler matrisine BACON algoritması uygulanarak sapan değerler belirlenir.

Robust uzaklıklar, sayısal olarak bize hangi gözlemlerin sapan değer olduğu hakkında bilgi vermekle beraber, bu gözlemler grafiksel olarak daha kolaylıkla sergilenebilir. Bu nedenle elde edilen yeni temel bileşenlere BACON algoritması uygulandıktan sonra elde edilen ortalama ve kovaryans matrisleri yardımıyla karşılık gelen Mahalanobis uzaklıkları hesaplanır. Bu uzaklıkların

- küp köküne ait Quantile-Quantile (Q-Q) veya
- klasik Mahalanobis uzaklıkların robust BACON uzaklıklarına karşı veya
- robust BACON uzaklıklarının indis

grafiklerine bakılarak tüm veri kümesi içerisinde sapan değer olan gözlemler kolaylıkla belirlenir. Ayrıca benzer olarak hangi gözlemlerin hangi temel bileşen üzerinde en fazla sapmaya neden olduğunun belirlenmesi; herbir temel bileşene ait Q-Q grafikleri yardımıyla yapılabilir.

5. UYGULAMA

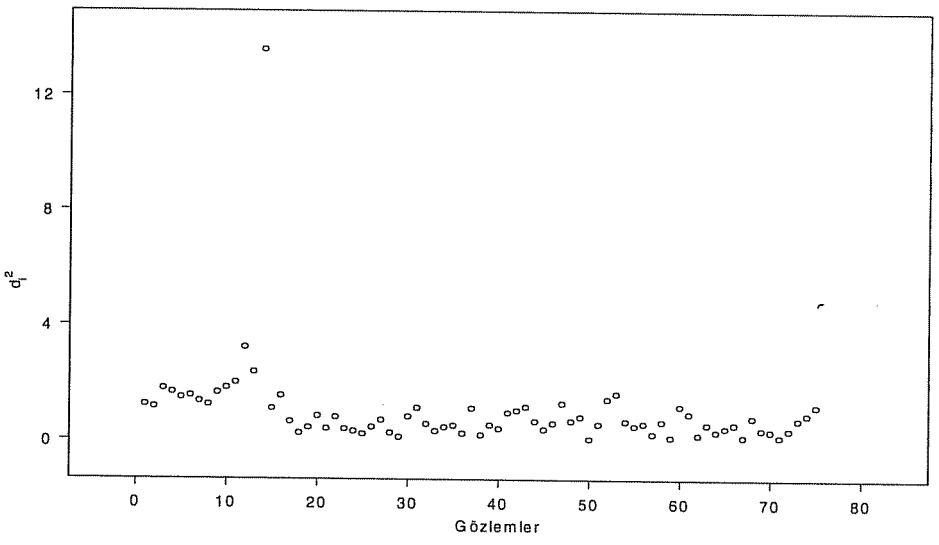
Önerilen yöntemin uygulanabilirliği iki veri kümesi üzerinde gösterilebilir. Bu veri kümelerinden birincisi, Hawkins-Bradu ve Kass (1984) sapan değerler hakkında ön bilgimiz olduğu için yöntemin performansını irdelememiz açısından çok kullanışlı bir veri kümesidir. Diğer veri kümesi de (Philips veri Kümesi, 1999) büyük bir veri kümesi olduğu için yine önerilen yöntemin büyük veri kümelerine uygulanması durumundaki başarısını göstermesi açısından ayrı bir öneme sahiptir.

Örnek 1. Hawkins-Bradu ve Kass veri kümesi (HBK)

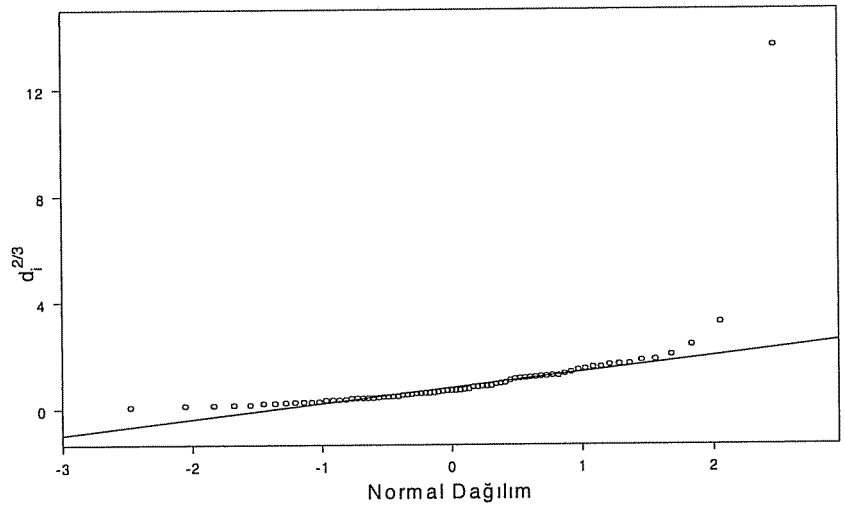
Bu veri kümesi Hawkins-Bradu ve Kass (1984) tarafından 1-14 arası gözlemlerin sapan değer olduğu bilinen, özel amaç için oluşturulmuş, $n=75$ ve $p=3$ olan bir veri kümesidir. Bu veri kümesi ile ilgili klasik Mahalanobis uzaklığına ait indis grafiği (Şekil 1) incelendiğinde gerçekten sapan değer olduğu bilinen

gözlemlerin çok azı sapan değer olarak görülmektedir (maskeleme problemi). Benzer bilgiler Mahalanobis uzaklığının küp köküne ait Q-Q grafiğinin (Şekil 2) incelenmesi ile de söylenebilir.

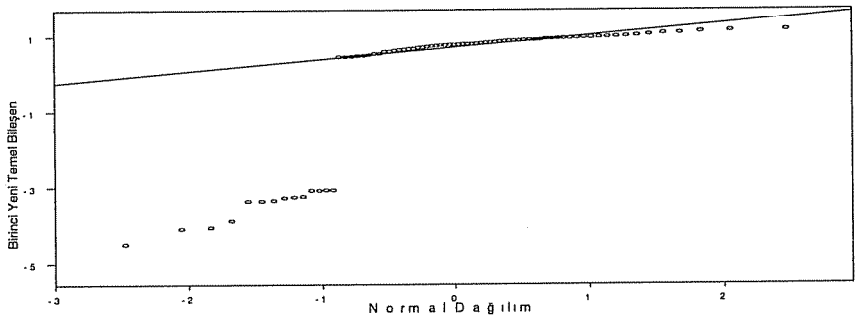
Bu bize klasik yöntemlerin sapan değerlerin varlığında sağlıklı sonuç vermediğini açık bir şekilde göstermektedir. Bu problemden kurtulabilmek için sapan değerlerden etkilenmeyecek kestiriciler kullanarak işlemlerin yapılması gerekmektedir. Örneğin robust M-kestiricisinin kullanımı ile tanımlı RTBA (Campbell, 1980) bu veri kümesine uygulandığında veri kümesindeki ilk 14 gözlemin sapan değer olduğu rahatlıkla söylenebilmektedir. Bu veri kümesine BTBA yöntemini uyguladığımızda elde ettiğimiz yeni temel bileşenlere ait Q-Q grafikleri Şekil 3 (a)-(c) de görülmektedir. Grafikler incelendiğinde 1. temel bileşen üzerinde ilk 14 gözlemin, 2. temel bileşen üzerinde 47, 52 nolu gözlemlerin ve 3. temel bileşen üzerinde de 13, 14, 53 nolu gözlemlerin etkili oldukları görülmektedir. Bu algoritma ile ilgili Mahalanobis uzaklığının indis (şekil 4) ve küp köküne ait Q-Q grafiği (Şekil 5) incelendiğinde de ilk 14 gözlemin ciddi anlamda problemli gözlemler oldukları hemen söylenebilmektedir.



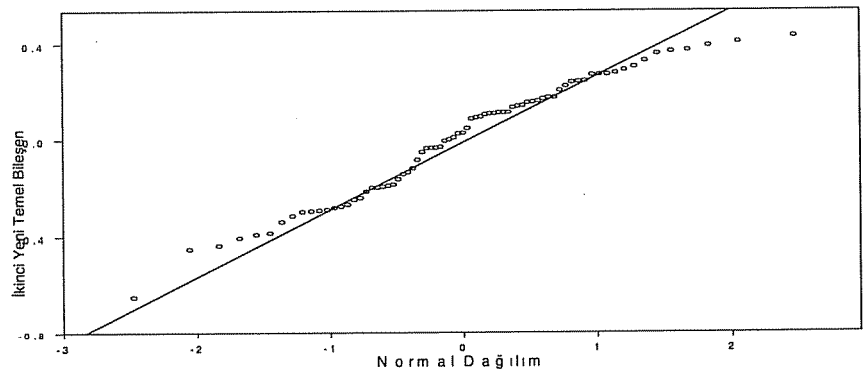
Şekil 1: Klasik Mahalanobis uzaklıkları için indis grafiği



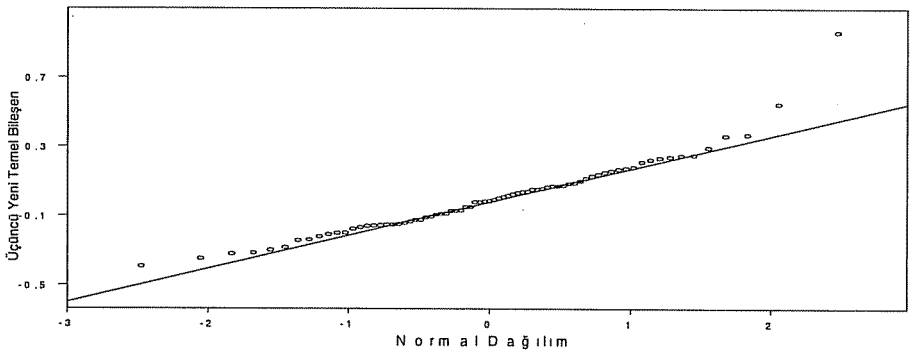
Şekil 2: Klasik Mahalanobis Uzaklığına ait Q-Q grafiği



(a)

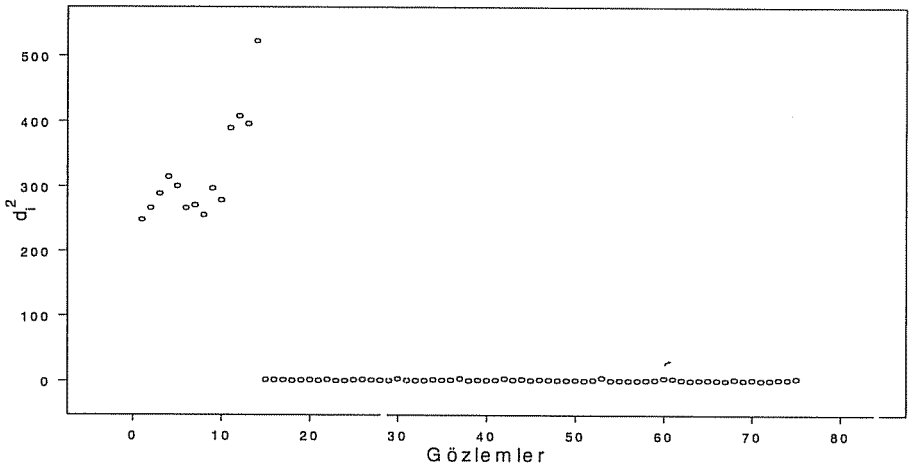


(b)

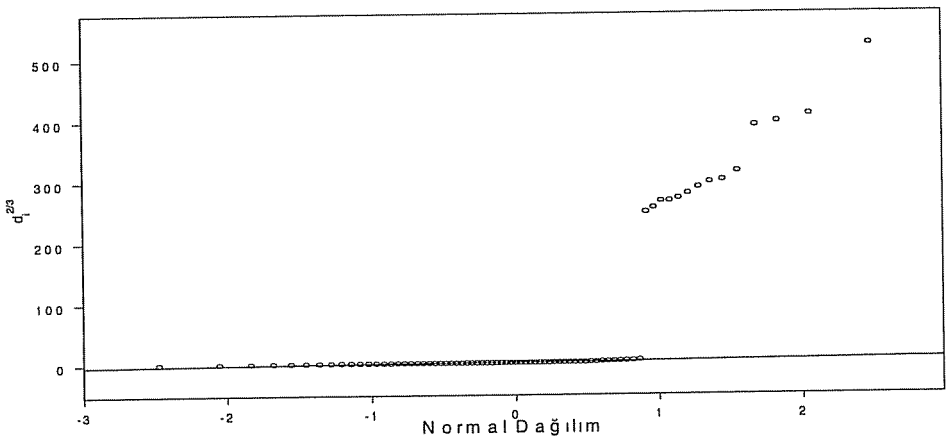


(c)

Şekil 3 (a,b,c):Hawkins-Bradu-Kass veri kümesinin her bir yeni temel bileşenine ait Q-Q grafikleri



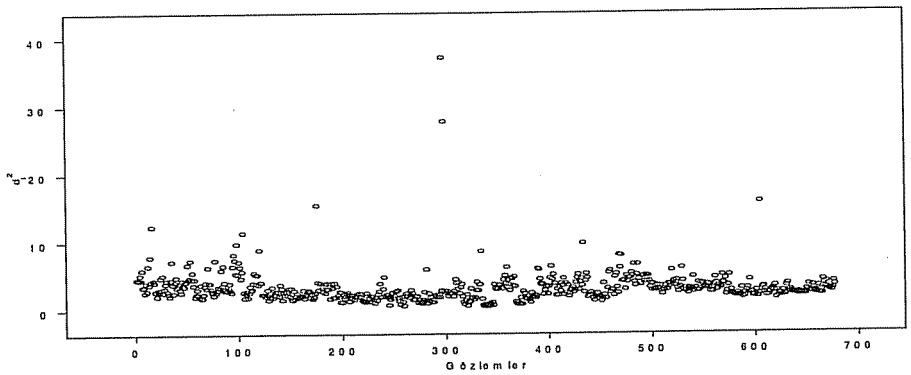
Şekil 4: BTBA den elde edilen robust Mahalanobis uzaklığına ait indis Grafiği



Şekil 5: BTBA den elde edilen robust Mahalanobis uzaklığına ait Q-Q grafiği

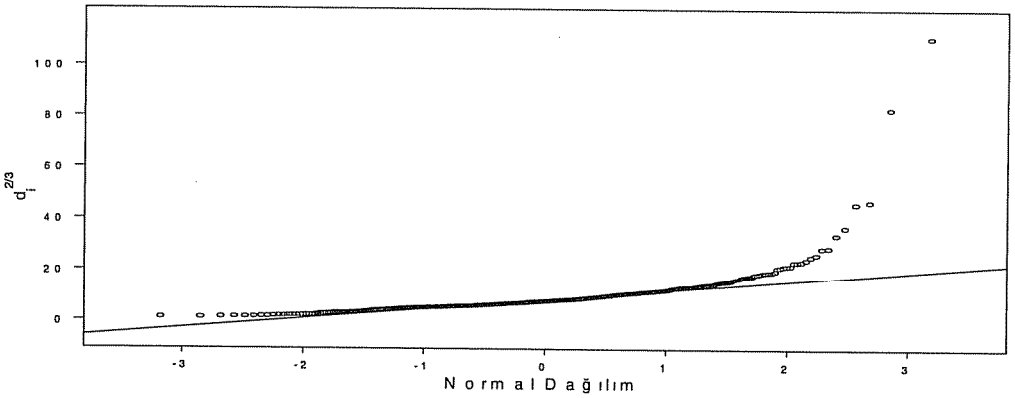
Örnek 2. Philips veri kümesi

Philips Mecoma (Hollanda), TV sehpası üretilen bir merkezde işlemler ince metal levhalarla yapılmakta ve presleme ile levhaya şekil verilmektedir. Yeni üretim tekniği ile çalışmalara başlanmış ve bu tekniğin kullanılabilirliği araştırılmak amacı ile 677 farklı parça üzerinde araştırma yapılmıştır (Rousseeuw ve van Driessen, 1999). Her bir parça üzerinde 9 farklı karakter için ölçümler yapılarak bu karakterler arasındaki ilişkinin olup olmadığı ve şekil bozukluklarının oluşup oluşmadığı araştırılmak istenmiştir.

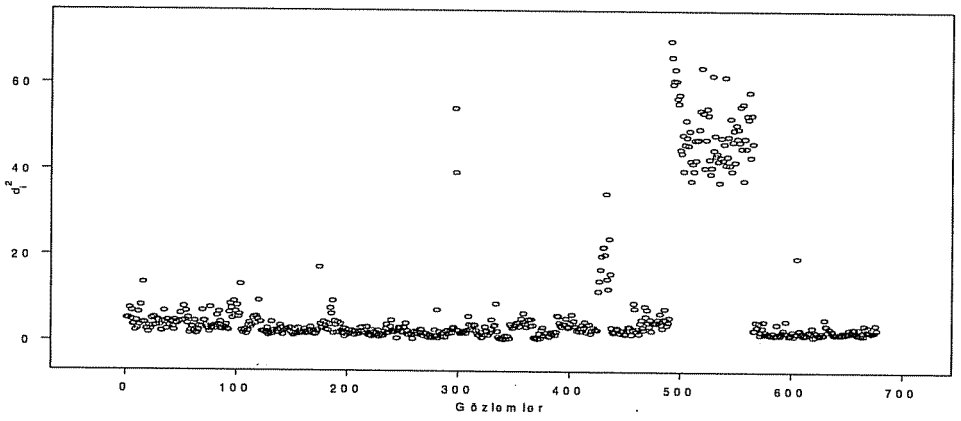


Şekil 6: Philips Veri Kümesinin Klasik Mahalanobis Uzaklığına ait İndis Grafiği

Örneğin; Rousseeuw ve van Driessen (1999) yapmış oldukları çalışmada MVEyöntemine dayalı robust uzaklıklar kullanarak çalışmalarını yapmış ve 491-565 nolu gözlemlerin grup halinde sapan değer olduklarını bulmuşlardır. Başka robust kestiriciler kullanılarak da benzer sonuçlar elde edilebilir. Ancak gözlem ve parametre sayısının fazla olması nedeniyle çoğu yöntemde hesapsal problemlerle karşılaşılması aşikardır. Hatta bazı yöntemlere uygulanması söz konusu dahi değildir. Campbell (1980) in RTBA yöntemi buna örnek olarak verilebilir.



Şekil 7: Philips Veri Kümesinin Klasik Mahalanobis Uzaklığına ait Q-Q grafiği



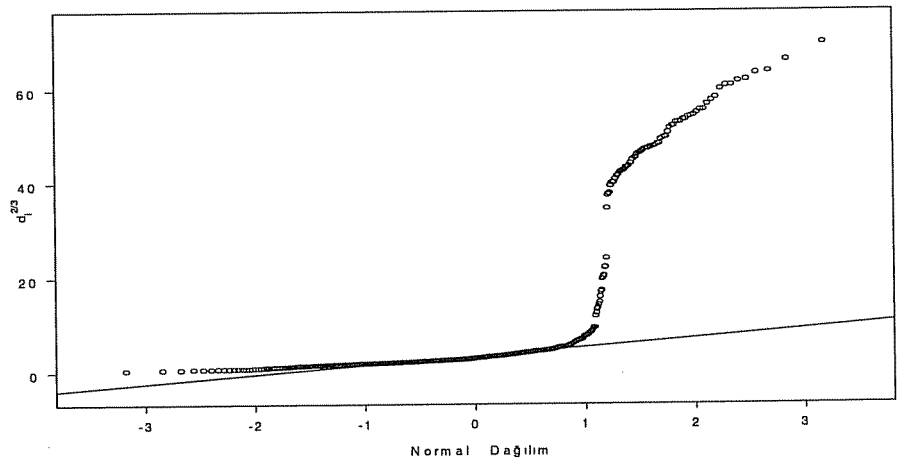
Şekil 8: Philips veri kümesine BTBA uygulandıktan sonra elde edilen Mahalanobis uzaklığının indis grafiği

Şimdi bu veri kümesini önerdiğimiz BTBA yöntemi kullanarak inceleyelim. İnceleme sonucu elde edilen yeni temel bileşenlere ait Q-Q grafikleri incelendiğinde

1. temel bileşen üzerinde16, 297, 298, 491-494, 500, 503, 507, 517, 519, 521, 523, 526, 550 nolu gözlemler,
2. temel bileşen üzerinde 16, 85, 83 nolu gözlemler,
3. temel bileşen üzerinde 16, 297, 298, 435, 521, 524 nolu gözlemler
4. temel bileşen üzerinde 16, 95, 96, 98, 104, 605 nolu gözlemler
5. temel bileşen üzerinde 116, 120 ve 175 nolu gözlemler
6. temel bileşen üzerinde 50 nolu gözlem
7. temel bileşen üzerinde 175, 426-437, 491-567, 605 nolu gözlemler
8. temel bileşen üzerinde 16, 297, 298, 426-437, 491-565, 605 nolu gözlemler
9. temel bileşen üzerinde 16,297,298,426-565,605 nolu gözlemler

etkili olarak görülmektedirler. BTBA yapıldığında elde edilen Mahalanobis uzaklığının indis ve küp köküne ait Q-Q grafikleri (Şekil 8-Şekil 9) incelendiğinde ise 491-565 nolu gözlemlerin ayrı bir grup olarak veriden ayrıldığı ve sapan değer olduğu hemen söylenebilir.

Buradan algoritmanın bizi robust yöntemler gibi sağlıklı sonuçlara ulaştırdığı söylenebilir.



Şekil 9: BTBA den elde edilen Mahalanobis uzaklığına ait Q-Q grafiği

Sonuç

Bu çalışmada robust temel bileşenler analizine dayalı olarak daha önce önerilen yöntemlere alternatif olabilecek bir algoritma verilmiştir. Çok değişkenli veri kümeleri içerisinde çoklu sapan değerleri bulmayı amaçlayan bu algoritma; büyük veri kümelerine (*1 milyon gözlem için bile*) uygulanabilmekte, model üzerinde

çok küçük etkisi olabilecek gözlemleri belirleyebilmekte, hesapsal problem içermemektedir. Bu nedenlerle şimdiye kadar yapılmış robust temel bileşenlerle ilgili yöntemlere alternatif olarak önerilmektedir.

KAYNAKLAR

1. Atkinson, A. C. ve Mulira H. M. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers", *Statistics and Computing*, 3, 27-35.
2. Atkinson, A., (1994). Fast very robust methods for the detection of multiple outliers. J. Amer. Statist. Assoc.89, 1329–1339.
3. Atkinson, A.C. (1986) "Masking Unmasked", Biometrica, 73,3,533-541
4. Atkinson, A.C., Riani, M., (2000). Robust Diagnostic Regression Analysis. New York; Springer
5. Bacon-Shone, J., and Fung, W.K. (1987), "A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data", *Journal of the Royal Statistical Society* (C), 36, No.2, 153-162.
6. Barnett, V. .and Lewis, T. (1994), *Outliers in Statistical Data* ,3rd edition, *New York: John Wiley and Sons.*
7. 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.
8. Butler, R. W., Davıes, P.L., Jhun, M., (1993). Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics, 21, 1385-1400.
9. Campbell, N. A. (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation", *Applied Statistics*, 29 , 231-237.

10. Caroni, C. and Prescott, P. (1992), "Sequential Application of Wilk's Multivariate Outlier Test", *Applied Statistics*, 41. 355-364.
11. Caroni, C. (2000) "Outlier Detection by Robust Principal Components Analysis", *Commun. Statist.-Simula.*, 29(1), 139-151.
12. Chatterjee, S., Hadı, A.S., (1988). Sensitivity Analysis in Linear Regression. New York: John Willey
13. Croux C., Ruiz-Gazen A. (1996) "A Fast Algorithm for Robust Principal Components Based on Projection Pursuit", COMPSTAT 96, Physica-Verlag, 211-216.
14. Croux, C. Fılzmoser P., (2001). A Projection-Pursuit based Measure of Association between two Multivariate Variables. In Preparation
15. Croux, C., Abd Ruiz-Gazen, A. (2000), " High Breakdown Estimators For Principal Components: The Projection-Pursuit Approach Revisited", Under Revision.
16. Croux, C., and Haesbroeck, G. (2000), "Principal Robust Estimators of the Covariance or Correlation Efficiencies," Biometrika, 87, 603-618.
17. Fung, W.K. (1993);"Unmasking Outliers and Leverage Points:A confirmation", J. Amer. Statist. Asso., 88, 515-519.
18. Gnanadesikan, R., Kettenring, J., (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28, 81–124.
19. Hadi , A. S. (1992), "Identifying Multiple Outliers in Multivariate Data" , *Journal of the Royal Statistical Society,* series(B), 54, 761-771.
20. Hadi, A. S. (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples", *Journal of the Royal Statistical Society*, series(B), 56, No. 2.
21. Hadi, A.S. and Simonoff, J. S. (1993), Procedures for the Identification of Multiple Outliers in Linear Models", *Journal of the American Statistical Association*, Vol. 88,414,1264-1272.
22. Hampel, F. R.,Ronchetti, E. M., Rousseeuw, P. J., and Stahel , W. A.(1986), *Robust Statistics: The Approach based on Influence Functions*, New York: John Wiley and Sons.

23. Hawkins, D. M. , Bradu, D. And Kass, G. V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets", Technometrics, 26,197-208.
24. Hawkins, D., (1980). Identi2cation of Outliers. Chapman & Hall, London.
25. Huber, P., (1981). Robust Statistics. New York: John Wiley and Sons
26. Hubert, M.,Rousseeuw, P.J., And Branden, K., V. (2001), "ROBPCA: A New Aprroach To Robust Principal Component Analysis" To appear in TechnometricsAvailableat http://www.wis.kuleuven.ac.be/stat/robust.html
27. Kıral G. ve Billor N. (2001). "BACON Temel Bileşenler Analizi" V. Ulusal Ekonometri ve İstatistik Sempozyumu,Ç.,Ü., Adana
28. Lı, G. And Chen, Z. (1985): Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. J. Amer. Statist. Assoc., 80, 759-766.
29. Lopuhaä, H.P. , Rousseeuw, P.J., (1991). Breakdown points of afine equavariant estimatos of multivariate location and covariance matrices. The Annals of Statistics, 19, 229-248.
30. Paul, S.R. and Fung, K. Y. (1991), "A Generalization Extreme Studentized Residual Multiple Outlier Detection Procedure in Linear Regression", Technometrics, 33,229-348.
31. Pıres, A.M. , Branco, J.A., (2001). Projection-Pursuit Approach for Robust Linear Discriminant Analysis. Preprint, Instituto Superior Tecnico, Dept. of Math
32. Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Journal of the American Statistical Association. 91, 1047–1061.
33. Rosner, B. (1975); "On the Detection of Many Outliers". Technometrics, 17,217-227.
34. Rousseeuw P.J.(1983),"Regression Techniques With High Breakdown Point", *The IMS Bulletin*, 12, 155.
35. Rousseeuw P.J.(1985), "Multivariate Estimation With High Breakdown Point in Mathematical Statistics and Applications", Vol B ,eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel, 283-297.

36. Rousseeuw, P. J. (1984), "Least Median of Squares Regression", *Journal of the American Statistical Association*, 79, 871-886.
37. Rousseeuw, P. J. ve van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points (with discussion)", *Journal of the American Statistical Association*, 85, 633.
38. Rousseeuw, P. J. ve Leroy, A. M. (1987), *Robust Regression and Outlier Detection, New York: John Wiley and Sons.*
39. Rousseeuw, P. J. ve van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Journal of the American Statistical Association,* 41, 212-223.
40. Simonoff, J. S. (1991), "General Approaches to Stepwise Identification of Unusual Values in Data Analysis", *Directions in Robust Statistics and Diagnostics:Part II*, W. Stahel and S. Weisberg, eds.,Springer –Verlag: New York, 223-242.
41. Steele J. M. And Steiger, W. L. (1986), " Algorithms and Complexity for Least Median of Squares Regression", Discrete Applied Mathematics, 13,509-517.
42. Wilks, S. S. (1963), "Multivariate Statistical Outliers", *Sankhya*, A25, 407-426.