ÇOK DEĞİŞKENLİ AYKIRI DEĞER TESPİTİ İÇİN KLASİK VE DAYANIKLI MAHALANOBİS UZAKLIK ÖLÇÜTLERİ: FİNANSAL VERİ İLE BİR UYGULAMA

M. Fevzi ESEN[1], Mehpare Timor[2]

Öz

Çok değişkenli veri setlerinde aykırı değerlerin varlığı anakütle parametre tahminini zorlaştırmakta ve hata varyansını arttırarak kullanılan istatistiki testin gücünü azaltmaktadır. Bu durum, değişkenlerin eşit varyansa ve çok değişkenli normal dağılıma sahip olduğu varsayımlarından sapmalara sebep olmaktadır. Çok değişkenli aykırı değer tespitinde kullanılan tekniklerden biri olan Mahalanobis uzaklığı, aykırı değişkenlere karşı hassas ölçütler olan çok değişkenli ortalamalar ve kovaryans matrisine dayalı olarak hesaplanmakta; çok değişkenli veri setlerinde aykırı gözlemlerin tespitinin engellenmesi veya normal gözlemlerin aykırı gözlem olarak tespit edilmesi problemlerine karşı dayanıklı ölçütlerle de kullanılmaktadır. Bu çalışmada, çok değişkenli aykırı değer tespitinde kullanılan klasik ve dayanıklı Mahalanobis ölçütlerinin aykırı gözlem tespitlerinin karşılaştırılması amaçlanmıştır. Uygulama verisi olarak, Ocak 2013 – Aralık 2017 döneminde New York ve NASDAQ borsasında yatırımcılar tarafından gerçekleştirilen 1.239.507 adet hisse senedi alım ve satım işlemi kullanılmıştır. Aykırı işlemlerin tespitinde miktar ve hacim değişkenleri ele alınarak, her bir işlem için klasik ve dayanıklı ölçütlere dayalı uzaklık skorları hesaplanarak, söz konusu teknikler karşılaştırılmıştır. Çalışma sonucunda, klasik Mahalanobis ölçütü ve En Küçük Hacimli Elipsoid ile tespit edilemeyen maskelenmiş aykırı gözlemlerin, Hızlı Minimum Kovaryans Determinant yöntemiyle tespit edilmiş olduğu; söz konusu yöntemin finans uygulama alanında çok değişkenli veri setlerinde aykırı gözlemlerin tespiti için kullanılabilecek etkin bir veri madenciliği yöntemi olduğu sonucuna ulaşılmıştır.

Anahtar Kelimeler: *Aykırı Değer Tespiti, Mahalanobis Uzaklığı, Dayanıklı Ölçütler*
Jel Sınıflandırması: *C1, C5*

CLASSICAL AND ROBUST MAHALANOBIS DISTANCE MEASURES FOR OUTLIER DETECTION: AN APPLICATION IN STOCK EXCHANGES

Abstract

The existence of outliers in multivariate data sets contaminates the parameter estimations and reduces the power of the statistical test by increasing the variance of the errors. This situation leads to deviations from the assumptions that the variables have equal variance and multivariate normal distribution. Mahalanobis distance is one of the techniques frequently used in multivariate outliers and it is calculated on the basis of multivariate location and covariance matrix, which are sensitive measures against outliers. In addition, due to the problems such as misidentification of a normal observation as an outlier and the presence of masking of an outlier, robust measures have been used. In this study, it is aimed to compare the performance of classical and robust Mahalanobis measures. 1.239.507 stock transactions executed by investors between the periods of January 2013 - December 2017 in New York Stock Exchange and NASDAQ are used for analysis. In order to determine outlying transactions, volume and value of trade have been analysed. Mahalanobis distances based on classical and robust measures have been calculated for each transaction and the measures are compared. As a result, the masked observations which cannot be detected by classical and robust Minimum Volume Ellipsoid measures, have been detected as outlying by Fast - Minimum Covariance Determinant (Fast MCD) measure. It has been concluded that Fast MCD can be used as an efficient estimator of multivariate location and scatter in presence of masked data for multivariate datasets in financial applications.

Key Words: *Outlier Analysis, Mahalanobis Distance, Robust Measures*
Jel Codes*: C1, C5*

[1] Dr. Öğretim Üyesi, İstanbul Medeniyet Üniversitesi, fevzi.esen@medeniyet.edu.tr, https://orcid.org/0000-0001-7823-0883
[2] Prof. Dr., İstanbul Üniversitesi, İşletme Fakültesi, Sayısal Yöntemler A.B.D., timorm@istanbul.edu.tr, https://orcid.org/0000-0002-9782-545X

DOI: 10.18092/ulikidince.579570
Makalenin Geliş Tarihi (Recieved Date): 18-06-2019
Yayına Kabul Tarihi (Acceptance Date): 21-08-2019

1. Giriş

Veri seti karakteristiğine uygun olmayan ve veri setindeki diğer gözlemlerden belirgin bir şekilde sapan gözlemler aykırı gözlem olarak adlandırılmaktadır (Hawkins, 1980). Veri setinin gözlenen veya teorik olarak tahmin edilen örüntüsünün aksine farklı örüntüler içeren aykırı gözlemler, verinin üretilme veya ölçüm sürecinden kaynaklandığı gibi, örneklem hatasından veya örneklem dağılımına ilişkin varsayımlardan da kaynaklanmaktadır. Veri setinin kalitesini olumsuz olarak etkileyen aykırı değerlerin veri setinden çıkartılması, gerçekleştirilecek istatistiki analizlerin sonuçlarının güvenilirliği açısından önemli olacağı için, veri setinde bulunan aykırı değerlerin doğru olarak tespit edilerek veri setinden çıkartılması, normalize edilmesi veya dönüştürülerek homojenliğin sağlanması önerilmektedir (Rousseeuw ve Leroy, 1987).

İstatistiksel modellerde parametre tahminlerinin iyileştirilmesi, çarpık dağılımlı veri setlerinin normalleştirilmesi ve varyansların homojen dağılımının sağlanması veya veri setine ilişkin tanımlayıcı istatistiklerin belirlenmesinde aykırı değerlerin tespiti sıklıkla yapılmaktadır. Bunun yanı sıra; literatürde aykırı değer analizi, veri setinde çoğu gözlemden uzak, uç değerlerde konumlanmış ve kolektif olarak değerlendirildiğinde anlamlı ve spesifik örüntüler içeren aykırı gözlemlerin tespit edilmesi gibi geniş bir uygulama alanına sahiptir.

Aykırı gözlemlerin tespitinde kullanılacak yöntemin seçimi, verinin hacmi, dağılımı ve değişken sayısına bağlı olarak değişmektedir. Veri setinin standart sapmasına bakarak veya kutu grafiği, Dixon testi, Weisberg t-testi, Walsh ve Grubbs t-testi gibi klasik yöntemlerin yanı sıra; regresyona dayalı Cook uzaklığı, uyumlar arasındaki fark (DFFITS) istatistiği, kaldıraç değerleri, S_i istatistiği gibi yöntemlerle aykırı değer tespiti yapılmaktadır. Sıfıra yakın veya negatif değerli gözlemlerin bulunduğu çok değişkenli ve yüksek hacimli verisetlerinde aykırı değerlerin tespiti için gözlemler arası ilişkiye dayalı Mahalanobis ölçütü önerilmektedir (Johnson ve Wichern, 2002).

Mahalanobis ölçütü, finansal piyasalarda hile tespiti ve iflas tahminlemesi (Cho vd., 2010; Pozollo vd., 2014; Stöckl & Hanke, 2014; Carminati vd., 2015; Vukovic, 2015), çok değişkenli modeller ile tahminleme (Jaffel vd., 2015; Pompella ve Dicanio, 2017; Qiu vd., 2017), karar destek sistemleri (Arteaga vd., 2016; Suo vd., 2018), hasta takip ve klinik acil uyarı sistemleri (Wang vd., 2011; Haldar vd., 2017), hata - arıza tespiti ve izleme (Carrato, 2018; Shang vd., 2018), kümeleme ve sınıflandırma problemleri (Xiang vd., 2008; Melynkov ve Melynkov, 2014; Shulgin vd., 2017; Nguyen vd., 2018; Ke vd., 2018) ve örüntü tanımlama (Chang, 2012; Fauvel vd., 2013; Wang vd., 2018) gibi bir çok uygulama alanında kullanılmaktadır.

Bu çalışmada, çok değişkenli verisetlerinde aykırı değer tespitinde kullanılan klasik ve dayanıklı Mahalanobis ölçütlerinin aykırı değerleri doğru olarak tespitindeki etkinlikleri karşılaştırılmıştır. Uygulama verisi olarak, New York Borsasında (NYSE) içeriden öğrenenler tarafından gerçekleştirilen hisse senedi alım ve satım işlemlerine ilişkin miktar ve hacim değişkenleri kullanılmıştır. Klasik uzaklık skorlarının yanı sıra, gözlemlerin sıkışması veya maskelenmesi problemlerine karşı kullanılan dayanıklı skorlar da hesaplanmıştır. R paket programı ile gerçekleştirilen analizde, klasik ve en küçük hacimli elipsoid (MVE) dayanıklı ölçütüyle aykırı gözlem olarak tespit edilemeyen (maskelenen) gözlemlerin, dayanıklı bir ölçüt olan hızlı minimum kovaryans determinant (h-MCD) tekniğiyle tespit edildiği; söz konusu tekniğin, verilerin maskelenmesi veya sıkışması durumunda kullanılabilecek etkili bir yöntem olduğu sonucuna ulaşılmıştır.

Çalışmada ikinci bölümde, klasik Mahalanobis ölçütü hakkında bilgi verilmiştir. Üçüncü bölümde ise, klasik Mahalanobis ölçütü ile tespit edilemeyen aykırı gözlemlerin tespiti için literatürde en sık kullanılan dayanıklı tahminleyiciler MVE ve h-MCD teknikleri incelenmiştir. Çalışmanın dördüncü bölümünde, finans uygulama alanından çok değişkenli bir veri seti kullanılarak klasik ve dayanıklı

ölçütlere bağlı uzaklık skorları hesaplanmıştır. Ayrıca, söz konusu tekniklerin aykırı değerlerin doğru tespitindeki performansları karşılaştırılmıştır. Son bölümde ise çalışma sonuçları değerlendirilerek önerilerde bulunulmuştur.

2. Klasik Mahalanobis Ölçütü

Değişkenler arası ilişkinin varlığı göz ardı edildiğinde öklit uzaklığının genelleştirilmiş bir şekli olan Mahalanobis ölçütü, çok değişkenli vektörlerin çok boyutlu uzaklıklarının ölçülmesi esasına dayalı olarak hesaplanmaktadır. Teknikte, merkezi parametre ölçütü ve varyans – kovaryans matrisi kullanılmaktadır (Aggrawal, 2013). Gözlemlerin öznitelikleri arasındaki ilişkinin (değişkenler arası korelasyon) ölçümü ve öznitelik kombinasyonlarının karşılaştırılmasının mümkün olduğu Mahalanobis ölçütü, çok değişkenli aykırı değer tespitinde kullanılan istatistiksel tekniklerin başında gelmektedir (Hodge ve Austin, 2004).

Çok değişkenli veri seti içerisindeki aykırı değerlerin tespitinde kullanılan Cook uzaklığı, kaldıraç noktası, DFFITS gibi tekniklerden farklı olarak Mahalanobis ölçütünde, parametre tahminleyicisi olarak kullanılan varyans - kovaryans matrisi ile gözlemlerin dağılımları dikkate alınmaktadır. Ayrıca, normal dağılan gözlemler ki-kare (χ^2) olasılık yoğunluk fonksiyonu kullanılarak olasılıklara dönüştürülmekte ve hesaplanan uzaklık değerleri χ^2 tablo değeriyle karşılaştırılıp aykırı gözlemler tespit edilmektedir (Johnson ve Wichern, 2002).

Çok değişkenli aykırı değerlerin tespitinde birçok teknik, ortalamalara dayalı olarak uzaklık ölçümü gerçekleştirirken, Mahalanobis ölçütünde konum parametreleri tahmincileri hem ortalamalara dayalı olarak hem de centroid, kırpılmış ortalamalar ve medyan ölçütlerine dayalı olarak gerçekleştirilebilmektedir. Ayrıca, hesaplanan mahalanobis skorunun (D) χ^2 dağılımına uygunluğunun aranması, χ^2 kritik değerinin anormal değerlerin tespitinde bir eşik değeri olarak kullanılmasını sağlamaktadır. Alternatif olarak, D^2 değeri ile Hotelling's T^2test istatistiği de karşılaştırılabilmektedir (Rousseeuw ve Zomeren, 1990).

Veri setinin merkezine uzak uç değerlerin kaldıraç noktası olduğu durumlarda, regresyon doğrusu söz konusu uç değerlere doğru kayacağı için, aykırı değerlerin tespitinde artık değerlerin en küçük kareler toplamına bakmak yeterli olmamakta ve aykırı olmayan veri noktalarının veri setinden atılması problemiyle karşılaşılmaktadır (Rousseeuw ve Leroy, 1987). Mahalanobis skoru hesaplanırken en küçük karelere bakılmadığı gibi, aykırı değişkenlerin centroid ve kovaryans matrisi üzerindeki negatif etkilerinin giderilmesinde medyan veya kırpılmış ortalama gibi ölçütlerin kullanılabiliyor olması, negatif etkileri azaltmaktadır. Centroid ve kovaryans matrisi, Mahalanobis ölçütünde serbestçe seçilmekte ve her bir gözlem için ayrı ayrı uzaklık skorları hesaplanmaktadır.

Mahalanobis ölçütünde varyans - kovaryans matrisi, değişkenler arası ilişkinin tespitini sağlamaktadır. Her bir değişken için grup ortalamaları ve varyansların hesaplanması ve öklid uzaklığında karşılaşılan ölçek farklılıkları ve korelasyon problemlerinin elimine edilmesi, Mahalanobis ölçütünün en önemli avantajları arasında yer almaktadır. Mahalanobis ölçütü, veri kümelerinin farklı büyüklük ve uzaklıkta olduğu durumlarda, lokal küme varyanslarının kullanılarak, hesaplanan korelasyonun yönü doğrultusunda uzaklık değerlerini ölçeklendirmektedir. Çok değişkenli aykırı değer analizinde Mahalanobis ölçütü kullanımının güçlü yanları:

1- Aykırı gözlemleri belirlemede istaitstiki grafik veya nümerik bir eşik değerine sahiptir,
2- Centroid ve kovaryans matrisi değerleri için dayanıklı ve bağımsız seçime izin vermektedir,
3- Veri setindeki aykırı gözlemlerin tespiti süresince, aykırı değerlerin veri setine etkisini azaltıcı tekniklerin kullanımına uygundur,
4- Veri seti içerisinde çok değişkenli gözlem yapmaya imkan sağladığından dolayı, veri seti içerisindeki anormal örüntülerin tespiti için elverişli bir yaklaşımdır,

5- Tahminlenecek kesin bir değer olmadığında, regresyon analizine alternatif sonuçlar sağlaması, olarak sıralanmaktadır (Warren et al., 2011).

$x_{i,1} ve\ x_{i,2}$değişkenleri için hesaplanan öklid uzaklığı:

$$ED_i = \sqrt{(x_{i,1} - x_{i,2})^2} \quad (1)$$

olarak ifade edilirse, her $n \in N$ için, R^n kümesinde tanımlanan $x_1 = (x_{i,1} \ldots x_{i,n})$ ve $x_2 = (x_{i,2} \ldots x_n)$ vektörleri için öklid uzaklığı,

$$ED_1 = \sqrt{(x_{i,1} - \bar{x}_1)^2 + (x_{i,2} - \bar{x}_2)^2} \quad (2)$$

olarak hesaplanmaktadır. p değişken için standardize edilmiş öklid uzaklığı ise,

$$SD = \sqrt{\sum_{j=1}^{p} \left(\frac{x_{i,j} - y_{i,j}}{\sigma_j} \right)^2} \quad (3)$$

şeklinde ifade edilmektedir. Değişkenler arası korelasyon dikkate alınmadığında (SD^2) skoru Mahalanobis uzaklığına eşittir. Değişkenlerarası korelasyon dikkate alındığında, n adet gözlem ve p adet değişkeni içeren X matrisi,

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix},$$

ve,

$$x_i = \begin{bmatrix} x_i \\ \vdots \\ x_n \end{bmatrix} ve\ \bar{x}_i = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$X_C = x_i - \bar{x}_i$$

şeklinde ifade edilirse, varyans- kovaryans matrisi,

$$C_X = \frac{1}{(n-1)} (X_c)^T (X_c), \quad (4)$$

gibi hesaplanmaktadır. Ayrıca, tanımlanan $x_1 ve\ x_2$ değişkenleri için varyans - kovaryans matrisi,

$$C_X = \begin{bmatrix} \sigma_1^2 & p_{1,2}\sigma_1\sigma_2 \\ p_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad (5)$$

ifade edilmektedir. $\sigma_1^2\ ve\ \sigma_2^2$ sırasıyla, $x_1 ve\ x_2$değişkenleri için varyans; $p_{1,2}\sigma_1\sigma_2$ değeri ise $x_1 ve\ x_2$arasındaki kovaryansı belirtmektedir. Her bir x_i için hesaplanan Mahalanobis skoru,

$$MD_i = \left(\sum_{i=1}^{n} (x_i - \bar{x})^T C_X^{-1} (x_i - \bar{x}) \right)^{\frac{1}{2}} \quad (6)$$

olarak hesaplanırken, varyans - kovaryans matrisinin tersi ise,

$$C_X^{-1} = \begin{bmatrix} \frac{\sigma_2^2}{\det(C_X)} & \frac{-p_{1,2}\sigma_1\sigma_2}{\det(C_X)} \\ \frac{-p_{1,2}\sigma_1\sigma_2}{\det(C_X)} & \frac{\sigma_1^2}{\det(C_X)} \end{bmatrix} \quad (7)$$

şeklinde tanımlanmaktadır. Mahalanobis skoru, $p-1$ serbestlik derecesinde χ^2 dağılımına uygun gözlemler normal kabul edilirken, Mahalanobis skoru kritik değerin üzerinde olan gözlemler aykırı olarak işaretlenmektedir (Rousseeuw ve Leroy, 1987).

3. Mahalanobis Dayanıklı Tahminleyiciler

Çok değişkenli aykırı değer analizinde veri setinin aykırı değer içerdiği durumlarda, aykırı gözlemin başka bir aykırı gözlemin tespitini engellemesi (masking) veya normal bir gözlemin aykırı gözlem olarak belirlenmesi (swamping) problemleri ortaya çıkmaktadır. Bir başka ifadeyle, veri kümesindeki kaldıraç noktaları ve etkin gözlemlerin teşhisi için konum parametresi $(\bar{x})$ ve dağılım matrisi $(\hat{C}_X)$ hesaplanmasında dayanıklı yöntemlerin kullanılması önerilmektedir (Rocke ve Woodruff, 1996; Cheng ve Feser, 2002). Nitekim, çok değişkenli aykırı değer tespitinde klasik Mahalanobis ölçütüne dayalı hesaplanan uzaklık skorlarının kritik değerden küçük olması, söz konusu gözlemin aykırı olarak nitelendirilemeyeceği manasına gelmemekte olup, aykırı olan fakat maskelenmiş gözlemin dayanaklı ölçütlerle açığa çıkarılması gerekmektedir (Daszykowski vd., 2007).

Rousseeuw (1985) konum parametreleri tahmininde ortalama değer, dağılım parametreleri tahmininde varyans-kovaryans matrisi yerine, dayanıklı konum ve dağılım ölçütleri önermektedir. Dayanaklı ölçümlerin, aykırı değer tespiti performansını arttırdığı ve maskeleme, veri sıkışması gibi uç değerlerin tespitini zorlaştıran problemlerin önüne geçtiği çeşitli çalışmalarda ifade edilmektedir (Thode, 2002; Willems vd., 2009).

En küçük hacimli elipsoid (MVE) dayanıklı uzaklık ölçütü, $h = veri\ setinde\ seçilen\ alt\ gözlem\ kümesini$ belirtmek üzere, $\frac{n}{2} \leq h < n$ önermesini sağlayan gözlemler içerisinde varyans-kovaryans matrisi determinantının en küçük olduğu elipsoid'e ait gözlemlerin tanımlanması ve bu kümeden hesaplanan değerlerin dayanıklı konum ve dağılım ölçütlerinin belirlenmesini sağlamaktadır (Rousseeuw, 1985). Buna göre, belirlenen tolerans elipsoidi dışında kalan gözlemler aykırı olarak işaretlenmektedir.

Çok değişkenli veri setlerinde, MVE yönteminin klasik tahminleyicilere göre daha etkin sonuçlar vermesine karşılık, $n\ ve\ p$ sayısının büyük olduğu veri setlerinde alt kümelerin belirlenip uzaklık skorlarının hesaplanması güçleşmektedir. Ayrıca, $p+1$ boyutlu bir matrisin her alt matrisinin p ranka sahip olduğu varsayımından hareketle, herhangi bir alt matrisin rankının p'den küçük olması, yani determinantının sıfır olması, hesaplanacak elipsoid hacminin de sıfır olacağı manasına gelmektedir. Bu sebeple, çok değişkenli büyük veri setlerinde aykırı gözlemlerin tespiti için Minimum Kovaryans Determinant (MCD) tekniği önerilmektedir (Rousseeuw ve Driessen, 1999).

MCD'nin konum ve dağılım tahminleyicilerinin belirlenmesinde yüksek kestirim gücüne sahip olduğu ve çok değişkenli aykırı değerlerin tespitinde kullanılan güvenilir bir yöntem olduğu belirtilmektedir (Coakley ve Hettmansperger, 1993). Yöntemde amaç, n adet gözlem içerisinden en küçük varyans - kovaryans matrisi determinantına sahip h adet gözlemin bulunarak konum ve dağılım parametrelerinin belirlenmesidir. Buna göre, $h = (n+p+1)/2$ veya $(n+p+1)/2 \leq h \leq n$ olmak üzere, h adet gözlemden elde edilen ortalama, konum tahmincisini, varyans-kovaryans matrisi de dağılım ölçüsünü belirtmektedir (Hardin ve Rocke, 2005). MCD tahminleyicisi sadece $h > p$ olduğu durumlarda hesaplanabilmektedir.

MCD'de $n\ ve\ p$ değerlerinin büyük olduğu durumlarda, her $h'lik$ örneğin kovaryans matrisinin hesaplanması zaman alıcı olmaktadır. Bu sebeple, MCD'ye alternatif olarak hesaplaması daha kolay olan hızlı minimum kovaryans determinantı (h-MCD) önerilmektedir (Hawkins ve Olive, 1999; Rousseeuw ve Van Driessen, 1999). Herbir gözlem çiftinin yer değiştirmesi yerine, daha fazla gözlemin yer değiştirmesi esasına dayalı olan h-MCD, işlem yükünü azaltarak hesaplama kolaylığı sağlamaktadır.

h-MCD tekniği adımları şu şekilde sıralanabilir (Rousseeuw ve Driessen, 1999):

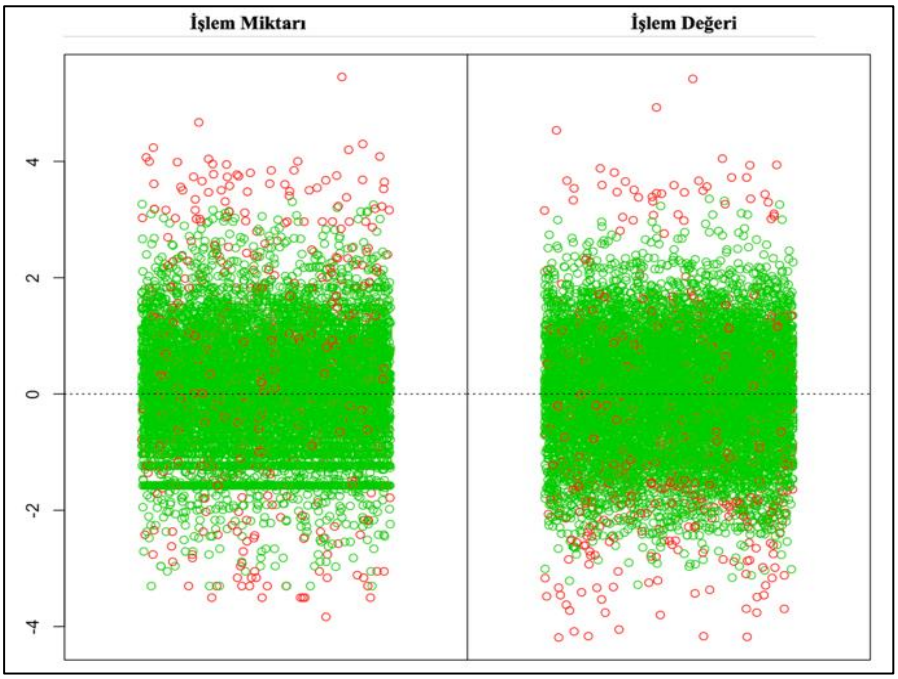
1. $\binom{n}{h}$ kombinasyonu kadar sayıda alt gözlem grupları oluşturulur,
2. Alt gözlem gruplarının herbiri için varyans-kovaryans matrisi determinantları $\det(h_n)$ kontrol edilir. Eğer $\det(h_n) = 0\ ise, \det(h_n) > 0$ oluncaya kadar rasgele olarak seçilen gözlemler alt gözlem gruplarına (h) eklenir. Herbir alt küme için determinant hesaplama işlemi tekrar edilir.
3. Her bir alt küme için mahalanobis skorları hesaplanarak sıralanır. En küçük uzaklık değerlerine sahip h adet gözlem ve en küçük determinanta sahip grubun tespiti işlemleri "C-adımı" olarak adlandırılmaktadır,
4. Determinantı en küçük olan alt gözlem grubu için $\bar{\hat{x}}$ istatistiği (ortalama vektörü) ve varyans-kovaryans matrisi $\hat{C}_X$ hesaplanır,
5. Herbir alt gözlem grubu için hesaplanan $\left|\hat{C}_X\right|$ değerinin en küçük olduğu alt grup için örneklem aritmetik ortalaması, konum parametresi tahmincisi $(\bar{x}_{MCD})$ ve dağılım tahmincisi olarak ise varyans-kovaryans matrisi MCD tahminleyicileri olarak belirlenir,
6. Bir önceki adımda tespit edilen alt gözlem grubu baz alınarak ortalama vektörü ve varyans-kovaryans matrisi üzerinden hesaplanan mahalanobis skorları ki-kare kritik değeriyle (χ^2) karşılaştırılarak, aykırı ve normal gözlemler tespit edilir.

h-MCD'yi MVE'den ayıran temel özelliklerden biri, başlangıç alt kümelerinin oluşturulmasına nasıl karar verileceğidir. h-MCD'de en az $p + 1$ gözlem alınarak seçilen alt kümelerin, hesaplanan varyans-kovaryans matrisi determinantının sıfırdan büyük olana kadar gözlem eklenmesiyle alt kümeler genişletilmekte ve en küçük determinanta sahip küme MCD tahminleyicisi olarak belirlenmektedir. Etkinlik fonksiyonuna sahip ve konum parametrelerinin tahmininde tanımlanan bozulma noktalarının en yüksek değere (%50) sahip olduğu h-MCD ölçütü, büyük boyutlu ve çok değişkenli veri setlerinde aykırı değer tespitinde dayanıklı bir ölçüt olarak tercih edilmektedir (Hubert ve Debruyne, 2010).

4. Veri Seti ve Uygulama

Çalışmada, Thomson Reuters veri tabanından elde edilen ve Ocak 2013 – Aralık 2017 döneminde New York ve NASDAQ borsalarında gerçekleştirilen 325.548 adet hisse senedi alım işlemi, 913.959 adet satım işlemi olmak üzere toplam 1.239.507 adet işlem verisi kullanılmıştır. İşlemler alım ve satım portföylerine bölünmüş olup, her bir işlem için miktar ve hacim değişkenleri incelenmiştir. Yıl içerisinde gerçekleştirilen menkul kıymet alım - satım miktarının menkul kıymetin emrin gerçekleştiği tarihteki dolar cinsinden düzeltilmiş kapanış fiyatının çarpımıyla toplam işlem hacmi hesaplanmıştır.

Grafik 1: **Alım İşlemleri için Dağılım Grafiği**



Not: Gösterim kolaylığı açısından, veri setinden rasgele olarak seçilen $n = 1000$ gözlemin MCD ölçütüne göre hesaplanan tek boyutlu uzaklık skorları gösterilmiştir. Normal gözlemler yeşil renkle, aykırı gözlemler ise kırmızı renkle işaretlenmiştir.

Grafik 1'de her bir değişken içerisindeki aykırı gözlemlerin tek boyutlu uzayda dağılım grafiğinde, her bir değişken içerisindeki gözlemlerin uzaklıkları verilmiştir. Buna göre, işlem miktarı açısından aykırı olarak tespit edilen bir gözlemin, işlem değeri açısından normal olarak etiketlendiği; söz konusu işlemin aykırı olup olmadığının tespit edilebilmesi için çok değişkenli konum ve dağılım tahminlerinin hesaplanması gerektiği anlaşılmaktadır.

Alım ve satım işlemleri için klasik ve dayanıklı Mahalanobis ölçütleriyle R programlama dilinde hesaplanan en yüksek uzaklık skorlarına sahip olan ilk on gözlem Tablo 1'de verilmiştir. Her bir ölçüt için çok değişkenli konum ve dağılım tahminicileri, örneklem ortalamaları ve varyans-kovaryans matrisine dayalı olarak hesaplanmıştır. Alım işlemleri için hesaplanan Mahalanobis uzaklığı skorları büyükten küçüğe doğru sıralandığında 319697, 319698, 319848 ve 319849 numaralı gözlemlerin her üç ölçütte de en büyük uzaklığa sahip aykırı gözlemler olduğu tespit edilmiştir. 325294, 325295, 325296, 325297, 314766 ve 314767 numaralı gözlemler klasik mahalanobis ölçütüne göre en yüksek skorlu aykırı gözlemler olarak tespit edilirken, söz konusu gözlemler dayanıklı ölçütlere göre kritik değerin altında kalmaktadır. h-MCD ve MVE'ye dayalı hesaplanan uzaklık skorları sıralandığında, en büyük uzaklık skorlarına sahip olan aykırı gözlemlerin benzer olduğu ve klasik ölçüte dayalı uzaklık skorlarından farklılaştığı göze çarpmaktadır. Alım işlemleri için klasik ölçütte 16.072 işlem, h-MCD ölçütünde 33.143 işlem, MVE ölçütünde ise 32.155 işlem aykırı olarak tespit edilmiştir.

Tablo 1: **Alım ve Satım İşlemleri için Farklı Ölçütlere göre Hesaplanmış Konum ve Dağılım Tahmincileri**

Teknik	İşlem Türü	Konum Parametresi Tahmincisi ($\overline{x}$)		Dağılım Tahmincisi Matrisi ($\widehat{C}_X$)	Normal/ Anormal Gözlem Sayısı	En Büyük Skora Sahip İlk On Aykırı Gözlem ve Uzaklık Skorları*
		İşlem Miktarı (log)	İşlem Hacmi (log)			
Klasik Mahalanobis Ölçütü	Alım	3,239	4,081	$\begin{bmatrix}1,181 & 0,826\\ 0,826 & 1,091\end{bmatrix}$	309476 / 16072	319697(10,42);319698(10,42);319848(8,35);319849(8,35);325294(7,15);325295(7,15);325296(7,15);325297(7,15);314766(7,01);314767(7,01)
	Satım	3,402	4,893	$\begin{bmatrix}0,893 & 0,772\\ 0,772 & 0,997\end{bmatrix}$	879641 / 34318	212530(9,82);197634(9,80);197635(9,80);639190(9,79);178293(9,77);178294(9,77);178295(9,77);43818(9,56);43819(9,56);43820(9,56)
Hızlı Minimum Kovaryans Determinant	Alım	2,963	4,035	$\begin{bmatrix}1,221 & 1,039\\ 1,039 & 1,034\end{bmatrix}$	292405 / 33143	319697(14,46);319698(14,46);319848(11,23);319849(11,23);325529(10,34);325530(10,34);325531(10,34);325532(10,34);325533(10,34);325527(10,33)
	Satım	3,385	4,915	$\begin{bmatrix}1,030 & 1,006\\ 1,006 & 1,167\end{bmatrix}$	837917 / 76042	212530(13,45);197634(13,44);197635(13,44);178293(13,42);178294(13,42);178295(13,42);43818(13,26);43819(13,26); 43820(13,26); 43821(13,26)
En Küçük Hacimli Elipsoid	Alım	3,091	4,095	$\begin{bmatrix}1,231 & 1,049\\ 1,049 & 1,128\end{bmatrix}$	293393 / 32155	319697(14,11);319698(14,11);319848(11,00);319849(11,00);325529(10,01);325530(10,01);325531(10,01);325532(10,01);325533(10,01);325527(10,00)
	Satım	3,387	4,913	$\begin{bmatrix}0,979 & 0,933\\ 0,933 & 1,085\end{bmatrix}$	842467 / 71492	212530(12,88);197634(12,87);197635(12,87);178293(12,85);178294(12,85);178295(12,85);43818(12,67);43819(12,67); 43820(12,67); 43821(12,67)

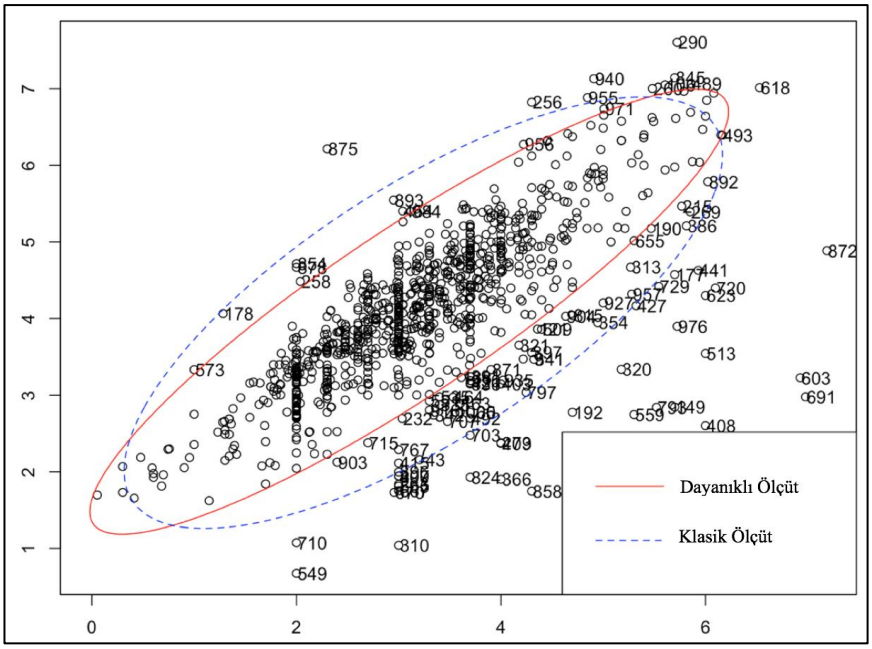
Not: Gözlemlere ilişkin Mahalanobis Skorları son sütunda parantez içerisinde gösterilmiştir. Tabloda, en yüksek uzaklık skoruna sahip olan ilk on gözlem verilmiştir. h-MCD için alım işlemlerinde, en küçük kovaryans matrisi determinantına sahip $h = 162.775$ adet gözlem, satım işlemleri için $h = 456.981$ adet gözlem belirlenmiştir. Ayrıca, parantez içerisinde ilgili gözlemlere ilişkin Mahalanobis skorları verilmiştir. $\sqrt{\chi^2_{2,.975}} = 2{,}716$ kritik (eşik) değeri olarak belirlenmiştir. Ayrıca aynı paket programda yapılan işlemlerde, alım ve satım veri setindeki her bir işlem için klasik Mahalanobis uzaklığının hesaplanmasında toplamda 136 saniye, MVE'ye dayalı uzaklık skorlarında 42 saniye, h-MCD için ise 8 saniye işlem süresi tespit edilmiştir.

Satım işlemleri değerlendirildiğinde ise, klasik ölçüte göre en yüksek Mahalanobis skoruna sahip aykırı gözlemlerden biri olan 639190 numaralı gözlemin, dayanıklı ölçütlerle tespit edilen en yüksek uzaklık skoruna sahip gözlemler arasında olmadığı görülmektedir. Ayrıca dayanıklı ölçütlerde hesaplanan uzaklık skorlarının klasik ölçüt skorlarından büyük olduğu; h-MCD ölçütünde 76.042, MVE

ölçütünde ise 71.492 işlemin aykırı gözlem olarak işaretlendiği görülmektedir. Konum parametresi tahmincilerine bakıldığında, alım ve satım işlemleri için işlem hacmi değişkeninde MVE ve klasik ölçüt konum ortalamalarının benzerliği göze çarpmaktadır.

Grafik 2'de, alım işlemleri için klasik ve dayanıklı uzaklık skorlarıyla oluşturulan tolerans elipsinde, sınırda ya da sınırın dışında bulunan gözlemlerin $\bar{x}$ konum parametresinden uzakta ve büyük sapmaya sahip olan aykırı gözlemler olduğu görülmektedir. Çok değişkenli normal dağılım varsayımı altında, Mahalanobis uzaklıklarının dağılımı ile klasik ve dayanıklı elipsler içerisinde kalan normal gözlemlere ilişkin Mahalanobis skorlarının dağılımı, $p-1$ serbestlik derecesinde χ^2 dağılımına uygundur. Klasik ve dayanıklı ölçütler için aykırı gözlem tespitinin farklılaşması, veri setinde aykırı bir gözlemin diğer aykırı bir gözlem tarafından maskelenmesi veya verideki aykırı gözlemlerin regresyon doğrusunu çekerek sorunsuz gözlemlerin aykırı olarak belirlenmesi gibi problemlere işaret etmektedir. Bu durumda, veri kümesinde regresyon doğrusunu en çok saptıran gözlemin silinerek, kalan gözlemler için parametre tahminlerinin yeniden hesaplanması gerekmektedir.

Grafik 2: **Alım İşlemlerine İlişkin Tolerans Elipsi**

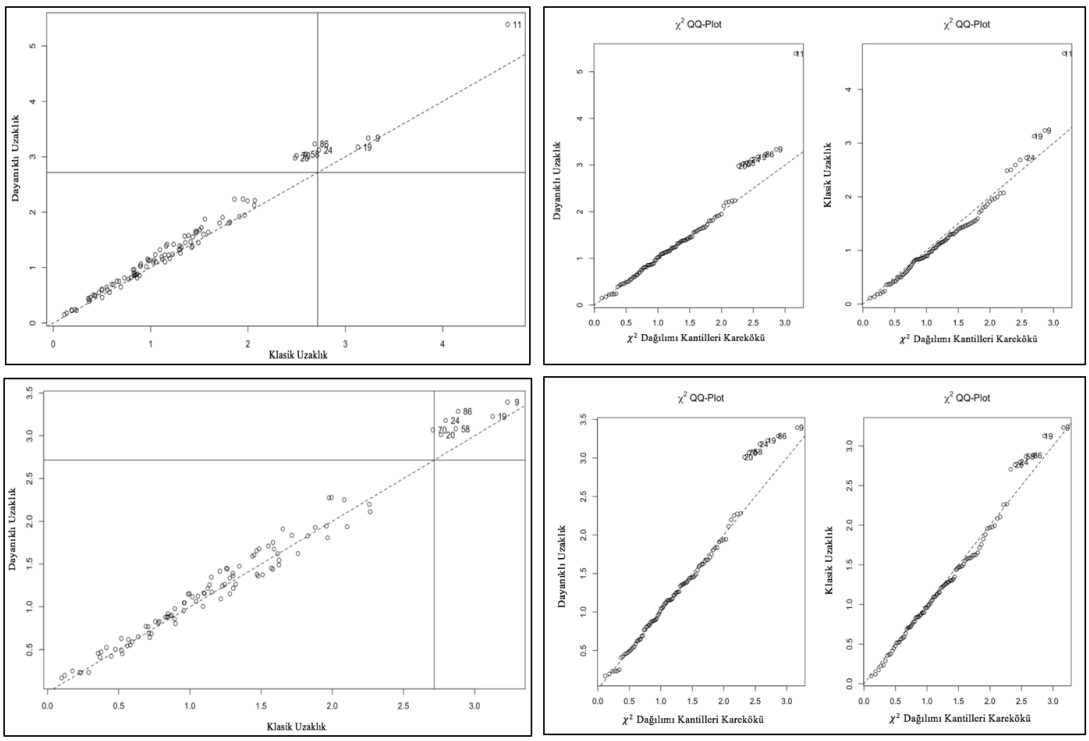


Not: Grafikte veri setine ilişkin $n = 1000$ gözlemin klasik ve dayanıklı ölçüte dayalı uzaklıkları hesaplanmış ve %97,5 güven elipsoidinde gösterilmiştir. x ekseni işlem hacmini, y ekseni ise işlem miktarını göstermektedir.

Grafik 3'de alım işlemlerinde, uzaklık skorları kritik değerin ($\sqrt{\chi^2_{2,.975}} = 2{,}716$) üzerinde olan ve klasik ve dayanıklı MVE ölçütünde aykırı olarak etiketlenen 9, 11, 19, 24 no'lu gözlemlere ek olarak, h-MCD ölçütü ile aykırı olarak tespit edilen 20, 58 ve 86 no'lu gözlemler yer almaktadır. Parametre tahminine etkisi olan ve en yüksek uzaklık skoruna sahip gözlemin (11 numaralı gözlem) veri setinden silinmesiyle yeniden hesaplanan klasik ve h-MCD skorlarına göre 9, 19, 20, 24, 58, 86 no'lu gözlemlerin kritik değerin üzerinde, bir başka ifadeyle, klasik ve dayanıklı ölçütlere göre aykırı gözlem olarak tespit edildiği anlaşılmaktadır. Bu durum, Tablo 2'de de gösterildiği üzere, veri setinde aykırı gözlemler bulunduğunda aykırı gözlemlerin örnek ortalamasına olan uzaklıklarının küçüldüğü ve aykırı

gözlemlerin normal gözlem olarak tespit edildiği bir durum olan maskeleme problemine işaret etmektedir (Tablo 2).

Grafik 3. **En Büyük Uzaklık Skoruna Sahip Alım İşleminin Veri Kümesinden Çıkartılması Öncesi ve Sonrası Mahalanabis Skorları ve Q-Q Plotu**



Not: Gösterim kolaylığı açısından $n = 100$ gözlem veri setinden rastgele olarak seçilmiştir.

Başlangıçta klasik ve dayanıklı MVE ölçütüyle tespit edilemeyen aykırı gözlemlerin, parametre tahminleri üzerinde etkiye sahip, en büyük uzaklık skorlu tek gözlemin silinmesiyle tespit edilebilir hale geldiği ve başlangıç çözümünde aykırı gözlemlerin h-MCD ile başarılı bir şekilde tespit edildiği görülmektedir.

Tablo 2: **Tek Gözlemin Silinmesi Öncesi ve Sonrası Aykırı Gözlemler**

Ölçüt	Aykırı Gözlemler	Tek Gözlem Silinmesi Sonrası Aykırı Gözlemler
Klasik	9(3,18), 11(4,32), 19(3,13), 24(2,73)	9(3,18),19(3,08),20(2,77),24(2,81),58(2,89),70(2,72),86 (2,92)
MVE	9(3,19), 11(5,03), 19(3,14), 24(2,80)	9(3,20),19(3,12),20(2,95),24(3,08),58(2,92),70(2,85),86 (3,08)
h-MCD	9(3,19), 11(5,24), 19(3,18), 20(2,92), 24 (3,09), 58(3,02), 70(2,95), 86(3,22)	9(3,30),19(3,15),20(3,01),24(3,13),58(3,07),70(3,10),86 (3,28)

Not: Aykırı gözlemlere ilişkin uzaklık skorları parantez içerisinde verilmiştir.

5. Sonuç

Bir veri setinde aykırı gözlemlerin varlığı, istatistiki açıdan yüksek hata oranları ve parametre tahminlerinden önemli ölçüde sapmalara sebep olmaktadır. Özellikle ekonomi ve finans alanında kullanılan veri setlerinde değişkenler arası korelasyonun olduğu durumlarda, tek bir aykırı gözlemin araştırılması yerine, söz konusu gözleme etki edebilecek diğer gözlemlerin tespiti, üzerinde durulması gereken bir probleme işaret etmketedir. Örneğin, hileli kredi kart işlemlerinin tespitinde işlemin hacmi,

miktarı, süresi ve türü, işlemin gerçekleştirildiği tarih gibi kullanıcı davranışlarını açıklayan birden fazla değişkenin varlığında, değişkenlerarası ilişkilerin de göz önünde bulundurulması, bir başka deyişle, çok değişkenli konum ve dağılım ölçütlerinin hesaplanması gerekmektedir. Çok değişkenli veri seti içerisinde gerek istatistiksel olarak veri setinin dağılımdan uzak tekil aykırı gözlemlerin tespiti, gerekse anormal bir paterni işaret edebilecek toplu aykırı gözlemlerin tespitinde Mahalanobis ölçütü geniş bir uygulama alanına sahiptir.

Aykırı değer tespitinde değişken ve gözlem sayısının arttığı durumlarda, kaldıraç değeri büyük olan bir gözlemin regresyon eğrisinin eğimini değiştirmesi, çok değişkenli gözlemlerin elipsoid serpilme diyagramının ağırlık merkezinden uzaklığının tespitini gerektirmektedir. Özellikle çok değişkenli veri setlerinde, aykırı gözlemin tespitinin bir başka gözlem tarafından engellenmesi veya aykırı olmayan bir gözlemin aykırı olarak sınıflandırılması en sık karşılaşılan problemlerden olup; bu gibi durumlarda klasik yöntemlerin kullanımı aykırı değerlerin tespitini engellemektedir. Ayrıca, veri setinin normal dağılmadığı durumlarda klasik doğrusal regresyon modellerinin aykırı değerlerin tespitine aşırı duyarlı olması, başka bir ifadeyle, aşırı değerlerin regresyon doğrusunu kaydırması, parametre tahmin değerlerini olumsuz olarak etkilemektedir. Bu sebeple çalışmada, çok değişkenli klasik konum ve dağılım tahminleyicisinin yanı sıra; hızlı minimum kovaryans determinant ve en küçük hacimli elipsoid dayanıklı yöntemleri kullanılarak, söz konusu tekniklerin aykırı gözlemlerin doğru olarak tespit edilmesindeki etkinlikleri karşılaştırılmıştır. Çalışmada, klasik ölçüte dayalı konum ve dağılım parametreleri kestirimlerinin aykırı gözlemlerden etkilendiği, aykırı gözlemlerin tespitini engellediği ve bu nedenle aykırı gözlemlerden etkilenmeyen dayanıklı yöntemlerin kullanılması gerektiği; gözlem ve değişken sayısının büyük olduğu veri setlerinde MVE ölçütü yerine h-MCD ölçütünün daha hızlı ve güvenilir sonuçlar verdiği sonucuna ulaşılmıştır. Çalışma sonucunda, birden fazla değişkenli, yüksek hacimli ve değişkenlerarası ilişkinin önemli olduğu veri setlerinde, özellikle de aykırı değerlerden şüphelenildiği durumlarda; χ^2 dağılımına uygun bir dağılım izlemesi, aykırı gözlemler için kritik değerin hesaplanabilmesi ve daha hızlı işlem süresine sahip olduğundan dolayı, dayanıklı h-MCD tekniğinin kullanılması önerilmektedir.

Ayrıca bu bağlamda, aykırı değer tespiti için gözlemlerin tekil veya kolektif olarak mı inceleneceği konusu, problemin hile tespiti kapsamında mı yoksa veri ön işleme sürecine yönelik mi ele alındığına göre değişmektedir. Çalışma kapsamında incelenen yöntemler, veri analitiği içerisinde finansal piyasalarda hile tespitinde kullanılabileceği gibi, bir veri setinin analize hazır hale getirilmesinde etkin olarak kullanılabilmektedir. Gözlemler arası ilişkilerin önemli olduğu işletme problemlerinde aykırı değerlerin tespitinde Mahalanobis ölçütüne alternatif olarak, derinlik tabanlı, en yakın komşu tabanlı ve kümeleme tabanlı ölçütlere ilişkin hesaplanacak uzaklık skorlarının karşılaştırılarak en uygun yöntemin kullanılması önerilmektedir.

Kaynakça

Aggarwal, Charu C., *Outlier Analysis*, Springer, 2013.

Arteaga, T.G., Alcantud, J.C.R., Calle, R.A. (2016). A cardinal dissensus measure based on the Mahalanobis distance, *European Journal of Operational Research*, 251(2), 575-585.

Carminati, M., Caron, R., Maggi, F., Epifani, I., Zanero, S. (2015). BankSealer: A decision support system for online banking fraud analysis and investigation, *Computers & Security*, 53, 175-186.

Carrato, R.G.H. (2018). Wind farm monitoring using Mahalanobis distance and fuzzy clustering, *Renewable Energy*, 123(C), 526-540.

Chang, C.C. (2012). A boosting approach for supervised Mahalanobis distance metric learning, *Pattern Recognition*, 45(2), 844-862.

Cheng, T. C. & Victoria-Feser, M. P. (2002). High breakdown estimation of multivariate mean and covariance with missing observations, *British Journal of Mathematical and Statistical Psychology*, 55, 317-335.

Cho, S., Hong, H., Ha, B.C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction, *Expert Systems with Applications*, 37(4), 3482-3488.

Coakley, C. W., Hettmansperger, T. P. (1993). A Bounded Influence, High Breakdown, Efficient Regression Estimator, *Journal of the American Statistical Association*, 88, 872-880.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis – a review: basic concepts. *Chemometrics and Intelligent Laboratory Systems*, 85(2), 203–219.

Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses*, Pattern Recognition*, 74, 406-421.

Fauvel, M., Chanussot, J., Benediktsson, J.A., Villa, A. (2013). Parsimonious Mahalanobis kernel for the classification of high dimensional data, *Pattern Recognition*, 46(3), 845-854.

Haldar, N., Khan, F., Ali, A., Abbas, H. (2016). Arrhythmia Classification using Mahalanobis Distance based Improved Fuzzy C-Means Clustering for Mobile Health Monitoring Systems. *Neurocomputing*, 220, 221-235.

Hardin, J. & Rocke, D.M. (2005). The Distributions of Robust Distances, *Journal of Computational and Graphical Statistics*, 14(4), 1-19.

Hawkins, D. (1980). *Identification of Outliers*, Chapman and Hall, 1980.

Hawkins, D.M., & Olive, D.J. (1999). Improved feasible solution algorithm for high breakdown estimation. Computational Statistics and Data Analysis, 30, 1-11.

Hodge, Victoria J., Austin, J. (2004). A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, 22(2), 85-126.

Hubert, M. & Debruyne, M. (2010). Minimum Covariance Determinant, *Computational Statistics*, 2(1), 36-43.

Jaffel, I., Taouali, O., Faouzi Harkat, M., Messaoud, H. (2015). A Fault Detection Index Using Principal Component Analysis And Mahalanobis Distance, *IFAC-PapersOnLine*, 48(21), 1397-1401.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5. Baskı). Prentice Hall, Upper Saddle River, NJ.

Ke, T., Lv, H., Sun, M., Zhang, L. (2018). A biased least squares support vector machine based on Mahalanobis distance for PU learning, *Physica A: Statistical Mechanics and its Applications*, 509, 422-438.

Leys, C., Klein, O., Dominicy, Y., Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance, *Journal of Experimental Social Psychology*, 74, 150-156.

Melnykov, I. & Melnykov, V. (2014). On K-means algorithm with the use of Mahalanobis distances, *Statistics & Probability Letters*, 84, 88-95.

Nguyen, B., Morell, C., Baets, B.D. (2018). Distance metric learning for ordinal classification based on triplet constraints, *Knowledge-Based Systems*, 142, 17-28.

Pompella, M. & Dicanio, A. (2017). Ratings based Inference and Credit Risk: Detecting likely-to-fail Banks with the PC-Mahalanobis Method, *Economic Modelling*, 67, 34-44.

Pozzolo, A.D., Caelen, O., Borgne, Y.L., Waterschoot, S., Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective, *Expert Systems with Applications*, 41(10), 4915-4928.

Qiu, Z., Zhou, B., Yuan, J. (2017). Protein–protein interaction site predictions with minimum covariance determinant and Mahalanobis distance, *Journal of Theoretical Biology*, 433, 57-63.

Rocke, D. M., Woodruff, D. L. (1996). Identification of Outliers in Multivariate Data, *Journal of the American Statistical Association*, 91, 1047-1061.

Rousseeuw, P.J. (1985). Multivariate Estimation With High Breakdown Point, *Mathematical Statistics and Applications*, 1, 283-297.

Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression & Outlier Detection*, Wiley&Sons, New Jersey.

Rousseeuw, P. J. & Zomeren, B. C. V. (1990). Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association*, 185(411), 633-634

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

Shang, J., Chen, M.Y., Zhang, H. (2018). Fault detection based on augmented kernel Mahalanobis distance for nonlinear dynamic processes, *Computers & Chemical Engineering*, 109, 311-312

Shulgin, S., Zinkina, J., Korotayev, A., Andreev, A. (2017). "Neighbors in values": A new dataset of cultural distances between countries based on individuals' values, and its application to the study of global trade, *Research in International Business and Finance*, 42, 966-985.

Stöckl, S. & Hanke, M. (2014). Financial Applications of the Mahalanobis Distance, *Applied Economics and Finance*, 1(2), 78-84.

Suo, M., Zhu, B., Zhang, Y., An, R., Li, S. (2018). Fuzzy Bayes risk based on Mahalanobis distance and Gaussian kernel for weight assignment in labeled multiple attribute decision making, *Journal of Knowledge-Based Systems*, 152(C), 26-39.

Thode, H.C. (2002). Testing for Normality, Marcel Dekker, New York.

Wang, P.C., Su, C.T., Chen, K.H., Chen, N.H. (2011). The application of rough set and Mahalanobis distance to enhance the quality of OSA diagnosis, *Expert Systems with Applications*, 38(6), 7828-7836,

Wang, Q., Wan, J., Yuan, Y. (2018). Locality constraint distance metric learning for traffic congestion detection, *Pattern Recognition*, 75, 272-281.

Warren, R. Smith, R., Cybenko, A. (2011). Use Of Mahalanobis Distance For Detecting Outliers And Outlier Clusters In Markedly Non-Normal Data: A Vehicular Traffic Example, *Air Force Research Laboratory Human Effectiveness Directorate Report*, 1-52.

Willems, G., Joe, H., Zamar, R. (2009). Diagnosing Multivariate Outliers Detected by Robust Estimators, *Journal of Computational and Graphical Statistics*, 18(1), 73-91

Xiang, S., Nie, F., Zhang, C. (2008). Learning a Mahalanobis distance metric for data clustering and classification, *Pattern Recognition*, 41(12), 3600-3612,

Vukovic, O. (2015). Analysing Bank Real Estate Portfolio Management by Using Impulse Response Function, Mahalanobis Distance and Financial Turbulence, *Procedia Economics and Finance*, 30, 932-938.

CLASSICAL AND ROBUST MAHALANOBIS DISTANCE MEASURES FOR OUTLIER DETECTION: AN APPLICATION IN STOCK EXCHANGES

Extended Abstract

Aim: The detection of outliers is a neccesity to increase the power of a statistical test. Due to parameter estimation, normalization of data sets and homogenous distribution of variances or determination of outliers for descriptive statistics, outlier analysis is frequently used. Besides that, in the literature, the outlier analysis has a wide range of applications, such as the detection of unusual patterns which is far from the other observations in the data set.

In multivariate outlier analysis, statistical techniques are based on the assumption that an outlying observation is located far from the center of the data distribution and the vector space is orthogonal. The use of many distance measures is not recommended in data sets where data size and type increase and the values are zero or negative. Mahalanobis distance is used to detect unusual patterns and outlying observations within the data set where the number of variables and the sample size are large. It considers the relationship between the observations. The application of Mahalanobis distance ranges from pattern recognition to classification and clustering problems, financial fraud detection, multivariate forecasting and fault detection – diagnosis.

Method: In this study, a total of 1,239,507 transactions are obtained from the Thomson Reuters database. Transactions are divided into purchase and sale portfolios. Value and volume of each transaction are used for outlier analysis. Classical and robust Mahalanobis distance scores are calculated for each transaction with R programming language.

Mahalanobis distance, which is a generalized form of euclidean distance, is calculated based on the measurement of multidimensional distances of multivariate vectors. While many statistical techniques perform distance measurements based on averages, Mahalanobis distance can be performed based on averages, centroid, cropped averages or median for detecting multivariate outliers. In addition, the normally distributed observations are converted to probabilities using the chi-square probability density function and the calculated distances are compared with the table values to determine outliers.

In particular, if the data set contains an outlier, it is necessary to prevent misclassification errors: masking (an undetected outlier) and swamping (detecting a nonoutlier as an outlier). It is recommended to use robust methods which calculate multivariate location parameter and scatter matrix for the diagnosis of leverage points and influential observations in the data set.

Minimum Volume Ellipsoid (MVE) measure is an equivariant robust estimator based on the ellipsoid with minimal volume. It allows to calculate robust multivariate location and scatter matrix that meet the $\frac{n}{2} \leq h < n$ criterion, where h indicates a subset of dataset which consists n observations. Since MVE has more effective results than classical estimators, it is difficult to calculate the distance scores in multivariate datasets, where the number of variables or observations is large. For this reason, Minimum Covariance Determinant (MCD) technique is recommended for detection of outliers in large multivariate datasets. In this technique, the aim is to determine the location and scatter parameters by finding the smallest variance - covariance matrix determinant within n observations. In the MCD technique, the calculation of the covariance matrix of each sample, containing h observations, is a time-consuming task. In such cases, the fastest minimum covariance determinant (h-MCD) has been proposed by Rousseeuw and Van Driessen (1999) as an alternative to MCD, which is easier to calculate. One of the key features that distinguish h-MCD from MVE is how to decide and create initial subsets.

Findings: It is observed that outlying observations with the largest robust distance scores are similar for robust MCD and MVE measures and the scores differ from the distance scores which are based on classical Mahalanobis measure. The number of outlying transactions are 16,072 for classical distance measure, 32,155 for MVE and 33,143 for h-MCD in purchases portfolio. Besides, 34,318 transactions for classical measures, 71,492 transactions for MVE and 76,042 transactions for h-MCD are detected as outliers for sales portfolio. This shows a potential masking problem or influential observations that affect the slope of regression. For this reason, we re-calculate the parameter estimates for the remaining observations in the data set by deleting the observation that distorts regression line. At the initial solution, while some observations are flagged as nonoutliers (based on classical and MVE measures), these observations are detected as outliers by using h-MCD. This situation indicates the masking problem, where the distances of outlying observations to the sample means are reduced and the outliers are determined as nonoutliers. It is shown that outliers are efficiently detected with h-MCD measure, compared to the other measures.

Conclusion: The use of classical Mahalanobis distance measure results in the prevention of the detection of outliers. In this study, it is shown that the estimates of location and scatter parameters based on classical and robust MVE measures are affected by outliers and therefore, the detection of outliers is prevented. We conclude that instead of MVE measure, h-MCD gives more reliable results in determining the outliers in the data sets where the number of observations and variables are large. Mahalanobis distance can be used effectively to determine multivariate deviations in the field of financial application. As an alternative to statistical techniques, it is recommended to use depth-based, nearest neighbor and clustering-based algorithms for outlier analysis.