

**ÇOK DEĞİŞKENLİ VERİDE AYKIRI DEĞERLERİN TESPİTİ İÇİN MVV
YÖNTEMİ VE DİĞER YÖNTEMLERLE KARŞILAŞTIRILMASI**

Kübra TURGUT

**YÜKSEK LİSANS TEZİ
İSTATİSTİK**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**HAZİRAN 2010
ANKARA**

Kübra TURGUT tarafından hazırlanan ÇOK DEĞİŞKENLİ VERİDE AYKIRI DEĞERLERİN TESPİTİ İÇİN MVV YÖNTEMİ VE DİĞER YÖNTEMLERLE KARŞILAŞTIRILMASI adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Ufuk EKİZ
Tez Danışmanı, İstatistik Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Hamza GAMGAM
İstatistik Anabilim Dalı, G.Ü.

Yrd. Doç. Dr. Ufuk EKİZ
İstatistik Anabilim Dalı, G.Ü.

Yrd. Doç. Dr. İhsan KARABULUT
İstatistik Anabilim Dalı, A.Ü.

Tarih : 22 / 06 / 2010

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Kübra TURGUT

**ÇOK DEĞİŞKENLİ VERİDE AYKIRI DEĞERLERİN TESPİTİ İÇİN MVV
YÖNTEMİ VE DİĞER YÖNTEMLERLE KARŞILAŞTIRILMASI
(Yüksek Lisans Tezi)**

Kübra TURGUT

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Haziran 2010**

ÖZET

Çok değişkenli veri setlerinde aykırı değerleri tespit etmek özellikle değişken sayısı ikiye geçtiğinde zor olabilmektedir. Bu nedenle konum ve kovaryans matrisinin sağlam tahminine dayanan çeşitli yöntemler önerilmiştir. Bu yöntemler etkili olmalarına rağmen geniş ve büyük boyutlu veri setleri için kullanışsızdırlar. Bu yöntemlerin hesaplama karmaşıklığı veri setlerinin boyutu arttığı zaman artmaktadır.

Bu çalışmanın amacı, çok değişkenli veride çoklu aykırı değerlerin belirlenmesinde kullanılan En Küçük Hacimli Elipsoid, En Küçük Kovaryans Determinantı ve Hızlı En Küçük Kovaryans Determinantı gibi yöntemlere alternatif olarak geliştirilen En Küçük Vektör Varyansı yöntemini tanıtmaktır.

Bu tezde öncelikle aykırı değer(outlier) ve bozulma noktası(breakdown point) kavramları ele alınmış, daha sonra yukarıda bahsedilen yöntemler tanıtılmış ve son olarak bir simülasyon çalışması ile En Küçük Vektör Varyansı ve diğer yöntemlerin aykırı değer belirleme oranı ve yöntemlerin hesaplama hızı bakımından karşılaştırması yapılmıştır.

Sonuçta, En Küçük Vektör Varyansı yönteminin geniş ve büyük boyutlu veri setlerine uygulanabilir olduğu görülmüştür ve bu algoritmanın hesaplama karmaşıklığı önemli derecede diğer yöntemlerden azdır.

Bilim Kodu : 205.1.066
Anahtar Kelimeler : Bozulma Noktası, Aykırı Değer Tespiti, Sağlam Tahmin Edici, Çok Değişkenli Konum ve Dağılım
Sayfa Adedi : 53
Tez Yöneticisi : Yrd. Doç. Dr. Ufuk EKİZ

**THE MVV METHOD FOR DETECTING OUTLIERS IN MULTIVARIATE
DATA AND A COMPARISON WITH THE OTHER METHODS**

(M.Sc. Thesis)

Kübra TURGUT

**GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY**

June 2010

ABSTRACT

Outliers in multivariate data sets can be hard to detect especially when the number of variables exceeds two. Therefore, various methods have been suggested that based on robust estimation of location and covariance matrix. Although these methods are influential, they are cumbersome for large and high dimension data sets. Computational complexity of these methods increases when the dimension of the data sets is getting increases.

The aim of this study is to present the Minimum Vector Variance method which is developed as an alternative to the methods like Minimum Volume Ellipsoid, Minimum Covariance Determinant and Fast Minimum Covariance Determinant that are used in identifying multiple outliers in multivariate data.

In this thesis, firstly outlier and breakdown point concepts have been given, secondly the above mentioned methods have been introduced and finally the Minimum Vector Variance method and other methods have been compared with a simulation study with respect to the ratio of outlier detection and the computation speed of the methods.

As a result, it has been found that the Minimum Vector Variance method is applicable to large and high dimension data sets and the computational complexity of this algorithm is significantly smaller than that of other methods.

Science Code : 205.1.066

**Key Words : Breakdown Point, Outlier Detection, Robust Estimation,
Multivariate Location and Scatter.**

Page Number: 53

Adviser : Yrd. Doç. Dr. Ufuk EKİZ

TEŞEKKÜR

Bu çalışmanın gerçekleşmesi sürecinde bilgi ve deneyimleriyle beni yönlendiren, çalışmalarının yoğunluğuna rağmen zaman ayırarak her konuda yardımlarını benden esirgemeyen hocam Yrd. Doç. Dr. Ufuk EKİZ'e, bugünkü bilgilerimi elde etmede katkısı olan tüm hocalarıma ve her koşulda benden desteğini esirgemeyen aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

ÖZET.....	iv
ABSTRACT.....	vi
TEŞEKKÜR.....	viii
İÇİNDEKİLER	ix
ÇİZELGELERİN LİSTESİ.....	x
ŞEKİLLERİN LİSTESİ	xi
SİMGELER VE KISALTMALAR.....	xii
1. GİRİŞ	1
2. AYKIRI DEĞER.....	3
3. KONUM PARAMETRESİ TAHMİN EDİCİLERİ.....	5
3.1. Budanmış Ortalama.....	5
4. BOZULMA NOKTASI	7
5. ÇOK DEĞİŞKENLİ DURUMDA AYKIRI DEĞER TESPİTİ İÇİN ÖNERİLEN METOTLAR.....	9
5.1. En Küçük Hacimli Elipsoid (MVE) Yöntemi.....	10
5.2. En Küçük Kovaryans Determinantı (MCD) Yöntemi.....	19
5.3. Hızlı En Küçük Kovaryans Determinantı (FMCD) Yöntemi.....	24
5.4. En Küçük Vektör Varyansı (MVV) Yöntemi.....	27
6. SİMÜLASYON ÇALIŞMASI.....	32
7. SONUÇ VE ÖNERİLER.....	40
KAYNAKLAR	41
EKLER.....	43
EK-1 MVE Yöntemi için MATLAB R2008a’da Hazırlanan Bilgisayar Kodları.....	44
EK-2 MCD Yöntemi için MATLAB R2008a’da Hazırlanan Bilgisayar Kodları	46
EK-3 FMCD Yöntemi için MATLAB R2008a’da Hazırlanan Bilgisayar Kodları...	48
EK-4 MVV Yöntemi için MATLAB R2008a’da Hazırlanan Bilgisayar Kodları.....	50
ÖZGEÇMİŞ.....	53

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Işık hızı verisi.....	3
Çizelge 4.1. Örnek çapı $n = 20$ iken bozulma sayısı m 'nin artması ile konum parametresi tahmin edicileri üzerindeki etkinin artması.....	7
Çizelge 5.1. 28 Canlı çeşidi için vücut ve beyin ağırlıkları verisi.....	15
Çizelge 5.2. 28 Canlı çeşidi için Mahalanobis uzaklıkları ve MVE için sağlam istatistiklere dayalı uzaklıklar.....	16
Çizelge 5.3. 28 Canlı çeşidi için Mahalanobis uzaklıkları ve MCD için sağlam istatistiklere dayalı uzaklıklar.....	20
Çizelge 5.4. Kovaryans determinantı ve vektör varyansı arasındaki hesaplama hızı oranları.....	31
Çizelge 6.1. $X_1 \sim N_2 \left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$ ve $X_2 \sim N_2 \left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$ olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri.....	33
Çizelge 6.2. $X_1 \sim N_2 \left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right)$ ve $X_2 \sim N_2 \left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \right)$ olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri.....	37
Çizelge 6.3. $X_1 \sim N_2 \left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right)$ ve $X_2 \sim N_2 \left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right)$ olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri.....	38

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 5.1. Mahalanobis uzaklığına ve MVE yöntemine göre oluşturulan elipsoidler.....	17
Şekil 5.2. Mahalanobis uzaklığına ve MCD yöntemine göre oluşturulan elipsoidler.....	22

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklama

h	Bozulma noktası
m	Bozulma sayısı
n	Örnek çapı
p	Değişken sayısı
α	Anlamlılık düzeyi
ε	Verinin bozulma oranı
[.]	Tam değer fonksiyonu
χ^2	Ki-kare dağılımı
Σ	Yığına ait kovaryans matrisi
λ_i	Özdeğer
$c^2_{n,p}$	Düzeltilme faktörü

Kısaltmalar

Açıklama

C	Örnek kovaryans matrisi
CD	Kovaryans determinanı
E	Beklenen değer
EKK	En küçük kareler
F	Dağılımların karışımı
FMCD	Hızlı en küçük kovaryans determinanı
H_i	Alt küme
MCD	En küçük kovaryans determinanı
MD_i	<i>i.</i> gözleme ilişkin Mahalanobis uzaklığı
MLE	En çok olabilirlik tahmin edicisi

Kısaltmalar**Açıklama****MVE**

En küçük hacimli elipsoid

MVV

En küçük vektör varyansı

N

Normal dağılım

R²

Belirleme katsayısı

T

Örnek aritmetik ortalaması vektörü

Tr

İz

VE

Elipsoid hacmi

vec

Vektör

VV

Vektör varyansı

1.GİRİŞ

Son yıllarda önemli istatistik çalışmalarından biri sağlam istatistiklerin elde edilme yöntemleri ve buna bağlı olarak aykırı değerlerin veri içinde tespit edilmesidir. Klasik istatistiksel yaklaşımlar veri içindeki aykırı değerlerden aşırı derecede etkilendikleri için klasik istatistiksel yaklaşımlarla elde edilen konum ve dağılış parametrelerinin tahminleri veriyi iyi temsil etmezler.

Konum ve dağılış parametrelerinin tahminlerini bulmak için kullanılan en eski ve en güvenilir yöntemlerden biri “En Çok Olabilirlik” (Maksimum Likelihood Estimator(MLE)) yöntemidir. Ancak bu yöntem aykırı değerlere karşı aşırı derecede duyarlı olduğundan veride aykırı değerler bulunması durumunda bu yöntemin kullanılması doğru değildir.

Sağlam istatistikler klasik istatistiksel yöntemlere alternatif bir yaklaşım verir. Klasik yöntemler uygulamada pek sağlanmayan varsayımlara aşırı derecede güvenirlir. Örneğin; hataların normal dağıldığı veya Merkezi Limit Teoremince tahmin edicilerin dağılımının normal olduğu varsayılır.

$X_1, X_2, \dots, X_n$ aynı dağılımlı ($X \sim N(\mu, \sigma^2)$) bağımsız, tesadüfi değişkenlerden oluşan bir tesadüfi örnek olsun. Klasik istatistiksel yöntemlere göre bu tesadüfi değişkenlerin konum ve dağılış parametrelerinin tahminleri MLE yöntemi ile bulunabilir. Ancak bu tesadüfi örnekteki kimi tesadüfi değişkenlerin bir kısmının varsayılanlardan farklı dağılım veya dağılımlardan geldiği düşünüldüğünde X tesadüfi değişkeni için varsayılan dağılımın konum ve dağılış parametresini tahmin etmek klasik istatistiksel yöntemlerle yapılamaz, çünkü elde edilen sonuçlar veriyi iyi temsil etmeyecektir.

$$F = (1 - \varepsilon)G + \varepsilon H \quad (1.1)$$

Eş. 1.1 bahsedilen tesadüfi örneğin dağılımını gösterebilirsin. Burada G , normal dağılıma sahip X tesadüfi değişkenlerinin yoğunluk fonksiyonu, H bu tesadüfi örnekte varsayılanlardan farklı başka bir dağılıma sahip tesadüfi değişkenlerin yoğunluk fonksiyonu ve ε sabit olmak üzere tesadüfi örnekte varsayılanlardan farklı başka bir dağılıma sahip tesadüfi değişkenlerin oranını ifade etmektedir. Bu nedenle dağılımların karışımı (mixture of distributions) F ile gösterilmiştir.

Böyle bir problemle karşılaşıldığında bu sorunu ortadan kaldırmak için birçok önemli istatistikçi sağlam yöntemler bulmuş ve bu yöntemleri veride tek ve birden fazla aykırı değer bulunması durumlarında uygulamaya çalışarak en sağlam yöntemi elde etmeye çalışmışlardır.

Sağlamlık kavramı dağılım ile ilişkilendirilebileceği gibi bundan başka bir varsayımla da ilişkilendirilebilir. Burada bir istatistiğin sağlam bir istatistik olarak nitelendirilmesinde amaç, o istatistiğin aykırı gözlemlerin sayısının artmasına karşın parametre için iyi tahminler üretmeye devam edebiliyor olmasıdır.

Konum ve kovaryans matrisinin sağlam tahmini için En Küçük Hacimli Elipsoid (Minimum Volume Elipsoid : MVE), En Küçük Kovaryans Determinantı (Minimum Covariance Determinant : MCD), Hızlı En Küçük Kovaryans Determinantı (Fast Minimum Covariance Determinant : FMCD) gibi birçok yöntem ileriye sürülmüş ve bu yöntemler çeşitli veri gruplarına uygulanarak sağlamlığı araştırılmıştır. Herwindiati ve ark. (2007) bu yöntemlere alternatif olarak En Küçük Vektör Varyansı (Minimum Vector Variance : MVV) adlı bir yöntem geliştirmiş ve bu yöntemin kullanımının bazı durumlarda daha avantajlı olduğunu ve daha sağlam sonuçlar elde edildiğini belirtmişlerdir. Çalışmamızın temelinde, son olarak geliştirilen bu yöntemin diğer yöntemlerden hangi durumlarda daha avantajlı olduğu uygulama aşamasında nedenleri ile birlikte irdelenecektir.

Öncelikle sağlamlığın anlaşılabilmesi için *Aykırı Değer* kavramının ne olduğunu tanımlamak gerekmektedir.

2. AYKIRI DEĞER

Aykırı değer verideki birimlerin çoğunluğuna uymayan, tipik olmayan gözlemler olarak adlandırılabilir. Birçok araştırmacı veride aykırı değer bulunması durumunda değeri kaydederken veya bir yerden başka bir yere geçirirken hata yapılabileceğini düşünerek orijinal veriye geri döner. Ancak burada amaç aykırı değerleri silmek değil, onları tespit edip göz önünde bulundurarak sağlam tahmin ediciler üretmektir.

Aykırı değerler birbirlerini gizleyerek birbirlerini etkileyebilirler. Örneğin; bir küçük ve bir büyük aykırı değer bulunduğ bir veri seti düşünelim. Tahmin edilen standart sapmanın büyük aykırı değer tarafından oldukça şişirilmesi sonucu, küçük aykırı değer sanki normal bir değere sahipmiş gibi görülmesi problemine literatürde *Maskeleye Problemi* denilmektedir. Büyük aykırı değer çıkarıldıktan veya düzeltildikten sonra tahmin edilen standart sapma küçülür ve küçük aykırı değer sıra dışı olduğu görülür.

2.1. Örnek

Simon Newcomb'un "*Işık hızı*" verisi aykırı değer tanımını kavramak için kullanılan ve 66 gözlemin içinde 2 aykırı değere sahip bir veridir [Gelman ve ark. 2004].

Çizelge 2.1. Işık hızı verisi

28	26	33	24	34	-44	27	16	40	-2
29	22	24	21	25	30	23	29	31	19
24	20	36	32	36	28	25	21	28	29
37	25	28	26	30	32	36	26	30	22
36	23	27	27	28	27	31	27	26	33
26	32	32	24	39	28	24	25	32	25
29	27	28	29	16	23				

Veride -44 ve -2 değerlerinden oluşan bir küçük ve bir büyük aykırı değer bulunmaktadır. -44 değeri örneğin standart sapmasını büyültecek ve -2 aykırı değer değilmiş gibi görülecektir. -2'nin aykırı değer olduğunu görebilmek için ise -44 değeri ya düzeltilmeli ya da çıkarılmalıdır. Oysa sağlam yöntemlerle dağılış parametresini tahmin ederken -44 değerini veriden çıkarmadan -2 değerinin de aykırı değer olduğu hemen görülecektir.

Aykırı değerlerin çoğunlukla “kötü” veri olduğu ve araştırmacıyı yanılttığı söylenmesinin aksine bu değerler araştırmacıya konu ile ilgili olarak beklenmeyen bilgiler verebilir.

Klasik istatistiksel yöntemlere göre elde edilen örnek ortalaması, örnek varyansı, örnek kovaryansları ile korelasyonları ve bunlara bağlı olarak regresyon modeline ilişkin En Küçük Kareler(EKK) tahminleri, güven aralıkları, belirleme katsayısı R^2 gibi istatistikler aykırı değerlerden aşırı derecede etkilenirler. Oysa sağlam yöntemler aykırı değerleri göz önünde bulundurarak bunlardan bu derecede etkilenmeyen tahminler elde etmeyi amaçlar.

Çok değişkenli veri analizinde aykırı değerlerin tespiti için konum parametresi tahmin edicileri üzerinde çalışmak dağılış parametresi tahmin edicileri üzerinde çalışmaktan daha önemlidir. Konum parametresi tahmininde meydana gelen herhangi bir değişikliğin doğrudan dağılış parametresi tahminine yansımaları ve dağılış parametresi tahmininde daha büyük bir değişikliğe yol açmasından dolayı aykırı değerlerin tespitinde kullanılan yöntemlerin tümünde konum parametresi tahmini ile ilgilenilmekte ve yorumlar konum parametresi tahmini üzerinden yapılmaktadır.

3. KONUM PARAMETRESİ TAHMİN EDİCİLERİ

Ortalama merkezi eğilimin sağlam bir ölçüsü olmadığı halde, medyan merkezi eğilimin sağlam bir ölçüsüdür. Çünkü aykırı değerler ortalama üzerinde büyük bir etki yaratırlar ve verinin merkezinden uzakta bulunurlar. Eğer ortalama, verinin merkezinin konum ölçüsü gibi düşünülürse, aykırı değerler mevcutken bulunan ortalama değeri sağlam olmayacaktır. Şayet veri, merkezde simetrik dağılmışsa ortalama ve medyan yaklaşık olarak eşit çıkar ancak tersi durumda bu geçerli değildir. Sonuç olarak medyanın, ortalamaya göre iyi ve sağlam bir alternatif olduğu söylenebilir.

2.1 Örneğine ait veride iki tane aykırı değer bulunması durumunda ortalama 26.21 bulunur. Aynı veriye ait medyan değerinin ise 27 olarak bulunmasından yola çıkarak medyanın ortalamaya göre aykırı değerlerden daha az etkilendiği ve bu değer merkezi eğilimin sağlam bir ölçüsü olduğu ortaya çıkmaktadır.

Konum parametresini tahmin etmede en çok kullanılan sağlam yöntem “Budlanmış ortalama” dır.

3.1. Budanmış Ortalama

Budanmış ortalama ilk en küçük m gözlem ve son en büyük m gözlemi veriden çıkarmak olarak tanımlanabilir ve Eş. 3.1 yardımıyla budanmış ortalama değeri bulunabilir.

$$\bar{X}_\alpha = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} X_{(i)} \quad (3.1)$$

Eş. 3.1’de $m = (n-1)\alpha$ ve $\alpha \in 0,1/2$ ‘dır. $[\cdot]$ ifadesi de tam değer fonksiyonudur. Budanmış ortalama aykırı değerlerden ortalamaya göre daha az etkilenmektedir.

Budanmış ortalamanın performansının görelî olarak iyi olmasına rağmen, daha iyi tahmin ediciler mevcuttur. Bu tahmin ediciler M-tahmin edicileri (Huber, 1973), R tahmin edicileri (Jureckova, 1971; Jaeckel, 1972), L tahmin edicileri (Bickel, 1973; Koenker ve Bassett, 1978) başlıkları altında toplanmaktadırlar.

Yukarıda genel isimleri verilen bu gruplardan herhangi birinde yer alan tahmin edicilerden hangisinin daha sağlam olduğunun iki temel ölçüsü bulunmaktadır. Bunlar “Etki Fonksiyonu ” ve “Bozulma Noktası”dır. Burada Bozulma Noktası ölçüsü tanımlanmakla yetinilecektir.

4. BOZULMA NOKTASI

$x_1, x_2 \dots x_n$ örneği Eş. 1.1'deki dağılımdan gözlenmiş olsun. Burada $G \sim N(\mu, \sigma^2)$ ve H herhangi bir dağılım olabilir. Yani H , elimizdeki varsayılan dağılımdan gelen “temiz” veriyi bozan birimlerin gözlemlendiği dağılımı ifade etsin ve ε verinin bozulma oranını göstere. Konum parametresine ait tahmin edicinin bu bozulma karşısındaki sağlamlığını ölçmek için Maronna ve ark.(2006) tarafından verilen aşağıdaki örneği inceleyelim.

$n = 20$ olmak üzere $X = X_1, X_2 \dots X_n$ tesadüfi örneği $X_i \sim (0,1)$ aynı dağılımlı ve bağımsız tesadüfi değişkenlerden oluşsun. $X_0 = 1000$ gibi bir değerin veriye m defa eklenmesiyle konum parametresi tahmin edicisinin nasıl değiştiğini görmek Çizelge 4.1 yardımıyla daha kolaylaşacaktır.

Çizelge 4.1. Örnek çapı $n = 20$ iken bozulma sayısı m 'nin artması ile konum parametresi tahmin edicileri üzerindeki etkinin artması

m	ε	Ortalama	Medyan	α -budanmış ort.
0	0	0.00	0.00	0.00
1	1/20	50	0.00	0.04
2	2/20	100	0.01	55.59
4	4/20	200	0.21	166.7
5	5/20	250	0.34	222.3
7	7/20	350	0.48	333.4
9	9/20	450	0.76	444.5
10	10/20	500	500.5	500.0

Burada Budanmış ortalamanın hesaplanması için $\alpha = 0.25$ olarak alınmıştır.

Çizelge 4.1'e göre $n = 20$ iken, dağılımı $N(0,1)$ olan aynı dağılımlı tesadüfi değişkenlerden oluşan örnek alındığında, bu dağılımın ortalaması 0 olur. Ancak

örneğe 1 tane 1000 değerinden oluşan veriyi bozan bir değer eklendiğinde ortalamanın 50'ye yükseldiği görülüyor. Medyan ise hala 0 ve %25'lik budanmış ortalama 0.04'tür.

Örneğe 2 tane 1000 değeri eklendiğinde ortalama 100'e ulaşır, medyan 0.01 ve budanmış ortalama 55.49'a yükselerek aşırı bir artış gösterir. $m = n/2$ olduğunda yani verinin yarısı 1000 değerinden oluştuğunda, bir başka deyişle verinin yarısı bozulduğunda ortalama 500, medyan 500.5 ve budanmış ortalama 500'e ulaşıyor.

Çizelge 4.1'de görüldüğü üzere her üç istatistik de bozulma sayısı m arttıkça sağlamlıktan uzaklaşmaktadır. $m > 0$ olduğu durumlar için ortalama istatistiği, $m \geq n/2$ içinse medyan istatistiği sağlamlıktan uzaklaşmaktadır. Medyan örneğe başka bir dağılımdan karışan gözlem sayısı $n/2$ (yani %50) oluncaya kadar sağlamlığını korumaktadır.

Sonuç olarak, tahmin edicinin yüksek bozulma noktasına sahip olması, onu daha sağlam yapmaktadır. Bozulma noktası %50'den büyük olduğunda, istatistik H dağılımından gözlenmiş gözlemlerden daha çok etkilenecek ve G yoğunluk fonksiyonunun konum parametresi hakkında çok sapmalı tahminler verecektir.

Araştırmacılar çok değişkenli durumda sağlam konum ve dağılım parametreleri tahminlerini elde etmede kullanacakları yöntem için bozulma noktası oranının mümkün olduğunca yüksek hatta %50 düzeyine sahip olmasını arzulamışlardır. Bu konu üzerinde yapılan ilk çalışmalarda çeşitli yöntemler bulunmuş ancak bozulma noktası oranının düşük değerlerde seyretmesinden dolayı bu yöntemler geçerliliğini yitirmişlerdir. Çünkü yukarıda bahsedildiği gibi bozulma noktası oranı ne kadar yüksekse yöntem o derecede sağlamdır.

Tüm bu anlatılanlara göre, "Bozulma Noktası"nın sağlamlığı ölçmek için önemli bir ölçüt olduğu açıktır.

5. ÇOK DEĞİŞKENLİ DURUMDA AYKIRI DEĞERLERİN TESPİTİ İÇİN ÖNERİLEN METOTLAR

Çok değişkenli veri setlerinde aykırı değerleri bulmak, değişken sayısı $p > 2$ olduğunda güç olabilmektedir ve birçok yöntem $\frac{n}{p+1}$ tane aykırı değer varlığında bozulmaktadır. Böyle bir durumda konum ve kovaryans matrisinin sağlam tahmini için Rousseeuw(1985) tarafından önce MVE ve MCD yöntemi, ilerleyen dönemde ise Rousseeuw ve van Driessen(1999) tarafından FMCD yöntemi geliştirilmiştir. Bu yöntemler finans, kimya, elektrik mühendisliği gibi birçok alanda kullanılmıştır.

Herwindiati ve ark. (2007) tarafından yapılan çalışmalarda MVE, MCD, FMCD yöntemlerinin geniş ve büyük boyutlu veri setleri için kullanışsız olduğu tespit edilmiş, büyük boyutlu veri setleriyle çalışılırken bu yöntemlerin zaman açısından dezavantaja sahip olduğu ve hesaplamaların karmaşık bir yapıya dönüştüğü söylenmektedir.

FMCD yönteminin MVE ve MCD yöntemine göre hesaplanabilirlik açısından daha üstün, buna rağmen geniş ve büyük boyutlu veri setleri için kullanışsız olduğu belirtilmiştir [Herwindiati ve ark. 2007]. Bu yöntem “Kovaryans Determinantı(CD)”nı minimize etmeye çalışır ve bu da hesaplama zamanı açısından bir olumsuzluk yaratır. Bu gibi olumsuzluklarla karşılaşmamak için kovaryans determinantını minimize etmek yerine “Vektör Varyansı(VV)” minimize eden MVV yöntemi diğer yöntemlere alternatif olarak geliştirilmiştir [Herwindiati ve ark. 2007].

MVV yönteminin hangi durumlarda avantajlı olduğunu belirtmek, kullanılmasındaki ve tercih edilmesindeki nedenleri ortaya koymak ve daha önceden öne sürülmüş diğer yöntemlerle karşılaştırmasını yapabilmek için öncelikle MVE, MCD ve FMCD yöntemlerinin üzerinde durulacaktır. Aşağıda çok değişkenli durumda aykırı değerlerin tespiti için geliştirilen yöntemler sırasıyla tanıtılmıştır.

5.1. En Küçük Hacimli Elipsoid (MVE) Yöntemi

Çok değişkenli veri kümesinde aykırı değerlerin bulunması, özellikle birkaç aykırı değer bulunması halinde oldukça önemlidir. Klasik aykırı değer belirleme yöntemi olarak kullanılan Mahalanobis Uzaklığı(MD_i) yönteminde çok değişkenli veri kümesinde bir veya iki aykırı değer bulunması halinde her bir gözlem için uzaklık hesaplanır.

$$X = (X_1, X_2 \dots X_n) = \left[\begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{p1} \end{pmatrix}, \dots, \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{pn} \end{pmatrix} \right] \text{ tesadüfî örnek olmak üzere } i. \text{ gözleme ilişkin}$$

Mahalanobis uzaklığı,

$$MD_i = \sqrt{(X_i - T(X))C^{-1}(X_i - T(X))^t} \quad (5.1)$$

ile bulunur.

Burada $T(X)$ ve $C(X)$ sırasıyla aritmetik ortalama vektörü ve örnek kovaryans matrisi olup,

$$T(X) = \frac{1}{n} \sum_{i=1}^n X_i \quad (5.2)$$

ve

$$C(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - T(X))(X_i - T(X))^t \quad (5.3)$$

olarak tanımlanır.

MD_i , her bir X_i noktasının veri kümesinin merkezine olan uzaklığını gösterir. MD_i , $\sqrt{\chi^2_{p,\alpha}}$ (*ki-kare*) dağılımına sahiptir. $MD_i > \sqrt{\chi^2_{p,\alpha}}$ olan gözlemler aykırı değer olarak tanımlanan gözlemlerdir.

Veride aykırı değerler bulunması halinde klasik En Çok Olabilirlik yöntemi gibi yöntemlerle bulunan konum ve kovaryans matrisinin tahmin edicileri olan $T(X)$ ve $C(X)$ sağlam olmadığından aykırı değerler her zaman büyük değerli MD_i 'ye sahip olmazlar ve böylece aykırı değer olarak görülmezler. Çünkü aykırı değerler $T(X)$ 'i önemli derecede etkiler ve $C(X)$ de $T(X)$ yönünde büyür.

Mahalanobis uzaklığı aykırı değerler tarafından etkilenen örnek ortalama ve kovaryans matrisine dayandığından dolayı veri kümesinde pek çok aykırı değer bulunması durumunda bu uzaklığın kullanılması uygun olmamakta ve hesaplama etkinliği zayıf kalmaktadır. Ayrıca çok değişkenli veri kümelerinde aykırı değer olan gözlemlerin aykırı değer olarak elde edilememesi durumunda “maskeleme” (masking) problemi veya aykırı değer olmayan gözlemlerin aykırı değer olarak bulunması durumunda “batma”(swamping) problemi meydana geldiğinden bu iki problemle karşılaşıldığında Mahalanobis uzaklığının kullanışsız bir yöntem olduğu çok iyi bilinmektedir.

Çok değişkenli veri kümelerinde çoklu aykırı değerlerin belirlenmesi problemi, X değişken matrisinin boyutunun büyümesi ile maskeleme ve batma problemlerini ortaya çıkarmış ve 1980'lerden itibaren bilgisayar teknolojisindeki hızlı ilerleme ile bu problemlerin üstesinden gelebilecek yöntemler geliştirilmiştir.

Maskleme ve batma problemlerinin varlığında aykırı değerlerden etkilenmeden çok değişkenli veri kümesi ile ilgili analizi yapabilmek için çoğu zaman sağlam yöntemlerden yararlanılmaktadır. Bu yöntemler konum ve kovaryansın sağlam

tahmin edicilerine dayanmaktadır. Bu sağlam tahmin edicilere dayanan uzaklık istatistiklerini kullanmak aykırı değerleri tespit etmede daha iyi bir yöntemdir.

Bu amaçla geliştirilen ilk yöntem En Küçük Hacimli Elipsoid'dir. Rousseeuw(1985) tarafından geliştirilen bu yöntemde, aykırı değerleri tespit etmek için veriden “şüphelenilenleri”(suspects) ayırarak kullanışlı bir yöntem olan “aykırı değer etiketlemesi” önerilmiş ve konum ile kovaryans yapısının sağlam tahminleri ilk olarak MVE yöntemi ile geliştirilmiştir. MVE yöntemine göre bulunan $T(X)$ ve $C(X)$ ile hesaplanan uzaklıklar daha sağlam sonuçlar vermektedir ve buna paralel olarak aykırı değerlerin daha güvenilir bir şekilde tespit edilmesi sağlanmaktadır.

Rousseeuw(1985) tarafından önerilen MVE yöntemi gözlemlerin en az yarısını içeren, çok değişkenli verideki aykırı değerlere karşı konum ve dağılış parametrelerinin sağlam tahminini veren, bozulma noktası yaklaşık olarak %50 olan yani gözlemlerin yarısı aykırı değerlerden oluşsa bile sağlamlığını koruyan bir tahmin edicidir. Bu yöntemde amaç n gözlem sayısı, p değişken sayısı olmak üzere

$X_{n \times p}$ veri matrisi için h birimden oluşan tüm alt kümeleri bularak $\binom{n}{h}$

kombinasyonu kadar alt küme için elipsoid hacimlerinin hesaplanıp en küçük hacme sahip elipsoidi oluşturan alt kümeyi belirlemektir. Bu “en iyi” alt kümedeki gözlemler temiz, diğer gözlemler ise aykırı değer olarak etiketlenir.

$h = \left\lceil \frac{n}{2} \right\rceil + 1$ olduğunda $T(X)$, X 'in en az h tane noktasını içeren en küçük hacimli elipsoidin merkezidir.

n çok büyük olduğunda $\binom{n}{h}$ sayıda alt küme için elipsoid hacimlerini hesaplamak zaman alıcıdır. Hesaplamadaki bu zorluğun üstesinden gelebilmek için Rousseeuw ile Leroy MVE yöntemini geliştirme yoluna gitmişlerdir. Temelleri

Rousseeuw(1984) tarafından atılan bu yöntemde $p+1$ tane farklı gözlemden oluşan birçok tesadüfi alt küme seçilir ve her bir J alt kümesi için,

$$d_i(T_J, C_J) = \sqrt{(x_i - T_J)C_J^{-1}(x_i - T_J)^t} \quad , \quad i = 1, \dots, n \quad (5.4)$$

değerleri hesaplanır. Burada $J = i_1, \dots, i_{p+1}$ alt kümesi için ortalama ve kovaryans matrisi sırasıyla,

$$T_J = \frac{1}{p+1} \sum_J x_i \quad (5.5)$$

ve

$$C_J = \frac{1}{p} \sum_J (x_i - T_J)(x_i - T_J)^t \quad (5.6)$$

ile hesaplanmaktadır.

Alt kümede h tane gözlem olana kadar yukarıdaki işlemlere devam edilir. $d_i(T_J, C_J)$ eşitliğinde n tane değer $100(h/n)$ inci yüzdeliği m_J olmak üzere,

$$m_J^2 = (x_i - T_J)C_J^{-1}(x_i - T_J)^t \quad_{h:n} \quad (5.7)$$

ile ifade edilmektedir.

T_J ve C_J değerlerine dayanan ve h tane gözlemi içeren elipsoidin hacmi,

$$(\det(m_J^2 C_J))^{1/2} = (\det(C_J))^{1/2} (m_J)^p \quad (5.8)$$

ile orantılıdır ve Eş.5.8'deki değerin en küçük olduğu elipsoid seçilerek “en iyi” J alt kümesi belirlenir. Böylelikle seçilen bu J alt kümesi için,

$$T(X) = T_J \quad (5.9)$$

ve

$$C(X) = (\chi_{p,\alpha}^2)^{-1} c_{n,p}^2 m_J^2 C_J \quad (5.10)$$

elde edilir.

Özdeğerler $\lambda_1, \lambda_2, \dots, \lambda_p$ olmak üzere,

$$\text{Elipsoid Hacmi (VE)} = \frac{\sqrt{\pi^p \cdot c^p}}{\Gamma\left(\frac{p}{2} + 1\right)} \sqrt{\lambda_1, \lambda_2, \dots, \lambda_p} \quad (5.11)$$

ile ifade edilmektedir. Burada c , pozitif bir sabit ve π , bilinen pi sayısı 3.14'tür

Lopuhaa ve Rousseeuw(1991) $h = \frac{n+p+1}{2}$ olduğunu belirtmişlerdir. Rousseeuw ve van Zomeren(1990) ise simülasyon çalışmalarının temelinde $p \leq 8$ için $c_{n,p}^2 = 1 + 15(n-p)^2$ ifadesini düzeltme faktörü olarak kullanmışlardır [Hadi, 1992].

Aşağıda Mahalanobis Uzaklığı ile MVE yönteminin Rousseeuw ile Leroy(2003)'ün “*Vücut ve Beyin Ağırlıkları*” verisi üzerinde uygulaması yapılmıştır.

5.1. Örnek

“*Vücut ve Beyin Ağırlıkları*” verisinde 28 canlı çeşidi için vücut ve beyin ağırlıkları yer almakta ve ağır bir vücudu yönetebilmek için büyük bir beynin gerekli olup olmadığı araştırılmaktadır [Rousseeuw ve Leroy, 2003].

$n = 28$ ve $p = 2$ için $p + 1 = 3$ olup mümkün 3276 tane alt örnek vardır. Bununla birlikte alt örneklerin sayısı artan n ve p ile birlikte artmakta ve büyük değerli n ve p için hesaplama zorlaşmaktadır.

Çizelge 5.1. 28 Canlı çeşidi için vücut ve beyin ağırlıkları verisi

(i)	Tür	Vücut Ağırlığı(kg) (x_i)	Beyin Ağırlığı(gr) (y_i)
1	Kunduz	1.350	8.100
2	İnek	465.000	423.000
3	Kurt	36.330	119.500
4	Keçi	27.660	115.000
5	Kobay	1.040	5.500
6	Diplodocus	11700.00	50.000
7	Asya Fili	2547.000	4603.000
8	Eşek	187.100	419.000
9	At	521.000	655.000
10	Potar Maymunu	10.000	115.000
11	Kedi	3.300	25.600
12	Zürafa	529.000	680.000
13	Goril	207.000	406.000
14	İnsan	62.000	1320.000
15	Afrika Fili	6654.000	5712.000
16	Triceratops	9400.000	70.000
17	Rhesus Maymunu	6.800	179.000
18	Kanguru	35.000	56.000
19	Hemster	0.120	1.000
20	Fare	0.023	0.400
21	Tavşan	2.500	12.100
22	Koyun	55.500	175.000
23	Jaguar	100.000	157.000
24	Şempanze	52.160	440.000
25	Brachiosaurus	87000.000	154.500
26	Sıçan	0.280	1.900
27	Kötebek	0.122	3.000
28	Domuz	192.000	180.000

Çizelge 5.2’de 28 canlı çeşidi için vücut ve beyin ağırlıkları verisine ait Mahalanobis uzaklıkları ve MVE yöntemi için elde edilen uzaklık değerleri verilmektedir [SAS Ins, 1999].

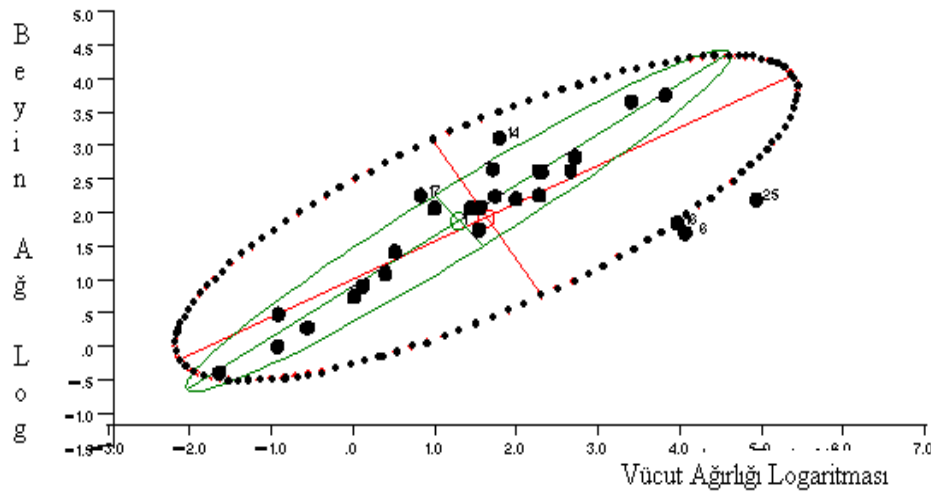
Çizelge 5.2. 28 Canlı çeşidi için Mahalanobis uzaklıkları ve MVE için sağlam istatistiklere dayalı uzaklıklar

N	Mahalanobis Uz.	MVE için Sağlam İstatistiklere Dayalı Uz.
1	1.006591	0.897076
2	0.695261	1.405302
3	0.300831	0.186726
4	0.380817	0.318701
5	1.146485	1.135697
6	2.644176	8.828036
7	1.708334	1.699233
8	0.706522	0.686680
9	0.858404	1.084163
10	0.798698	1.580835
11	0.686485	0.693346
12	0.874349	1.071492
13	0.677791	0.717545
14	1.721526	3.398698
15	1.761947	1.762703
16	2.369473	7.999472
17	1.222253	2.805954
18	0.203178	1.207332
19	1.855201	1.773317
20	2.266268	2.074971
21	0.831416	0.785954
22	0.416158	0.342200
23	0.264182	0.918383
24	1.046120	1.782334
25	2.911101	9.565443
26	1.586458	1.543748
27	1.582124	1.808423
28	0.394664	1.523235

Mahalanobis uzaklığına göre sadece 25. gözlem $\sqrt{\chi^2_{2,025}} = 2.72$ değerini aşarak aykırı değer olarak tespit edilmektedir (Çizelge 5.2). Buradan anlaşılabacağı gibi 28 birimlik örnekte sadece bir tane aykırı değer vardır ve bu değer beyin ağırlığı vücut ağırlığına göre çok küçük bir değere sahip olan Brachiosaurus dinazoruna aittir. Oysa MVE yönteminde kullanılan uzaklıklara göre 6, 14, 16, 17 ve 25. gözlemler aykırı

değer olarak tespit edilmektedir (Çizelge 5.2). Bu değerler beyin ağırlığı vücut ağırlığına göre küçük olan 3 dinazor çeşidi ve beyin ağırlığı vücut ağırlığına göre büyük olan insan ve rhesus maymununa aittir.

Şekil 5.1’de verilen Mahalanobis uzaklığına göre çizilen elipsoid ve MVE yöntemine göre elde edilen elipsoid incelenirse bu iki yöntemin aykırı değerleri tespit etmede gösterdikleri başarı arasındaki fark daha iyi anlaşılacaktır. Burada veriye logaritmik dönüşüm yapılmıştır çünkü orijinal veride çok büyük ve çok küçük değerler bulunduğundan logaritmik dönüşüm yapılmadığı takdirde veriye ait grafiği oluşturmak oldukça zor olmaktadır.



Şekil 5.1. Mahalanobis uzaklığına ve MVE yöntemine göre oluşturulan elipsoidler

Klasik yöntemlerle elde edilen tahmin ediciler bu aykırı değerler tarafından genişletilir ve buna bağlı olarak Mahalanobis uzaklığına göre oluşturulmuş %97.5’lik tolerans elipsoidi(kesikli çizgiyle gösterilen) sadece 25.gözlemi elipsoid dışında bırakarak tek bir aykırı değer tespit eder.

MVE yöntemine göre oluşturulan elipsoide göre(kesiksiz çizgiyle gösterilen) 6, 14, 16, 17 ve 25. gözlemler elipsoidin dışında kalarak 5 tane aykırı değer tespit edilmiş olur (Şekil 5.1).

Burada MVE yöntemi mümkün tüm alt kümeleri belirlemiş, sağlam uzaklıkları hesaplamış ve elipsoidi oluşturmuştur. Oysa n ve p büyük değerlere sahip olduğunda tüm mümkün alt kümeleri belirlemek, uzaklıkları hesaplamak ve elipsoidi oluşturmak oldukça güç ve zaman alıcıdır. Rousseeuw ve van Zomeren(1990) bu zorluktan kaçınmak için aynı veri setini kullanarak sadece 1500 alt örnek belirlemiş ve uzaklıkları bu 1500 alt örneğe göre hesaplamıştır. Seçilen 1500 alt örnekte 1,2,28 alt kümesinin minimum hacme sahip olduğu görülmüş ve elipsoid oluşturulmuştur. 3276 mümkün alt örneğin içinde 2,9,12 , 4,12,22 , 8,19,21 , 11,17,21 alt kümelerinin kovaryans matrisleri 10^{-8} 'den daha küçük olduğu için yukarıdaki çalışmada bu 4 alt küme ihmal edilip geriye kalan 3272 alt örnek üzerinden MVE hesaplanmış ve 1,22,28 alt kümesinin minimum hacme sahip olduğu görülmüştür.

İki boyutlu örnekler için verinin grafiğine bakılması büyük ölçüde fayda sağlamaktadır. Ancak uygulamada genellikle büyük boyutlu örnekler ile çalışıldığından verinin grafiğinin oluşturulması zorlaşmaktadır. Bu nedenle görsel anlayışa güvenemeyeceğimizden aykırı değerlerin tespiti için uzaklıkların hesaplanması gerekmektedir.

MVE yönteminin klasik yöntemle göre daha sağlam sonuçlar vermesine ve aykırı değerlerin tespit edilmesinde daha güvenilir olmasına rağmen bu yöntemin kullanılmasında bazı zorluklarla karşılaşmaktadır. Bu zorluklar aşağıdaki gibi sıralanabilir,

- 1- Alt örneklerin sayısına karar verilmesi
- 2- Bu yöntem $p + 1$ boyutlu her alt örneğin p ranka sahip olduğu varsayımı altında gerçekleşir. Oysa uygulamada bir alt kümenin rankının p 'den küçük olması durumunda determinant 0'a eşit olur ve bu nedenle karşılık gelen elipsoidin hacmi 0 olur. Rousseeuw ve van Zomeren(1990) bazı tekil kovaryans matrisine sahip alt örnekleri tamamen ihmal ederek bu zorluktan kaçınmışlardır.

3- $p+1$ boyutlu alt örnekler p ranka sahip olsa bile bazı alt örneklerin kovaryans matrisleri yaklaşık olarak 0 determinanta sahip olur, bu nedenle karşılık gelen elipsoidler 0 hacme sahip olur.

MVE yönteminde n veya p 'nin büyük olması bu yöntemin uygulanabilirliğini zorlaştırmakta ve hesaplanmasındaki karmaşıklık nedeniyle zaman alıcı olmaktadır. Bu nedenle MVE yönteminin sadece küçük boyutlu veri setlerine uygulanabilmesi bu yöntemin en büyük dezavantajıdır. Oysa uygulamada genellikle büyük boyutlu veri setleriyle çalışılmaktadır. MVE yönteminin genel olarak $\frac{n}{p} > 5$ olması

durumunda veriye uygulanması hesaplamayı zorlaştırmamaktadır. Bu yönteminin yukarıda bahsedilen dezavantajlara sahip olmasından dolayı Rousseeuw(1985) bu yönteme alternatif olarak MCD yöntemini geliştirmiş ve geliştirilen bu yeni yöntemde bahsedilen problemlerle karşılaşmamayı amaçlamıştır.

5.2. En Küçük Kovaryans Determinantı (MCD) Yöntemi

Rousseeuw'un MCD yöntemi çok değişkenli konum ve dağılış parametrelerinin oldukça sağlam bir tahmin edicisidir ve MVE yöntemi gibi %50'lik bozulma noktasına sahiptir. Bu yöntemin amacı, n tane gözlem üzerinden kovaryans matrisinin determinantı en küçük olan alt kümeleri ve h tane gözlemi araştırarak, kovaryans matrisinin determinantını en küçük yapacak olan h tane gözlemden oluşan alt kümeyi bulmaktır.

MCD yönteminde $h = \frac{n+p+1}{2}$ olarak alınır. Elde edilecek olan h tane gözlemden hesaplanan ortalama MCD' nin konum parametreleri tahminleri, aynı gözlemlerden elde edilen varyans-kovaryans matrisi de dağılış parametreleri tahminleri olacaktır.

MCD yönteminin algoritması aşağıdaki gibidir,

- 1- $\binom{n}{h}$ sayıda alt örnek oluşturulur.
- 2- $\binom{n}{h}$ sayıda alt örneğin her biri için $T(x)$ ve $C(x)$ istatistikleri hesaplanır.
- 3- $|C(x)|$ değeri en küçük olan alt örnek temel veri alt örneğini oluşturur ve bu alt örnek üzerinden elde edilen $T(x)$ ve $C(x)$ en uygun tahmin edicilerdir.
- 4- Bu temel alt örnek etrafında ki-kare dağılımına göre kritik değer tespit edilir ve bu değerden büyük değerli olan noktalar aykırı değer olarak tespit edilir.

Bu yöntemde $T(x)$, kovaryans matrisinin determinantı en küçük olan $x = x_1, x_2, \dots, x_n$ örneğindeki h tane noktanın ortalamasıdır.

MVE yönteminde kullanılan Rousseeuw ile Leroy'un 28 canlı çeşidi için vücut ve beyin ağırlıklarını içeren ve bu 28 çeşit canlı için beyin ağırlıkları ile vücut ağırlıklarını karşılaştıran "*Vücut ve Beyin Ağırlıkları*" verisi MCD yönteminin uygulanması için de kullanılmıştır [SAS Ins, 1999].

Çizelge 5.3'de $n=28$ ve $p=2$ olan "*Vücut ve Beyin Ağırlıkları*" verisi için Mahalanobis uzaklıkları ile MCD yöntemi için elde edilen uzaklık değerleri verilmektedir.

Çizelge 5.3. 28 Canlı çeşidi için Mahalanobis uzaklıkları ve MCD için sağlam istatistiklere dayalı uzaklıklar

N	Mahalanobis Uz.	MCD için Sağlam İstatistiklere Dayalı Uz.
1	1.006591	0.855347
2	0.695261	1.477050
3	0.300831	0.239828
4	0.380817	0.517719
5	1.146485	1.108362
6	2.644176	10.599341
7	1.708334	1.808455
8	0.706522	0.690099
9	0.858404	1.052423

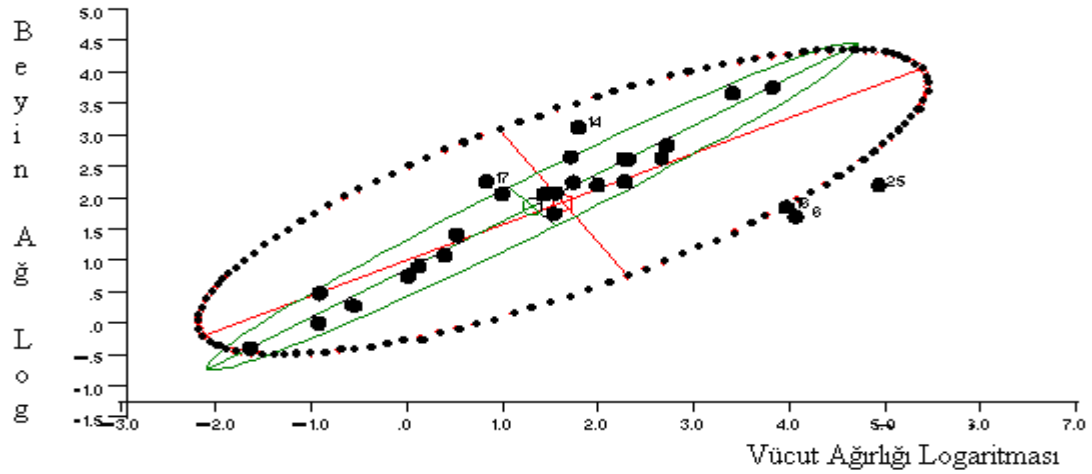
Çizelge 5.3. (Devam) 28 Canlı çeşidi için Mahalanobis uzaklıkları ve MCD için sağlam istatistiklere dayalı uzaklıklar

N	Mahalanobis Uz.	MCD için Sağlam İstatistiklere Dayalı Uz.
10	0.798698	2.077131
11	0.686485	0.888545
12	0.874349	1.035824
13	0.677791	0.683978
14	1.721526	4.257963
15	1.761947	1.716065
16	2.369473	9.584992
17	1.222253	3.571700
18	0.203178	1.323783
19	1.855201	1.741064
20	2.266268	2.026528
21	0.831416	0.743545
22	0.416158	0.419923
23	0.264182	0.944610
24	1.046120	2.289334
25	2.911101	11.471953
26	1.586458	1.518721
27	1.582124	2.054593
28	0.394664	1.675651

Mahalanobis uzaklığına göre sadece 25. gözlem $\sqrt{\chi^2_{2,.025}} = 2.72$ değerini aşarak aykırı değer olarak tespit edilmektedir ve Mahalanobis uzaklığı ile 28 birimlik örnekte sadece bir tane aykırı değer tespit edilebilmiştir (Çizelge 5.3).

Bununla birlikte MCD yönteminin uygulanması sonucunda 6, 14, 16, 17 ve 25. gözlemler $\sqrt{\chi^2_{2,.025}} = 2.72$ kritik değerini aşarak aykırı değer olarak tespit edilmiş, kullanılan yöntemin aykırı değerleri tespit etmede sağlam bir metot olduğu görülmüştür (Çizelge 5.3).

Şekil 5.2’de Mahalanobis uzaklığına göre çizilen elipsoid ve MCD yöntemine göre elde edilen elipsoid gösterilmektedir. Şekil 5.2 yardımıyla yukarıda bahsedilen sonuçlar görsel olarak ortaya konmakta ve her iki elipsoid arasındaki farklılık açık bir şekilde gösterilmektedir.



Şekil 5.2. Mahalanobis Uzaklığına ve MCD yöntemine göre oluşturulan elipsoidler

Aykırı değerler klasik yöntemler kullanılarak elde edilen tahmin edicileri önemli derecede etkiler ve sonuç olarak Mahalanobis uzaklığına göre oluşturulmuş %97.5’lik tolerans elipsoidi(kesikli çizgiyle gösterilen) tek bir gözlemi dışarıda bırakır. MCD yöntemine göre elde edilen elipsoid ise (kesiksiz çizgiyle gösterilen) 5 tane gözlemi elipsoid dışında bırakmaktadır (Şekil 5.2).

Rousseeuw ile Leroy ’un “*Vücut ve Beyin Ağırlıkları*” verisine her iki yöntemin de uygulanıp aynı sonuçların elde edilmesine rağmen MVE yönteminin n ve p büyük olduğunda veriye uygulanması MCD yöntemine göre daha zor olduğundan Rousseeuw(1985), istatistiksel etkinlik ve hesaplama açısından MVE’ye göre daha üstün olan bu yöntemin kullanılmasını önermiştir.

Genel olarak MVE ve MCD tahmin edicilerinin asimptotik özelliklerini belirlemek zordur. $p=1$ olan özel bir durum göz önünde bulundurulduğunda MVE ve MCD tahminlerinin hesaplanması için kesin bir algoritma ortaya koymak oldukça kolaydır. Öncelikle veri $x_1 < x_2 \leq x_3 \leq \dots \leq x_n$ şeklinde sıralanır. Bu örneğin sadece $n-h+1$ “yarımları (halves)” hesaplanmaya ihtiyaç duyar. Yani,

$$H_1 = x_1, \dots, x_n, \quad H_2 = x_2, \dots, x_{n+1}, \quad \dots, \quad H_{n-h+1} = x_{n-h+1}, \dots, x_n$$

olacak şekilde tanımlanır.

Her bir H_i için,

$$OrtaNokta(i) = \frac{1}{2}(x_i + x_{i+h-1}) \quad (5.12)$$

$$Uzunluk(i) = x_{i+h-1} - x_i \quad (5.13)$$

$$Ortalama(i) = \frac{1}{h} \sum_{j=i}^{i+h-1} x_j \quad (5.14)$$

$$Varyans(i) = \frac{1}{h-1} \sum_{j=1}^{i+h-1} (x_j - ort(i))^2 \quad (5.15)$$

olarak tanımlanır [Rousseeuw, 1985].

MVE yöntemi için $Uzunluk(i)$ 'si en küçük olan H_i seçilir ve $T(x):OrtaNokta(i)$ alınır. Birden fazla $Uzunluk(i)$ 'nin en küçük olduğu durumda, $T(x)$ karşılık gelen $OrtaNokta(i)$ 'lerin ortalaması alınarak bulunur.

MCD yöntemi için $Varyans(i)$ 'si en küçük olan H_i seçilir ve $T(x):Ortalama(i)$ alınır. Şayet birçok çözüm varsa ortalamaları alınarak bulunur.

MCD tahmin edicisinin hesaplanmasında ortalama ve kovaryans matrisi n elemanlı bir kümede en küçük determinanta sahip $h < n$ gözleme dayalı olarak hesaplanmaktadır. p , veri kümesindeki değişken sayısı olduğuna göre MCD tahmin edicisi $p < h$ iken hesaplanabilmektedir. Aksi takdirde h alt kümenin kovaryans matrisinin determinantı sıfır çıkacaktır.

MCD yönteminde birden çok alt küme ile hesaplama yapmak gerektiğinden n sayısı büyük olduğunda MCD' nin hesaplanması zorlaşmaktadır ve kesin sonuca ulaşmak imkansız olmaktadır. MCD tahmin edicisinin belirttiğimiz dezavantajlarından dolayı bu yöntem günümüzde nadir olarak kullanılmaktadır.

5.3. Hızlı En Küçük Kovaryans Determinantı (FMCD) Yöntemi

MCD yönteminin uygulanması geniş ve büyük boyutlu veri setleri için hesaplanabilirlik açısından karmaşık ve zaman alıcı bir yapıya sahiptir. Bu problemin üstesinden gelebilmek için Rousseeuw ve Van Driessen(1999) kısa sürede güvenilir sonuçlar veren ve diğer iki yöntemle benzer şekilde %50 bozulma noktasına sahip yeni bir yöntem geliştirip bu yöntemi MCD tahmin edicisindeki gibi kovaryans determinantını minimize etme prensibine dayandırmışlardır.

“*C-adımı*” olarak adlandırılan bir adıma dayanan FMCD tahmin edicisi büyük veri kümelerine rahatlıkla uygulanabilir, daha hızlı ve MCD’ye göre daha iyi olan bir yöntemdir.

n gözlem sayısı ve p değişken sayısı olmak üzere, $X = X_1, X_2, \dots, X_n$ tesadüfi bir örnek ve $|H_1| = h$ ve $H_1 \subset 1, \dots, n$ olsun.

$$T_1 = \frac{1}{h} \sum_{i \in H_1} X_i \quad (5.16)$$

ve

$$C_1 = \frac{1}{h} \sum_{i \in H_1} (X_i - T_1)(X_i - T_1)^t \quad (5.17)$$

olmak üzere, $\det(C_1) \neq 0$ ise uzaklıklar

$$d_1(i) = \sqrt{(X_i - T_1)C_1^{-1}(X_i - T_1)^t} \quad , \quad i = 1, \dots, n \quad (5.18)$$

ile hesaplanır.

$(d_1)_{1:n} \leq (d_1)_{2:n} \leq \dots \leq (d_1)_{n:n}$ sıralı uzaklıkları ifade etsin. Buna göre $d_1(i); i \in H_2 = (d_1)_{1:n}, \dots, (d_1)_{h:n}$ olan H_2 alınır ve H_2 'ye ait T_2 ile C_2 değerleri hesaplanır. Buradan $\det(C_2) \leq \det(C_1)$ sonucu elde edilir ve en küçük uzaklık değerli h tane gözleme ve daha küçük determinanta sahip olduğu için C_2 'ye C_1 'e göre daha fazla yoğunlaşıldığından (Concentration) bu işlemlerin tümüne “C-adımı” denilmektedir.

FMCD yönteminin algoritması aşağıdaki gibidir,

1- $h = \left\lceil \frac{n+p+1}{2} \right\rceil$ olmak üzere, H_{old} , h sayıda nokta içeren tesadüfi bir alt küme olsun.

2- H_{old} 'a ait olan tüm gözlemlerin ortalama vektörü $\bar{X}_{H_{old}}$ ve kovaryans matrisi $C_{H_{old}}$ hesaplanır. Daha sonra $d^2_{H_{old}}(i) = (X_i - \bar{X}_{H_{old}})C_{H_{old}}^{-1}(X_i - \bar{X}_{H_{old}})^t$ uzaklıkları $i = 1, \dots, n$ için hesaplanır.

3- Artan sıraya göre kareli uzaklıklar sıralanır.

$$d^2_{H_{old}}(\pi(1)) \leq d^2_{H_{old}}(\pi(2)) \leq \dots \leq d^2_{H_{old}}(\pi(n))$$

Burada $\pi, 1, 2, \dots, n$ için permütasyon anlamına gelir.

4- $H_{new} = X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(h)}$ tanımlanır.

5- $\bar{X}_{H_{new}}, C_{H_{new}}$ ve $d^2_{H_{new}}(i)$ değerleri hesaplanır.

6- Şayet $\det(C_{H_{new}}) = 0$ ise 1-5 arasındaki adımlar tekrarlanır. Şayet $\det(C_{H_{new}}) = \det(C_{H_{old}})$ ise süreç durdurulur. Aksi takdirde $\det(C_{H_k}) = \det(C_{H_{k+1}})$ ise süreç k . tekrarlama kadar devam eder. Böylece $\det(C_{H_1}) \geq \det(C_{H_2}) \geq \dots \geq \det(C_{H_k}) = \det(C_{H_{k+1}})$ olur.

$T_{MCD} = \bar{X}_{H_k}$ ve $C_{MCD} = C_{H_k}$ bu süreçten elde edilen konum ve kovaryans matrisi olsun. FMCD'ye dayalı olan Sağlam kareli Mahalanobis uzaklığı,

$$d_{MCD}^2(X_i, T_{MCD}) = (X_i - T_{MCD})C_{MCD}^{-1}(X_i - T_{MCD})^t, \quad i = 1, \dots, n \quad (5.19)$$

şeklinde tanımlanır. $d_{MCD}(X_i, T_{MCD})$ değeri $\sqrt{\chi_{p,\alpha}^2}$ değerinden büyük olan gözlemler aykırı değer olarak etiketlenir.

C-adımı FMCD algoritması için H_1 'in birçok başlangıç seçimlerinin alınması, $\det(C_1) \geq \det(C_2) \geq \dots \geq \det(C_k) = \det(C_{k+1})$ olana dek C-adımının uygulanması ve en küçük determinanta sahip çözümün alınması şeklinde bir fikir oluşturmaktadır. Bu fikrin uygulanabilmesi için başlangıç alt kümeleri H_1 'lerin nasıl oluşturulması gerektiğine karar vermek oldukça önemli bir noktadır. MVE algoritmasındaki gibi $p+1$ tane gözlem alınarak J alt kümesi oluşturulur ve $T_0 = \text{ort}(J)$ ve $C_0 = \text{cov}(J)$ değerleri hesaplanır. Şayet $\det(C_0) = 0$ ise J alt kümesine $\det(C_0) > 0$ olana kadar gözlemler birer birer eklenerek alt küme genişletilir.

Daha sonra $d_0^2(i) = (x_i - T_0)C_0^{-1}(x_i - T_0)^t$ uzaklıkları $i = 1, \dots, n$ için hesaplanır ve $d_0(\pi(1)) \leq \dots \leq d_0(\pi(n))$ şeklinde uzaklıklar sıralanıp $H_1 = \pi(1), \dots, \pi(h)$ alınır. Burada dikkat edilmesi gereken husus $p+1$ tane noktadan daha az sayıda nokta alınmaması gerektiği aksi takdirde C_0 tekil çıkacaktır.

FMCD algoritması MCD tahmin edicisini iyileştirici özellikler taşıdığından dolayı MVE ile MCD yöntemlerine göre daha hızlı çalışan, hesaplanabilirlik açısından daha iyi olan ve daha doğru sonuçlar veren bir algoritmadır. Bu yöntem S-Plus, R, SAS gibi istatistiksel paket programlarında yer almaktadır.

Literatürde konum ve dağılım parametrelerinin tahmini için sıkça kullanılan ve en popüler olan sağlam parametrik yöntem FMCD algoritmasıdır. Buna rağmen geniş ve büyük boyutlu veri setleri için kovaryans matrisinin tersini almayı gerektiren Mahalanobis uzaklığını kullandığı için ve amaç fonksiyonu MCD algoritmasındaki gibi kovaryans determinantını minimize etmek olduğundan zaman açısından bir

olumsuzlukla karşılaşılır ve böylece FMCD yöntemi kullanışlılığını kaybeder. Sonuç olarak bu yöntemin hesaplama karmaşıklığı veri setlerinin boyutu arttığında üstel olarak artmaktadır ve hesaplama etkinliği azalmaktadır. Oysa Angiulli ve Pizzuti(2005) hesaplama etkinliğinin herhangi bir algoritmanın aykırı değeri bulmadaki verimliliği kadar önemli olduğunu belirtmiştir. Bu nedenle kovaryans matrisinin determinantını minimize etmek yerine vektör varyansını minimize eden ve daha kısa sürede FMCD kadar iyi sonuçlar ortaya koyan yeni bir tahmin yöntemi geliştirilmiştir.

5.4. En Küçük Vektör Varyansı (MVV) Yöntemi

Önceki bölümlerde verilen bilgilere göre görüyoruz ki çok değişkenli veri analizinde aykırı değerleri tespit etmede kullanılan MVE, MCD ve FMCD yöntemlerinin en büyük dezavantajı geniş ve büyük boyutlu veri setlerinde p değişken sayısı arttıkça bu yöntemlerin kullanılmasının güçleşmesi ve zaman açısından analizciyi sıkıntıya sürüklemesidir. Herwindiati ve ark. (2007) bu probleme karşılık diğer yöntemler kadar sağlam, %50'lik bozulma noktasına sahip, geniş ve büyük boyutlu veri setlerine kolaylıkla uygulanabilen, karmaşıklığı az ve hesaplanabilirlik açısından diğer yöntemlere göre daha üstün olan MVV algoritmasını geliştirmişlerdir. Bu yöntem FMCD yöntemindeki gibi kovaryans determinantını minimize etmek yerine vektör varyansını minimize etmeye çalıştığı için p değişken sayısı arttıkça vektör varyansını hesaplamak, kovaryans determinantını hesaplamaya göre daha kısa sürede gerçekleşmektedir.

$X = \begin{pmatrix} X_{(1)}^t & X_{(2)}^t \end{pmatrix}^t$ tesadüfi bir vektör olsun. X 'in kovaryans matrisi,

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \quad (5.20)$$

olmak üzere,

$$C_{ij} = E (X_{(i)} - \mu_{(i)})(X_{(j)} - \mu_{(j)})^t \quad (5.21)$$

$$\mu_{(i)} = E X_{(i)} \quad , \quad i, j = 1, 2. \quad (5.22)$$

şeklinde ifade edilir.

Kovaryans matrisi çok değişkenli analizde bilginin temel kaynağıdır ve analizcilerin öncelikle ilgilendikleri kovaryans matrisinin tahminidir. Önceki bölümlerde bahsedilen sağlam yöntemlerin tümü çok değişkenli dağılış ölçüsü olan kovaryans determinantına dayanmaktadır. Djauhari(2007) çok değişkenli dağılış ölçüsünün daha iyi anlaşılabilmesi için kovaryans determinantına alternatif olarak vektör varyansının kullanılmasını önermiştir.

$Tr C_{12} C_{21}$, $X_{(1)}$ ve $X_{(2)}$ tesadüfi vektörleri arasındaki lineer ilişkinin ölçüsünü belirtir ve vektör kovaryansı olarak adlandırılır. Vektör kovaryansı “iz” veya $C_{12} C_{21}$ ’in tüm köşegen elemanlarının toplamına eşittir. Vektör kovaryansı vec operatörü kullanılarak $\langle vec C_{12} , vec C_{12} \rangle$ veya eşdeğer olarak $\|vec C_{12}\|^2$ şeklinde ifade edilebilir. Bu nedenle $X_{(1)}$ ve $X_{(2)}$ ’ in vektör varyansları sırasıyla $\|vec C_{11}\|^2 = Tr C_{11}^2$ ve $\|vec C_{22}\|^2 = Tr C_{22}^2$ olarak ifade edilir.

X , kovaryans matrisi C olan p boyutlu tesadüfi bir vektör olsun. X ’in vektör varyansı $Tr C^2$, C^2 ’nin köşegen elemanlarının karelerinin toplamına eşittir. Vektör varyansının büyük değeri ortalama vektörü etrafında daha fazla dağılış olduğunu ifade eder. $Tr C^2$ ’nin birkaç önemli özelliği vardır. Bunlardan ilki, kovaryans determinantının tersine hesaplanmasının basit olmasıdır. İkinci özelliği, C ’nin pozitif tanımlı olmasına gerek duymamasıdır. Bu nedenle $Tr C^2$ tekil olmayan kovaryans matrisleri ve küçük boyutlu veri setleri için kullanılabilir gibi bir kısıtlamaya sahip değildir. Oysa kovaryans determinantını kullanan diğer yöntemlerde determinant sürekli olarak 0 çıkabileceğinden bu yöntemler sadece tekil

olmayan kovaryans matrisleri için kullanılabilir. MVV yöntemi ise geniş ve büyük boyutlu veri setleri için tekil kovaryans matrisine sahip olunması durumunda bile kullanılabilen bir yöntemdir. Üçüncü özelliği ise p küçük olduğunda vektör varyansına dayalı testler ile kovaryans determinantına dayalı testler benzer duyarlılığa sahiptir. Ancak n ve p büyük değere sahip olduğunda vektör varyansına dayalı testler kovaryans determinantına dayalı testlere göre daha duyarlıdır.

MVV yönteminin algoritması aşağıdaki gibidir,

1- $h = \left\lceil \frac{n+p+1}{2} \right\rceil$ olmak üzere, H_{old} , h sayıda nokta içeren tesadüfi bir alt küme olsun.

2- H_{old} 'a ait olan tüm gözlemlerin ortalama vektörü $\bar{X}_{H_{old}}$ ve kovaryans matrisi $C_{H_{old}}$ hesaplanır. Daha sonra $d^2_{H_{old}}(i) = (X_i - \bar{X}_{H_{old}})C_{H_{old}}^{-1}(X_i - \bar{X}_{H_{old}})^t$ uzaklıkları $i = 1, \dots, n$ için hesaplanır.

3- Artan sıraya göre kareli uzaklıklar sıralanır.

$$d^2_{H_{old}}(\pi(1)) \leq d^2_{H_{old}}(\pi(2)) \leq \dots \leq d^2_{H_{old}}(\pi(n))$$

Burada π , $1, 2, \dots, n$ için permütasyon anlamına gelir.

4- $H_{new} = X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(h)}$ tanımlanır.

5- $\bar{X}_{H_{new}}$, $C_{H_{new}}$ ve $d^2_{H_{new}}(i)$ değerleri hesaplanır.

6- Şayet $Tr C_{H_{new}}^2 = 0$ ise 1-5 arasındaki adımlar tekrarlanır. $Tr C_{H_{new}}^2 = Tr C_{H_{old}}^2$

ise süreç durdurulur. Aksi takdirde süreç k .tekrarlamaya kadar devam eder.

$$Tr C_{H_k}^2 = Tr C_{H_{k+1}}^2 \quad (5.23)$$

ifadesi doğru olur ve böylece $Tr C_{H_1}^2 \geq Tr C_{H_2}^2 \geq \dots \geq Tr C_{H_k}^2 = Tr C_{H_{k+1}}^2$ ifadesi elde edilir.

$T_{MVV} = \bar{X}_{H_k}$ ve $C_{MVV} = C_{H_k}$ MVV algoritmasından elde edilen konum ve kovaryans matrisi olsun. MVV yöntemine dayalı sağlam Mahalanobis uzaklığı,

$$d_{MVV}^2(X_i, T_{MVV}) = (X_i - T_{MVV})^T C_{MVV}^{-1} (X_i - T_{MVV}), \quad i = 1, 2, \dots, n \quad (5.24)$$

olarak tanımlanır. $d_{MVV}(X_i, T_{MVV})$ değeri $\sqrt{\chi_{p,\alpha}^2}$ değerinden büyük olan gözlemler aykırı değer olarak etiketlenir.

Bu algoritma FMCD yöntemi ile karşılaştırıldığında daha az hesaplama karmaşıklığına sahiptir. p 'nin büyük değerleri için Cholesky ayrışımına dayanan kovaryans determinantının hesaplanmasındaki işlem sayısı,

$$p + p(p-1) + p(p-1) \sum_{i=1}^{p-1} (p-i-1) = p^2(p-1) \quad (5.25)$$

ile belirlenir.

MVV tahmin edicileri kovaryans matrisinin tüm özdeğerlerinin fonksiyonu olduklarından MVE ve MCD yöntemlerindeki gibi aynı bozulma noktasına sahiptir.

Özdeğerler $\lambda_1, \lambda_2, \dots, \lambda_p$ olmak üzere,

$$VV = \sum_{i=1}^p \lambda_i^2 \text{ 'dir.} \quad (5.26)$$

Çizelge 5.4 p değişken sayısının birkaç değeri için kovaryans determinantının ve vektör varyansının hesaplama hızları açısından farklılıklarının karşılaştırılmasını göstermektedir [Djahuri, 2008]. Kovaryans determinantı Cholesky ayrışımı kullanılarak hesaplanmıştır.

Çizelge 5.4. Kovaryans determinanı ve vektör varyansı arasındaki hesaplama hızı oranları

p	Hesaplama Hızı Oranı(CD:VV)
10	6:1
25	13:1
50	34:1
75	67:1
100	95:1
150	127:1
200	231:1
250	326:1
300	443:1

Çizelge 5.4 çok değişkenli dağılış ölçüsü için vektör varyansının kullanılması sonucunda sonuçlara daha kısa sürede ulaşıldığını açıkça ortaya koymaktadır. p değişken sayısı arttıkça kovaryans determinantını hesaplama zamanı uzarken, vektör varyansını hesaplama zamanı aynı kalmaktadır.

6. SİMÜLASYON ÇALIŞMASI

Çalışmamızın bu bölümünde MVV yöntemi ile bir önceki bölümde açıklanan diğer yöntemlerin aykırı değerleri doğru tespit etme oranları ve sonuca ulaşma hızları bakımından karşılaştırılması yapılmıştır. Bu amaçla yapılan simülasyon çalışması için MATLAB R2008a paket programı kullanılmıştır.

Bu simülasyon çalışmasında öncelikle her bir yöntem için örnek hacmi n sırasıyla 5,10,15 ve 20 olarak ele alınmış ve örnekteki aykırı gözlem sayısı artırılarak örnek

seçme işlemi 1000 defa tekrarlanmıştır. $X_1 \sim N_2\left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ ve

$X_2 \sim N_2\left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ olduğu durumda MVE, MCD, FMCD ve MVV

yöntemlerinin aykırı değerleri doğru tespit etme oranları ve sonuca ulaşma süreleri saniye bakımından Çizelge 6.1’de verilmiştir.

Daha sonra Σ_1 ve Σ_2 matrisleri farklı iken her bir yöntem için örnek hacmi n sırasıyla 5 ve 10 olarak ele alınmış ve örnekte farklı miktarda aykırı gözlem bulunması durumu için, örnek seçme işlemi 1000 defa tekrarlanmıştır. Kovaryans matrisleri farklı iken her bir yöntem için gözlenen aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri sırasıyla Çizelge 6.2 ve 6.3’de verilmiştir.

Çizelge 6.1. $X_1 \sim N_2\left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ ve $X_2 \sim N_2\left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$

olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri

TEKRAR=1000 , $p = 2$										
n	n_1	n_2	MVE		MCD		FMCD		MVV	
			oran	zaman	oran	zaman	oran	zaman	oran	zaman
5	4	1	0.5370	18.93	0.5750	13.09	0.5650	15.89	0.5840	14.66
	3	2	0.0140	15.87	0.0980	10.77	0.2330	15.97	0.2450	14.99
10	9	1	0.5280	268.81	0.1990	65.21	0.3260	38.64	0.3120	37.16
	8	2	0.5780	270.87	0.3080	66.36	0.3770	38.83	0.3560	37.34
	7	3	0.5060	268.94	0.5100	66.00	0.4640	38.75	0.4730	37.85
	6	4	0.2600	268.58	0.5780	68.32	0.5970	38.82	0.6100	37.91
15	13	2	0.6460	12517	0.3900	4986.71	0.2700	454.19	0.2880	379.20
	11	4	0.5810	12789	0.5860	4680.5	0.3810	428.38	0.3770	382.30
	9	6	0.1380	12952	0.5480	5743.8	0.5430	368.62	0.5520	377.46
20	18	2	---	---	---	---	0.1930	11615	0.2090	14134
	16	4	---	---	---	---	0.2480	11074	0.2510	11808
	14	6	---	---	---	---	0.3210	12273	0.3270	11924

Burada, örnek hacmi n , ortalama vektörü $\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ve kovaryans matrisi

$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ olan X_1 yığınınından elde edilen gözlemlerin sayısı n_1 , ortalama vektörü

$\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$ ve kovaryans matrisi $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ olan X_2 yığınınından elde edilen

gözlemlerin sayısı n_2 ve değişken sayısı da p ile ifade edilmiştir. Çizelge 6.1'e göre aşağıdaki bulgular elde edilmiştir.

Çok değişkenli veride aykırı değerleri tespit etmek için MVE yöntemi kullanıldığında örnek hacmi 5 ve veride tek bir aykırı değer mevcut iken bu aykırı değerleri tespit etmede MVE yönteminin 0.5370 oranında başarı gösterdiği ve 18.93 saniyede sonuca ulaştığı görülmüştür. Örnek içindeki aykırı değer sayısı 2 olduğunda ise MVE yönteminin aykırı değerleri tespit etme oranında önemli derecede düşüş olduğu ve 0.0140 düzeyine indiği gözlenmiştir.

Örnek hacmi 5 ve veride tek bir aykırı değer mevcut iken MCD yönteminin kullanılması durumunda aykırı değeri tespit etmede 0.5750 oranında başarı elde edilmiş ve sonuca kısa sürede ulaşıldığı görülmüştür. Örnek içindeki aykırı değer sayısı artırıldığında ise tıpkı MVE yöntemindeki gibi aykırı değerleri tespit etme oranında önemli derecede düşüş olduğu gözlenmiştir.

FMCD yönteminin örnek hacmi 5 ve veride tek bir aykırı değer bulunması durumunda aykırı değerleri tespit etmede 0.5650 oranında başarı gösterdiği ve sonuca kısa bir süre içinde ulaştığı görülmüştür. Örnek içindeki aykırı değer sayısı artırıldığında aykırı değerleri tespit etme oranında bir düşüş olmasına rağmen buradaki düşüş MVE ve MCD yöntemlerinin veri içindeki aykırı değer sayısı artırıldığında bu aykırı değerleri tespit etmede gösterdikleri başarı oranındaki düşüşten daha az olduğu söylenebilir.

Örnek hacmi 5 ve veride tek bir aykırı değer olduğu durumda MVV yönteminin veriye uygulanmasıyla aykırı değeri tespit etme oranı 0.5840 bulunarak en yüksek değer elde edilmiş ve işlem süresi 14.66 olarak gözlenmiştir. Veri içindeki aykırı değer sayısı artırıldığında ise bu aykırı değerleri tespit etmede düşüş meydana geldiği görülmüş ancak FMCD yöntemindeki gibi diğer iki yöntemin göstermiş olduğu düşüşten daha az oranda bir azalma olduğu gözlenmiştir.

Çok değişkenli veride aykırı değerleri tespit etmek için kullanılan yöntemlerin örnek hacmi büyüdüğünde nasıl davrandıklarını anlayabilmek için örnek hacmi 10 olarak ele alınırsa, MVE yönteminin veri içindeki aykırı değerleri tespit etmede yüksek oranda başarıya sahip olduğu görülmesine rağmen aykırı değer sayısı arttıkça bu oranda düşüş meydana geldiği gözlenmekle birlikte sonuca ulaşma süresinin de arttığı açıkça görülmektedir.

Verideki aykırı değerleri tespit etmek için MCD yöntemi kullanıldığında verideki aykırı gözlem sayısı artırılırsa yöntemin aykırı değerleri tespit etmede gösterdiği başarının artmakta ve aynı zamanda aykırı değerleri tespit etme süresinin MVE yöntemine göre çok daha az olduğu gözlenmiştir.

FMCD yöntemi verideki aykırı değerleri tespit etmede yüksek seviyede başarı gösterirken veri içindeki aykırı değer sayısı artsa bile bu seviyenin gittikçe artış gösterdiği ve FMCD yönteminin MVE ile MCD yöntemlerine göre daha kısa zamanda aykırı değerleri tespit ettiği gözlenmiştir.

MVV yönteminin verideki aykırı değerleri tespit etme oranının aykırı değer sayısı artarken artış gösterdiği ve bununla birlikte en kısa sürede aykırı değerleri tespit etmeyi gerçekleştirdiği gözlenmiştir. Veride 4 tane aykırı gözlem bulunması durumunda MVV yöntemi diğer üç yöntemle göre daha yüksek oranda başarı göstermiş ve FMCD yöntemiyle aralarında ufak bir fark olmasına rağmen daha kısa sürede hesaplamayı tamamladığı görülmüştür.

Verideki aykırı gözlem sayısının artması durumunda MVE yönteminin aykırı değerleri tespit etme oranında azalma gözlenirken, diğer üç yöntemin aykırı değerleri tespit etmede gösterdikleri başarıda sürekli bir artış olduğu gözlenmiştir.

Örnek hacmi 15 olduğunda, MVE yönteminin veriye uygulanmasıyla aykırı değerleri tespit etme oranının verideki aykırı gözlemlerin artış göstermesiyle birlikte önemli derecede düşüş gösterdiği ayrıca aykırı değerleri tespit etme süresinin de uzayarak 3.5 saate yakın bir sürede hesaplamanın sona erdiği gözlenmiştir.

MCD yönteminin aykırı değerleri tespit etmede yüksek seviyede başarı gösterdiği gözlenmekle birlikte MVE yönteminin aykırı değerleri tespit ettiği süreden daha kısa sürede tespit etmeyi gerçekleştirdiği görülmüştür. Ancak MCD, FMCD ve MVV ile karşılaştırılırsa bu sürenin uzun bir süre olduğu söylenebilir.

FMCD yönteminin aykırı değerleri tespit etme oranı verideki aykırı değer sayısı arttığında artış gösterirken MVE ile MCD yöntemine göre oldukça az zamanda aykırı değerleri tespit ettiği gözlenmiştir.

MVV yönteminin hem aykırı değerleri tespit etmede yüksek orana sahip olduğu hem de diğer üç yöntemle nazaran daha kısa sürede sonuca ulaşmayı başardığı

gözlenmiştir. Ayrıca verideki aykırı değer sayısı en üst seviyede iken MVV yönteminin aykırı değerleri tespit etme oranının diğer yöntemlere göre daha yüksek olduğu gözlenmiştir.

Son olarak örnek hacmi 20 olduğunda, MVE ile MCD yöntemlerinin örnek hacmi büyük iken kullanılmaları durumunda hesaplama günlerce sürebildiğinden dolayı bu iki yöntemin aykırı değerleri tespit etmede tamamen başarısız olduğu ve zaman açısından oldukça büyük bir dezavantaj oluşturdukları gözlenmiştir.

Örnek hacmi büyük olsa dahi FMCD yönteminin aykırı değerleri tespit etmede MVE ile MCD yöntemlerine oranla daha hızlı olduğu ve yaklaşık 3.5 saat gibi bir sürede hesaplamayı bitirebildiği gözlenmiştir. Üstelik veride aykırı gözlem sayısı arttıkça FMCD yönteminin aykırı değerleri tespit etme oranının da artış gösterdiği görülmüştür.

MVV yönteminde veride aykırı değer sayısı 2 iken aykırı değerleri tespit etme oranının 0.2090 olarak belirlendiği ve bu değer FMCD yönteminin verideki aykırı değer sayısı 2 olduğu durumda ulaştığı orandan daha büyük bir oran olduğu görülmektedir. Aykırı değer sayısı 4 ve 6 olduğu durumda da MVV yönteminin FMCD yöntemine göre aykırı değerleri tespit etme oranlarının daha yüksek olduğu gözlemlendiğinden MVV yönteminin örnek hacmi büyük iken FMCD yöntemine göre daha başarılı olduğu söylenebilir.

Çizelge 6.2. $X_1 \sim N_2\left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}\right)$ ve $X_2 \sim N_2\left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$

olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri

TEKRAR=1000 , $p = 2$										
n	n_1	n_2	MVE		MCD		FMCD		MVV	
			oran	zaman	oran	zaman	oran	zaman	oran	zaman
5	4	1	0.0700	15.97	0.0700	11.30	0.1830	17.66	0.1950	18.01
	3	2	0.0020	15.63	0.0070	10.67	0.0370	17.36	0.0260	17.09
10	9	1	0.0300	278.10	0.0040	73.09	0.0040	42.64	0.0040	40.27
	8	2	0.0450	275.44	0.0060	62.68	0.0170	42.99	0.0160	40.80
	7	3	0.0260	274.40	0.0150	63.28	0.0330	45.75	0.0370	41.44
	6	4	0.0160	291.17	0.0720	63.50	0.1020	42.55	0.1070	40.92

Kovaryans matrisine ait değerler değiştirildiğinde yöntemlerin aykırı değerleri tespit etme oranlarında bir düşüş meydana geldiği gözlenmiştir.

Örnek hacmi 5 iken, MVE ile MCD yöntemlerinin aykırı değerleri tespit etme oranlarının oldukça düştüğü, örnek içindeki aykırı değer sayısı arttıkça bu aykırı gözlemleri belirlemenin zorlaştığı söylenebilir. FMCD ile MVV yöntemlerinin aykırı değerleri tespit etme oranları kısmen düşük olsa dahi, MVE ile MCD yöntemlerine göre daha yüksek tespit etme oranlarına sahip oldukları gözlenmiştir.

Örnek hacmi 10 olduğunda, MVE yönteminin aykırı değerleri tespit etme oranları çok düşükken bununla birlikte diğer yöntemlere nazaran aykırı değerleri tespit etme sürelerinin de fazla olduğu gözlenmiştir.

MCD yöntemine ait aykırı değer tespit etme oranları da oldukça düşük olmasına karşın verideki aykırı değer sayısı artarken, oranların da buna paralel bir şekilde artış gösterdiği görülmektedir. Oysa MVE yönteminde böyle bir durumun söz konusu olmadığı gözlenmiştir. Örnek içindeki aykırı değer sayısı 4 iken MVE yöntemi aykırı değerleri belirlemede daha uzun sürede daha az başarı gösterirken, MCD yönteminin daha kısa sürede daha yüksek oranda belirlemeyi gerçekleştirdiği söylenebilir.

FMCD yönteminin aykırı değer sayısı az olduğunda aykırı değerleri belirlemede düşük oranlara sahip olduğu ancak aykırı değer sayısı artarken belirleme oranlarında artış meydana geldiği gözlenmiştir. Ayrıca MVE ile MCD yöntemlerinden daha kısa sürede aykırı değerleri tespit ettiği gözlenmiştir.

MVV yönteminin de aykırı değerleri tespit etme oranları örnekteki aykırı değer sayısı arttıkça artış gösterirken, aykırı değer sayısı en çok olduğunda diğer üç yöntem arasında en iyi yöntem olarak belirlenen FMCD yönteminden bile daha yüksek oranda aykırı değerleri tespit etmeyi gerçekleştirdiği görülmüştür. Ayrıca en kısa sürede en yüksek düzeyde tespit etmeyi gerçekleştirebildiğinden kovaryans matrisi farklılaştırıldığında bile MVV yönteminin diğer üç yöntemden daha iyi bir yöntem olduğu söylenebilir.

Çizelge 6.3. $X_1 \sim N_2 \left(\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right)$ ve $X_2 \sim N_2 \left(\mu_2 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right)$

olduğu durumda yöntemlerin aykırı değerleri tespit etme oranları ve sonuca ulaşma süreleri

TEKRAR=1000 , $p = 2$										
			MVE		MCD		FMCD		MVV	
n	n_1	n_2	oran	zaman	oran	zaman	oran	zaman	oran	zaman
5	4	1	0.0760	17.82	0.0560	11.80	0.2160	16.18	0.2180	19.81
	3	2	0.0140	18.99	0.1190	11.10	0.2320	16.79	0.2460	17.64
10	9	1	0.0230	289.99	0.0030	70.03	0.0050	41.34	0.0090	41.45
	8	2	0.0380	282.54	0.0050	69.03	0.0130	42.52	0.0160	41.59
	7	3	0.0500	267.85	0.0150	70.27	0.0370	42.78	0.0370	41.40
	6	4	0.0430	306.77	0.0740	70.45	0.0920	37.51	0.1070	42.02

Örnek hacmi 5 olduğunda; MVE ile MCD yöntemlerinin aykırı değerleri tespit etmede FMCD ile MVV yöntemlerine göre daha başarısız olduğu bir kez daha gözlenmiştir. FMCD ile MVV yöntemi karşılaştırılacak olursa örnek içinde 1 ve 2 aykırı değer olması durumlarının her ikisinde de MVV yönteminin aykırı değerleri belirleme oranlarının daha yüksek olduğu gözlenmiştir.

Örnek hacmi 10 olduğunda, MVE yönteminin uzun sürede aykırı değerleri belirleyebildiği gözlenmiştir. Bu yöntemin aykırı değerleri tespit etme oranları MCD

yönteminin tespit etme oranlarından daha büyük olsa da bu değerlerin oldukça düşük olduğu gözlenmiştir.

MCD yöntemi aykırı değerleri MVE yönteminden daha kısa sürede tespit etse dahi aykırı değerleri belirleme oranlarının oldukça düşük olduğu gözlenmiştir.

FMCD yönteminin MCD yöntemine nazaran aykırı değerleri daha kısa sürede, daha yüksek oranda belirleyebildiği görülmüştür. Ancak MVV yönteminin FMCD yöntemiyle hemen hemen aynı sürede daha yüksek oranda aykırı değerleri belirlediği gözlenmiştir.

Kovaryans matrisinin farklılaştırılmasıyla sözü edilen yöntemlerin aykırı değerleri tespit etmede gösterdikleri başarıda önemli derecede düşüş olduğu gözlenmiştir. Bu durumun temel nedeni oluşturulan alt kümelerin sayısı arttıkça gözlemlerin hangi örnekten geldiğinin ayırt edilmesinde azalış meydana gelmesidir.

7. SONUÇ VE ÖNERİLER

Bu tez çalışmasında çok değişkenli veri analizinde aykırı değerlerin bulunması durumunda konum ve kovaryans matrisinin sağlam tahminlerini elde edebilmek için aykırı değerleri tespit etmede kullanılan MVE, MCD, FMCD gibi yöntemler ile bu yöntemlere alternatif olarak geliştirilen MVV yöntemi üzerinde durulmuştur. Çalışmanın amacı, MVV yönteminin diğer yöntemlerden hangi durumlarda daha avantajlı olduğunu nedenleriyle birlikte açıklamaktır. Bu nedenle MATLAB R2008a paket programı yardımıyla bir simülasyon çalışması yapılmıştır.

Örnek hacmi ve veri içindeki aykırı değer sayısı arttıkça hesaplamanın uzun sürmesinden dolayı MVE ve MCD yöntemlerinin kullanılmasının uygun olmadığı görülmüştür. Üstelik değişken sayısı arttıkça hesaplama daha uzun süreceğinden bu yöntemlerin kullanılabilirlikleri azalacaktır. FMCD yöntemi ile MVV yönteminin düşük örnek hacimlerinde benzer özellikler göstermesine rağmen örnek hacmi arttıkça MVV yönteminin aykırı değerleri tespit etmede daha başarılı olduğu görülmüştür. Genel olarak veri içindeki aykırı değer sayısı en çok iken MVV yönteminin diğer üç yönteme nazaran daha yüksek oranda aykırı değerleri belirleyebildiği söylenebilir.

Sonuç olarak, MVV yönteminin çok değişkenli veri analizinde aykırı değerleri tespit etmede kullanılan birçok yöntemden daha yüksek orandaki sonuçlara daha kısa sürede ulaştığı gözlemlendiği için bu yöntemin birçok durumda kullanılmasının uygun olacağı düşünülmektedir.

KAYNAKLAR

- Angiulli, F. And Pizzuti, C., “Outlier mining and large high-dimensional data sets”, *IEEE Transactions on Knowledge and Data Engineering*, 17 (2): 203-215 (2005).
- Bickel, P. J., “On some analogues to linear combination of order statistics in the linear model”, *The Annals of Statistics*, 1 (4): 597-616 (1973).
- Djauhari, M. A., “ A measure of multivariate data concentration”, *Journal of Applied Probability and Statistics*, 2(2): 139-155 (2007).
- Djauhari, M. A., “ A Robust Estimation of Location and Scatter”, *Malaysian Journal of Mathematical Sciences*, 2 (1): 1-24 (2008).
- Gelman, A., Carlin, J. B., Stern, H. S., And Rubin, D. B., “Bayesian Data Analysis 2nd ed. ”, *Chapman & Hall/CRC*, Boca Raton, 77 (2004).
- Hadi, A. S., “Identifying Multiple Outliers in Multivariate Data”, *Journal of Royal Statist. Soc. Series B*, 53 (3): 761-771 (1992).
- Herwindiati, D. E., Djauhari, M. A. And Mashuri, M., “Robust Multivariate Outlier Labeling”, *Communications in Statistics-Simulation and Computation*, 36 (6): 1287-1294 (2007).
- Huber, P. J., “Robust Regression: Asymptotics, Conjectures And Monte Carlo”, *The Annals of Statistics*, 1 (5): 799-821 (1973).
- İnternet : SAS Institute Inc. “Robust Regression Examples”
<http://support.sas.com/rnd/app/da/iml/robustreg.html> (1999).
- Jaeckel, L. A., “Estimating Regression Coefficients By Minimizing The Dispersion Of The Residuals”, *The Annals of Mathematical Statistics*, 43 (5): 1449-1458 (1972).
- Jureckova, J., “Nonparametric Estimate Of Regression Coefficients”, *The Annals of Mathematical Statistics*, 42 (4): 1328-1338 (1971).
- Koenker, R. And Bassett, G. J., “Regression Quantiles”, *Econometrica*, 46 (1): 33-50 (1978).
- Lopuhaa, H. P. And Rousseeuw, P.J., “Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices”, *The Annals of Statistics*, 19 (1): 229-248 (1991).
- Maronna, R. A., Martin, R. D. And Yohai, V. J., “Robust Statistics: Theory and Methods”, *John Wiley & Sons*, England, 1-55 (2006).

Rousseeuw, P. J., “Least Median of Squares Regression”, *Journal of the American Statistical Association*, 79 (388): 871-880 (1984).

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

Rousseeuw, P. J. And van Zomeren, B. C., “Unmasking Multivariate Outliers and Leverage Points”, *Journal of the American Statistical Association*, 85 (411): 633-639 (1990).

Rousseeuw, P. J. And van Driessen, K., “A Fast Algorithm for the Minimum Covariance Determinant Estimator”, *Technometrics*, 41 (3): 212-223 (1999).

Rousseeuw, P. J. And Leroy, A. M., “Robust Regression and Outlier Detection”, *John Wiley & Sons*, New Jersey, 57-60 (2003).

EKLER

EK-1 MVE Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

tic
tekrar=1000 %döngü sayısı
p=2 % değişken sayısı
alfa=0.975
mu1=[1 1]
mu2=[6 6]
%var1=[0.1,0.05;0.05,0.1]
var1=[1,0;0,1]
var2=[1,0;0,1]
n=5
n2=2
n1=n-n2
dogru=0
for i=1:tekrar
x1=mvnrnd(mu1,var1,n1);
x2=mvnrnd(mu2,var2,n2);
x=[x1;x2];

for h=floor((n+p+1)/2):(n+p+1)/2 % floor tamsayıya yuvarlar
c=(chi2inv(h/n,h-1));
%c=(chi2inv(alfa,h-1));
t=factorial(n)/(factorial(n-h)*factorial(h)); %t n'in h'lı kombinasyonlarının
sayısı: h elemanlı mümkün alt küme sayısı.
y=combntrns(1:n,h); % 1'den n'ye kadar olan sayıların h lık kombinasyonları.
hacim1=[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0];
for r=1:t
z=y(r,:); % z: n'nin h'lık tüm kombinasyonlarından y nin r.inci sırasındaki
kombinasyon
x; % mixture dağılımdan üretilmiş veri matrisi
yy=[1,1];
for ii=1:h
xx=x(z(ii,:),:); % x veri matrisinde yer alan elemanlardan h tanesi ile oluşturulmuş bir
kombinasyon
yy=[yy;xx] ;
end
yy(1,:)=[];
zz=[yy(:,1);yy(:,2)];
D=[1];
for jj=1:h
F=(yy(jj)-mean(yy))*inv(cov(yy))*(yy(jj)-mean(yy))';
D=[D;F];
end
D(1,:)=[];
dd=max(D);
zz';
hacim=[zz',sqrt((det((dd/c)*cov(yy))))];

```

EK-1 (Devam) MVE Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

hacim1=[hacim1;hacim];

end
hacim1(1,:)=[];
H= sortrows(hacim1, 2*h+1); %en küçük hacimli h gözlemden oluşan alt örnek
A=H(1,:); % en küçük hacime sahip h birimden oluşan alt örnek ve hacim değeri.
A(:,23)=[]; %en küçük hacime sahip h birimden oluşan alt örnek.
k1=A;
k2=A;
for m=1:h
k1(:,h+1)=[];
k2(:,1)=[];
end

Q=[k1',k2']; %en küçük hacime sahip h birimden oluşan alt örnek.
ortQ=mean(Q);
varQ=cov(Q);
ay=[1,1,1];
for mm=1:n
DD2=sqrt((x(mm)-ortQ)*inv(varQ)*(x(mm)-ortQ)');
if DD2>sqrt(chi2inv(alfa,n-1))
ay=[ay;mm,x(mm,:)];
end
end
ay(1,:)=[];
ay(:,1)=[];
x2;
if sum(sum(ay))==sum(sum(x2))
dogru=dogru+1
end
%bu Q da yer alan gözlemler üzerinden belirlenen uzaklık değerleri
%ki-kare kritik değeri ile karşılaştırılıp aykırı gözlemler
%belirlenecek. Bu aykırı olarak tespit edilen gözlemler eğer
%gerçekte bizim veriye yerleştirdiğimiz aykırı gözlemlerle aynı ise
%bir çetele atılacak ve döngü birdaha tekrarlanacak. Böylece
%yöntemin doğru tespit etme oranı elde edilmiş olacak.

end
end
oran=dogru/tekrar
t=toc

```

EK-2 MCD Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

tic
tekrar=1000 %döngü sayısı
p=2 % değişken sayısı
alfa=0.975
mu1=[1 1]
mu2=[6 6]
%var1=[0.1,0.05;0.05,0.1]
var1=[1,0;0,1]
var2=[1,0;0,1]
n=15
n2=6
n1=n-n2
dogru=0
for i=1:tekrar
    x1=mvnrnd(mu1,var1,n1);
    x2=mvnrnd(mu2,var2,n2);
    x=[x1;x2];

%-----
for h=floor((n+p+1)/2):(n+p+1)/2 % floor tamsayıya yuvarlar
    %c=(chi2inv(h/n,h-1));
    t=factorial(n)/(factorial(n-h)*factorial(h)); %t n'in h'lı kombinasyonlarının
    sayısı: h elemanlı mümkün alt küme sayısı.
    y=combntrns(1:n,h); % 1'den n'ye kadar olan sayıların h lık kombinasyonları.
    determ1=[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0];
    for r=1:t
        z=y(r,:); % z: n'nin h'lık tüm kombinasyonlarından y nin r.inci sırasındaki
        kombinasyon
        x; % mixture dağılımdan üretilmiş veri matrisi
        yy=[1,1];
        for ii=1:h
            xx=x(z(ii),:); % x veri matrisinde yer alan elemanlardan h tanesi ile oluşturulmuş bir
            kombinasyon
            yy=[yy;xx];
        end
        yy(1,:)=[];

    fff=det(cov(yy));
    zz=[yy(:,1);yy(:,2)]; %yy matrisinin sütunlarını tek bir satıra yan yana yazar.
    determ=[zz',fff];
    determ1=[determ1;determ];
    end
    determ1(1,:)=[];
    H= sortrows(determ1, 2*h+1);

```

EK-2 (Devam) MCD Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

A=H(1,:); % en küçük kovaryans determinantına sahip h birimden oluşan alt örnek ve hacim değeri.

A(:,19)=[]; %en küçük kovaryans determinantına sahip h birimden oluşan alt örnek.

k1=A;

k2=A;

for m=1:h

k1(:,h+1)=[];

k2(:,1)=[];

end

Q=[k1',k2']; %en küçük hacime sahip h birimden oluşan alt örnek.

ortQ=mean(Q);

varQ=cov(Q);

ay=[1,1,1];

for mm=1:n

DD2=sqrt((x(mm)-ortQ)*inv(varQ)*(x(mm)-ortQ)');

if DD2>sqrt(chi2inv(alfa,n-1))

ay=[ay;mm,x(mm,:)];

end

end

ay(1,:)=[];

ay(:,1)=[];

x2;

if sum(sum(ay))==sum(sum(x2))

dogru=dogru+1

end

end

end

oran=dogru/tekrar

t=toc

EK-3 FMCD Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

tic
tekrar=1000 %döngü sayısı
p=2 % değişken sayısı
alfa=0.975
mu1=[1 1]
mu2=[6 6]
var1=[1,0;0,1]
var2=[1,0;0,1]
n=15
n2=6
n1=n-n2
dogru=0
for i=1:tekrar
x1=mvnrnd(mu1,var1,n1);
x2=mvnrnd(mu2,var2,n2);
x=[x1;x2];
ss=x;
q=1:n;
q';
x=[q',x];

%-----
for h=floor((n+p+1)/2):floor(n+p+1)/2 % floor tamsayıya yuvarlar
t=factorial(n)/(factorial(n-h)*factorial(h)); %t n'in h'lı kombinasyonlarının
sayısı: h elemanlı mümkün alt küme sayısı.
y=combntrns(1:n,h); % 1'den n'ye kadar olan sayıların h lık kombinasyonları.
alt = 1; ust = t+1;
r =floor (alt + (ust-alt) * rand(1,1));
z=y(r,:); % z: n'nin h'lık tüm kombinasyonları olan y nin r.inci sırasındaki
kombinasyon
D=-1;
uu=10;

for kk=1:uu
D1=D;
yy=[1,1];
K=x;
K(:,1)=[];
for ii=1:h
xx=K(z(ii),:); % x veri matrisinde yer alan elemanlardan h tanesi ile oluşturulmuş
bir kombinasyon
yy=[yy;xx];
end
yy(1,:)=[];
ortyy=mean(yy);
varyyy=cov(yy);

```

EK-3 (Devam) FMCD Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

D=det(varyy);
kk;
if D==D1
i;
'hey';
kk=uu ;
ay=[1,1,1];
for ee=1:n
DD2=[ee,((ss(ee)-ortyy)*inv(varyy)*(ss(ee)-ortyy)')];
if DD2>sqrt(chi2inv(alfa,n-1))
ay=[ay;ee,ss(ee,:)];
end
end
ay(1,:)=[];
ay(:,1)=[];
if sum(sum(ay))==sum(sum(x2))
dogru=dogru+1;
end
break
else
d2=[1,1];
for mm=1:n
DD2=[mm,((x(mm)-ortyy)*inv(varyy)*(x(mm)-ortyy)')];
d2=[d2;DD2];
end
d2(1,:)=[]; %yy matrisinden hesaplanan ortalama ve ör.varyansı üzerinden
hesaplanan cook uzaklıkları.
Hd2= sortrows(d2,2);
for tt=1:n-h
Hd2(h+1,:)=[];
end
Hd2;
z=Hd2;
z(:,2)=[];

z';.
end
end
end
oran=dogru/tekrar
end
t=toc

```

EK-4 MVV Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

tic
tekrar=1000 %döngü sayısı
p=2 % değişken sayısı
alfa=0.975
mu1=[1 1]
mu2=[6 6]
var1=[1,0;0,1]
var2=[1,0;0,1]
n=20
n2=6
n1=n-n2
dogru=0
for i=1:tekrar
x1=mvnrnd(mu1,var1,n1);
x2=mvnrnd(mu2,var2,n2);
x=[x1;x2];
ss=x;
q=1:n;
q';
x=[q',x];

%-----
for h=floor((n+p+1)/2):floor(n+p+1)/2 % floor tamsayıya yuvarlar
t=factorial(n)/(factorial(n-h)*factorial(h)); %t n'in h'lı kombinasyonlarının
sayısı: h elemanlı mümkün alt küme sayısı.
y=combntrns(1:n,h); % 1'den n'ye kadar olan sayıların h lık kombinasyonları.
alt = 1; ust = t+1;
r =floor (alt + (ust-alt) * rand(1,1));
z=y(r,:); % z: n'nin h'lık tüm kombinasyonları olan y nin r.inci sırasındaki
kombinasyon
D=-1;
uu=10;

for kk=1:uu
D1=D;
yy=[1,1];
K=x;
K(:,1)=[];
for ii=1:h
xx=K(z(ii),:); % x veri matrisinde yer alan elemanlardan h tanesi ile oluşturulmuş
bir kombinasyon
yy=[yy;xx];
end
yy(1,:)=[];
ortyy=mean(yy);
varyyy=cov(yy);

```

EK-4 (Devam) MVV Yöntemi için MATLAB R2008a'da Hazırlanan Bilgisayar Kodları

```

D=trace(varyy)'*(varyy);
kk;
if D==D1
i;
'hey';
kk=uu ;
ay=[1,1,1];
for ee=1:n
DD2=[ee,((ss(ee)-ortyy)*inv(varyy)*(ss(ee)-ortyy)')];
if DD2>sqrt(chi2inv(alfa,n-1))
ay=[ay;ee,ss(ee,:)];
end
end
ay(1,:)=[];
ay(:,1)=[];
if sum(sum(ay))==sum(sum(x2))
dogru=dogru+1;
end
break
else
d2=[1,1];
for mm=1:n
DD2=[mm,((x(mm)-ortyy)*inv(varyy)*(x(mm)-ortyy)')];
d2=[d2;DD2];
end
d2(1,:)=[]; %yy matrisinden hesaplanan ortalama ve ör.varyansı üzerinden
hesaplanan cook uzaklıkları.
Hd2= sortrows(d2,2); %d2 matrisindeki uzaklıklar bakımından gözlemler sıraya
dizilir. Sonra bu matrisin
% birinci sütununda yer alan gözlemlerin
% üretilme numaralarının yeni yerinin ne
% olduğu, hangi gözlemlerin
% uzaklıklarının en küçük olduğunun tespiti
% için önemli.
for tt=1:n-h
Hd2(h+1,:)=[];
end
Hd2;
z=Hd2;
z(:,2)=[];
z';

```


EK-4 (Devam) MVV Yöntemi için MATLAB R2008a’da Hazırlanan Bilgisayar
Kodları

```
end  
end  
end  
oran=dogru/tekrar  
end
```

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : TURGUT, Kübra
Uyruğu : T.C.
Doğum tarihi ve yeri : 24.02.1984 Samsun
Medeni hali : Bekar
Telefon : 0 (362) 231 07 46
e-mail : kubra_turgut@yahoo.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek Lisans	Gazi Üniversitesi/ İstatistik Bölümü	2010
Lisans	Gazi Üniversitesi/ İstatistik Bölümü	2007
Lise	Samsun Atatürk Anadolu Lisesi	2002

Yabancı Dil

İngilizce

Hobiler

Yüzme, Resim, Fotoğraf Çekmek