TC
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS

AYKIRI DEĞERLERİN TESPİTİ İÇİN KULLANILAN DAYANIKLI UZAKLIK YÖNTEMLERİNİN KARŞILAŞTIRILMASI

TUBA ÇELEBİ

İSTATİSTİK ANABİLİM DALI

SAMSUN
2019

TEZ ONAYI

Tuba ÇELEBİ tarafından hazırlanan "Aykırı Değerlerin Tespiti için Kullanılan Dayanıklı Uzaklık Yöntemlerinin Karşılaştırılması" adlı tez çalışması 09/07/2019 tarihinde aşağıdaki jüri tarafından Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman Doç. Dr. Pelin KASAP
İstatistik Anabilim Dalı

Jüri Üyeleri

Başkan Prof. Dr. Vedide Rezan USLU
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Pelin KASAP
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Dr. Öğretim Üyesi Tuba KOÇ
Çankırı Karatekin Üniversitesi
İstatistik Anabilim Dalı

Yukarıdaki sonucu onaylarım. .../.../2019

Prof. Dr. Bahtiyar ÖZTÜRK
Enstitü Müdürü

ETİK BEYAN

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

09/07/2019

Tuba ÇELEBİ

ÖZET

Yüksek Lisans Tezi

AYKIRI DEĞERLERİN TESPİTİ İÇİN KULLANILAN DAYANIKLI UZAKLIK YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Tuba Çelebi

Ondokuz Mayıs Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Teorisi Anabilim Dalı

Danışman: Doç. Dr. Pelin Kasap

Aykırı değer konusu, en eski istatistiksel ilgi alanlarından biridir ve birçok veri kümesi değişen miktarlarda aykırı değerler içerdiğinden, güncel konulardan biri olmaya devam etmektedir. Veride aykırı değerler olması istatistiksel analizleri olumsuz yönde etkilemektedir. Bu nedenle aykırı değerlerin tespiti istatistikte önemli bir yere sahiptir. Aykırı değerlerin tespitinde sıklıkla kullanılan yöntemlerden biri Mahalanobis uzaklığıdır. Ancak bu uzaklık aykırı değerleri tespit ederken aykırı değerlerin varlığından oldukça fazla etkilenen klasik konum ve ölçek parametrelerinin tahmin edicilerini kullanır. Aykırı değerlerin tespitini daha güvenilir yapmak için Mahalanobis uzaklıklarının hesaplanmasında klasik tahmin ediciler yerine dayanıklı konum ve ölçek parametrelerinin tahmin edicileri kullanılabilir. Bu amaçla, bu tez çalışmasında, dayanıklı tahmin edicilerden hızlı-en küçük kovaryans determinant (FMCD), en küçük hacimli elipsoit (MVE), M-tahmin edicileri (MEST), Stahel-Donoho tahmin edicisi (SDE), dikey gnanadesikan-kettenring (OGK) ve parçalı uyarlanabilir hesaplama yönünden etkin aykırı gözlem belirleyicisi (BACON) yöntemleri kullanılmıştır. Bu yöntemler normal dağılıma sahip veriler ve aykırı değerler içeren verileri modellemede sıklıkla kullanılan bir dağılım olan uzun kuyruklu simetrik (LTS) dağılıma sahip veriler için Monte-Carlo benzetim çalışması ile karşılaştırılmıştır. Karşılaştırma için performans kriteri olarak bu yöntemlerin aykırı değerleri tespit etmedeki başarı oranları kullanılmıştır. Ayrıca, tüm yöntemler için yanlış tespit sayıları ve aykırı değerleri tespit etme hızları da hesaplanmıştır. Hem normal dağılıma sahip veriler için hem de LTS dağılımına sahip veriler için dayanıklı yöntemlerin klasik yöntemden daha iyi başarı oranına sahip olduğu görülmüştür. Dayanıklı yöntemler arasında birçok durumda en yüksek başarı oranını ise OGK yöntemi vermektedir.

Temmuz 2019, 67 sayfa

Anahtar Kelimeler: Mahalanobis uzaklığı, Dayanıklı Tahmin, Aykırı değer tespiti, Uzun-kuyruklu simetrik dağılım

ABSTRACT

Master's Thesis

THE COMPARISON OF ROBUST DISTANCE METHODS FOR OUTLIERS DETECTION

Tuba Çelebi

Ondokuz Mayis University
Graduate School of Sciences
Department of Statistics

Supervisor: Assoc. Prof. Dr. Pelin Kasap

The issue of an outlier is one of the oldest statistical interests, and since many data sets contain varying amounts of the outliers, they remain one of the current issues. Outliers in the data adversely affect the statistical analysis. Therefore, the detection of outliers have an important place in statistics. Mahalanobis distance is one of the methods commonly used in the detection of outliers. However, this distance uses the estimators of classical location and scale parameters, which are highly influenced by the presence of outliers, when detection of outliers. To make the detection of outliers more reliable, estimators of robust location and scale parameters can be used instead of classical estimators to calculate Mahalanobis distances. For this purpose, in this thesis, Fast-Minimum Covariance Determinant (FMCD), Minimun Volume Ellipsoid (MVE), M-Estimators (MEST), Stahel-Donoho Estimator (SDE), Orthogonalized Gnanadesikan-Kettenring (OGK) and Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) methods have used. These methods have been compared with the Monte-Carlo simulation study for data with normal distribution and data with Long-Tailed Symmetric (LTS) distribution which is a distribution commonly used in modeling data including outliers. For comparison, the success rates of these methods in determining outliers have used as performance criteria. In addition, false detection rates and the time to detect outliers have been calculated. For both normal distribution data and LTS distribution data, robust methods have been found to have a better success rate than the classical method. Among the robust methods, OGK method has given the highest success rate in many cases.

July 2019, 67 pages

Key Words**:** Mahalanobis distance, Robust estimation, Outlier detection, Long-tailed symmetric distribution

ÖNSÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan yüksek lisans tez çalışmalarım boyunca yardım ve desteğini benden esirgemeyen danışman hocam Sayın Doç. Dr. Pelin KASAP'a teşekkürü bir borç bilirim.

Beni bugünlere getirmek için hiçbir fedakârlıktan kaçınmayan aileme sonsuz teşekkür ederim.

Temmuz 2019, Samsun

Tuba Çelebi

İÇİNDEKİLER DİZİNİ

KISALTMALAR

BACON	Parçalı Uyarlanabilir Hesaplama Yönünden Etkin Aykırı Gözlem Belirleyicisi (Blocked Adaptive Computationally Efficient Outlier Nominators)
FMCD	Hızlı En Küçük Kovaryans Determinant (Fast Minimum Covariance Determinant)
LTS	Uzun Kuyruklu Simetrik (Long-Tailed Symmetric)
MD	Mahalanobis Uzaklığı (Mahalanobis Distance)
MEST	M-Tahmin Edicileri (M-Estimation)
MVE	En Küçük Hacimli Elipsoit (Minimum Volume Elipsoid)
OGK	Dikey Gnanadesikan-Kettenring (Orthogonal Gnanadesikan-Kettenring)
SDE	Stahel-Donoho Tahmin Edicisi (Stahel-Donoho Estimator)

ÇİZELGELER DİZİNİ

1. GİRİŞ

Aykırı değer konusu, en eski istatistiksel ilgi alanlarından biridir ve neredeyse tüm veri kümeleri, değişen miktarlarda aykırı değerler içerdiğinden, en önemli konulardan biri olmaya devam etmektedir (Werner 2003). Aykırı değerlerin doğru tanımlanması istatistiksel analizde önemli bir rol oynar. Klasik istatistiksel modeller, verilerde bulunan aykırı değerler göz ardı edilerek uygulanırsa, sonuçlar yanıltıcı olabilir (Reza ve Ruhi, 2015).

Aykırı değerleri tespit etmek için geliştirilen klasik tahminleme yöntemleri, aykırı değerlerden etkilendiklerinden dolayı başarısız olabilirler. Çok değişkenli veri yapılarında aykırı değer durumu tek değişkenli veri yapılarına göre daha karmaşıktır. Çünkü bir gözlem değişkenler tek tek incelendiğinde aykırılık göstermezken, tüm değişkenler aynı anda ele alındığında aykırı olabilir. Örneğin, 20 yaşında ve 3 kez boşanmış bir birey düşünülsün. 3 kez boşanmak boşanma sayısı değişkeni bakımından aykırı değildir. 20 yaşında olmak da evlilik için aşırı bir yaş değildir. Ancak bu iki değişken birlikte incelendiğinde, 20 yaşında 3 kez boşanmış bir bireyin toplumdaki genel eğilime nazaran oldukça sıra dışı olduğu söylenebilir (Alpar, 2013).

Aykırı değerin tanımı birkaç kaynağa göre aşağıdaki gibi sıralanabilir:

i. Aykırı değer, içinde bulunduğu örneğin diğer üyelerinden belirgin bir şekilde sapan gibi gözüken bir gözlemdir (Grubbs, 1969).
ii. Aykırı değer başka bir mekanizma tarafından üretildiğine dair şüphe uyandırmak için diğer gözlemlerden çok fazla sapan bir gözlemdir (Hawkins, 1980).
iii. Aykırı değer, veri kümesinin geri kalanıyla tutarsız görünen bir gözlem veya gözlem alt kümesidir (Barnett ve Lewis, 1994).

Aykırı değerler, anormal davranışların altında yatan yararlı bilgiler içerebildiği için genellikle normal değerlerden daha ilginçtir (Fawzy vd, 2013).

Aykırı değerler, mekanik hatalar, sistem davranışlarındaki değişiklikler, hileli davranışlar, insan hatası, cihaz hatası veya basitçe popülasyonlardaki doğal sapmalar nedeniyle ortaya çıkar (Gaber, 2007).

İstatistik alanında çoğu analiz öncesinde verilerin normal dağıldığı varsayılmaktadır (Shu, 1978). Ancak çoğu zaman gerçek hayat verileri aykırı değer içerdiğinden bu varsayıma uymamaktadır. Bu durumda Normal dağılım yerine aykırı değerlerin modellenmesinde sıklıkla kullanılan Uzun kuyruklu simetrik (Long-Tailed Symmetric - LTS) dağılımın tercih edilmesi daha kullanışlı olabilir. Bu sebeplerden dolayı bu tez çalışmasında aykırı değerlerin tespitinde kullanılan klasik ve dayanıklı yöntemleri karşılaştırırken hem Normal dağılımına sahip veriler hem de LTS dağılımına sahip veriler üzerinde analizler gerçekleştirilmiş ve sonuçlar yorumlanmıştır. Literatürde LTS dağılıma sahip çok değişkenli veri kümelerinde aykırı değer tespiti için çalışmaya rastlanmamıştır.

Bu tez çalışmasının ikinci bölümünde geçmiş yıllarda aykırı değer tespit yöntemleri ile ilgili olan çalışmaların özetlerine yer verilmiştir. Üçüncü bölümde aykırı değerlerin tespitinde klasik konum ve ölçek parametreleri ile hesaplanan Mahalanobis uzaklığı verilmiştir. Dördüncü bölümde dayanıklı konum ve ölçek parametrelerinin tahmininde sıklıkla kullanılan tahmin ediciler tanıtılmıştır. Beşinci bölümde aykırı değer içeren verileri modellemede sıklıkla kullanılan uzun kuyruklu simetrik dağılıma yer verilmiştir. Altıncı bölümde aykırı değer tespiti için ele alınan Mahalanobis uzaklığı-MD, hızlı en küçük kovaryans determinant-FMCD, en küçük hacimli elipsoit-MVE, M-tahmin edicileri-MEST, Stahel-Donoho tahmin edicisi-SDE, dikey gnanadesikan-kettenring-OGK ve parçalı uyarlanabilir hesaplama yönünden etkin aykırı gözlem belirleyicisi-BACON yöntemlerini karşılaştırmak amacıyla Monte-Carlo benzetim çalışması yapılmıştır. Son bölümde ise bu çalışma sonucunda elde edilen bulgular yorumlanmıştır.

2. KAYNAK ÖZETLERİ

Pena ve Prieto (2001), kovaryans matrisi için dayanıklı bir tahmin edici ve çok değişkenli aykırı değer tespit yöntemleri üzerinde çalışmışlardır. Aykırı değer tespit yöntemlerinden FMCD, SDE ve Kurtosis yöntemlerinin performansını benzetim çalışması yoluyla analiz etmişlerdir.

Riani vd (2009), çok değişkenli normal dağılıma sahip bir veri setinde aykırı değerleri tespit ederken Mahalanobis uzaklıklarının dayanıklı olması için ileri arama yöntemini tanıtmışlardır. Önerdikleri ileri arama yöntemini ve diğer dayanıklı aykırı değer tespit yöntemlerinden MCD'nin ağırlıklandırılmış versiyonlarını karşılaştırmışlardır.

Alameddine vd (2010), su kalitesi verilerini kullanarak dayanıklı yöntemlerden MCD, MVE ve MEST ile aykırı değerleri tespit etmiş ve bu yöntemleri karşılaştırmışlardır.

Chen vd (2010), toplanan trafik verisinden önemli ve değerli bilgileri elde etmek ve trafik veri kümelerindeki aykırı değerleri tespit etmek için trafik mühendisliği alanında aykırı değer madenciliği tekniğini tanıtmışlardır.

Fallahi vd (2011), çalışmalarında küresel konumlandırma sistemi (Global Positioning System-GPS) dayanıklı tahmin edicilerden MEST, L_1 tahmin edicisi, $L_1 - L_2$ tahmin edicisi ve örneklem medyan tahmin edicisini kullanmışlardır.

Sajesh ve Srinivasan (2013), çalışmasında tek değişkenli veri setlerinde aykırı değer tespitine kısaca değinmiş, çok değişkenli veri setlerinde aykırı değer tespiti yöntemlerini kapsamlı olarak ele almışlardır. Bu yöntemleri dayanıklı uzaklık tabanlı yöntemler ve geleneksel olmayan yöntemler olarak iki ana başlıkta incelemişlerdir. Dayanıklı uzaklık tabanlı yöntemler olan Comedian, Kurtosis, FMCD ve OGK yöntemlerini benzetim çalışması yoluyla karşılaştırmışlardır.

Fawzy vd (2013), kablosuz algılayıcı ağlardaki (wireless sensor networks) aykırı değerleri tespit etmek için bir yöntem geliştirmişlerdir. Geliştirdikleri bu yöntemi literatürde hazır bulunan bir veri setine ve kendilerinin üretmiş oldukları veri setine uygulamışlardır. Bu yöntemin aykırı değerleri tespit etmedeki başarı oranlarını, yanlış tespit oranlarını ve aykırı değerleri tespit sürelerini incelemişlerdir.

Reza ve Ruhi (2015), çalışmalarında bağımsız bileşen analizi hakkında bilgi vermiş ve bu yöntemi çok değişkenli aykırı değer tespitinde kullanmak üzere literatürde hazır bulunan veri setine ve benzetim çalışması ile ürettiği veri setine uygulamıştır.

Uzabacı vd (2018), çok değişkenli veri setindeki aykırı değerleri tespit etmek için dayanıklı yöntemlerden BACON, FMCD ile konum ve ölçek parametrelerinin tahmininde kullandıkları en küçük kovaryans determinant tahmin edicilerle hesapladıkları Mahalanobis uzaklıklarını kullanmışlardır. Bu yöntemleri çok değişkenli simetrik dağılımlardan Normal, Laplace ve Cauchy dağılımlarıyla ürettiği veri setlerine uygulayarak karşılaştırmışlardır.

Leys vd (2018), Monte-Carlo benzetim çalışması ile ürettiği çok değişkenli veri seti üzerinde dayanıklı yöntemlerden MCD ile elde ettiği Mahalanobis uzaklıklarıyla aykırı değerleri tespit etmiş ve bu yöntemin kullanımını önermişlerdir.

Cabana vd (2019), çok değişkenli aykırı değer tespiti için dayanıklı Mahalanobis uzaklıklarını hesaplamışlardır. Dayanıklı Mahalanobis uzaklıklarını hesaplarken konum ve ölçek parametrelerinin tahmininde Shirinkage yöntemini kullanmışlardır. Shirinkage yöntemi ile Kurtosis, MCD ve ağırlıklandırılmış MCD yöntemlerini hem benzetim çalışması ile hem de gerçek veri seti üzerinde test etmişlerdir

3. AYKIRI DEĞER TESPİTİ

Tüm istatistiksel yöntemler, bir takım varsayımlara açıkça veya dolaylı olarak dayanmaktadır. En yaygın olarak kullanılan model biçimselleştirilmesi, gözlemlenen verilerin normal dağılıma sahip olduğu varsayımıdır. Bu varsayım iki asırdır istatistik alanında mevcuttur ve regresyondaki tüm klasik metotların, varyans analizinin ve çok değişkenli analizin iskeletini oluşturmaktadır (Maronna vd, 2006).

Pratikte, üzerinde çalışılan gözlemlerin normal dağılıma sahip olması istenmektedir. Ancak bazı gözlemler, farklı bir modele uyar ya da hiç bir modele uymazlar. Bu tür gözlemler, tüm veri analizi yöntemlerinde ve istatistiksel modelleme uygulamalarında yaygın olarak karşımıza çıkmaktadır. Bu tür gözlemlere aykırı değer adı verilir ve tek bir aykırı değerin varlığı, normallik varsayımı altında en iyi olan klasik istatistiksel yöntem üzerinde büyük bir bozucu etkiye sahip olabilir (Maronna vd, 2006).

Hemen hemen tüm istatistiksel analizlerde, eksik değerler ve aykırı değerler problem yaratır. Çok değişkenli veri kümelerinin çoğu, aykırı değerler içerir. Aykırı değerler, verideki birimlerin çoğunluğuna uymayan sıra dışı noktalar olarak tanımlanır. Verideki aykırı değerler, yanlış sonuçlara ve yanlış tahminlere yol açabilir. Çünkü bir modelin parametrelerinin tahminlerini önemli bir şekilde etkilerler. Bu nedenle, aykırı değerlerin varlığını belirlemek önemlidir (Alkan vd, 2015).

Büyük hatalar genellikle kendilerini aykırı değer olarak gösterir, ancak tüm aykırı değerler büyük hatalar değildir. Bazı aykırı değerler gerçektir ve örneklemin en önemli gözlemleri olabilir. Örneğin, jeodezik bir nokta aniden farklı bir pozisyonda görünüyorsa, bu bir tür büyük bir hata anlamına gelebilir ya da yeraltının değişmesi anlamına gelebilir ve bu olasılıkları ayırt etmek için deneyim gerekir (Hampel, 2001).

Eğer veriler aykırı değerler içermiyorsa, dayanıklı yöntem klasik yöntemle yaklaşık olarak aynı sonuçları verirken, aykırı değerlerin küçük bir oranı dahi mevcutsa, dayanıklı yöntem aykırı değer içermeyen verilere uygulanan klasik yöntemle yaklaşık olarak aynı sonuçları verir. Verilerin büyük bir kısmının uyum

sağlaması sonucunda, dayanıklı yöntemler, yüksek-boyutlu çok değişkenli durumlarda bile, aykırı değerleri saptamak için çok güvenilir yöntemlerdir (Maronna vd, 2006).

Aykırı değerlere sahip çok değişkenli gözlem vektörleri, varyansı mümkün olduğunca şişirme ve korelasyonu düşürme eğilimindedirler (Campbell, 1980).

Çok değişkenli veri setlerinde aykırı değerleri tespit etmek için çeşitli yöntemler bulunmaktadır. Bu yöntemler aşağıdaki gibidir:

- Mahalanobis uzaklıkları hesaplanarak aykırı değerlere karar vermek
- Chernoff yüzleri ya da ikon grafikleri gibi grafiklerden yararlanmak
- Genelleştirilmiş varyans oranına göre karar vermek (Alpar, 2013).

3.1. Mahalanobis Uzaklığı

Çok değişkenli verilerin şekli ve büyüklüğü kovaryans matrisi tarafından incelenir. Kovaryans matrisini dikkate alan iyi bilinen bir uzaklık ölçümü, Mahalanobis uzaklığıdır. Bir p boyutlu çok değişkenli örneklem $x_i\,(i = 1, \dots, n)$ için Mahalanobis uzaklığı şöyle tanımlanır:

$$MD_i = \sqrt{(x_i - \hat{\mu})' \widehat{\sum}^{-1} (x_i - \hat{\mu})} \qquad i = 1, \dots, n. \qquad \textbf{(3.1)}$$

$\sqrt{\chi^2_{p;\alpha}}$ değerinden büyük olan gözlemler aykırı gözlem olarak adlandırılırlar. Burada, $\chi^2_{p;\alpha}$, α kritik değerli ve p serbestlik dereceli χ^2 dağılımını gösterir. p veri setindeki değişken sayısını ifade etmektedir (Hubert ve Debruyne, 2010; Hubert vd, 2012). $\hat{\mu}$, konum parametresinin tahmin edicisi olan ortalama vektörünü ve $\widehat{\sum}$ ise ölçek parametresinin tahmin edicisi olan varyans-kovaryans matrisini ifade eder.

Mahalanobis uzaklığı, gözlemlenen rasgele bir örneklem $x_1, \dots, x_n$ 'in çok değişkenli normal dağılımdan geldiğini test etmek için de kullanılabilir (McLachlan, 1999).

Mahalanobis uzaklığının (3.1) eşitliğindeki kullanımı aykırı değerlere karşı çok duyarlıdır. Çünkü hesaplanması klasik tahmin ediciler olan aritmetik ortalama ve kovaryans matrisine bağlıdır (Maesschalck vd, 2000). Bu sebepten Mahalanobis uzaklığını aykırı değer tespitinde daha güvenilir yapmak için (3.1) eşitliğinde konum

ve ölçek parametrelerinin tahmini için klasik tahmin ediciler yerine dayanıklı tahmin ediciler kullanılabilir.

4. DAYANIKLI KONUM VE ÖLÇEK TAHMİN EDİCİLERİ

Dayanıklılık, Simon Newcomb ile en az 19. yy'ın sonlarına kadar takip edilebilen uzun bir geçmişe sahiptir. Ama ilk büyük adımlar Tukey (1960,1962), Huber (1964,1967) ve Hampel (1971,1974)'in temel çalışmaları ile 1960'larda ve 1970'lerin başında atılmıştır. Bu araştırmacılar tarafından önerilen dayanıklılığın uygulanabilirliği, bilgisayarların artan hız ve erişilebilirliği ile mümkün olmuştur.

Dayanıklılık istatistik alanında son kırk yılda hem araştırma alanlarında hem de kanıtlanmış yayınlanan makalelerde önemli gelişmeler yaşamıştır. Huber (1981), Hampel vd (1986), Rousseeuw ve Leroy (1987) ve Staudle ve Sheather (1990) tarafından etkili kitaplar yazılmıştır (Maronna vd, 2006).

Dayanıklı istatistik, istatistiksel prosedürlerin dayanıklılık teorisidir. Sistematik olarak, bilinen prosedürler üzerindeki modelleme varsayımlarından sapmaların etkilerini inceler ve gerekirse yeni ve daha iyi prosedürler geliştirir (Hampel, 2001).

Örnek ortalama vektörü ve örnek kovaryans matrisi, klasik çok değişkenli analizin yapı taşıdır. Çok değişkenli varyans analizi, temel bileşenler analizi, faktör analizi, kanonik korelasyon analizi, diskriminant analizi ve sınıflandırması ve kümelenmesi dahil olmak üzere çok sayıda çok değişkenli veri analizi yöntemleri için çok önemlidir. Çok değişkenli normal modellerde, konum ve ölçek parametrelerinin en uygun tahmin edicileridir. Bununla birlikte, bu klasik konum ve ölçek parametre tahmin edicilerinin, olağandışı gözlemlere karşı son derece hassas olduğu ve verilerdeki küçük sapmalara duyarlı olduğu iyi bilinmektedir (Zuo, 2006).

Mahalanobis uzaklığı hesaplanırken kullanılan klasik konum ve ölçek parametre tahmin edicilerinin aykırı değerlerden olumsuz yönde etkilenmesi dayanıklı konum ve ölçek parametre tahmin edicilerinin kullanımını popüler hale getirmiştir. Dayanıklı konum ve ölçek parametrelerinin sıklıkla kullanılan tahmin edicileri aşağıda açıklanmaktadır. Bu tahmin ediciler (3.1) eşitliğinde yerine yazılarak dayanıklı Mahalanobis uzaklıkları elde edilir ve $\chi^2_{p;\alpha}$ kritik değeri ile karşılaştırılarak aykırı değerler tespit edilir.

4.1. M-Tahmin Edicisi

Huber (1964) 'ın dönem ödevinden esinlenen Maronna (1976), ilk olarak çok değişkenli konum ve ölçek parametrelerinin genel M-tahmin edicilerini tanıtmıştır (Zuo, 2006).

M-tahmin ediciler, klasik tahmin edicilerin basit bir değişimi olarak düşünülebilir; Verilerin ana gövdesinden geldiği düşünülen gözlemlere tam ağırlık verir, ancak kirletilmiş dağılımın kuyruklarından gelen gözlemlere daha az ağırlık veya etki verir (Campbell, 1980).

$x_1, x_2, \ldots, x_n$ p- değişkenli bir örneklem olsun. Bu örneklemin konum ve ölçek parametrelerinin MEST tahmin edicileri şöyle hesaplanır;

$$\hat{\mu}_{MEST} = \frac{\sum_{i=1}^{n} w_1\{d(x_i, t_n, C_n)\}x_i}{\sum_{i=1}^{n} w_1\{d(x_i, t_n, C_n)\}} \quad \textbf{(4.1)}$$

$$\hat{\sum}_{MEST} = \frac{1}{n}\sum_{i=1}^{n} w_2\{d^2(x_i, t_n, C_n)\}(x_i - t_n)(x_i - t_n)' \quad \textbf{(4.2)}$$

Burada $d(x_i, t, C) = \{(x_i - t)'C^{-1}(x_i - t)\}^{1/2}$, x_i ve t arasındaki istatistiksel mesafedir, C ise pozitif tanımlı matristir. w_1 ve w_2 özel ağırlık fonksiyonlarıdır (Maronna, 1976; Croux ve Haesbroeck 1999).

4.2. Stahel-Donoho Tahmin Edicisi

Stahel-Donoho tahmin edicisi (SDE), Stahel (1981) ve Donoho (1982) tarafından bağımsız olarak önerilmiştir. SDE'nin amacı, çok değişkenli konum ve ölçek parametrelerinin klasik tahmininde aykırı gözlemlerin etkisini azaltmaktır. SDE yönteminde her bir gözlem tek değişkenli uzaya yansıtılır ve aykırılığın ölçüsü hesaplanır. Çok değişkenli uzaydan tek değişkenli uzaya sonsuz sayıda yansıma yönü olduğu için aykırılığın sonsuz sayıda değeri elde edilir. Bu nedenle amaç mümkün olan tüm yansıma yönlerinden en küçük üst sınırı (supremum) bulmaktır (Filzmoser vd, 2009).

x_i gözlemi için "aykırılık" değeri aşağıdaki gibi tanımlanır:

$$u_i = sup_{\|v\|} \frac{|v'x_i - med_j(v'x_j)|}{med_k|v'x_k - med_j(v'x_j)|}. \quad \textbf{(4.3)}$$

Burada v, p-boyutlu izdüşüm vektörüdür. Tüm gözlemler için u_i değerleri belirlendikten sonra konum ve ölçek parametrelerinin tahminleri şöyle hesaplanır;

$$\hat{\mu}_{SDE} = \frac{\sum_{i=1}^{n} w(u_i)x_i}{\sum_{i=1}^{n} w(u_i)} \quad \textbf{(4.4)}$$

$$\hat{\sum}_{SDE} = \frac{\sum_{i=1}^{n} w(u_i)(x_i - t_n)\left((x_i - t_n)\right)'}{\sum_{i=1}^{n} w(u_i)} \quad \textbf{(4.5)}$$

Burada $w(u_i)$, pozitif artan ağırlık fonksiyonudur.

Bu yöntem diğer dayanıklı tahmin edicilere kıyasla çok fazla ilgi görmemiştir. Bu muhtemelen hesaplamalarının uzun süre almasından kaynaklanmaktadır. Bir başka sebep ise Zuo vd (2004) 'nin makalesinde kısmen çözülmüş olan asimptotik teorinin eksikliğidir (Gervini, 2002).

SDE, aykırı değerlerin tespiti için iyi bir tahmin edicidir. Ancak örneklem boyutu çok büyükse, çok yavaş olabilmektedir (Hadi vd, 2009; Todorov ve Filzmoser, 2009).

4.3. En Küçük Hacimli Elipsoit Tahmin Edicisi

Rousseeuw(1984,1985) tarafından ortaya konan en küçük hacimli elipsoit (MVE), çok değişkenli konum ve ölçek parametre tahmininde yüksek kırılma noktasına sahip dayanıklı tahmin edicilerinin ilkidir. MVE, yaygın kullanımı ve aykırı değer tespiti için onu güvenilir yapan aykırı değerlere karşı yüksek direnci sayesinde popüler olmuştur.

MVE tahmin edicisi, n gözlemin h birimini kapsayan en küçük elipsoit hacmine dayanır. Bu tahmin edici düşük yanlılığı ve dayanıklı uzaklıklara dayalı kullanımı nedeniyle çok değişkenli veride aykırı değer tespiti için oldukça kullanışlıdır.

p boyutlu ve n gözlemli veri seti X_n=$\{x_1, \dots, x_n\}$ olsun. X_n veri setinin çok değişkenli konum ve ölçek parametrelerinin MVE tahmin edicisi, X_n'in en az h noktasını kapsayan en küçük hacmine sahip elipsoidin merkezi ve kovaryans yapısı olarak tanımlanır. Burada sözü edilen h, $n/2 + 1$ ile n arasından seçilebilir.

MVE konum ve ölçek parametrelerinin tahminleri;

$$\{i; (x_i - t)' C^{-1} (x_i - t) \leq c^2\} \geq h \quad \textbf{(4.6)}$$

koşuluna uyan h gözlemin ortalaması ve kovaryans matrisidir (Van Aelst ve Rousseeuw, 2009).

4.4. Hızlı-En Küçük Kovaryans Determinant Tahmin Edicisi

Çok değişkenli istatistikte çoğu yöntemde kovaryans ve korelasyon matrisi önemli bir rol oynar. Örneğin onlar, temel bileşenler analizinin, diskriminant analizinin, faktör analizinin ve daha birçok analizin ortak noktalarıdır (Mardia ve Kent, 1979; Alkan vd, 2015).

Hızlı en küçük kovaryans determinant (FMCD) tahmin edicisi, çok değişkenli konum ve ölçek parametrelerinin oldukça dayanıklı bir tahmin edicisidir (Hubert ve Debruyne, 2010; Rousseeuw ve Driessen, 1999). Kovaryans matrisi tahmini, birçok çok değişkenli istatistiksel yöntemin temel taşıyken; FMCD ise çok değişkenli tekniklerin etkili hesaplanmasını ve dayanıklılığını geliştirmek için kullanılmıştır (Hubert ve Debruyne, 2010). FMCD; finans, tıp, kalite kontrolü, görüntü analizi, kimya gibi çok sayıda araştırma alanına uygulanmıştır (Verdonck ve Hubert, 2011).

FMCD, en düşük determinanta sahip klasik kovaryans matrisi için veri setinin genişliği olan n üzerinden seçilen h gözlem bulmayı amaçlar (Rousseeuw ve Driessen, 1999; Hubert vd, 2005). Burada sözü edilen h gözlem $[(n+p+1)/2] \leq h \leq n$ aralığında olmalıdır. h'ın seçimi için $\binom{n}{h}$ kombinasyonu kadar deneme yapılarak en küçük determinanta sahip olan h gözlem seçilir. Konumun FMCD tahmin edicisi bu h gözlemin ortalaması ve ölçeğin FMCD tahmin edicisi ise h gözlemin kovaryans matrisidir (Rousseeuw ve Driessen, 1999).

$(x_1, x_2, \dots, x_n)$ veri setinin FMCD tahmin edicisi, mümkün olan tüm h genişliğindeki alt kümelerinin arasından en küçük determinanta sahip olan kovaryans matrisinin ait olduğu $(x_{i1}, x_{i2}, \dots, x_{ih})$ h birimlik alt kümesi kullanılarak elde edilir. FMCD konum ve ölçek parametrelerinin tahmin edicileri;

$$\hat{\mu}_{FMCD} = \frac{1}{h}\sum_{j=1}^{h} x_{ij} \qquad \textbf{(4.7)}$$

$$\hat{\Sigma}_{FMCD} = \left[\frac{1}{h-1}\sum_{j=1}^{h}\left(x_{ij} - \hat{\mu}_{FMCD}\right)\left(x_{ij} - \hat{\mu}_{FMCD}\right)'\right] \qquad \textbf{(4.8)}$$

olarak verilir (Rousseeuw ve Driessen, 1999; Todorov ve Filzmoser, 2009; Hubert vd, 2012).

4.5. Dikey Gnanadesikan-Kettenring Tahmin Edicisi

Kovaryans matrisinin afin eş değişim gereksiniminin bırakılması durumunda yüksek kırılma noktası olan çok daha hızlı tahminler hesaplanabilmektedir (Todorov ve Filzmoser, 2009). Afin eş değişim özelliği herhangi bir doğrusal dönüşüm altında sonuçların değişmeden kalmasının istendiği bazı durumlarda gerekli olan bir özelliktir. (Maronna vd, 2006). Afin eş değişim hakkında daha ayrıntılı bilgi, Lopuhaa ve Rousseeuw (1991) tarafından verilmiştir. Maronna ve Zamar (2002) tarafından pozitif tanımlı ve yaklaşık olarak afin eş değişimli dayanıklı kovaryans matrisi elde etmek için genel bir yöntem önermiştir. Bu yöntem Gnanadesikan ve Kettenring (1972) tarafından dayanıklı kovaryans matrisi tahminine uygulanmıştır. Çok değişkenli konum ve ölçek parametrelerinin Dikey Gnanadesikan-Kettenring (OGK) tahmin edicisi p-değişkenli veri seti için aşağıda verilen adımlarla hesaplanır:

(1) $m(.)$ ve $s(.)$ sırasıyla dayanıklı tek değişkenli konum ve ölçek parametrelerinin tahmin edicileri olsun.

(2) $D = diag((s(X_1), \dots, s(X_p))$ olmak üzere $y_i = D^{-1}x_i, \quad i = 1, \dots, n$ hesaplansın. $Y = (y_1, \dots, y_n)$ standartlaştırılmış veri matrisidir.

(3) U=(u_{jk}) olmak üzere eğer $j \neq k$ ise $u_{jk} = {}^1\!/_4\left(s\left(\frac{Y_j}{s(Y_j)} + \frac{Y_k}{s(Y_k)}\right)^2 - s\left(\frac{Y_j}{s(Y_j)} + \frac{Y_k}{s(Y_k)}\right)^2\right)$, $j = k$ ise $u_{jk} = 1$ olarak alınır. Burada $Y_l, l = 1, \dots, p$' dir.

(4) $\Lambda = diag(\Lambda_1, \dots, \Lambda_1)$ olmak üzere $U = E\Lambda E^T$ şeklinde yazılsın. E, U matrisinin özvektör matrisidir.

(5) $z_i = E^T y_i = E^T D^{-1} x_i$ ve $A = DE$ olsun.

(6) $m = m(z_i) = \left(m(Z_1), \dots, m\left(Z_p\right)\right)$ ve Γ=$diag(s(Z_1)^2, \dots, s\left(Z_p\right)^2)$ olmak üzere OGK konum tahmin edicisi

$$\hat{\mu}_{OGK} = Am \tag{4.9}$$

pozitif tanımlı matris olan OGK ölçek tahmin edicisi

$$\hat{\sum}_{OGK} = [A\Gamma A'] \tag{4.10}$$

dir (Todorov ve Filzmoser, 2009; Hubert vd, 2012).

4.6. Parçalı Uyarlanabilir Hesaplama Yönünden Etkin Aykırı Gözlem Belirleyicisi Tahmin Edicisi

Parçalı Uyarlanabilir Hesaplama Yönünden Etkin Aykırı Gözlem Belirleyicisi (BACON), Billor vd (2000) tarafından önerilmiştir. Hadi vd (2009), dayanıklı konum ve ölçek tahmin edicilerini elde etmek için, yinelemeli tahminlerle çalışmayı önermişler ve yöntemlerini Hadi (1992, 1994)'nin çalışmasına dayandırmışlardır. Hadi'nin yöntemine göre, veri seti içerisinde aykırı değer içermeyen küçük bir alt küme oluşturulduktan sonra, bu temiz alt kümenin aykırı değerler olarak gösterilmeyen tüm veri değerlerini içerene kadar yavaş yavaş büyümesine izin verilir (Billor vd, 2000).

Bir veri matrisi X'in çok değişkenli konum ve ölçek BACON tahmin edicisi aşağıdaki gibi hesaplanır:

(1) Çok değişkenli nxp boyutlarına sahip veri matrisi X ve ilk temel alt kümeye dahil edilecek gözlem sayısı m olsun (Burada $m > p$ olmalıdır).

(2) $i = 1, \dots, n$ için (3.1) eşitliği yardımıyla Mahalanobis uzaklıkları (MD_i) hesaplansın.

(3) $m = cp$ değerleri bir önceki adımda hesaplanan en küçük MD_i değerleriyle tanımlansın. Bunlar potansiyel bir temel alt küme olarak adlandırılsın. c keyfi olarak belirlenen bir tamsayıdır.

(4) $\hat{\mu}_b$ ve $\widehat{\sum}_b$ potansiyel temel alt kümenin konum ve ölçek tahmin edicileri olsun. Eğer $\widehat{\sum}_b$ tam ranka ulaşmazsa tam ranka ulaşana kadar en küçük uzaklıklara sahip gözlemler eklenerek temel alt küme ve alt kümeyi tam ranklı yapmak için eklenen gözlem sayısı ile m arttırılsın.

Bu yöntemin avantajı afin eş değişim özelliğine sahip olması ve düşük hesaplama maliyetidir. Hesaplama açısından kolay olan bu yöntem diğer yöntemler için uzun hesaplama süresi gerektirebilen yüzbinlerce veri kümesine kolayca uygulanabilmektedir (Billor vd, 2000).

5. UZUN KUYRUKLU SİMETRİK DAĞILIM

Uzun kuyruklu simetrik (LTS) dağılım, sıklıkla kullanılan normal olmayan bir simetrik dağılımdır (Tiku vd, 2001; Kasap vd, 2016). Dayanıklılık araştırmalarına verilen önemin etkisi uzun kuyruklu simetrik dağılımların normal dağılımın yerini doldurduğunda görülmektedir. İstatistik alanında yapılan birçok analiz öncesinde verilerin normal dağıldığı varsayılmaktadır. Bir örneklemdeki en küçük veya en büyük birkaç gözlemin varlığı nedeniyle böyle bir dağılım varsayımının sağlanamayacağı açıktır (Shu 1978).

$LTS(\ t, \sigma^2)$ dağılımının olasılık yoğunluk fonksiyonu aşağıdaki gibidir.

$$\boldsymbol{f(x) = \frac{1}{\sqrt{k}\beta(1/2, t-1/2)\sigma}\left\{1 + \frac{(x-\mu)^2}{k\sigma^2}\right\}^{-t}, k = 2t-3; t \geq 2, -\infty < x < \infty} \qquad \textbf{(5.1)}$$

Burada t şekil parametresidir ve $\beta(.,.)$ Beta fonksiyonudur. LTS dağılımının ortalama ve varyansı $E(X) = \mu$ ve $Var(X) = \sigma^2$ şeklindedir. $y = \sqrt{v/k}\,(X/\sigma)$ şeklinde tanımlanan rasgele değişken, $v = 2t - 1$ serbestlik dereceli Student t dağılımına sahiptir. LTS dağılımlar ailesi, içerisinde aykırı değer bulunan örneklemleri modellemek için uygundur (Tiku ve Akkaya 2004).

LTS'nin basıklığı $\beta_2 = 3(p - 3/2)/(p - 5/2)$ ile tanımlanır ve basıklık her zaman 3'ten daha büyük ya da 3'e eşittir. Şekil parametresi t'nin farklı değerleri için dağılımın şekli de farklı olur. LTS dağılımında $t = 1$ olduğunda Cauchy dağılımı, $t = \infty$ olduğu durumda ise normal dağılıma dönüşür (Kasap, 2011). Şekil parametresinin birkaç farklı değeri için LTS dağılımının basıklık değeri aşağıdaki gibidir (Tiku, 2004; Kasap vd, 2016):

Çizelge 5.1. Şekil parametresi t'nin farklı değerleri için LTS dağılımının basıklık değerleri

t	2.5	3.5	5	10	∞
β_2	∞	9	4.2	3.4	3

Bu çalışmada LTS dağılımının kullanılmasının nedeni normal dağılımın makul bir alternatifi olması ve aykırı değerlerin modellenmesinde sıklıkla kullanılan bir

dağılım olmasıdır. Daha önce yapılan çalışmalar incelendiğinde LTS dağılımının daha çok dayanıklılık çalışmalarında parametre tahmini için kullanıldığı görülmektedir (Tiku vd, 2001; Altın ve Şenoğlu, 2008; Kasap vd, 2016). Bu çalışmada ise LTS dağılımı aykırı değerler içeren veriyi modellemede uygun bir dağılım olduğundan kullanılmıştır.

6. UYGULAMA

Tezin uygulama bölümünde 3. bölümde tanıtılan klasik Mahalanobis uzaklığı (MD) ve 4. bölümde tanıtılan Hızlı-en küçük kovaryans determinant (FMCD), En küçük hacimli elipsoit (MVE), M-tahmin edicisi (MEST), Stahel-Donoho (SDE), Dikey Gnanadesikan-Kettenring (OGK) ve BACON yöntemleri ile elde edilen dayanıklı konum ve ölçek tahmin edicileri kullanılmıştır. Mahalanobis uzaklıklıkları hesaplanırken klasik tahmin edicilerle elde edilen MD'nin yanısıra dayanıklı tahmin yöntemleri ile elde edilen konum ve ölçek tahmin edicileri (3.1) eşitliğinde verilen Mahalanobis uzaklığı formülünde yerine yazılarak dayanıklı uzaklıklar da elde edilmiştir. Klasik ve dayanıklı tahmin ediciler için elde edilen Mahalanobis uzaklıkları, normal dağılım ve normal dağılımın makul bir alternatifi olan uzun kuyruklu simetrik dağılım kullanılarak Monte-Carlo benzetim çalışması ile ayrı ayrı incelenmiştir.

6.1. Monte-Carlo Benzetim Çalışması

Monte-Carlo benzetim çalışması yoluyla MD ve yukarıda bahsedilen dayanıklı yöntemlerin performansları başlık (6.1.1) de çok değişkenli normal dağılım için, başlık (6.1.2) de uzun kuyruklu simetrik dağılım için incelenmiştir. Yöntemlerin aykırı değerleri belirlemedeki performansları

$$Başarı\ Oranı = \frac{Doğru\ Tespit\ Edilen\ Aykırı\ Değer\ Sayısı}{Veri\ Setine\ Eklenen\ Aykırı\ Değer\ Sayısı} * 100 \qquad \textbf{(6.1)}$$

formülü ile elde edilen başarı oranları ile değerlendirilmiştir (Chen vd, 2010). Başarı oranı, yöntemlerin veri setine eklenen aykırı değerleri bulma oranı olarak ifade edilmektedir. Ayrıca yöntemler, bazen aykırı değer olarak ifade edilmeyen verileri de aykırı değer olarak belirleyebilir. Bu çalışmada aykırı değer olmayan gözlemlerin aykırı değer olarak belirlenmesi yanlış tespit sayısı olarak ifade edilmiş ve yanlış tespit sayıları da ayrı tablolar halinde verilmiştir. Yanlış tespit sayılarını belirlemek için m iterasyon düşünülsün ve m_i, $i-inci$ iterasyondaki aykırı değer olarak tespit edilen aykırı değer olmayan gözlem sayısı olsun. Bu durumda, yanlış tespit sayısı $Max(m_i)$, yani m iterasyonda elde edilen sayıların en büyüğü olarak alınır. Bu çalışmada $m = 1000$ olarak alınmıştır ve $i = 1, \dots, m$'dir (Sajesh ve Srinivasan,

2013). Bir başka deyişle, başarı oranı veri setine aykırı değer olarak eklenen gözlemleri bulma oranını ifade ederken, yanlış tespit sayısı ise yöntemlerin, aykırı değer olarak eklenen veri yerine aykırı değer olmayan veriyi aykırı değer olarak tespit etme sayısını gösterir. Yöntemlerin aykırı değerleri belirlemesinde programların sonuç verme süreleri de büyük n değerleri için saniye cinsinden değerlendirilmiştir.

Aykırı değer tespitinde kullanılan yöntemler karşılaştırılırken Monte-Carlo benzetim çalışmasında her bir durum için iterasyon 1000 kez tekrarlanmıştır.

Benzetim çalışması R-3.5.1 programında "rrcov", "robustX", "LaplacesDemon" ve "tictoc" paketleri yardımıyla gerçekleştirilmiştir. R-Paket programında Mahalanobis uzaklıkları "mahalanobis" komutu ile dayanıklı yöntemlerin konum ve ölçek tahminleri ise FMCD yöntemi için "CovMcd" komutu, MVE yöntemi için "CovMve" komutu, MEST yöntemi için "CovMest" komutu, SDE yöntemi için "CovSde" komutu, OGK yöntemi için "CovOgk" komutu ve BACON yöntemi için "BACON" komutu ile hesaplanmıştır (Todorov, 2009, Stahel ve Maechler, 2019). Benzetim çalışmasında kullanılan diğer komutlar ve yazılan programlar Ek bölümünde detaylı olarak yer verilmiştir.

6.1.1. Çok değişkenli normal dağılım için Monte-Carlo benzetim çalışması

Bu bölümde, Monte-Carlo benzetim çalışması yoluyla örneklem genişliği n= 50, 100, 200, 1000 olarak belirlenen ve değişken sayısı p=5, 10, 20 olan çok değişkenli normal dağılımdan 0 ortalamalı ve 1 varyanslı veri seti üretilmiştir. Bu veri setleri belli oranlarda kirletilerek aykırı değerler oluşturulmuştur. Veri setine eklenen aykırı değerlerin oranı θ-kirlilik seviyesi ile gösterilmiştir. Bu çalışmada θ - kirlilik seviyesi 0.1 ve 0.2 olarak belirlenmiştir. θ kirlilik seviyesinin 0.1 olması veri setinin % 10 unun kirletilerek aykırı değer oluşturulması anlamına gelmektedir. Aykırı değerler tespit edilirken $\sqrt{\chi^2_{p;0.975}}$ kritik değerinden büyük olan gözlemler aykırı gözlem olarak adlandırılmıştır. Bu kritik değer değişken sayısı p=5, 10, 20 için ayrı ayrı hesaplanmıştır.

θ kirlilik seviyesi için $n(1-\theta)$ gözlem p-değişkenli $N(\mathbf{0}, \boldsymbol{I_p})$ dağılımından üretilmiştir ve burada $\boldsymbol{I_p}$, pxp boyutlu birim matrisi; $\mathbf{0} = [\mathbf{0}, \mathbf{0}, \dots, \mathbf{0}]$ ise $1xp$ boyutlu ortalama vektörünü ifade etmektedir. Veri setine eklenen $n\theta$ aykırı değerler ise

$N(\mathbf{0}, \xi^2 \boldsymbol{I_p})$ dağılımından üretilmiştir. Burada, farklı varyans değerlerine sahip olan aykırı değerler üretebilmek için ξ=0.5, 0.75, 2, 5 ve 10 olarak seçilmiştir. Buna göre, θ kirlilik seviyesi için kirletilmiş model,

$$n(1-\theta)\, N(\mathbf{0}, \boldsymbol{I_p}) + n\theta\, N(\mathbf{0}, \xi^2 \boldsymbol{I_p}) \qquad \textbf{(6.2)}$$

şeklinde ifade edilebilir (Pena ve Prieto, 2001; Sajesh ve Srinivasan, 2013).

Çizelge 6.1. Çok değişkenli normal dağılıma sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.90)\ N(\mathbf{0}, \boldsymbol{I_p}) + n(0.10)\ N(\mathbf{0}, \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.000	0.180	0.280	0.080	0.120	0.040	0.000	0.000	0.940	0.940	1.500	0.080	0.000	0.000	0.000	1.440	2.080	1.960	0.000	0.140
	100	0.000	0.010	0.010	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.020	0.070	0.120	0.000	0.000
	200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1000	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.75	50	0.040	2.040	2.220	2.140	1.920	1.300	0.180	0.020	8.080	8.200	11.220	4.340	1.080	0.020	0.000	9.860	11.620	10.940	0.620	7.900
	100	0.120	0.360	0.290	0.3200	0.410	0.920	0.100	0.010	0.720	0.770	0.610	0.990	0.460	0.010	0.000	3.250	3.460	4.680	0.230	0.000
	200	0.095	0.130	0.165	0.185	0.205	0.535	0.105	0.025	0.070	0.075	0.115	0.140	0.270	0.000	0.000	0.110	0.120	0.085	0.110	0.000
	1000	0.062	0.055	0.056	0.072	0.074	0.337	0.052	0.013	0.020	0.022	0.018	0.026	0.103	0.010	0.003	0.003	0.003	0.005	0.021	0.000
2	50	43.92	64.18	66.900	68.88	70.940	73.900	53.820	58.220	89.420	87.340	89.880	93.020	91.000	65.120	48.760	88.860	85.240	85.460	98.240	98.460
	100	48.13	61.120	61.920	65.930	66.140	73.430	58.870	69.300	85.890	86.470	89.350	90.720	92.260	80.140	87.170	98.700	94.860	96.810	98.890	93.030
	200	51.05	60.720	60.560	64.065	64.230	73.825	60.840	72.860	85.275	84.975	87.680	88.525	91.960	84.000	92.345	98.630	97.840	98.795	99.320	97.130
	1000	52.147	60.436	59.813	62.780	62.843	72.381	60.986	75.191	84.749	84.334	86.266	86.241	91.304	85.842	94.012	98.318	98.150	98.517	99.220	98.441
5	50	91.800	98.860	99.060	99.100	99.260	99.560	98.260	99.480	99.980	99.940	99.940	99.980	99.980	99.940	99.980	100	99.840	99.840	100	100
	100	90.510	98.760	98.820	99.150	99.050	99.500	98.840	99.390	99.970	99.990	99.990	99.990	100	99.980	100	100	100	100	100	100
	200	89.405	98.710	98.875	99.020	99.105	99.345	99.040	99.000	99.995	99.970	99.975	99.990	100	99.995	99.990	100	100	100	100	100
	1000	88.569	98.840	98.791	98.927	98.914	99.386	99.126	98.664	99.994	99.987	99.991	99.986	99.997	99.993	99.989	100	100	100	100	100
10	50	98.380	99.960	99.920	99.980	99.980	99.980	99.880	100	100	100	100	100	100	100	100	100	100	100	100	100
	100	95.630	99.970	99.980	99.970	99.960	99.970	99.920	99.940	100	100	100	100	100	100	100	100	100	100	100	100
	200	94.155	99.975	99.970	99.965	99.970	99.980	99.980	99.740	100	100	100	100	100	100	100	100	100	100	100	100
	1000	92.742	99.947	99.963	99.966	99.971	99.980	99.969	99.521	100	100	100	100	100	100	99.998	100	100	100	100	100

Çizelge 6.1'de çok değişkenli normal dağılıma sahip bir veri seti için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.1 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ'nin büyümesinin, eklenen aykırı değerlerin varyansının da büyümesi anlamına geldiği açıktır. Dolayısıyla varyansı büyüyen aykırı değerlerin tespit edilme oranlarının artması da beklenen bir sonuçtur. ξ=2 olduğu durumda değişken sayısı p =5, 10 ve 20 iken en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 ve 10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi, birbirlerine yakın ve oldukça iyi başarı oranlarına sahiptir. Değişken sayısının yüksek olduğu p=10 ve 20 durumlarında ise klasik yöntem de dahil olmak üzere tüm yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin çok uzun olması sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.2. Çok değişkenli normal dağılıma sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.90)\ N(\mathbf{0}, \boldsymbol{I_p}) + n(0.10)\ N(\mathbf{0}, \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	4	18	19	18	19	17	19	3	20	19	20	23	15	16	2	15	15	15	18	25
	100	8	17	19	20	19	26	20	7	31	27	28	28	26	10	7	39	37	39	25	8
	200	13	19	19	24	23	40	15	13	29	28	27	30	43	15	13	55	40	40	42	15
	1000	50	55	53	63	68	132	51	49	65	57	65	54	129	56	60	85	77	66	145	58
0.75	50	4	17	19	20	16	14	14	3	20	19	20	22	14	12	2	15	15	15	15	25
	100	8	17	19	15	20	23	8	7	24	23	25	25	24	9	5	37	35	37	25	7
	200	11	15	17	21	19	37	14	12	21	22	25	26	35	12	11	39	37	40	40	13
	1000	41	45	47	53	54	105	44	44	51	47	61	60	102	45	47	56	57	66	119	50
2	50	3	15	16	17	12	12	17	2	17	16	17	18	11	9	1	13	14	13	11	21
	100	4	10	11	12	12	16	6	3	23	20	18	20	18	7	3	27	26	27	17	5
	200	5	11	9	14	14	26	10	4	12	12	15	16	26	10	4	23	24	28	28	9
	1000	13	24	22	31	28	66	24	12	25	27	33	34	69	28	7	30	31	38	79	31
5	50	1	15	14	16	12	11	18	2	15	14	15	17	11	9	1	10	11	11	12	20
	100	1	10	16	11	12	19	8	1	21	19	19	17	18	9	1	28	28	28	18	4
	200	1	10	12	12	12	27	12	1	15	15	15	14	27	10	1	25	28	24	30	9
	1000	0	26	21	31	29	76	33	0	25	26	34	31	66	36	0	29	28	37	70	37
10	50	1	13	14	16	11	11	12	1	15	14	15	17	11	9	1	10	10	10	12	20
	100	1	12	11	12	10	19	10	1	20	22	20	17	16	6	1	29	28	27	17	5
	200	0	15	10	12	13	27	10	0	13	12	15	14	24	10	0	29	22	25	26	9
	1000	0	22	21	29	36	70	35	0	26	23	34	32	68	35	0	27	29	38	75	38

Çizelge 6.2'de çok değişkenli normal dağılıma sahip bir veri seti için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.2 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. p=5 olduğu durumda BACON yöntemi FMCD, MVE, MEST, SDE ve OGK yöntemleri ile benzer sonuçları verirken, p=10 ve 20 için FMCD, MVE, MEST, SDE ve OGK yöntemlerine göre daha az yanlış tespit sayısına sahiptir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009).

Çizelge 6.3. Çok değişkenli normal dağılıma sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.80)\ N(\mathbf{0}, \boldsymbol{I_p}) + n(0.20)\ N(\mathbf{0}, \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.000	0.300	0.280	0.290	0.170	0.040	0.050	0.000	1.380	1.100	1.950	0.190	0.010	0.020	0.000	1.260	1.850	2.780	0.000	0.280
	100	0.000	0.005	0.010	0.005	0.005	0.005	0.005	0.000	0.035	0.020	0.005	0.010	0.000	0.000	0.000	0.100	0.130	0.080	0.000	0.000
	200	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.75	50	0.140	2.550	3.340	3.160	2.700	1.870	0.210	0.050	9.570	9.110	12.300	5.180	1.420	0.100	0.000	10.660	12.470	12.810	1.010	10.210
	100	0.135	0.435	0.635	0.545	0.685	1.295	0.170	0.040	0.890	1.075	0.865	1.120	0.690	0.030	0.005	4.115	4.500	5.240	0.325	0.000
	200	0.058	0.190	0.228	0.235	0.295	0.800	0.088	0.038	0.125	0.205	0.148	0.235	0.443	0.008	0.018	0.198	0.235	0.208	0.145	0.008
	1000	0.087	0.102	0.112	0.142	0.136	0.534	0.100	0.027	0.036	0.030	0.037	0.055	0.190	0.029	0.004	0.011	0.008	0.007	0.051	0.006
2	50	34.120	57.810	60.320	64.310	67.060	70.460	48.690	43.700	86.650	83.770	85.590	90.600	88.600	60.400	34.630	80.900	78.340	78.410	97.610	98.660
	100	37.095	54.830	55.455	60.650	62.345	70.560	54.045	54.970	83.375	82.690	85.520	88.335	90.335	76.855	74.885	98.215	91.820	94.960	98.710	91.020
	200	38.838	54.125	54.080	58.903	59.533	70.743	55.338	58.370	81.550	80.853	84.710	85.013	90.238	80.918	81.470	98.153	96.738	98.118	99.123	96.700
	1000	39.705	53.394	52.811	57.374	57.553	69.910	52.292	60.525	80.920	80.052	82.842	82.906	90.242	82.172	84.224	97.719	97.420	97.963	99.146	98.253
5	50	74.000	98.510	98.480	98.850	98.910	99.240	98.440	94.870	99.960	99.800	99.910	99.980	99.980	99.920	99.710	98.910	99.720	94.330	100	100
	100	72.725	98.380	98.465	98.760	98.770	99.315	99.080	93.515	99.990	99.990	100	99.990	100	99.990	99.855	100	100	100	100	100
	200	71.350	98.380	98.538	98.580	98.730	99.403	99.078	92.015	99.980	99.988	99.990	99.988	99.993	99.998	99.620	100	100	100	100	100
	1000	70.534	98.460	98.439	98.643	98.706	99.286	99.139	91.073	99.985	99.985	99.990	99.988	99.995	99.995	99.342	100	100	100	100	100
10	50	82.480	99.930	99.930	99.970	99.960	99.950	100	99.020	100	100	100	100	100	100	100	100	100	95.860	100	100
	100	78.390	99.945	99.890	99.975	99.985	99.980	100	96.785	100	100	100	100	100	100	100	100	100	100	100	100
	200	76.905	99.950	99.930	99.948	99.963	99.960	100	95.333	100	100	100	100	100	100	99.903	100	100	100	100	100
	1000	75.460	99.950	99.945	99.946	99.951	99.974	100	93.938	100	100	100	100	100	100	99.708	100	100	100	100	100

Çizelge 6.3'te çok değişkenli normal dağılıma sahip bir veri seti için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.3 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ=2 olduğu durumda değişken sayısı p=5 iken ve n=50, 100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=2 olduğu durumda değişken sayısı p=10 ve 20 iken ve n=100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 ve 10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranları vermektedir. Yine ξ=5 ve 10 olduğu ve değişken sayısının p=10 durumda dayanıklı yöntemlerin yüksek başarı oranına sahip olduğu, p=20 olduğu durumda ise klasik yöntem de dahil olmak üzere tüm yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir.

Çizelge 6.1 ve Çizelge 6.3 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla p=5 iken ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranları artarken, ξ=2, 5 ve 10 için tüm yöntemlerin başarı oranları azalmıştır. p=10 ve 20 iken θ kirlilik seviyesinin artması ξ=2 olduğu durumda tüm yöntemlerin başarı oranlarını azaltmıştır. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.4. Çok değişkenli normal dağılıma sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.80)\ N(\mathbf{0}, \boldsymbol{I_p}) + n(0.20)\ N(\mathbf{0}, \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	6	20	18	17	18	17	27	5	20	19	20	23	17	21	3	15	15	15	20	25
	100	9	22	22	24	24	31	16	9	38	32	28	35	34	42	9	40	38	39	31	36
	200	15	26	25	30	32	50	19	16	43	34	37	40	50	20	20	83	57	49	54	25
	1000	61	74	78	95	88	173	62	69	100	86	85	87	187	78	88	179	110	99	209	94
0.75	50	4	16	17	20	16	17	16	4	20	19	20	22	16	17	2	15	15	15	19	25
	100	8	17	18	17	18	21	10	7	30	24	26	25	26	11	6	35	37	37	26	9
	200	12	17	20	28	23	37	14	14	25	23	30	27	41	14	15	50	41	40	44	15
	1000	47	53	52	59	61	104	47	51	61	55	72	77	121	52	56	84	69	77	153	56
2	50	2	14	12	13	12	8	10	1	13	12	13	13	9	9	1	10	10	11	10	16
	100	2	10	8	8	8	12	5	1	16	12	13	14	14	5	1	19	18	16	12	6
	200	3	7	7	9	11	18	8	3	9	10	13	12	20	8	2	18	15	17	22	8
	1000	5	12	13	19	19	50	12	3	16	13	19	17	54	18	2	21	22	23	56	27
5	50	0	11	10	12	8	9	14	1	10	9	10	11	8	9	1	6	11	7	9	15
	100	0	6	9	9	7	14	8	0	13	13	11	9	13	6	0	19	27	19	12	6
	200	0	6	7	7	8	23	11	0	11	11	10	9	19	9	0	19	16	17	21	9
	1000	0	13	11	16	15	56	35	0	16	15	21	19	56	32	0	21	19	26	64	31
10	50	0	10	8	11	8	10	14	0	10	9	10	10	8	8	1	6	10	6	9	15
	100	0	9	7	7	7	13	7	0	14	12	14	10	13	7	0	18	27	18	14	4
	200	0	7	7	10	10	20	9	0	9	10	10	8	21	10	0	17	21	17	19	8
	1000	0	11	12	16	18	58	33	0	15	15	20	18	55	30	0	20	29	23	57	35

Çizelge 6.4'de çok değişkenli normal dağılıma sahip bir veri seti için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5,\ 0.75,\ 2,\ 5,\ 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.4 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının genel olarak azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009). Çizelge 6.2 ile Çizelge 6.4 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla genel olarak tüm yöntemlerin yanış tespit sayılarında azalma meydana gelmiştir.

6.1.2. Uzun kuyruklu simetrik dağılım için Monte-Carlo benzetim çalışması

Bu bölümde, Monte-Carlo benzetim çalışması yoluyla örneklem genişliği n= 50, 100, 200, 1000 olarak belirlenen ve değişken sayısı p=5, 10, 20 olan uzun kuyruklu simetrik dağılımdan farklı şekil parametreli ve farklı varyanslı bir veri seti üretilmiştir. Bu veri setleri belli oranlarda kirletilerek aykırı değerler oluşturulmuştur. Veri setine eklenen aykırı değerlerin oranı θ -kirlilik seviyesi ile gösterilmiştir. Bu çalışmada θ- kirlilik seviyesi 0.1 ve 0.2 olarak belirlenmiştir. θ kirlilik seviyesinin 0.1 olması veri setinin % 10 unun kirletilerek aykırı değer oluşturulması anlamına gelmektedir. LTS dağılımı için şekil parametresi $\mathrm{t} = 3.5$ ve 5 olarak alınmıştır. Aykırı değerler tespit edilirken $\sqrt{\chi^2_{p;0.975}}$ kritik değerinden büyük olan gözlemler aykırı gözlem olarak adlandırılmıştır. Bu kritik değer değişken sayısı p=5, 10, 20 için ayrı ayrı hesaplanmıştır.

θ kirlilik seviyesi için $n(1-\theta)$ gözlem p-değişkenli $LTS(\boldsymbol{t}, \boldsymbol{I_p})$ dağılımından üretilmiştir. Burada $\boldsymbol{I_p}$, pxp boyutlu birim matrisi ifade etmektedir. Veri setine

eklenen $n\theta$ sayıda aykırı değerler ise $LTS(\boldsymbol{t}, \xi^2 \boldsymbol{I_p})$ dağılımından üretilmiştir. Burada, farklı varyans değerlerine sahip olan aykırı değerler üretebilmek için ξ=0.5, 0.75, 2, 5 ve 10 olarak seçilmiştir. Buna göre, θ kirlilik seviyesi için kirletilmiş model,

$$n(1-\theta)\ LTS(\mathbf{0}, \boldsymbol{I_p}) + n\theta\ LTS(\mathbf{0}, \xi^2 \boldsymbol{I_p}) \qquad \textbf{(6.3)}$$

şeklinde ifade edilebilir.

Çizelge 6.5. Şekil parametresi $t = 3.5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.90)\ LTS(t = 3.5\ , \boldsymbol{I_p})$ + $n(0.10)\ LTS(t = 3.5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.060	0.740	0.640	0.940	0.820	0.460	0.200	0.060	1.820	2.060	2.220	0.660	0.340	0.080	0.000	1.860	2.460	2.760	0.080	0.780
	100	0.100	0.260	0.320	0.400	0.290	0.530	0.190	0.070	0.320	0.140	0.120	0.250	0.230	0.090	0.010	0.300	0.350	0.530	0.100	0.030
	200	0.110	0.190	0.145	0.250	0.250	0.415	0.110	0.070	0.125	0.080	0.125	0.135	0.210	0.095	0.015	0.155	0.065	0.035	0.125	0.045
	1000	0.094	0.168	0.163	0.183	0.186	0.385	0.150	0.059	0.108	0.093	0.078	0.105	0.168	0.077	0.028	0.054	0.044	0.042	0.068	0.035
0.75	50	0.820	3.980	4.500	5.220	4.900	4.560	1.900	0.540	9.920	10.600	14.180	7.760	4.240	0.600	0.080	10.780	11.940	12.080	3.800	12.520
	100	0.920	2.240	2.430	2.680	2.980	4.360	1.390	0.870	2.620	3.210	2.840	4.110	3.470	0.960	0.340	6.760	5.500	6.740	2.620	0.440
	200	1.130	1.715	1.845	2.085	2.350	4.260	1.640	0.830	1.850	1.855	2.005	2.220	3.250	1.100	0.650	2.600	1.985	1.805	2.230	0.695
	1000	1.142	1.707	1.582	1.927	1.933	3.535	1.484	0.927	1.405	1.249	1.457	1.447	2.712	1.173	0.662	1.086	0.918	1.000	1.875	0.796
2	50	40.240	60.900	62.680	66.760	69.040	71.840	53.520	53.080	87.460	84.500	87.500	90.040	87.480	63.420	46.320	86.200	82.980	83.380	96.540	98.460
	100	43.050	59.320	59.630	63.820	63.870	71.950	56.020	62.020	84.110	83.150	85.590	88.120	88.850	75.460	81.220	97.400	92.150	94.720	97.820	88.750
	200	43.810	58.135	58.305	61.705	61.220	71.985	57.995	64.570	81.790	80.805	83.280	84.200	89.135	78.910	85.785	97.610	96.110	96.930	98.290	94.530
	1000	44.400	57.226	56.549	60.770	60.579	71.464	56.143	65.596	80.252	79.357	82.030	81.805	89.122	79.951	87.449	96.527	96.044	96.600	98.408	96.121
5	50	88.460	98.180	98.580	98.760	98.940	99.020	98.140	98.580	99.980	99.920	99.820	100	100	99.860	99.960	99.920	99.720	99.540	100	100
	100	86.760	98.160	98.110	98.780	98.790	99.060	98.380	98.480	99.970	99.990	99.940	100	99.990	99.950	99.990	100	100	100	100	100
	200	84.575	98.285	98.225	98.500	98.675	99.260	98.580	97.850	99.980	99.970	99.960	99.960	99.995	99.975	99.980	100	100	100	100	100
	1000	82.584	98.288	98.242	98.519	98.459	99.219	98.529	96.878	99.981	99.967	99.978	99.971	99.989	99.976	99.920	100	100	100	100	100
10	50	96.820	99.940	99.960	99.980	99.900	99.980	99.920	100	100	100	100	100	100	100	100	100	100	100	100	100
	100	93.510	99.960	99.950	99.960	99.930	99.980	99.960	99.860	100	100	100	100	100	100	100	100	100	100	100	100
	200	91.115	99.920	99.940	99.925	99.970	99.960	99.955	99.430	100	100	100	100	100	100	100	100	100	100	100	100
	1000	88.318	99.921	99.935	99.938	99.944	99.965	99.951	98.639	100	100	100	100	100	100	99.990	100	100	100	100	100

Çizelge 6.5'te şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.5 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ'nin büyümesinin, eklenen aykırı değerlerin varyansının da büyümesi anlamına geldiği açıktır. Dolayısıyla varyansı büyüyen aykırı değerlerin tespit edilme oranlarının artması da beklenen bir sonuçtur. ξ=2 olduğu durumda değişken sayısı p=5 iken ve n=50, 100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=2 olduğu durumda değişken sayısı p=10 ve 20 iken ve n=100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 ve 10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranları vermektedir. Değişken sayısının yüksek olduğu p=10 ve 20 durumlarında ise klasik yöntem de dahil olmak üzere tüm yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir.

Çizelge 6.1 ve Çizelge 6.5 karşılaştırıldığında, Çizelge 6.1'de aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli normal dağılım yerine Çizelge 6.5'te şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması ile p=5 olduğunda ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranlarında artış gözlenmiştir. Varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yötemlerde başarı oranının artması, büyük ξ=2, 5 ve 10 değerlerine göre aykırı değerlerin bulunmasının zor olduğu durumda bile şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının normal dağılıma göre aykırı değerlerin tespitinde daha kullanışlı olduğunu göstermektedir. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin uzun sürmesi sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.6. Şekil parametresi $t = 3.5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.90)\ LTS(t = 3.5\ , \boldsymbol{I_p}) + n(0.10)\ LTS(t = 3.5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	6	20	18	19	19	18	27	6	20	19	20	23	18	19	4	15	15	15	19	25
	100	11	24	24	24	27	30	20	11	38	31	31	35	38	17	11	40	37	38	35	17
	200	20	33	30	35	34	53	27	23	42	42	41	43	58	31	25	69	55	54	59	34
	1000	74	108	113	122	122	207	99	92	156	126	138	140	221	112	105	182	148	149	240	132
0.75	50	5	18	17	18	18	17	18	6	20	19	20	23	18	17	3	15	15	15	18	25
	100	10	21	19	24	25	28	19	12	30	25	28	33	29	18	11	38	36	40	31	12
	200	17	27	28	35	34	51	26	18	34	34	42	40	53	29	21	60	49	49	55	27
	1000	70	93	92	107	115	177	87	84	113	107	125	121	194	102	91	141	136	143	221	114
2	50	5	15	17	18	14	15	21	4	17	17	17	18	15	15	3	13	13	15	15	21
	100	7	21	16	20	17	23	13	7	28	21	22	27	23	15	6	29	27	28	26	13
	200	11	21	20	25	25	38	22	12	27	24	31	31	38	24	12	44	34	39	41	22
	1000	38	67	65	82	81	138	61	39	80	74	85	90	155	72	35	101	95	99	166	87
5	50	3	14	14	15	14	13	19	3	15	14	15	18	14	18	2	11	11	11	13	20
	100	3	16	15	16	17	23	17	3	25	27	22	23	25	18	4	29	28	27	23	13
	200	3	20	22	26	23	37	23	3	29	25	29	30	40	23	4	44	34	38	44	21
	1000	6	67	62	77	78	140	76	5	82	74	92	90	155	88	4	100	91	103	165	93
10	50	2	17	14	17	13	13	16	3	15	14	15	18	14	13	2	10	11	10	14	20
	100	1	18	15	17	15	23	15	2	23	22	23	22	23	24	3	30	28	27	25	10
	200	2	21	23	24	26	37	22	1	28	27	28	29	41	27	2	46	34	38	41	24
	1000	1	65	61	75	76	145	77	1	83	76	86	82	149	87	1	96	89	104	67	93

Çizelge 6.6'da şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.6 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının genel olarak azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını p=20 ve ξ=10 olduğu durum hariç OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009).

Çizelge 6.2 ve Çizelge 6.6 karşılaştırıldığında, Çizelge 6.2'de aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli normal dağılım yerine Çizelge 6.6'da şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması ile tüm yöntemler için yanlış tespit sayısının arttığı görülmüştür.

Çizelge 6.7. Şekil parametresi $t = 3.5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.80)\ LTS(t = 3.5\ , \boldsymbol{I_p})$ + $n(0.20)\ LTS(t = 3.5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.100	1.250	1.130	1.140	1.250	0.890	0.480	0.090	2.560	1.830	2.650	0.960	0.500	0.210	0.000	1.790	2.600	2.730	0.260	1.270
	100	0.120	0.440	0.390	0.435	0.465	0.720	0.175	0.055	0.600	0.365	0.350	0.410	0.415	0.125	0.020	0.505	0.535	0.605	0.195	0.055
	200	0.128	0.313	0.290	0.350	0.295	0.773	0.178	0.075	0.225	0.188	0.165	0.225	0.338	0.113	0.033	0.250	0.105	0.113	0.150	0.053
	1000	0.144	0.231	0.227	0.253	0.286	0.648	0.211	0.084	0.168	0.128	0.137	0.130	0.293	0.110	0.037	0.095	0.059	0.048	0.120	0.060
0.75	50	0.900	5.090	5.160	5.900	6.350	5.570	2.190	0.730	11.700	12.320	14.330	9.870	5.340	0.940	0.070	12.050	13.320	13.360	4.190	14.440
	100	1.245	2.460	2.800	3.140	3.415	5.245	1.980	0.915	3.775	3.940	3.795	4.725	4.530	1.185	0.465	7.745	7.390	7.715	3.475	0.680
	200	1.288	2.250	2.308	2.488	2.733	4.883	1.853	1.020	2.203	2.203	2.400	2.680	4.115	1.440	0.680	3.028	2.315	2.355	3.035	0.833
	1000	13.140	1.904	1.789	2.236	2.266	4.290	1.744	1.095	1.667	1.545	1.850	1.848	3.503	1.399	0.814	1.391	1.197	1.291	2.444	1.014
2	50	30.710	57.370	58.600	61.860	64.060	66.690	52.610	41.360	83.080	79.960	82.380	88.330	84.090	60.880	35.180	79.110	75.900	75.760	95.010	97.800
	100	33.255	54.360	54.170	57.900	59.355	68.265	52.530	49.360	80.825	78.120	80.505	84.470	85.505	71.930	67.440	96.780	89.445	91.410	96.680	86.160
	200	33.755	52.753	51.465	56.133	57.008	68.270	52.713	51.090	78.626	76.310	78.983	80.915	86.780	75.875	72.813	96.825	93.940	95.435	97.505	93.470
	1000	34.189	51.735	50.462	55.389	55.525	68.219	49.157	51.968	76.508	74.626	77.382	77.543	87.080	75.444	74.883	95.463	94.441	95.150	97.893	95.481
5	50	69.540	97.850	97.590	98.210	98.510	98.770	97.770	92.010	99.940	99.460	99.690	99.990	99.980	99.870	99.200	98.720	93.420	93.800	100	100
	100	66.025	97.560	97.730	98.295	98.140	98.970	98.315	89.265	99.985	99.905	99.965	99.970	99.995	99.935	99.535	100	100	100	100	100
	200	63.673	97.673	97.485	98.065	97.998	98.938	98.510	86.890	99.968	99.905	99.968	99.968	99.975	99.968	98.955	100	100	100	100	100
	1000	61.965	97.665	97.570	97.949	97.929	99.049	98.471	84.291	99.968	99.963	99.958	99.964	99.900	99.976	97.727	100	100	100	100	100
10	50	78.550	100	100	100	100	100	100	98.590	100	100	100	100	100	100	100	100	95.250	95.510	100	100
	100	72.860	100	100	100	100	100	100	94.595	100	100	100	100	100	100	99.960	100	100	100	100	100
	200	70.115	100	100	100	100	100	100	91.460	100	100	100	100	100	100	99.695	100	100	100	100	100
	1000	67.077	100	100	100	100	100	100	88.514	100	100	100	100	100	100	98.858	100	100	100	100	100

Çizelge 6.7'de şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.7 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ=2 olduğu durumda değişken sayısı p=5 iken ve n=50, 100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=2 olduğu durumda değişken sayısı p=10 ve 20 iken ve n=100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 ve 10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranları vermektedir. Değişken sayısının yüksek olduğu p=10 ve 20 durumlarında ise klasik yöntem de dahil olmak üzere tüm yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir.

Çizelge 6.3 ve Çizelge 6.7 karşılaştırıldığında, Çizelge 6.3'te aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.7'de şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması ile p=5 olduğunda ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranlarında artış gözlenmiştir. Varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yötemlerde başarı oranının artması, büyük ξ=2, 5 ve 10 değerlerine göre aykırı değerlerin bulunmasının zor olduğu durumda bile şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının normal dağılıma göre aykırı değerlerin tespitinde daha kullanışlı olduğunu göstermektedir. Çizelge 6.5 ve Çizelge 6.7 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla p=5 iken ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranları artarken, ξ=2, 5 ve 10 için tüm yöntemlerin başarı oranları azalmıştır. p=10 ve 20 iken θ kirlilik seviyesinin artması ξ=2 olduğu durumda tüm yöntemlerin başarı oranlarını azaltmıştır. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.8. Şekil parametresi $t = 3.5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.80)\ LTS(t = 3.5\ , \boldsymbol{I_p}) + n(0.20)\ LTS(t = 3.5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	7	21	18	19	20	18	23	7	20	19	20	23	19	23	5	15	15	15	19	25
	100	11	26	25	27	30	34	34	13	38	33	33	35	35	37	13	40	38	38	36	42
	200	20	40	40	43	41	62	32	24	58	46	50	54	67	43	29	84	62	60	69	54
	1000	79	134	122	147	140	234	108	101	174	147	157	164	270	133	125	283	181	174	305	167
0.75	50	5	18	17	19	17	17	18	6	20	19	20	21	18	19	5	15	15	15	19	25
	100	11	21	22	22	25	28	22	11	32	29	32	30	30	18	11	38	36	36	30	19
	200	17	26	28	34	34	47	27	22	39	34	39	45	55	28	23	64	48	58	57	31
	1000	71	98	94	109	114	176	87	83	127	110	131	126	204	104	97	157	144	159	237	125
2	50	3	14	12	12	12	11	18	3	15	12	15	13	11	13	2	10	11	11	11	16
	100	5	11	11	18	14	19	16	5	19	17	16	17	18	16	4	21	20	22	21	14
	200	10	15	15	19	22	29	15	7	23	18	23	22	30	19	6	37	24	26	35	21
	1000	23	52	45	56	58	113	41	19	60	53	65	64	117	61	19	73	65	70	122	75
5	50	2	13	10	11	10	11	15	2	10	10	10	13	11	11	1	6	7	7	12	15
	100	2	12	11	15	13	17	17	1	18	16	15	15	18	19	3	20	17	19	19	18
	200	2	19	16	16	17	29	20	2	22	18	20	22	36	21	2	37	28	24	33	22
	1000	2	46	44	55	51	106	66	2	63	52	61	56	119	79	2	73	62	70	125	82
10	50	1	11	11	11	11	12	14	1	10	9	10	12	11	12	1	6	6	6	11	15
	100	0	15	11	13	12	17	17	1	20	15	16	14	20	14	1	20	16	19	21	10
	200	1	16	15	18	19	53	23	1	21	20	23	18	31	21	1	35	23	26	33	20
	1000	1	45	49	55	51	114	74	1	61	53	70	56	118	83	1	72	63	79	123	84

Çizelge 6.8'de şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.8 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının genel olarak azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını p=20 ve ξ=10 olduğu durum hariç OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009).

Çizelge 6.4 ve Çizelge 6.8 karşılaştırıldığında, Çizelge 6.4'te aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.8'de şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması ile tüm yönemler için yanlış tespit sayısının arttığı görülmüştür. Çizelge 6.6 ve Çizelge 6.8 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla p=5 iken ξ=0.5 ve 0.75 için tüm yöntemlerin yanlış tespit sayıları artarken, ξ=2, 5 ve 10 için tüm yöntemlerin yanlış tespit sayıları azalmıştır. p=10 ve 20 iken θ kirlilik seviyesinin artması ξ=2 olduğu durumda tüm yöntemlerin başarı oranlarını azaltmıştır.

Çizelge 6.9. Şekil parametresi $t = 5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.90)\ LTS(t = 5\ , \boldsymbol{I_p}) + n(0.10)\ LTS(t = 5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.000	0.460	0.460	0.400	0.500	0.180	0.160	0.000	1.300	1.460	1.980	0.520	0.120	0.020	0.000	1.460	2.440	2.220	0.060	0.360
	100	0.020	0.090	0.050	0.050	0.080	0.220	0.020	0.010	0.040	0.120	0.010	0.040	0.060	0.020	0.000	0.140	0.220	0.120	0.010	0.000
	200	0.030	0.040	0.060	0.025	0.015	0.080	0.020	0.000	0.010	0.015	0.005	0.040	0.020	0.005	0.000	0.000	0.000	0.005	0.005	0.000
	1000	0.026	0.029	0.029	0.037	0.034	0.092	0.033	0.009	0.011	0.011	0.011	0.016	0.021	0.011	0.000	0.001	0.003	0.001	0.004	0.002
0.75	50	0.340	3.300	4.140	4.620	4.000	3.380	0.860	0.200	8.660	9.400	12.180	5.600	2.860	0.280	0.020	10.000	11.220	12.760	1.960	9.760
	100	0.460	1.250	1.320	1.730	1.730	3.200	0.770	0.350	1.680	1.890	2.040	2.250	2.300	0.490	0.130	4.770	4.720	5.830	1.170	0.080
	200	0.550	1.025	1.030	1.220	1.360	2.155	0.880	0.034	0.640	0.855	0.960	1.040	1.695	0.425	0.180	0.790	0.785	0.855	1.015	0.200
	1000	0.586	0.798	0.769	0.952	0.948	2.105	0.700	0.406	0.575	0.539	0.600	0.575	1.258	0.444	0.183	0.286	0.248	0.296	0.663	0.219
2	50	40.900	62.000	64.240	68.860	69.680	72.600	52.780	53.320	88.080	85.460	87.380	91.060	89.140	63.560	47.680	86.760	85.360	84.780	97.700	98.440
	100	45.820	59.450	60.260	63.160	65.800	73.050	56.750	65.180	84.610	84.520	85.730	88.100	90.060	77.280	83.030	97.870	93.260	95.470	98.610	90.060
	200	47.070	58.545	58.510	63.240	62.945	72.670	58.435	68.230	83.120	82.410	84.960	85.375	90.015	80.185	88.720	98.035	96.420	97.710	98.730	95.465
	1000	47.321	57.946	57.465	61.443	61.672	71.913	57.633	69.760	81.770	81.150	83.430	83.396	89.850	81.770	90.396	97.224	96.765	97.455	98.735	97.017
5	50	89.960	98.460	98.500	98.700	99.180	99.200	98.100	98.860	100	99.900	99.980	100	99.980	99.880	99.960	99.960	99.700	99.680	100	100
	100	87.740	98.560	98.470	98.750	98.790	99.200	98.590	98.720	99.980	99.980	99.980	99.990	100	99.970	100	100	100	100	100	100
	200	86.745	98.530	98.465	98.680	98.605	99.190	98.465	98.440	99.985	99.980	99.980	99.980	99.990	99.985	99.975	100	100	100	100	100
	1000	85.158	98.468	98.397	98.708	98.761	99.265	98.793	97.709	99.984	99.979	99.979	99.983	99.997	99.987	99.958	100	100	100	100	100
10	50	97.000	100	99.860	99.920	100	100	100	99.980	100	100	100	100	100	100	100	100	100	100	100	100
	100	94.580	100	100	100	100	100	100	99.880	100	100	100	100	100	100	100	100	100	100	100	100
	200	92.270	100	100	100	100	100	100	99.575	100	100	100	100	100	100	100	100	100	100	100	100
	1000	90.250	100	100	100	100	100	100	99.132	100	100	100	100	100	100	99.997	100	100	100	100	100

Çizelge 6.9'da şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.9 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ=2 olduğu durumda değişken sayısı p=5 iken ve n=50, 100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=2 olduğu durumda değişken sayısı p=10 ve 20 iken ve n=100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 ve 10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranları vermektedir. Değişken sayısının p=10, 20 ve ξ=5 olduğu durumda klasik yöntem de dahil tüm çok iyi sonuç verdiği görülmektedir.

Çizelge 6.1 ve Çizelge 6.9 karşılaştırıldığında, Çizelge 6.1'de aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.9'da şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanılması ile p=5 olduğunda ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranlarında artış gözlenmiştir. Varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yötemlerde başarı oranının artması, büyük ξ=2, 5 ve 10 değerlerine göre aykırı değerlerin bulunmasının zor olduğu durumda bile şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının normal dağılıma göre aykırı değerlerin tespitinde daha kullanışlı olduğunu göstermektedir. Çizelge 6.5 ve Çizelge 6.9 karşılaştırıldığında ise şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması, ξ=0.5 ve 0.75 değerleri için şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanımına göre daha başarılı sonuçlar vermiştir. Yani şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanılması aykırı değer tespitini zorlaştıran küçük varyans-kovaryans katsayısı için daha etkilidir. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.10. Şekil parametresi $t = 5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.90)\ LTS(t = 5\ , \boldsymbol{I_p}) + n(0.10)\ LTS(t = 5\ , \xi^2 \boldsymbol{I_p})$

					p=5							p=10							p=20		
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	6	19	21	18	19	18	20	5	20	19	20	22	20	18	3	15	15	15	18	25
	100	10	21	23	21	27	30	22	10	34	31	28	31	31	45	11	39	38	38	31	11
	200	18	27	30	33	32	51	23	18	37	36	37	40	51	28	20	69	53	47	52	29
	1000	65	99	87	96	103	180	79	81	108	97	108	107	192	88	91	143	119	114	208	101
0.75	50	5	20	17	18	16	15	19	4	20	19	20	21	19	17	2	15	15	15	18	25
	100	8	17	20	23	20	26	14	10	30	29	27	29	27	14	9	38	37	36	29	10
	200	15	25	26	31	30	42	24	18	35	29	34	37	49	22	17	54	43	44	54	23
	1000	61	77	74	89	94	146	71	67	93	84	98	101	161	79	77	108	100	116	184	84
2	50	4	18	14	16	14	14	20	4	16	16	17	18	13	15	2	14	13	14	14	21
	100	7	13	14	15	15	22	13	6	25	24	23	24	20	11	4	30	29	27	22	8
	200	9	19	19	23	22	33	17	9	21	23	26	28	35	17	8	36	32	32	37	17
	1000	28	52	49	61	72	115	46	28	57	55	69	67	123	54	22	65	63	83	139	62
5	50	2	16	12	17	14	13	16	2	15	14	15	18	12	13	2	11	11	11	14	20
	100	2	12	16	14	15	18	13	2	25	20	21	20	21	12	3	28	27	28	22	8
	200	3	18	21	21	20	36	18	2	21	20	24	25	35	19	2	36	29	37	38	17
	1000	3	48	46	60	61	112	61	3	65	57	70	71	129	67	2	73	65	73	130	74
10	50	2	13	13	14	13	12	17	2	15	14	15	19	14	14	1	10	10	10	13	20
	100	2	14	13	18	14	21	13	2	20	20	21	19	22	12	3	28	26	28	21	10
	200	1	16	17	25	21	35	20	1	21	22	23	25	38	20	1	39	28	30	38	17
	1000	1	53	50	67	58	116	65	1	59	58	73	77	123	66	1	71	67	81	131	66

Çizelge 6.10' da şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.1 olduğu, yani verilerin %10 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.10 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının genel olarak azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını p=5, 10 ve 20 için OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009).

Çizelge 6.2 ve Çizelge 6.10 karşılaştırıldığında, Çizelge 6.2'de aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.10'da şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımı kullanılmıştır. Normal dağılım yerine şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanılması ile tüm yöntemler için yanlış tespit sayısının arttığı görülmüştür. Çizelge 6.10 ve Çizelge 6.6 karşılaştırıldığında ise şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanılması şekil parametresi $\mathrm{t} = 3.5$ olan LTS dağılımının kullanımına göre daha az yanlış tespit sayısına sahiptir.

Çizelge 6.11. Şekil parametresi $t = 5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin başarı oranları; $n(0.80)\ LTS(t = 5\ , \boldsymbol{I_p}) + n(0.20)\ LTS(t = 5\ , \xi^2 \boldsymbol{I_p})$

		p=5							p=10							p=20					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	0.010	0.540	0.610	0.670	0.490	0.460	0.200	0.000	1.660	1.740	2.090	0.710	0.110	0.050	0.000	1.530	2.250	2.700	0.110	0.490
	100	0.025	0.160	0.110	0.175	0.165	0.295	0.070	0.000	0.155	0.125	0.095	0.130	0.080	0.010	0.000	0.355	0.275	0.290	0.015	0.005
	200	0.033	0.090	0.073	0.090	0.080	0.273	0.053	0.010	0.055	0.025	0.023	0.033	0.073	0.008	0.003	0.075	0.038	0.013	0.013	0.005
	1000	0.027	0.055	0.047	0.058	0.063	0.192	0.037	0.008	0.027	0.012	0.017	0.015	0.050	0.013	0.005	0.009	0.006	0.003	0.009	0.005
0.75	50	0.570	3.760	4.210	4.800	4.870	4.260	1.100	0.200	10.420	10.340	13.220	8.030	3.690	0.410	0.030	11.350	12.550	12.510	2.680	11.910
	100	0.775	1.550	1.685	1.895	2.200	3.500	1.015	0.330	2.225	2.610	2.230	3.195	2.725	0.410	0.090	5.945	6.160	6.350	1.710	0.155
	200	0.733	1.138	1.198	1.593	1.420	3.078	0.920	0.385	0.980	1.078	1.198	1.313	2.295	0.528	0.173	1.503	1.170	1.118	1.455	0.288
	1000	0.731	0.973	0.962	1.203	1.212	2.589	0.915	0.487	0.738	0.655	0.780	0.797	1.728	0.588	0.246	0.468	0.355	0.446	1.003	0.301
2	50	32.550	57.030	59.550	64.080	64.660	67.080	50.760	41.970	85.730	81.190	83.900	89.340	85.990	59.970	34.860	79.830	76.630	77.130	96.260	98.500
	100	34.710	54.210	55.110	59.285	60.500	68.905	52.685	51.840	81.320	79.550	82.420	85.770	87.400	72.750	70.775	97.110	90.335	92.690	97.615	87.620
	200	35.770	53.120	52.708	57.175	57.818	68.620	53.030	54.060	79.490	78.048	81.018	81.710	88.290	77.238	76.493	97.305	95.208	96.358	98.185	94.920
	1000	36.494	52.051	51.045	56.204	56.275	68.778	50.013	55.325	77.972	76.409	79.305	79.384	88.196	77.639	78.854	96.285	95.644	96.370	98.479	96.630
5	50	71.260	97.780	98.070	98.730	98.640	98.880	98.040	93.070	99.980	99.670	99.820	100	99.960	99.800	99.430	98.550	94.150	93.970	100	100
	100	68.750	97.935	97.850	98.355	98.230	99.100	98.595	91.250	99.975	99.915	99.990	99.985	99.995	99.970	99.715	100	100	100	100	100
	200	66.913	98.075	97.950	98.355	98.375	99.103	98.638	89.228	99.985	99.953	99.978	99.973	99.980	99.970	99.268	100	100	100	100	100
	1000	65.301	97.988	97.962	98.279	98.224	99.139	98.756	87.157	99.970	99.969	99.974	99.972	99.991	99.987	98.526	100	100	100	100	100
10	50	80.040	99.890	99.940	99.950	99.950	99.960	99.910	98.670	100	99.990	100	100	100	100	100	100	95.690	95.930	100	100
	100	75.290	99.945	99.905	99.945	99.965	99.940	99.945	95.370	100	100	100	100	100	100	99.985	100	100	100	100	100
	200	72.730	99.923	99.890	99.948	99.928	99.975	99.943	93.128	100	100	100	100	100	100	99.798	100	100	100	100	100
	1000	70.456	99.921	99.912	99.926	99.927	99.972	99.951	90.857	100	99.999	99.999	100	100	100	99.311	100	100	100	100	100

Çizelge 6.11'de şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları verilmiştir. Ayrıca başarı oranları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.11 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ=0.5 ve 0.75 iken tüm yöntemlerin başarı oranları oldukça düşük olduğu görülmüştür. ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ=2 olduğu durumda değişken sayısı p=5 iken ve n=50, 100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=2 olduğu durumda değişken sayısı p=10 ve 20 iken ve n=100, 200 ve 1000 için en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 olduğunda p=5 ve 10 için en iyi başarı oranlarını yine OGK yöntemi vermektedir. ξ=10 ve p=5 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranları vermektedir. Değişken sayısının yüksek olduğu p=10 ve 20 durumlarında ve ξ=10 için ise klasik yöntem de dahil olmak üzere tüm yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir.

Çizelge 6.3 ve Çizelge 6.11 karşılaştırıldığında, Çizelge 6.3'te aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.11'de şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanılması ile p=5 olduğunda ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranlarında artış gözlenmiştir. ξ=0.5 ve 0.75 iken tüm yöntemlerde başarı oranının artması, ξ=2, 5 ve 10 değerlerine göre aykırı değerlerin bulunmasının zor olduğu durumda bile şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının normal dağılıma göre aykırı değerlerin tespitinde daha kullanışlı olduğunu göstermektedir. Çizelge 6.5 ve Çizelge 6.11 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla p=5 iken ξ=0.5 ve 0.75 için tüm yöntemlerin başarı oranları artarken, ξ=2, 5 ve 10 için tüm yöntemlerin başarı oranları azalmıştır. p=10 ve 20 iken θ kirlilik seviyesinin artması ξ=2 olduğu durumda tüm yöntemlerin başarı oranlarını azaltmıştır. Çizelge 6.9 ve Çizelge 6.11 karşılaştırıldığında LTS dağılımının şekil parametresinin artmasıyla ξ=0.5 ve 0.75 için yöntemlerin başarı oranları düşerken, ξ=2, 5 ve 10 için yöntemlerin başarı oranlarında artış meydana

gelmiştir. SDE yönteminin başarı oranı değişken sayısının artmasıyla sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için hesaplanmamıştır (Todorov, 2009).

Çizelge 6.12. Şekil parametresi $t = 5$ olan LTS dağılımına sahip veriler için aykırı değer tespit yöntemlerinin yanlış tespit sayıları; $n(0.80)\ LTS(t = 5\ , \boldsymbol{I_p}) + n(0.20)\ LTS(t = 5\ , \xi^2 \boldsymbol{I_p})$

		***p*=5**							***p*=10**							***p*=20**					
ξ	n	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	SDE	OGK	BACON	MD	FMCD	MVE	MEST	OGK	BACON
0.5	50	6	22	19	22	21	21	23	6	20	19	20	22	18	23	4	15	15	15	19	25
	100	11	24	23	26	26	32	23	11	39	31	33	34	33	34	11	40	39	39	35	44
	200	19	35	36	37	37	61	27	22	53	42	42	48	60	35	26	83	58	55	67	36
	1000	74	109	100	122	117	213	88	91	151	124	134	139	230	113	110	240	161	144	263	135
0.75	50	6	20	17	19	17	18	19	5	20	19	20	21	18	14	3	15	15	15	17	25
	100	10	20	24	23	21	25	13	12	27	27	26	29	28	14	9	38	35	35	31	12
	200	15	24	24	33	31	41	23	19	37	31	36	36	48	25	19	56	44	48	54	25
	1000	62	79	79	99	96	156	74	72	96	99	109	110	171	81	85	128	110	134	204	98
2	50	3	13	11	13	12	10	16	2	14	14	13	14	10	12	2	11	11	11	10	16
	100	4	11	11	12	12	16	10	3	16	15	15	15	16	13	2	20	19	22	17	11
	200	6	14	12	14	15	28	15	4	16	15	19	17	26	14	6	30	21	23	26	15
	1000	17	35	30	47	44	100	31	12	41	38	48	46	100	38	11	55	44	49	99	52
5	50	1	10	9	12	9	10	16	1	10	10	10	12	10	11	1	6	7	7	11	15
	100	1	9	14	12	11	17	13	1	17	14	15	13	18	10	1	20	17	18	17	14
	200	1	16	13	14	15	27	17	1	15	15	19	16	27	17	1	28	21	22	25	14
	1000	1	31	31	37	36	89	54	1	39	37	45	38	95	59	1	50	43	48	97	62
10	50	0	11	9	12	9	10	16	1	10	9	10	13	11	8	1	6	6	6	11	15
	100	0	12	12	12	11	15	15	1	17	12	14	11	16	11	0	20	16	18	18	9
	200	0	13	13	13	15	29	18	0	19	15	19	15	26	18	0	29	20	20	29	17
	1000	0	34	32	39	43	91	57	0	40	40	44	40	95	59	0	50	45	51	99	64

Çizelge 6.12'de şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımına sahip veriler için θ kirlilik seviyesinin 0.2 olduğu, yani verilerin %20 oranında kirletildiği durum için örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin yanlış tespit sayıları verilmiştir. Ayrıca yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Çizelge 6.12 incelendiğinde, veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin artmasıyla tüm yöntemlerin yanlış tespit sayılarının genel olarak azaldığı görülmüştür. p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir. n=1000 olduğunda en çok yanlış tespit sayısını p=20 ve ξ=10 olduğu durum hariç OGK yöntemi vermektedir. Değişken sayısının p=5 ve 10 olduğu durumlarda n=50 ve 100 için FMCD, MVE, MEST ve SDE yöntemlerinin benzer yanlış tespit sayılarına sahip olduğu görülmektedir. SDE yöntemi için sonuç verme süresinin çok uzun sürmesi sebebiyle p=20 için yanlış tespit sayısı hesaplanmamıştır (Todorov, 2009).

Çizelge 6.4 ve Çizelge 6.12 karşılaştırıldığında, Çizelge 6.4'te aykırı değer tespitinde kullanılan verilerin dağılımı için çok değişkenli Normal dağılım yerine Çizelge 6.12'de şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımı kullanılmıştır. Şekil parametresi $\mathrm{t} = 5$ olan LTS dağılımının kullanılması ile tüm yöntemler için yanlış tespit sayısının arttığı görülmüştür. Çizelge 6.6 ve Çizelge 6.12 karşılaştırıldığında θ kirlilik seviyesinin artmasıyla p=5 iken ξ=0.5 ve 0.75 için tüm yöntemlerin yanlış tespit sayıları artarken, ξ=2, 5 ve 10 için tüm yöntemlerin yanlış tespit sayıları azalmıştır. p=10 ve 20 iken θ kirlilik seviyesinin artması ξ=2 olduğu durumda tüm yöntemlerin başarı oranlarını azaltmıştır. Çizelge 6.10 ve Çizelge 6.12 karşılaştırıldığında LTS dağılımının şekil parametresinin artmasıyla yöntemlerin aykırı değerleri tespit sayısında genel olarak bir azalma görülmektedir.

Çizelge 6.13. Aykırı değer tespit yöntemleri için benzetim çalışmasının sonuç verme süreleri (saniye cinsinden)

	θ	p	n	**MD**	**FMCD**	**MVE**	**MEST**	**SDE**	**OGK**	**BACON**
Çok değişkenli normal	0.1	5	200	0.75	41.11	26.25	31.02	40.05	4.45	4.15
			1000	1.39	85.02	97.27	106.25	266.4	9.27	6.46
		10	200	1.12	92.78	50.78	52.22	381.81	7.59	4.12
			1000	2.45	193.46	211.94	209.25	2923.97	25.23	8
		20	200	1.56	308.86	115.39	122.11	-	19.74	5.35
			1000	4.66	603.99	492.14	508.2	-	83.25	13.11
	0.2	5	200	0.78	40.48	22.86	30.11	40.31	4.33	4.75
			1000	1.56	83.28	89.67	106.24	270.63	9.49	7.42
		10	200	1	92.25	47.09	56.97	380.32	7.88	4.32
			1000	2.51	188.57	199.23	207.3	2584.83	25.08	9.42
		20	200	1.62	309.91	116.58	124.16	-	20.62	5.56
			1000	4.44	627.78	504.77	515.14	-	83.61	14.9
$t = 3.5$ şekil parametreli LTS	0.1	5	200	1.48	41.39	23.36	30.91	40.89	5.03	4.9
			1000	3.5	84.73	96.58	118.77	268.75	11.23	8.51
		10	200	2.18	106.79	48.27	60.56	384.97	8.5	5.54
			1000	5.56	206.7	204.11	224.85	2651.65	27.91	12.03
		20	200	3.33	311.17	119.59	123.89	-	22	7.42
			1000	10.13	612.34	502.77	512.76	-	94.25	19.22
	0.2	5	200	1.5	41.16	23.55	30.95	40.69	4.86	4.76
			1000	3.51	84.11	91.87	110.19	268.86	10.9	8.86
		10	200	2.12	92.69	48.02	53.31	382.35	8.57	5.56
			1000	5.92	188.7	200.9	211.12	2584.31	27.67	12.79
		20	200	3.27	301.44	117.04	122.6	-	21.37	7.67
			1000	10.58	591.77	502	511.05	-	88.5	19.37
$t = 5$ şekil parametreli LTS	0.1	5	200	1.4	43.03	24.28	31.35	40.88	4.77	4.65
			1000	2.87	100.86	95.08	109.36	268.59	10.58	8.27
		10	200	2.2	92.61	48.33	53.22	383.43	8.98	5
			1000	5.3	190.75	202.11	210.71	2689.81	29.68	11.1
		20	200	3.02	310.75	116.92	122.99	-	21.13	7.22
			1000	9.11	616.23	501.71	511.09	-	87.64	18.43
	0.2	5	200	1.47	44.53	23.94	31.49	40.74	4.77	4.75
			1000	2.83	91.66	105.32	108.89	268.37	10.49	8.38
		10	200	1.83	100.68	47.92	52.54	381.83	8.18	5.32
			1000	5.03	203.09	204.05	210.13	2554.11	26.89	12.11
		20	200	3.01	301.79	116.7	124.4	-	21.04	7.07
			1000	9.23	594.67	498.31	508.93	-	87.3	18.62

Çizelge 6.13 incelendiğinde sonuç verme sürelerinin kirlilik seviyesinden çok az etkilendiği söylenebilir. Değişken sayısı p ve örnek genişliği n'in artması tüm yöntemlerin sonuç verme sürelerini de arttırmıştır. Aykırı değer tespiti için kullandığımız yöntemlerin sonuç verme sürelerinin verilerin sahip olduğu dağılımdan etkilenmediği söylenebilir. Bu yöntemlerden MD, OGK ve BACON'un bu başarı oranlarına diğer yöntemlere göre daha kısa sürede ulaştıkları görülmektedir. Sde yönteminde p'nin artmasıyla sonuç verme süresinin çok uzun olması dikkati çekmektedir (Todorov, 2009; Hadi vd, 2009).

7. SONUÇ

Bu çalışmada klasik Mahalanobis uzaklığı (MD) ve Hızlı-en küçük kovaryans determinant (FMCD), En küçük hacimli elipsoit (MVE), M-tahmin edicisi (MEST), Stahel-Donoho tahmin edicisi (SDE), Dikey Gnanadesikan-Kettenring (OGK) ve BACON yöntemleri ile elde edilen dayanıklı konum ve ölçek tahmin edicileri kullanılmıştır. Mahalanobis uzaklığı hesaplanırken dayanıklı tahmin yöntemleri ile elde edilen konum ve ölçek tahmin edicileri (3.1) eşitliğinde verilen Mahalanobis uzaklığı formülünde yerine yazılarak dayanıklı uzaklıklar elde edilmiştir. Klasik ve dayanıklı tahmin ediciler için elde edilen Mahalanobis uzaklıkları, normal dağılım ve normal dağılımın makul bir alternatifi olan uzun kuyruklu simetrik dağılım kullanılarak Monte-Carlo benzetim çalışması ile ayrı ayrı incelenmiştir. Çok değişkenli normal dağılıma sahip bir veri seti ve t= 3.5 ve t= 5 şekil parametreli LTS dağılımına sahip veri seti için θ kirlilik seviyesinin 0.1 ve 0.2 olduğu, yani verilerin %10 ve %20 oranında kirletildiği durum için, örneklem genişliği n=50, 100, 200 ve 1000 ve değişken sayısının ise p=5, 10, 20 olduğu durumlar için yöntemlerin başarı oranları, yanlış tespit sayıları ve bu yöntemlerin veri setine eklenen aykırı değerleri bulma süreleri saniye cinsinden elde edilmiştir. Ayrıca başarı oranları ve yanlış tespit sayıları, $\xi = 0.5, 0.75, 2, 5, 10$ katsayıları ile farklı varyanslar için de değerlendirilmiştir. Veri setine eklenen aykırı değerlerin varyans-kovaryans matrisinin katsayısı olan ξ'nin büyümesiyle başarı oranlarında da ciddi artışlar meydana gelmiştir. ξ'nin büyümesinin, eklenen aykırı değerlerin varyansının da büyümesi anlamına geldiği açıktır. Dolayısıyla varyansı büyüyen aykırı değerlerin tespit edilme oranlarının artması da beklenen bir sonuçtur.

Normal dağılım ve uzun kuyruklu simetrik dağılım (t= 3.5 ve t= 5) için ξ=2 olduğu durumda p=5, 10 ve 20 iken en yüksek başarı oranını OGK yöntemi vermektedir. ξ=5 olduğunda ve değişken sayısı p=5 iken her iki dağılımın tüm durumları için de OGK yöntemi yine en yüksek başarı oranına sahiptir. ξ=5 ve p=10 olduğunda tüm dayanıklı yöntemler klasik yöntemden daha iyi ve birbirlerine yakın oldukça iyi başarı oranlarına sahiptir. Değişken sayısının yüksek olduğu p=20 durumunda ve ξ=5 ve ξ=10 olduğunda klasik yöntem de dahil olmak üzere tüm

yöntemlerin başarı oranlarının çok iyi olduğu görülmektedir. Değişken sayısı p=5, 10 ve 20 iken en az yanlış tespit sayısını MD yöntemi vermektedir.

Tüm yöntemlerin aykırı değerleri tespit etmedeki program sonuç verme süreleri saniye cinsinde hesaplanmıştır. En hızlı sonuç veren yöntemin MD yöntemi, en geç sonuç veren yöntemin ise SDE yöntemi olduğu görülmüştür. SDE yönteminde programın çalışma süresinin çok uzun sürmesi nedeniyle p=20 için hesaplamalar yapılamamıştır. Hem normal dağılım için hem de uzun kuyruklu simetrik dağılım için programların sonuç verme süreleri de dikkate alındığında aykırı değer tespitinde OGK yönteminin oldukça başarılı sonuçlar verdiği söylenebilir. Aykırı değer içeren verileri modellemede sıklıkla kullanılan LTS dağılımı normal dağılıma alternatif olarak tercih edilebilir.

KAYNAKLAR

Alameddine, I., Kenney, M. A., Gosnell, R. J. and Reckhow, K. H. 2010. Robust multivariate outlier detection methods for environmental data. *Journal of environmental engineering*, 136(11), 1299-1304.

Alkan, B.B., Atakan. V. and Alkan, N. 2015. A comparison of Different Procedures for Principal Component Analysis in the Presence of Outliers. *Journal of Applied Statistics*. 42(8), 1716-1722.

Alpar, R. 2013. *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler.* Detay Yayıncılık, Ankara

Altın, A. ve Şenoğlu, B. 2008. Konum Parametresinin Bazı Sağlam Tahmin Edicilerinin Örnekleme Alanında Kullanılması Ve Bir Tarım Uygulaması. *AKÜ Fen Bilimleri Dergisi*. 8(1), 291-306.

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*. John Wiley, New York.

Billor, N., Hadi, A. S. and Velleman, P. F. 2000. BACON: blocked adaptive computationally efficient outlier nominators. *Computational statistics & data analysis*, 34(3), 279-298.

Cabana, E., Lillo, R. E. and Laniado, H. 2019. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *arXiv preprint arXiv:1904.02596*.

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

Chen, S., Wang, W. and Van Zuylen, H. 2010. A comparison of outlier detection algorithms for ITS data. *Expert Systems with Applications*, 37(2), 1169-1178.

Croux, C. and Haesbroeck, G. 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2), 161-190.

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D. L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1), 1-18.

Donoho, D.L. 1982. Breakdown Properties of Multivariate Location Estimators. Doctoral Dissertation, Harvard University, Boston, USA

Fallahi, K., Cheng, C. T. and Fattouche, M. 2011. Robust positioning systems in the presence of outliers under weak GPS signal conditions. *IEEE Systems Journal*, 6(3), 401-413.

Fawzy, A., Mokhtar, H. M. and Hegazy, O. 2013. Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, 14(2), 157-164.

Gaber, M. M. 2007. *Data stream processing in sensor networks.* Springer, 41-48, Heidelberg, Berlin.

Gervini, D. 2002. The influence function of the Stahel–Donoho estimator of multivariate location and scatter. *Statistics & probability letters*, 60(4), 425-435.

Gnanadesikan, R. and Kettenring, J. R. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81-124.

Grubbs, F. E. 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11: 1–21.

Filzmoser, P., Serneels, S., Maronna, R. and Van Espen, P. J. 2009. *Robust multivariate methods in chemometrics*.

Hadi, A. S. 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3), 761-771.

Hadi, A. S. 1994. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2), 393-396.

Hadi, A. S., Imon, A. R. and Werner, M. 2009. Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), 57-70.

Hampel, F. R. 1971. A general definition of qualitative robustness. *The Annals of Mathematical Statistics*, 42, 1887-1896.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346), 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. 1986. *Robust statistics: The Approach Basedon Influence Functions.* John Wiley & Sons, New York.

Hampel, F. R. 2001. *Robust statistics: A brief introduction and overview* (Report No. 94). Seminar für Statistik, Eidgenössische Technische Hochschule.

Hawkins, D. M. 1980. *Identification of outliers.* Chapman and Hall, 11, London.

Huber, P.J. 1964. Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, 35, 73-101.

Huber, P.J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 21 June – 18 July, Book of Abstracts, 221-233, Berkeley, America.

Huber, P.J. 1981. *Robust Statistics*. John Wiley & Sons, New York.

Hubert M. , Rousseuw P.J. and Brandan K.V. 2005. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47(1), 64-79.

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

Hubert, M., Rousseeuw, P. J. and Verdonck, T. 2012. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3), 618-637.

Kasap, P., Senoglu B. and Arslan, O. 2016. Stochastic analysis of covariance when the error distribution is long-tailed symmetric. *Journal of Applied Statistics*, 43(11), 1977-1997.

Kasap, P. 2011. Stokastik ANCOVA: İstatistiksel sonuç çıkarımı. Doktora tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, 151, Ankara

Leys, C., Klein, O., Dominicy, Y. and Ley, C. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of experimental social psychology*, 74, 150-156.

Lopuhaa, H. P. and Rousseeuw, P. J. 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1), 229-248.

Maesschalck, R.D., Jouan-Rimbaud, D. and Massart, D.L. 2000. The Mahalanobis Distance. *Chemometrics and intelligent laboratory systems*, 50(1), 1-18.

Mardia, K.V., Kent, J.T. and Bibby J.M. 1979. *Multivariate Analysis.* Academic Press, London.

Maronna, R. A. 1976. Robust M-estimators of multivariate location and scatter. *The annals of statistics*, 51-67.

Maronna, R. A. and Zamar, R. H. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.

Maronna, R. A., Martin, R. D. and Yohai, V.J. 2006. *Robust Statistics: Theory and Methods.* John Wiley & Sons, England.

McLachlan, G. J. 1999. Mahalanobis distance. *Resonance*, 4(6), 20-26.

Peña, D. and Prieto, F. J. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3), 286-310.

Reza, M. S. and Ruhi, S. 2015. Multivariate outlier detection using independent component analysis. *Science Journal of Applied Mathematics and Statistics, Science Publishing Group, USA*, 3(4), 171-176.

Riani, M., Atkinson, A. C. and Cerioli, A. 2009. Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 71(2), 447-466.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American statistical association*, 79(388), 871-880.

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297), 37.

Rousseeuw, P.J. and Leroy, A.M. 1987. *Robust regression and outlier detection.* John Wiley & Sons, New York.

Rousseeuw P.J. and Driessen K.V. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3), 212-223.

Sajesh, T. A. and Srinivasan, M. R. 2013. An overview of multiple outliers in multidimensional data. *Sri Lankan Journal of Applied Statistics*, 14(2), 87-120.

Shu, V. S. 1978. Robust estimation of a location parameter in the presence of outliers. Doctoral Dissertation, Graduate School of the Iowa State University, 190, Ames, Iowa.

Stahel, W.A. 1981. *Breakdown of covariance estimators* (Report No. 31). Fachgruppe für Statistik, ETH. Zurich.

Stahel, W., Maechler, M. 2019. Package 'robustX'. *R package version 0.5-03, URL http://CRAN. R-project. org/package= robustX*.

Staudte, R. and Sheather, S. 1990. *Robust estimation and testing*. Wiley, 357, New York.

Tiku, M. L., Islam, M. Q. and Selçuk, A. S. 2001. Nonnormal regression. II. Symmetric distributions. *Communications in Statistics-Theory and Methods*, 30(6), 1021-1045.

Tiku, M. L. 2004. *Robust estimation and hypothesis testing*. New Age International.

Tiku, M.L. and A.D. Akkaya. 2004. *Robust estimation and hypothesis testing*. New Age International.

Todorov, V. 2009. rrcov: Scalable robust estimators with high breakdown point. *R package version 0.5-03, URL http://CRAN. R-project. org/package= rrcov*.

Todorov, V. and Filzmoser, P. 2009. An object-oriented framework for robust multivariate analysis.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 448-485.

Tukey, J.W. 1962. The future of data analysis. *The Annals of Mathematical Statistics* 33, 1-67.

Uzabaci, E., Ercan, I. and Alpu, O. 2018. Evaluation of outlier detection method performance in symmetric multivariate distributions. *Communications in Statistics-Simulation and Computation*, 1-16.

Van Aelst, S. and Rousseeuw, P. 2009. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), 71-82.

Verdonck, T. and Hubert, M. 2011. Robust covariance estimation for financial applications. In EURANDOM-ISI Workshop on Actuarial and Financial Statistics, 29-30 August, Eindhoven.

Werner, M. 2003. Identification of multivariate outliers in large data sets. Doctoral Dissertation, University of Colorado, 241, Denver.

Zuo, Y. 2006. Robust location and scatter estimators in multivariate analysis. *In Frontiers In Statistics,* 467-490.

Zuo, Y., Cui, H. and He, X. 2004. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *The Annals of Statistics*, 32(1), 167-188.

EKLER

EK 1 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN KLASİK YÖNTEM İLE TESPİTİ

EK 2 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN FMCD YÖNTEMİ İLE TESPİTİ

EK 3 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN MVE YÖNTEMİ İLE TESPİTİ

EK 4 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN MEST YÖNTEMİ İLE TESPİTİ

EK 5 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN SDE YÖNTEMİ İLE TESPİTİ

EK 6 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN OGK YÖNTEMİ İLE TESPİTİ

EK 7 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN BACON YÖNTEMİ İLE TESPİTİ

EK 8 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN KLASİK YÖNTEM İLE TESPİTİ (*t*=3.5 İÇİN)

EK 9 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN FMCD YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 10 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN MVE YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 10 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN MVE YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 11 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN MEST YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 12 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN SDE YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 13 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN OGK YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 14 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN BACON YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

EK 1 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN KLASİK YÖNTEM İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000 #iterasyon sayısı
p=5 #değişken sayısı
n=50 #veri seti genişliği
müa=0 #veri setine eklenen aykırı değerlerin ortalaması
sda=0.5 # veri setine eklenen aykırı değerlerin standart sapması (0.5, 0.75, 2, 5 ve 10
değerleri alınmıştır.)
teta=0.1#kirlilik seviyesi(0.1 ve 0.2 değerleri alınmıştır.)
r=teta*n #veri setine eklenen aykırı değer sayısı
kritik<-qchisq(.975, df=p) #aykırı değerleri belirlemede kullanılan kesim noktası
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda))) #normal dağılıma sahip veri üretme
   ort<-colMeans(data)
   cov<-cov(data)
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1]) #yanlış tespit sayısı
meano<-(mean(sonuç[,2])/r)*100 #başarı oranı
son<-c(maxi,meano)
son
toc()
```

EK 2 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN FMCD YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
teta=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   mcd<-CovMcd(data)
   ort<-mcd@center
   cov<-mcd@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 3 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN MVE YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
teta=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   mve<-CovMve(data)
   ort<-mve@center
   cov<-mve@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 4 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN MEST YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
teta=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   mest<-CovMest(data)
   ort<-mest@center
   cov<-mest@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 5 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN SDE YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
teta=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   sde<-CovSde(data)
   ort<-sde@center
   cov<-sde@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 6 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN OGK YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
alfa=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   ogk<-CovOgk(data)
   ort<-ogk@center
   cov<-ogk@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 7 R-PAKET PROGRAMINDA NORMAL DAĞILIMA SAHİP VERİLERDE AYKIRI DEĞERLERİN BACON YÖNTEMİ İLE TESPİTİ

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.5
teta=0.1
r=teta*n
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   data<-replicate(p, c(rnorm(n-r,0,1),rnorm(r,müa,sda)))
   bacon<-BACON(data)
   ort<-bacon$center
   cov<-bacon$cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(maxi,meano)
son
toc()
```

EK 8 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN KLASİK YÖNTEM İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000 #iterasyon sayısı
p=5 #değişken sayısı
n=50 #veri seti genişliği
müa=0 #veri setine eklenen aykırı değerlerin ortalaması
sda=0.5 # veri setine eklenen aykırı değerlerin standart sapması (0.5, 0.75, 2, 5 ve 10 değerleri alınmıştır.)
teta=0.1#kirlilik seviyesi(0.1 ve 0.2 değerleri alınmıştır.)
r=teta*n #veri setine eklenen aykırı değer sayısı
t=3.5 #LTS dağılımının şekil parametresi
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p) #aykırı değerleri belirlemede kullanılan kesim noktası
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast #LTS dağılımına sahip veri üretme
   cov<-cov(data)
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1]) #yanlış tespit sayısı
meano<-(mean(sonuç[,2])/r)*100 #başarı oranı
son<-c(maxi,meano)
son
toc()
```

EK 9 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN FMCD YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   mcd<-CovMcd(data)
   ort<-mcd@center
   cov<-mcd@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

EK 10 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN MVE YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   mve<-CovMve(data)
   ort<-mve@center
   cov<-mve@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

EK 11 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN MEST YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   mest<-CovMest(data)
   ort<-mest@center
   cov<-mest@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

EK 12 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN SDE YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   sde<-CovSde(data)
   ort<-sde@center
   cov<-sde@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

EK 13 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN OGK YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   ogk<-CovOgk(data)
   ort<-ogk@center
   cov<-ogk@cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

EK 14 R-PAKET PROGRAMINDA LTS DAĞILIMINA SAHİP VERİLERDE AYKIRI DEĞERLERİN BACON YÖNTEMİ İLE TESPİTİ (*t*=3.5 İÇİN)

```
rm(list=ls(all=TRUE))
tic("total")
m=1000
p=5
n=50
müa=0
sda=0.1
teta=0.1
r=teta*n
t=3.5
nu=2*t-1
k=2*t-3
katsayi=1/sqrt(nu/k)
kritik<-qchisq(.975, df=p)
a<-matrix(nrow=m,ncol=1)
b<-matrix(nrow=m,ncol=1)
for (i in 1:m) {
   datast<-replicate(p, c(rst(n-r,0,1,nu),rst(r,müa,sda,nu)))
   data<-katsayi*datast
   bacon<-BACON(data)
   ort<-bacon$center
   cov<-bacon$cov
   distance<-mahalanobis(data,ort,cov)
   outliers<-ifelse(distance>kritik,1,0)
   inlier<-outliers[1:(n-r)]
   a[i]<-sum(inlier)
   outlier<-outliers[(n-r+1):n]
   b[i]<-sum(outlier)
   sonuç<-cbind(a,b)
}
maxi<-max(sonuç[,1])
meano<-(mean(sonuç[,2])/r)*100
son<-c(meano,maxi)
son
toc()
```

ÖZGEÇMİŞ

Adı ve Soyadı: Tuba ÇELEBİ

Doğum Yeri: Samsun

Doğum Tarihi: 20.04.1991

Yabancı Dili: İngilizce

Eğitim Durumu

Lise : Recep Tanrıverdi Lisesi (2009)

Lisans: Sinop Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü (2013)

Çalıştığı Kurum

T.C. Cumhurbaşkanlığı Savunma Sanayii Başkanlığı (2019- …)

Yayınlar

Kasap, P., Çelebi, T. Ve Acıtaş, Ş. (2016). Temel Bileşenler Analizinde Bazı Dayanıklı Kovaryans Matrisi Tahminleri ve Bir Uygulama, X. International Statistics Days Conference , 07-09 October 2016, Giresun, Turkey

Kasap, P., Çorba, B.Ş. ve Çelebi, T. (2017). Data Mining Methods and Their Application in Health Insurance, International Conference on Computational and Statistical Methods in Applied Sciences (COSTAS), 9-11 November 2017, Samsun, Turkey

Kasap, P., Çorba, B.Ş. ve Çelebi, T. (2018). Comparison of Artificial Neural Networks and Decision Trees Methods on Health Insurance Data, International Journal For Scientific Research & Development, Vol. 6, Issue 10, pp. 252-257.